BenchCouncil Transactions
on Benchmarks, Standards and Evaluations

TBench
Volume 5, Issue 4
2025

Original Articles

ISSN: 2772-4859

BenchCouncil Transactions on Benchmarks, Standards and Evaluations (TBench) is an open-access multi-disciplinary journal dedicated to benchmarks, standards, evaluations, optimizations, and data sets. This journal is a peer-reviewed, subsidized open access journal where The International Open Benchmark Council pays the OA fee. Authors do not have to pay any open access publication fee. However, at least one of the authors must register BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench) (https://www.benchcouncil.org/bench/) and present their work. It seeks a fast-track publication with an average turnaround time of one month.

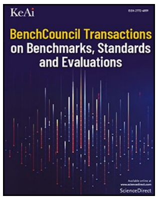
Contents

Contents

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Expert consensus and reliability validation of the portfolio assessment guideline for Chinese practical writing: An empirical study based on fleiss' kappa

Ying Wang (王莹)[a,b,1,*], Ibnatul Jalilah Yusof[b]

[a] *Department of Humanities and Arts Education, Hebei Finance University, Lianchi District, Baoding, Hebei Province, China*
[b] *Faculty of Educational Science and Technology, Universiti Teknologi Malaysia, Johor Bahru, Malaysia*

ARTICLE INFO

Keywords:
Assessment guideline
Chinese practical writing
Fleiss' kappa
Portfolio assessment
Reliability validation

ABSTRACT

Purpose: Portfolio assessment has been increasingly recognized as an effective approach to fostering comprehensive writing ability. However, its application in Chinese practical writing remains limited. The lack of standardized evaluation criteria has hindered its reliability and broader implementation. This study aimed to systematically develop and validate a Portfolio Assessment Guideline for Chinese practical writing, focusing on inter-rater reliability and coverage across four core dimensions: content, logical structure, language, and format.
Methods: Five higher education experts with extensive experience in practical writing instruction and research independently rated the guideline and its scoring rubrics for two key genres (summary and official notice), and inter-rater agreement was assessed using Fleiss' Kappa coefficients.
Findings: The Kappa values for six core modules ranged from 0.79 to 1.00, with an overall Kappa of 0.87 across 14 sub-dimensions, indicating "almost perfect" agreement. Genre-specific analysis showed high overall consistency for summary (κ=0.87) and official notice (κ=0.89), with the summary's "logical structure" dimension achieving "substantial agreement" (κ=0.68). Based on expert feedback, descriptive indicators were refined without altering the core framework.
Value: The findings provide robust evidence for the psychometric quality of the guideline, supporting its potential application in higher education and professional training for enhancing Chinese practical writing abilities.

1. Introduction

Portfolio assessment, as a continuous paradigm, integrates the assessment of both learning processes and competency development by systematically collecting learners' multiple outputs over a given period (e.g., drafts, revisions, reflective records). It has been widely applied in the field of language education, particularly in writing assessment [1,2]. Compared with traditional standardized testing, portfolio assessment offers the advantage of capturing learners' performance in authentic contexts, thereby balancing both product-oriented and process-oriented values [3,4]. This approach is particularly well aligned with the nature of practical writing. As functional texts, Chinese practical writings (e.g., plans, notices, requests) emphasize genre conventions, situational appropriateness, and accuracy of expression. The development of such ability relies on multiple rounds of writing practice and revision based on feedback, which calls for an evaluation tool that values both process and reflection [5].

In China's higher education system, practical writing courses serve as a core vehicle for cultivating students' professional communication ability, covering a wide range of genres such as governmental documents, administrative writings, and news reports [6]. However, existing research indicates two major limitations in current approaches to assessing Chinese practical writing: first, the lack of a specialized evaluation framework, as most assessments rely heavily on instructors' subjective judgment with blurred definitions of key dimensions such as "thematic content" and "formatting conventions"; and second, insufficient validation of assessment tools, particularly regarding inter-rater reliability, which undermines the objectivity and fairness of the results

Peer review under the responsibility of The International Open Benchmark Council.

[*] Corresponding author: Department of Humanities and Arts Education, He Finance University, Lianchi District, Baoding, Hebei Province, China.
E-mail address: ying20@graduate.utm.my (Y. Wang).

[1] This research was completed while the first author, Wang Ying, was pursuing her doctoral degree at Universiti Teknologi Malaysia.

https://doi.org/10.1016/j.tbench.2025.100248

Available online 2 December 2025

[5]. Inter-rater reliability, as a core indicator of the quality of any assessment tool, essentially ensures that different raters' judgments of the same performance are not unduly influenced by individual biases. It is therefore a prerequisite for the widespread application of assessment instruments [7].

Fleiss' Kappa, an extension of Cohen's Kappa, is specifically designed to test the consistency of categorical ratings among multiple raters (≥3). By controlling for chance agreement, it has been established as a reliable metric for measuring inter-rater agreement in fields such as medical diagnosis and educational measurement [8–10]. However, its application in the field of Chinese writing assessment remains rare, particularly in reliability testing of portfolio-based assessments of practical writing. Existing studies mostly adopt Pearson correlation coefficients or Cohen's Kappa, which are insufficient to address the reliability requirements of multi-rater, multi-dimensional evaluations.

Against this backdrop, the present study systematically examines the inter-rater reliability of the Portfolio Assessment Guideline for Chinese Practical Writing developed in earlier research, using Fleiss' Kappa as the statistical method. Specifically, the study addresses the following research questions:

Do the six core modules of the guideline demonstrate reliable inter-rater consistency when evaluating four key abilities: thematic content, logical structure, linguistic expression, and formatting conventions?

For the two representative genres, summary and official notice, how consistent are raters' judgments with regard to the specific assessment indicators defined in the guideline?

Based on the consistency analysis, what directions for refinement and implications for wider application can be derived for the guideline?

2. Literature review

2.1. Portfolio assessment in writing education

The value of portfolio assessment in writing education has been widely recognized. Its core advantage lies in overcoming the limitations of traditional tests, which capture only "one-shot performances." By longitudinally tracking learners' textual production (e.g., drafts, revisions, reflective journals), portfolio assessment provides a dynamic picture of competency development [1,2]. In higher education contexts, portfolio assessment serves both formative and summative purposes: the former enhances learning improvement through continuous feedback, while the latter enables a holistic judgment of competency achievement based on multiple sources of evidence [11]. Empirical studies have shown that this approach effectively improves students' metacognitive awareness, writing autonomy, and learning motivation [12].

From an international perspective, portfolio assessment has yielded abundant findings in ESL/EFL writing. It not only helps address specific weaknesses of learners at different proficiency levels (e.g., sentence construction among lower-proficiency learners) [13], but also, through innovative forms such as e-portfolios, strengthens feedback interaction and peer review, thereby promoting the comprehensive development of writing skills ranging from basic to advanced [14,15]. Comparative studies further confirm its advantages over traditional summative assessments in enhancing learners' writing complexity, accuracy, fluency, and self-efficacy [16].

In the Chinese writing domain, exploration of portfolio assessment began relatively late [17]. Early studies primarily focused on basic education, such as Yu [18], who proposed an implementation procedure for "composition growth portfolios" in high school, and Wei [19], who designed a primary school e-portfolio on the Moodle platform. Although some studies (e.g., [20]) have attempted to apply portfolio assessment to teaching Chinese as a second language, confirming its positive effects on promoting learners' autonomy and reflective ability, the field still faces clear limitations: (i) a scarcity of empirical research, with most studies confined to theoretical frameworks or process design [21]; (ii) limited application in higher education, especially in practical genres such as functional writing; and (iii) insufficient scientific validation of assessment outcomes, as the lack of quantitative analysis makes it difficult to establish reliability and validity for broader adoption [22].

Overall, while existing research has predominantly focused on the impact of portfolio assessment on learning outcomes, little attention has been given to the reliability mechanisms (e.g., inter-rater consistency) and standardization pathways of portfolio-based assessment in Chinese practical writing. This research gap provides the central entry point for the present study.

2.2. Assessment of practical / functional writing

Practical or functional writing distinguishes itself from literary writing by its explicit communicative purposes, specific pragmatic contexts, and standardized textual structures [6]. The core of this type of writing lies in "situational appropriateness," requiring writers to select suitable genre structures, linguistic styles, and formatting conventions according to communicative needs [23]. In Chinese higher education, practical writing courses cover a variety of genres such as official notices, summaries, and reports. Their assessment must move beyond the "literary orientation" often found in general writing evaluation, instead focusing on functional indicators such as accuracy of information transmission, compliance with formatting standards, and pragmatic appropriateness [5].

Scholars have reached consensus regarding the core components of practical writing ability. Unlike literary writing, which emphasizes creativity, practical writing focuses on four major dimensions: thematic content (relevance and completeness of information), logical structure (conformity to genre frameworks and coherence), linguistic expression (appropriateness and accuracy of register), and formatting conventions (standardization of layout and presentation) [24,25]. This categorization not only echoes the general writing elements such as "form" and "mechanics" proposed by Diederich et al. [26], but also reflects the genre-specific characteristics of practical writing, aligning with the requirements for functional writing in China's college entrance examination marking criteria.

Nevertheless, current practices in assessing practical writing reveal significant shortcomings. Recent studies [27] have pointed out the issue of "academic transplantation," whereby existing assessment tools simply borrow dimensions from academic writing (e.g., originality, depth of argumentation), thereby neglecting the functional essence of practical writing. More critically, specialized assessment tools for Chinese practical writing remain scarce, and their reliability, particularly regarding inter-rater consistency in multi-rater settings, has not been rigorously validated from a psychometric perspective. This deficiency undermines the objectivity and comparability of evaluation results [5]. The situation highlights the pressing need to develop standardized assessment guidelines and to rigorously validate their reliability.

2.3. Inter-Rater reliability in writing assessment

Inter-rater reliability is a core indicator of the scientific rigor of any assessment tool. It essentially examines whether different raters can make consistent and stable judgments about the same evaluation object [28]. In writing assessment, the rating process is inevitably influenced by subjective factors such as raters' prior experience and their varied interpretations of standards [29]. Therefore, reliability validation becomes a prerequisite for both the development and application of assessment tools. Only when raters can achieve a high level of consensus based on the same criteria can assessment results be regarded as objective and valid.

The choice of consistency testing methods depends on the type of data and the number of raters involved [30]:

Pearson correlation coefficient is applicable to continuous data (e.g., scores on a 0–100 scale). It measures the linear association between two raters' scores, but it cannot distinguish between absolute agreement and trend agreement, nor can it account for chance agreement.

Cohen's Kappa is designed for two raters evaluating unordered categorical data (e.g., "excellent / good / average"). By controlling for chance agreement, it provides more accurate results. However, it is limited to two-rater contexts [29].

Intraclass Correlation Coefficient (ICC) can be extended to multiple raters and is suitable for continuous or ordinal data. Yet, its applicability to categorical data is weaker, and its results can vary considerably depending on whether an absolute agreement or relative agreement model is selected.

In multi-rater contexts (more than two raters), the limitations of these methods become particularly evident: correlation coefficients cannot control for chance agreement, and Cohen's Kappa requires multiple pairwise comparisons to indirectly infer overall consistency. Both approaches are inadequate for complex evaluation scenarios, such as portfolio assessment, where multiple experts provide categorical ratings across multiple dimensions.

A critical research gap lies in the limited application of appropriate reliability analysis methods in the evaluation of Chinese practical writing. Specifically, there has been insufficient systematic examination of inter-rater agreement on detailed criteria such as "thematic content" and "formatting conventions." As a result, the scientific robustness of existing evaluation tools remains under-validated.

2.4. Application of fleiss' kappa in educational assessment

Fleiss' Kappa, an extension of Cohen's Kappa, is specifically designed to measure the consistency of categorical data involving multiple raters ($m \geq 2$), multiple objects ($n \geq 1$), and multiple categories ($k \geq 2$) [8]. Its core advantages are as follows:

By calculating the ratio of observed agreement to expected agreement (chance probability), it effectively eliminates the influence of chance agreement, thereby addressing the limitation of correlation coefficients in controlling for random consistency [31].

It is suitable for scenarios with unequal numbers of raters (e.g., some objects rated by three raters, others by five), making it more flexible than the intraclass correlation coefficient (ICC).

It demonstrates stronger adaptability to categorical data (e.g., "meets requirements / partially meets requirements / does not meet requirements"), which makes it particularly well suited to writing assessment contexts that involve multiple dimensions and multiple rating levels [32].

The interpretive standards proposed by Fleiss [33] have been widely adopted in academic research.Table 1

In the field of educational assessment, Fleiss' Kappa has been widely used for testing inter-rater reliability in language assessment, such as validating the effectiveness of writing scoring rubrics [32] and evaluating the impact of rater training in speaking assessments. Its value in measuring multi-rater agreement on categorical data has also been well established in fields such as medicine and psychology [10].

However, its application in Chinese writing assessment remains rare. In particular, within the complex context of portfolio assessment, which involves multiple text versions and multiple competency dimensions, no studies to date have systematically employed this method to examine the inter-rater reliability of assessment guidelines. This gap leaves the psychometric properties (e.g., reliability) of Chinese practical writing portfolio assessment tools without sufficient empirical evidence, thus constraining their standardization and broader implementation.

3. Research methodology

This section elaborates on the research design, participants, assessment instrument, data collection procedures, and analytical methods. The aim is to address the central question: "What is the inter-rater reliability of the Chinese Practical Writing Portfolio Assessment Guideline?" and to ensure the scientific rigor and replicability of the study.

3.1. Participants

Content validity is a fundamental criterion in the development of assessment instruments. Its essence lies in verifying, through expert judgment, the degree of alignment between the instrument's content and the intended evaluation objectives. Specifically, this includes the appropriateness of tasks, clarity of standards, alignment with curricular goals, and feasibility of implementation [34,35]. Although no consensus exists regarding the optimal number of experts, Lynn [36] suggested a range of 3–10, while Polit and Beck [37] considered 2–15 to be acceptable. The key requirement is that experts must possess sufficient professional expertise related to the assessment object [38].

In line with these principles, five experts were recruited as raters in this study. All were associate professors or above in the field of Chinese language education at domestic universities, and they met the following qualifications:

More than ten years of teaching experience in practical writing courses, with a solid understanding of the pedagogical characteristics of key genres such as summary and official notice;

Experience in leading or contributing to the compilation of practical writing textbooks, thus possessing systematic knowledge of assessment standards;

Familiarity with the development of writing assessment tools and a sound understanding of the basic logic of reliability testing.

All raters had previously received training in the use of analytic scoring rubrics. However, they had no prior exposure to the specific portfolio assessment guideline under investigation, in order to minimize potential bias and ensure objectivity in scoring.

3.2. Assessment instrument

The development of the Portfolio Assessment Guideline for Chinese Practical Writing was driven by the need to address the lack of standardized criteria in the assessment of Chinese practical writing portfolios. The guideline integrates Moya's systemic assessment framework [39], Lam's theory of self-regulated learning [40], and Delett's structured design principles [41], thereby encompassing three essential dimensions: assessment, teaching, and learning.

3.2.1. Portfolio assessment guideline for Chinese practical writing

The development of the guideline followed a "goal-setting – content selection – standard formulation – implementation validation" sequence, with the following core steps:

Goal setting: The guideline emphasizes three main goals: (i) enhancing students' ability in four major areas of practical writing, namely content relevance and completeness, organizational structure, linguistic appropriateness and accuracy, and formatting conventions;; (ii) cultivating a self-regulated learning cycle of planning, monitoring, evaluating through reflective journals and multi-source feedback; (iii) providing teachers with multidimensional data to inform pedagogical

Table 1
Interpretation of Kappa values for Inter-rater agreement.

Kappa Value Range	Level of Agreement	Interpretation
Kappa < 0.40	Poor agreement (Unacceptable)	Indicates significant divergence in raters' understanding of the evaluation criteria.
$0.40 \leq$ Kappa ≤ 0.75	Intermediate to good agreement (Acceptable)	Suggests that the standards are generally clear but still have room for refinement.
Kappa > 0.75	Excellent agreement (Highly reliable)	Indicates strong consensus among raters and highly reliable results.

adjustments.

Content structure: Drawing on Moya's "process–product" evaluation and Lam's SRL cycle, the portfolio consists of three types of materials: writing samples (including annotated drafts, revised drafts with justification, and final drafts with format checklists), reflective journals (records of learning strategies and self-evaluations of goals), and feedback records (self-assessment checklists, peer review forms, and teacher diagnostic reports).

Scoring system: Based on the standards of the Chinese College Entrance Examination and the specific features of practical writing, the scoring rubric comprises four dimensions and eleven indicators, supported by a three-level rating scale (with descriptors for each level).

Implementation management: In accordance with Delett's principle of "task–standard alignment," the teaching activities were designed as a cyclical sequence of input, output, reflection, revision. The portfolio adopts a "genre–category" filing system, organized by weekly timeline, to ensure systematic collection and management of materials.

3.2.2. Expert rating form

To examine the inter-rater reliability of the guideline, an Expert Rating Form was designed based on its core components. The form specifies six modules: assessment goals, portfolio content, evaluation standards, instructional procedures, management guidelines, and implementation recommendations, which were further operationalized into 14 rating dimensions (e.g., "clarity of indicator descriptions," "practicality of instructional procedures").

Each dimension employed a 3-point ordinal scale (1 = not applicable/inappropriate, 2 = partially applicable/problematic, 3 = applicable/appropriate). This design enabled the collection of expert judgments on the consistency of each dimension, in alignment with the analytical requirements of Fleiss' Kappa.

3.3. Research procedures

3.3.1. Rater training

A two-hour standardized workshop was conducted to ensure consistency among raters. The training included: an overview of the guideline's development background and core objectives; operational definitions of the four scoring dimensions and eleven indicators; instructions on the use of the three-level analytic rubric.

3.3.2. Independent rating and feedback collection

Within one week, raters independently evaluated the Portfolio Assessment Guideline using the Expert Rating Form. All ratings were recorded in electronic forms, and raters were prohibited from any communication during the process. Upon completion, both the rating forms and qualitative feedback from the experts were collected to support subsequent data analysis and guideline refinement.

3.4. Data analysis

Data analysis was conducted using Excel, following these steps:

Data organization: Expert ratings were converted into a "rater–dimension" matrix (rows = raters, columns = rating dimensions, cells = rating levels).

Reliability calculation: Fleiss' Kappa values were computed for (a) the six core modules, (b) the fourteen rating dimensions as a whole, and (c) the four primary writing dimensions (content, structure, language, format) across two key genres (summary and official notice).

Result interpretation: Consistency levels were evaluated according to Fleiss' criteria (<0.40 = poor, 0.40–0.75 = fair to good, >0.75 = excellent) [33]. Dimensions with low agreement were further analyzed in light of the experts' qualitative feedback to identify directions for refinement.

4. Results

4.1. Overall consistency of the assessment guideline

Fleiss' Kappa analysis revealed that the inter-rater consistency coefficients across the six core modules of the guideline ranged from 0.79 to 1.00, with an overall Kappa value of 0.87 across the 14 dimensions (Table 2). According to Fleiss' standard (≥0.75 = excellent), this indicates excellent consistency [33]. Notably, the "Implementation Objectives" and "Instructional Procedures" modules achieved perfect agreement (κ = 1.00), suggesting a strong consensus among raters regarding the evaluation purposes and operational procedures of the guideline. The "Assessment Criteria" module (κ = 0.84) demonstrated slightly higher consistency compared to modules such as "Portfolio Content" and "Teacher Toolkit" (κ = 0.79), indicating that raters' understanding was more stable for the core evaluation dimensions.

4.2. Consistency of scoring rubrics for specific genres

To evaluate the effectiveness of the guideline across different genres of practical writing, two types of scoring rubrics: summary and official notice, were examined. Both genres demonstrated high inter-rater consistency (see Table 3). The overall Fleiss' Kappa for official notices (κ = 0.89) was slightly higher than that for summaries (κ = 0.87).

At the dimension level, the "Structure" dimension in summaries showed the lowest agreement (κ = 0.68, good), whereas all other dimensions achieved excellent agreement (κ ≥ 0.79). For official notices, both the "Language" and "Format" dimensions reached perfect agreement (κ = 1.00), while "Content" and "Structure" achieved excellent agreement (κ = 0.79).

Table 2
Fleiss' Kappa values for each module of the assessment guideline.

Module	Scoring Dimensions (Items)	No. of Items	κ	Interpretation
Implementation Objectives	1. Clarity of objective statements 2. Comprehensiveness of ability cultivation	2	1.00	Excellent
Portfolio Content	3. Relevance of material types to ability assessment 4. Rationality of genre requirements	2	0.79	Excellent
Assessment Criteria	5. Genre-specific adaptability of rubrics 6. Operability of indicator descriptions 7. Differentiation of three-level scoring system 8. Appropriateness of four-dimension weighting	4	0.84	Excellent
Teacher Toolkit	9. Scientific rigor of dynamic tracking sheets 10. Articulation between rubrics and tracking sheets	2	0.79	Excellent
Instructional Procedures	11. Guidance effectiveness for ability development 12. Validity of multi-source feedback	2	1.00	Excellent
Management Norms	13. Feasibility of timeline 14. Scientific categorization of portfolios	2	0.79	Excellent
Overall		14	0.87	Excellent

Table 3
Fleiss' Kappa Values for Specific Genre Scoring Rubrics.

Genre	Module	Scoring Dimensions (Items)	No. of Items	κ	Interpretation
Summary					
	Content	1. Task Coverage 2. Problem Objectivity 3. Experience Abstraction Level 4. Data Support 5. Causal Attribution Logic 6. Improvement Feasibility	6	0.93	Excellent
	Structure	7. Framework Coherence 8. Hierarchy Clarity	2	0.68	Good
	Language	9. Expression Objectivity 10. Graphic Assistance	2	0.79	Excellent
	Format	11. Element Completeness	1	1.00	Excellent
	Overall		11	0.87	Excellent
Official Notice					
	Content	1. Core Elements 2. Information Completeness 3. Information Conciseness	3	0.79	Excellent
	Structure	4. Framework Coherence 5. Hierarchy Clarity	2	0.79	Excellent
	Language	6.Terminology Standardization 7. Expression Precision	2	1.00	Excellent
	Format	8. Letterhead 9. Body 10. Closing Elements	3	1.00	Excellent
	Overall		10	0.89	Excellent

4.3. *Evaluation of guideline applicability*

Experts reached perfect consensus across the four core indicators of guideline applicability (κ = 1.00), confirming that the guideline: (i) demonstrates a high degree of alignment between assessment dimensions and key abilities; (ii) provides scoring criteria with strong operability; (iii) is suitable for both undergraduate education and professional training contexts; and (iv) holds potential for large-scale implementation.

4.4. *Analysis of expert feedback*

While the core framework of the guideline received strong recognition, expert feedback indicated that certain dimensions required further refinement.

4.4.1. *Core issues requiring optimization and revision measures*

Within the assessment criteria module, the item “operability of indicator descriptions” was rated “2″ by two experts. For instance, in the “summary” genre, the original standard for “degree of experience extraction” merely referred to “transferable methodology” without specifying “transfer contexts.” After revision, the indicator was refined into a three-level description: “transferable principles (including applicable contexts) → superficial experience (context not specified) → non-transferable,” supplemented with concrete examples such as “small- and medium-scale campus events.”

The original weight of “degree of experience extraction” in the “summary” genre was 5 points, which all five experts considered too low; for the “official notice” genre, the original weight of “degree of information conciseness” was 5 points, with three experts recommending an increase. Following revision, the weight of “degree of experience extraction” was adjusted to 10 points (while maintaining the total weight of the content module), and “degree of information conciseness” was adjusted to 8 points.

The original standard for “clarity of hierarchy” in the “summary” genre mentioned only “graded headings,” without covering intra-paragraph logic; in the “official notice” genre, the original description of “loose structure” lacked quantitative benchmarks. After revision, “clarity of hierarchy” was expanded into a composite standard of “graded headings + paragraph coherence + intra-paragraph logic,” with explicit penalty criteria such as “absence of a purpose paragraph” and “disordered sequence of key elements.”

4.4.2. *Points of divergence and resolution*

One expert recommended adding genres such as “plans” and “reports” to the portfolio. The research did not adopt this suggestion in order to avoid excessive burden on teachers and students, but plans to address it later by adopting a “core genres + optional genres” model.

One expert argued that “degree of chart/visual aid use” in the “summary” genre is non-essential for purely textual summaries and suggested reducing its weight. In response, the research renamed this dimension “clarity of information presentation,” retained its weight, but broadened the scope of evaluation to include textual expression.

5. Discussion

5.1. *Reliability characteristics of the assessment guideline*

This study found that the Chinese Practical Writing Portfolio Assessment Guideline demonstrated high inter-rater reliability (overall κ = 0.87), indicating that the dimensional design and descriptors of the guideline are highly clear and operationalizable. This finding is consistent with the conclusions of Lloyd et al. [10], who, in their study of short-answer tasks in large-scale statistics courses, enhanced inter-rater consistency (Fleiss' κ = 0.68) by refining the behavioral descriptors of their rating scale (e.g., a three-tiered operational definition of “step completeness”). Their results confirmed the widely accepted consensus that explicitly defined descriptors are the core prerequisite for reliability. A further comparison suggests that the higher reliability observed in this study (0.87 > 0.68) may be attributed to the guideline's concretized definitions of writing-specific dimensions such as format conventions and logical structure (e.g., “the header must include the issuing number and the authorizing officer”), which reduced the room for raters' subjective interpretation [42].

Regarding differences across modules, the “Implementation Objectives” and “Instructional Procedures” modules achieved perfect agreement (κ = 1.00), not only supporting Rezaei and Lovorn's [29] argument that “clarity of rating criteria determines reliability,” but also echoing the core finding of Klenowski et al. [43] that the reliability of portfolio assessment primarily depends on the explicitness of its purposes and processes. These modules focus on the how-to aspects (e.g., “submission of the revised draft in Week 6″), which constitute low-ambiguity tasks requiring minimal subjective inference from raters. By contrast, the κ values for the “Portfolio Content” and “Operational Norms” modules (κ = 0.79) were slightly lower, likely because these modules involve more context-dependent judgments (e.g., “representativeness of portfolio samples”). Such contextual dependency, as highlighted by Eckes [42], is a core source of rater divergence, which indicates that even after training, raters' interpretations of “contextual appropriateness” may still be shaped by their individual experiences.

5.2. Genre specificity and rating consistency

The rating consistency of official notices was extremely high (overall κ = 0.89, with both the language and format dimensions achieving κ = 1.00). This reflects the fact that the high degree of conventionalization within this genre reduces raters' judgmental difficulty. Such a result aligns with the effect of genre-specific evaluation criteria: research has indicated that compared with general dimensions, genre-specific rubrics (e.g., explicit textual structure and formatting) are more conducive to rater consensus [44].

In contrast, the structural dimension of summary writing yielded a relatively lower κ value (0.68), suggesting that flexible conventions (e. g., logical coherence) are more susceptible to raters' subjective interpretations. This finding echoes Weigle's classic assertion that rating reliability tends to decline in more open-ended genres lacking a unified paradigm [45]. Therefore, future descriptors for such genres should incorporate prototypical exemplars to provide raters with a shared frame of reference.

5.3. Implications for practical writing assessment

First, genre-adaptiveness of scoring rubrics is crucial. The findings of this study corroborate the claim of Lloyd et al. [10]: raters achieved a Fleiss' Kappa of 0.68 when employing detailed rubrics in short-answer tasks, with even higher consistency observed in aligned genres (e.g., texts with explicit formats). The high consistency in official notices illustrates that criteria grounded in genre features (e.g., "authoritativeness" and "format completeness") can effectively enhance reliability.

Second, rater training should be genre-specific. For instance, in the structural dimension of summary writing, training should include "prototypical paradigm comparisons" to unify raters' evaluative standards through case-based learning. This aligns with Rezaei & Lovorn's [29] conclusion that training combined with operationalized descriptors is a key pathway to improving reliability.

Finally, assessment tools need to allow for dynamic refinement. Lloyd et al. [10] demonstrated that the continuous revision and supplementation of scoring exemplars can further reduce rating discrepancies. The expert feedback in this study, which emphasized the need to "add positive and negative cases and refine ambiguous descriptors," is highly consistent with this perspective.

5.4. Limitations and future research

The limitations of this study include: (i) a relatively small sample of raters ($n = 5$), which may restrict generalizability; (ii) the examination of only two types of practical writing, excluding other common genres such as reports and contracts; and (iii) the absence of further validation of structural validity and discriminant capacity. Future research could: (i) incorporate a larger and more diverse group of raters to test consistency; (ii) expand the genre coverage to build a rubric repository for practical writing; and (iii) adopt mixed methods (e.g., cognitive interviews) to uncover the decision-making mechanisms underlying rating discrepancies.

6. Conclusion

Through Fleiss' Kappa analysis, this study confirms that the Portfolio Assessment Guideline for Chinese Practical Writing demonstrates excellent inter-rater reliability, with consistency across its six core modules and two genre-specific scoring criteria reaching or approaching the level of "excellent." This finding provides empirical support for the psychometric quality of the guideline and offers a practical framework for the standardization of Chinese practical writing assessment.

The results further indicate that the degree of rigidity in genre conventions significantly influences rating consistency, highlighting the need for a balanced approach between a "general framework" and "genre-specific indicators" in practical writing assessment. Revisions to the guideline based on expert feedback have enhanced its applicability, making it not only suitable for higher education but also a valuable reference for writing evaluation in professional training contexts.

Future research should validate the stability of the guideline with larger samples and a wider range of genres, and integrate evidence of validity to construct a more comprehensive assessment system. Ultimately, such efforts will advance the evaluation of Chinese practical writing from "experience-based judgment" toward "evidence-based assessment."

Funding

This paper has no funding support.

CRediT authorship contribution statement

Ying Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Ibnatul Jalilah Yusof:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

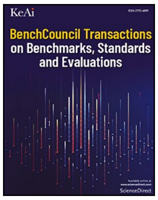
[1] R. Lam, Portfolio Assessment For the Teaching and Learning of Writing, Springer, Singapore, 2018.
[2] L. Hamp-Lyons, W. Con, Assessing the Portfolio Principles For practice, Theory and Research, Hampton Press, 2000.
[3] J.S. Barrot, Effects of Facebook-based e-portfolio on ESL learners ' writing performance, Lang. Cult. Curric. 34 (1) (2021) 95–111.
[4] P. Zhang, G. Tur, A systematic review of e-portfolio use during the pandemic: inspiration for post-COVID-19 practices, Open. Prax. 16 (3) (2024) 429–444.
[5] H.T. Huang, Master 's thesis, Nanning Normal University, 2023, https://doi.org/10.27037/d.cnki.ggxsc.2023.000911.
[6] Y.Y. Wang, Practical Writing Skills and Standards, People 's Posts and Telecommunications Press, 2022.
[7] Räz, T. (2023). Inter-rater reliability is individual fairness. arXiv preprint arXiv: 2308.05458. https://doi.org/10.48550/arXiv.2308.05458.
[8] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull 76 (5) (1971) 378.
[9] A. Zapf, S. Castell, L. Morawietz, A. Karch, Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC. Med. Res. Methodol 16 (93) (2016) 1–10, https://doi.org/10.1186/s12874-016-0200-9.
[10] S. Lloyd, M. Beckman, D. Pearl, R. Passonneau, Z. Li, Z. Wang, Foundations for NLP-assisted formative assessment feedback for short-answer tasks in large-enrollment classes, arXiv preprint arXiv:2205.02829, https://arxiv.org/abs/2205.02829, 2022.
[11] V. Klenowski, Developing Portfolios For Learning and assessment: Processes and Principles, Routledge, 2002.
[12] T. Burner, The potential formative benefits of portfolio assessment in second and foreign language writing contexts: a review of the literature, Stud. Educ. Eval. 43 (2014) 139–149.
[13] I. Listiana, F.N. Yusuf, S.M. Isman, Portfolio assessment: benefits for students At different writing proficiency level, J. Pendidik. Bhs. Dan. Sastra 20 (2) (2021) 243–256.
[14] W. Ngui, V. Pang, W. Hiew, K.W. Lee, Exploring the impact of e- portfolio on ESL students'writing skills through the lenses of Malaysian undergraduates, Comput.-Assist. Lang. Learn. Electron. J. 21 (3) (2020) 105–121.
[15] N. Pourdana, K. Tavassoli, Differential impacts of e-portfolio assessment on language learners ' engagement modes and genre-based writing improvement, Lang. Test. Asia 12 (1) (2022) 7.
[16] B.O.S. Al-Hawamdeh, N. Hussen, N.S.G. Abdelrasheed, Portfolio vs. summative assessment: impacts on EFL learners ' writing complexity, accuracy, and fluency (CAF); self-efficacy; learning anxiety; and autonomy, Lang. Test. Asia 13 (1) (2023) 12.
[17] J. Yang, Master 's thesis, Lanzhou University, 2020, https://doi.org/10.27204/d.cnki.glzhu.2020.003088.
[18] Y. Yu, Establishing a "composition growth portfolio": an attempt to effectively improve middle school students ' writing levels, Chin. Lang. Teach. Middle. Sch. (12) (2007) 39–41. CNKI:SUN:ZYJX.0.2007-12-017.

[19] Y.M. Wei, Master 's thesis, Northeast Normal University, 2012.
[20] Y. Liu, Master's thesis, Shandong Normal University, 2018.
[21] J.F. Mo, The application of'portfolio evaluation. Chinese Language Teaching, Henan Education (Basic Education Edition), 2021, pp. 65–66.
[22] X.H. Wang, Master 's thesis, Central China Normal University, 2019.
[23] W. Grabe, R. Kaplan, Theory and Practice of Writing, 1996. London and New York.
[24] S.Q. Lu, Master 's thesis, Shaanxi Normal University, 2015.
[25] X.X. Zuo, Master 's thesis, Luoyang Normal University, 2023.
[26] P.B. Diederich, J.W. French, S.T. Carlton, Factors in judgments of writing ability, ETS. Res. Bull. Ser. 1961 (2) (1961) i–93.
[27] E.A. Shabani, J. Panahi, Examining consistency among different rubrics for assessing writing, Lang. Test. Asia 10 (1) (2020) 12.
[28] G.T.L. Brown, H.L. Andrade, F. Chen, Accuracy in student self-assessment: directions and cautions for research, Assess. Educ.:. Princ. Policy. Pract. 22 (4) (2014) 444–457, https://doi.org/10.1080/0969594X.2014.996523.
[29] A.R. Rezaei, M. Lovorn, Reliability and validity of rubrics for assessment through writing, Assess. Writ. 15 (1) (2010) 18–39.
[30] T.F. McNamara, Language Testing, Oxford University Press, 2000.
[31] N. Gisev, J.S. Bell, T.F. Chen, Interrater agreement and interrater reliability: key concepts, approaches, and applications, Res. Soc. Adm. Pharm. 9 (3) (2013) 330–338.
[32] J.Y. Lee, Doctoral dissertation, Pennsylvania State University, 2022.
[33] J.L. Fleiss, Statistical Methods For Rates Andproportions, 2ndEd, John Wiley, New York, 1981, pp. 38–46.
[34] E. Almanasreh, R. Moles, T.F. Chen, Evaluation of methods used for estimating content validity, Res. Soc. Adm. Pharm. 15 (2) (2019) 214–221.
[35] S.J. Osterlind, What is Constructing Test items? Springer, Netherlands, 1998, pp. 1–16.
[36] M.R. Lynn, Determination and quantification of content validity, Nurs. Res 35 (6) (1986) 382–386.
[37] D.F. Polit, C.T. Beck, The content validity index: are you sure you know what's being reported? Critique and recommendations, Res. Nurs. Health 29 (5) (2006) 489–497.
[38] C. Welch, Item and prompt development in performance testing, Handb. Test. Dev. (2006) 303–327.
[39] S.S. Moya, J.M. O'malley, A portfolio assessment model for ESL, J. Educ. Issues. Lang. Minor. Stud. 13 (1) (1994) 13–36.
[40] R. Lam, Promoting self-regulated learning through portfolio assessment: testimony and recommendations, Assess. Eval. High. Educ. 39 (6) (2014) 699–714.
[41] J.S. Delett, S. Barnhardt, J.A. Kevorkian, A framework for portfolio assessment in the foreign language classroom, Foreign. Lang. Ann. 34 (6) (2001) 559–568.
[42] T. Eckes, Operational rater types in writing assessment: linking rater cognition to rater behavior, Lang. Assess. Q. 9 (3) (2012) 270–292.
[43] V. Klenowski, S. Askew, E. Carnell, Portfolios for learning, assessment and professional development in higher education, Assess. Eval. High. Educ. 31 (3) (2006) 267–286.
[44] Z.A. Philippakos, C.A. MacArthur, The use of genre-specific evaluation criteria for revision, Lang. Lit. Spectr. 26 (2016) 41–52.
[45] S.C. Weigle, Assessing Writing, Cambridge University Press, 2002.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Full Length Article

Evaluating barriers to establish digital trust in industry 4.0 for supply chain resilience in the Indian manufacturing industry

Vaibhav Sharma [a], Rajeev Agrawal [a,*], Anbesh Jamwal [b], Vijaya Kumar Manupati [c], Vikas Kumar [d]

[a] *Department of Mechanical Engineering, Malaviya National Institute of Technology Jaipur Rajasthan 302017, India*
[b] *Operations and Decision Sciences, Jaipuria Institute of Management, Jaipur, India*
[c] *Operations & Supply Chain Management, Indian Institute of Management Mumbai, Mumbai, India*
[d] *University of Portsmouth, Winston Churchill Avenue, Portsmouth, Hampshire PO1 2UP, United Kingdom*

ARTICLE INFO

Keywords:
Digital trust
Supply chain 4.0
Resiliency
Information-sharing
Digital transformation

ABSTRACT

Recent developments in Industry 4.0 technologies have led the manufacturing industry to implement them in its supply chains. The current state of lack of trust in digital systems has made organizations eager to build resilient systems to cope with uncertain circumstances. However, the challenges with handling stakeholder data with transparency, visibility, and accountability still persist. This transition demands the establishment of digital trust for secure information sharing and mitigating risks related to cybersecurity, data privacy, and potential misuse. Through a systematic literature review, this study identifies 17 barriers to establishing digital trust and applies exploratory factor analysis to group them into key dimensions. Further, a case-based analysis in the emerging Indian manufacturing economy's context employing Pythagorean Fuzzy Analytic Hierarchy Process-Decision-Making Trial and Evaluation Laboratory is conducted to prioritize these barriers and explore their interrelationships. The findings reveal that 'Top management commitment' and 'Cybersecurity' are the most influential barriers to be taken care of to promote collaboration and responsiveness in a digitally enabled supply chain environment. The study contributes by guiding practitioners and researchers working on the digital transformations for supply chains, highlighting digital trust as a foundational capability for achieving resiliency in Supply Chain 4.0. Being less explored in the field of supply chain digitalization, this study is a first step forward to explore digital trust in the Supply Chain 4.0 for resilience.

1. Introduction

Rising supply chain (SC) disruptions from disasters, instability, technological advancements, and shifting customer expectations exude the need for a resilient SC to cope with uncertainties, prevent delivery-supply delays, unmet demand, revenue loss, and business goodwill [1,2]. In the current era of digital transformation, resiliency can be built through digital trust (DT) in Industry 4.0 (I4.0), with a specific focus on transparency, legitimacy, and effectiveness for digital enterprise[1] The integration of I4.0 technologies, such as Artificial Intelligence (AI), Internet of Things (IoT), Big Data analytics, and blockchain (BC), has an impact on the resilience of different SCs [3–7]. These technologies enhance decision-making, responsiveness, and agility across the SC [8,9]. SC integrated with I4.0 can enhance flexibility through real-time asset tracking, enabling inventory, transportation, and distribution management [2]. Even leading consulting firms suggest that both data analytics and DT are important in business1.

It is well discussed in the literature that I4.0 potential is non-differentiable to transform and reshape SC practices, organizations, and their individual [10,11], which puts pressure on organizations to shift from traditional SC to Supply Chain 4.0 (SC4.0) [12]. Traditional SC refers to the flow of physical items and information through physical distribution channels, which consist of suppliers, warehouses, manufacturers, distributors, and customers [13]. However, the traditional SC

Peer review under the responsibility of The International Open Benchmark Council.

* Corresponding author.

E-mail addresses: vaib.me81@gmail.com (V. Sharma), ragrawal.mech@mnit.ac.in (R. Agrawal), anbesh.jamwal@jaipuria.ac.in (A. Jamwal), manupativijay@iimmumbai.ac.in (V.K. Manupati), vikas.kumar@port.ac.uk (V. Kumar).

[1] https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf

https://doi.org/10.1016/j.tbench.2025.100247

Available online 3 December 2025

lacks real-time information flow, which lacks visibility, transparency, and integration, leading to inefficiencies in coordination, duty delays, and collaborative decision-making [9]. To address modern business challenges, the adoption of I4.0 technologies [14], leading transition toward SC4.0 or smart SCs [12]. In SC4.0 operations, there is a need to build a more efficient, flexible, and resilient system as it works on real-time data exchanges and process automation. This results in improved forecasting, reduced stockouts, and overproduction, and enables stronger collaboration and coordination among SC partners, with increased intelligent decision-making capabilities [15–17]. With the availability of these advantages, SC4.0 is still confronted with challenges of data integrity, cybersecurity, and trust in the I4.0 technologies' operations [18,19]. This underscores the need to explore DT in SC4.0.

In the cxtant literature, DT has been discussed as the confidence of stakeholders in an organization's ethical practices for the security, reliability in handling their sensitive and market-competitive information [20]. This lack of confidence can create a sense of hesitation among participants, particularly for organizations that operate or want to scale advanced digital technologies. This is evident in the report by James [21] that discusses eight recent cyberattacks on manufacturing, including data breaches at Volswagon's (April 2024), Nexperia's, Duvel Moortgat's, and Hoya Corporation's (March 2024), highlighting the vulnerability of interconnected systems. These incidents disrupt the end-to-end digital operations. Also, damaging the stakeholders' confidence and calling for an urgency to explore trust in digital transformation, making cybersecurity an SC risk, not a mere IT concern [18]. Further, in a global survey by PricewaterhouseCoopers [22] it is revealed that 58 % manufacturers anticipate software-level incidents in their devices, and 63 % of them foresaw a rise in third-party related cyber-threats. This necessitates that organizations revise policies and contracts with high-risk vendors and extend cybersecurity support through the DT framework and governance under unified compliance.

In an emerging economy like India, where government-led initiatives like 'Digital India' and 'Make in India' are creating a sense of motivation and pressure on firms to transform their traditional SC operations to SC4.0 operations to remain competitive [23,24]. However, many firms, being in their nascent stages, struggle to cope with digital transformation [25]. Although organizations may be aware of the cybersecurity risks and the importance of being resilient in the current uncertain environment, the aspect of DT is less explored [26]. This motivates us to explore the barriers to DT as a prerequisite for a resilient SC4.0

SC4.0 takes advantage of I4.0 technologies to act in uncertain and cyber-attack vulnerable environments [4]. This can instigate organizations to achieve resiliency in SC4.0 operations through transparency in real-time information exchange, which can enable swift recovery [6,7]. For instance, I4.0 technologies such as digital twin can capture dynamic demand patterns through real-time information sharing to forecast and predict inventory levels through analytics, thereby improving flexibility and resilience [3,27]. As visibility in SC 4.0 relies on information sharing about the product, location, and identity of upstream and downstream suppliers [5]. AI improves predictive maintenance, and BC ensures transaction transparency [28], and cloud platforms support seamless data sharing [29]. The confidence of SC stakeholders depends on the security of these platforms and the asset tracking mechanisms [30]. The extant literature has depicted the potential of I4.0 technologies in SC to build a resilient system. However, the effective security of information shared through these technologies relies on the building of DT among stakeholders in the digital ecosystem utilized by the organizations. This fear lies underneath due to risks of data breaches, identity theft, and unauthorized access to confidential business data [31]. Because integrity, security, and reliability of digital systems are the central column to build DT [32,33].

Despite the growing concerns over the security of data, DT has remained less explored, particularly in the fastly emerging economies like India. Most studies are focused on covering the technological and implementation aspects of SC4.0, subtly discussing DT in manufacturing SC of an emerging economy. For instance, Strazzullo [20] has explored DT as an internal phenomenon within manufacturing companies, highlighting the roles of both the individual and the organization. This study has given a specific focus on examining SC4.0 through the lens of DT for resiliency through the empirical validation of barriers to DT. Moreover, integrated multi-method approaches to capture priority and interdependence of barriers through industry perspectives in the horizontal value chain are still less explored. Therefore, this study aims to fill this gap by exploring, validating, and analyzing the DT barriers for resilient SC4.0 and pursues the following objectives:

- To identify and empirically validate the DT barriers.
- To prioritize and determine the causal relationship among the selected barriers.

To fulfill these objectives, the barriers were identified from literature and validated via an industry survey followed by Exploratory Factor Analysis (EFA) to uncover underlying dimensions. Further, a case study in an emerging economy's manufacturing organization used expert inputs analyzed using Pythagorean Fuzzy (PF) sets integrated with the Analytic Hierarchy Process (AHP) for prioritization, and the Decision-Making Trial and Evaluation Laboratory (DEMATEL) technique was applied to determine causal relationships among the barriers.

The study is composed of six sections. Section 2 presents a discussion of the background literature. Section 3 discusses the Methodology and its application. Section 4 presents discussions. Section 5 provides the implications. Section 6 discusses the conclusion and future research with practical limitations.

2. Background literature

This section develops the context and theoretical background of the study through the available literature on SC 4.0, DT, and their relationship to resiliency.

2.1. Supply chain 4.0 for resiliency

SC4.0 synchronizes operations with suppliers and customers [34] through collaborative actions and real-time information sharing [35], for the common perceived benefits such as inventory optimization, reduced delivery lead times, and enhanced SC agility and responsiveness [36,37]. In the horizontal value chain, integrating digital technologies enhances decision-making, risk management, visibility, transparency, and accountability throughout the SC [8,38,39]. Recently, the above issues have been highlighted by Patil, Srivastava [40] on digital twins for SC transparency, and suggested that DT can increase sustainable organizational performance. These digital transformations allow SC to reconfigure sustainability practices at the structural, process, and plant levels [41]. I4.0 technologies can provide greater flexibility and end-to-end visibility with a stakeholder-focused objective through real-time data exchanges as an organization's commitment to enhance its SC capabilities [42]. These commitments can be in terms of taking actions to build a robust cyber-secured infrastructure, data integrity measures for improving SC performance [43]. These measures are crucial to motivating organizations to establish a digitally trusted ecosystem for all SC stakeholders, fostering innovation and flexibility in the product value creation process [10]. In this regard, a cloud-based platform facilitates collaboration across the SC by enabling the sharing of data and information [44]. Therefore, I4.0 enables data-driven decision-making through collaboration, promoting transparency and addressing issues of disruptions, such as cyberattacks [10,38].

2.2. Theoretical background

Trust is regarded as a foundational concept in relational exchanges, building cooperation and reducing uncertainty in both traditional and digital environments. According to the Trust Theory by McAllister [45], trust is conceptualised as one party's confidence in the reliability and integrity of another party. Process-based trust, in this context, reflects collecting detailed information to enhance interpersonal trust, whether between individuals or organizations. Rather than forming instantaneously, such trust evolves progressively across multiple interactions, wherein stakeholders actively evaluate available evidence to assess the credibility and dependability of their counterparts [46]. In the specific focus of SC, whereby the trustor's trust in one target transfers to another associated target, which implies that trust in SC practices corresponds to institutional trust, and its consequences can be transferred to particular products (interpersonal trust), thereby instilling the catalyst effect in operating on digital technologies [47]. Long established in the social sciences, Trust theory frames trust as a multi-dimensional construct encompassing cognitive, emotional, and moral components, and anchored perceptions of vulnerability, competence, and reliability between parties [48,49]. The traditional frameworks differentiate competence-based, integrity-based, relational, institutional, and system-based trust, which incubate cooperation and mitigate opportunism in complex organizational environments [50]. System-based trust corresponds to the confidence derived from institutional (organizational) management and arrangements to put regulations into action to reduce uncertainty in exchanges. The rational evaluations of competence and reliability based on available evidence put forth the cognition-based trust. Affect-based trust precludes emotional perceptions of the product and its information which correspond to the goodwill that goes beyond purely rational assessments. Institutional trust provides a broader spectrum consisting of societal norms, laws, and certifications that signal the legitimacy and compliance with international standards and frameworks [51]. Extending this view Lin and Lin [52] has utilised Commitment–Trust Theory in the purview of cloud SC adoption and outlined that a trusted relationship exists not only between people but also between people and computing systems, encompassing persistent trust in infrastructures, dynamic trust in specific situations, and persistent social-based trust that bridges social and technological confidence. Morgan and Hunt [53] extended this view through Commitment–Trust Theory (CTT) and reinforced that trust, coupled with commitment, is central to sustaining long-term relationships, a principle increasingly critical in digital ecosystems. In the SC4.0 environment, DT is built on the system's security, reliability, and stability, as well as the provider's credibility. These foundations serve as catalysts for commitment, which in turn strengthens long-term relational trust. This corresponds to the delicate nature of trust, which, although dynamic and evolving in context, interactions, and perceived risks, can deteriorate rapidly when breaches occur [54].

Extending the notion of trust to digital ecosystems, as depicted in Fig. 1, the trust in I4.0 is referred to as DT [55] can be understood as the willingness of stakeholders to rely on I4.0 technologies and platforms [56] under conditions of uncertainty, where confidence in data integrity, safety, system reliability, transparency, and institutional safeguards substitutes for traditional professional assurances. While DT is grounded in the same theoretical foundations as conventional trust, it is operationalized through mechanisms such as cybersecurity, regulatory compliance, and governance frameworks that ensure safety resiliency, transparency, interoperability, and ethical alignment in technologically mediated relationships [51]. System-based Trust in the digital ecosystem is expressed through cybersecurity protocols, BC verification, and governance structures that provide the robust structural assurance needed for secure information exchange and transaction integrity in SC4.0. Cognition-based Trust corresponds to confidence built on algorithmic transparency, data accuracy, and demonstrable technological competence of AI, IoT, and analytics systems for informed SC4.0 decision-making. Affect-based trust reflects user perception of fairness, ethical alignment, and confidence in digital platforms and automated systems. Institutional Trust in I4.0 context is anchored in adherence to international standards, regulatory frameworks, and certifications (e.g., ISO (International Organization for Standardization), GDPR (General Data Protection Regulation), NIST (National Institute of Standards and Technology)), which reassure compliance, accountability, and ethical practices across SC4.0.

2.3. Digital trust

DT has been the topic of discussion since the 90 s. However, there has

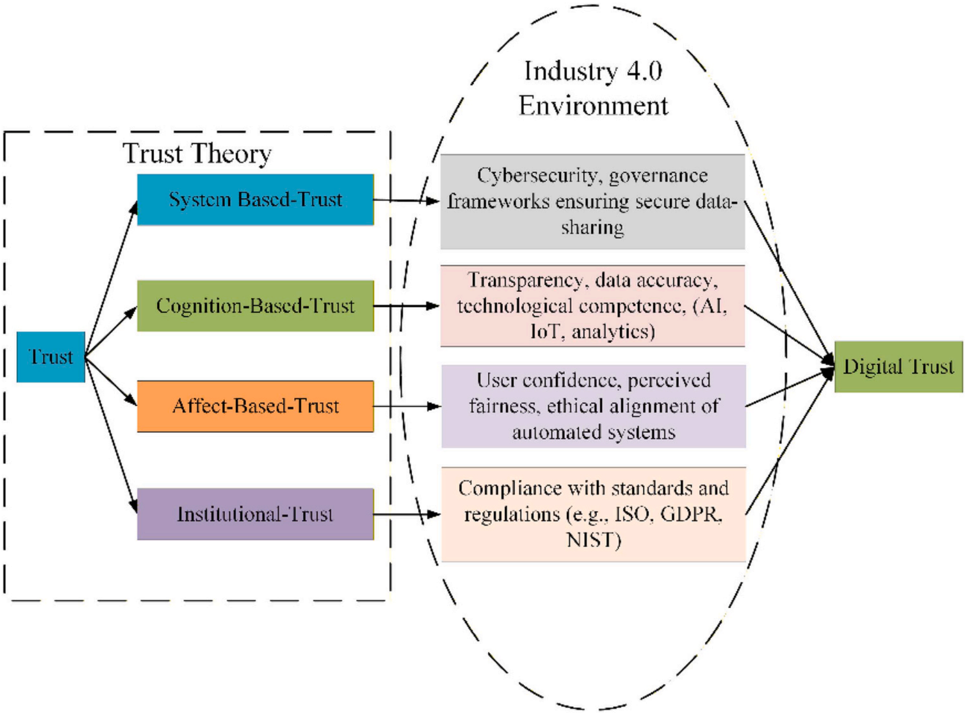

Industry 4.0
Environment
Trust Theory
System Based-Trust
Cognition-Based-Trust
Affect-Based-Trust
Institutional-Trust
Trust
Cybersecurity, governance frameworks ensuring secure data-sharing
Transparency, data accuracy, technological competence, (AI, IoT, analytics)
User confidence, perceived fairness, ethical alignment of automated systems
Compliance with standards and regulations (e.g., ISO, GDPR, NIST)
Digital Trust

Fig. 1. Trust and Digital Trust.

been substantive discussion since 2016. It is characterized by a trustor, a trustee, trust in online merchants, vulnerability to trust violations by organizations, characteristics of an individual, and organizational responsibility [33].

DT is well-defined by Pietrzak and Takala [33] as:

> *"Digital trust is the measure of confidence that workers, consumers/buyers, partners, and other stakeholders have in an organization's ability to protect data and the privacy of individuals."*

According to the World Economic Forum [57]

> *"Digital trust is individuals' expectation that digital technologies and services – and the organizations providing them – will protect all stakeholders' interests and uphold societal expectations and values."*

Within the purview of the above definition, the literature also explains DT, which encompasses security, identifiability, traceability, and reliability [33]. DT is the stakeholders' confidence in maintaining the integrity of relationships, interactions, and transactions among the participants of the digital ecosystem [58]. The security and legitimacy of connected devices in the I4.0 environment are the pertinent requirements for DT [59]. So that the stakeholders can confidently engage in online transactions. At the same time, their data and information are secured [60]. A survey within the emerging economy's business landscape by Sivarama Krishnan [61] highlights that leaders' major focus is on cybersecurity. With these cybersecurity threats, an attack could result in a loss of customer data. User satisfaction with using digital technologies indirectly influences DT based on its perception [62].

2.4. *Digital trust and supply chain 4.0*

DT in SC4.0 lies much in altering the exchange and processing of information and fostering formal or informal business relations [55,63]. I4.0 technologies can help with transparency, better decision-making, asset utilization, and lower SC risks with reduced warehouse, transportation, and inventory costs [12]. The coupling of I4.0 technologies in the SCs opens up ways for information sharing as a key aspect among SC stakeholders, which is based on a trust mechanism [64]. This enables SC4.0 to access the stakeholders' real-time information sharing with full data transparency across the multi-tier as well [65]. However, the data protection of stakeholders in an SC4.0 environment depends on data integrity, privacy, and compliance management [66]. Furthermore, the issues of security, transparency, privacy, integrity, and reliability can be addressed by a robust digital ecosystem, enabling an efficient, resilient, and stakeholder-centric SC [67]. This generates an interest in exploring DT in SC4.0 for resiliency.

2.4.1. *Digital trust and supply chain resiliency*

The current digital environment creates a lacuna of trust among SC stakeholders, making them resistant to collaboration and agreement on specific data sharing, such as inventory and demand forecast policies [68]. For effective SC operations with I4.0, information sharing among partners is required, where trust between suppliers, distributors, and logistics providers is essential [69]. SC resilience corresponds to identifying flexibility, collaboration, agility, and trust within the network, building capabilities and enhanced risk management with reduced disruptions for trustworthy collaborations. I4.0 technology, such as the BC, enhances SC resilience by increasing SC trust with increased operational efficiency, cybersecurity, and order fulfilment with secured information sharing resistant to cyberattacks, enabling trusted transactions with reduced risk [14,70]. The more resilient the SC, the more trust in the organization handling data through BC [71]. However, the role of trust in addressing resiliency issues is not only related to the BC, but also trust in I4.0, as the umbrella is still to be addressed for SC resiliency. This study aims to fill this gap by identifying barriers to DT in SC4.0 for resiliency.

2.5. *Research gaps*

The extant literature within the context of the I4.0 environment for SC requires further exploration from the perspective of DT for resilient SC4.0. DT foundational requirements include security, identifiability, traceability, accountability, and fairness, which are well-articulated by [33,59,60]. Also, digital transformations have been linked to the sustainability of SC4.0 [41]. Lastly, stakeholder-centric initiatives by organizations with limited insights, aligning with DT, have been presented, where diverse stakeholders' expectations are critical to addressing cybersecurity challenges [58–60,62]. Despite studies highlighting challenges in adopting I4.0 technologies [12,34]. Similarly, the trade-offs between real-time information transparency and privacy concern the stakeholders' trust in the digital SC [10,17,65]. The role of digital systems and collaborative platforms in enhancing DT among the stakeholders is under-researched, with limited empirical investigations [10,44]. However, there is also a lack of exploration of a comprehensive assessment of the barriers to DT for SC 4.0 in the dynamic and emerging economies. Furthermore, the specific interplay of barriers to DT for resilient SC 4.0 remains less explored. Addressing these gaps is essential to advancing theoretical and practical knowledge of DT in SC 4.0, enabling robust, secure, and resilient SC 4.0.

3. Methodology

The strategic approach to analyzing the barriers to DT in SC4.0 in the emerging economy context is outlined through the methodology presented in Fig. 2. This approach extends the foundations led by previous studies (Table 1), which have demonstrated the effectiveness of mixed-method analysis designs in capturing complex I4.0 and SC problems. For instance Yadav and Singh [72] employed a three-phased methodology incorporating a literature review to identify 39 variables, Principal Component Analysis for reducing them into 12 factors, and fuzzy DEMATEL to examine cause-effect relationships. Similarly, Shayganmehr, Gupta [73] have utilized a two-module hybrid methodology combining EFA to categorize 29 critical success factors into five smaller constructs with a Hierarchical Fuzzy Expert System to assess the readiness of "swift trust" and "coordination" and to prescribe the most appropriate I4.0 tools through a case study. Other studies have utilised PF-AHP-DEMATEL for evaluating barriers to circular SC implementation [74], and EFA with AHP to evaluate challenges associated with I4.0 initiatives in the context of sustainable SC in emerging economies [75]. These applications reinforce the robustness of EFA, fuzzy logic, and case-based techniques in identifying and analyzing barriers. Guided by the literature, sixteen barriers were identified from the available literature and the consultation of area experts (demographics in Table 2) with rich experience in the fields of manufacturing, sustainability, I4.0, and SC management and their interrelations. Then, the industry experts were contacted through email, followed by discussions to understand the issues of DT in SC4.0. The identified barriers were shared with the experts, and mutual sharing of thoughts and experiences took place over the telephone and in person, resulting in the experts adding 'User Experience and Usability for Technology' as one more barrier and finalizing a total of seventeen barriers, as shown in Table 3 for empirical investigations.

3.1. *Barriers to digital trust in supply chain 4.0 for resiliency*

Through a three-step literature review, PRISMA is shown in Fig. 3. First, a search in the Scopus and Web of Science databases with keywords related to 'Supply Chain resiliency', 'Supply Chain', 'Trust', and 'Industry 4.0' was conducted. The articles were funneled down based on duplicity, research context, and studies focusing on the SC4.0 scenario. The study identified articles relevant to collecting key barriers to DT in SC4.0 for resiliency, resulting in sixteen barriers. Finally, the expert panel examined these barriers for confirmation, adding one more,

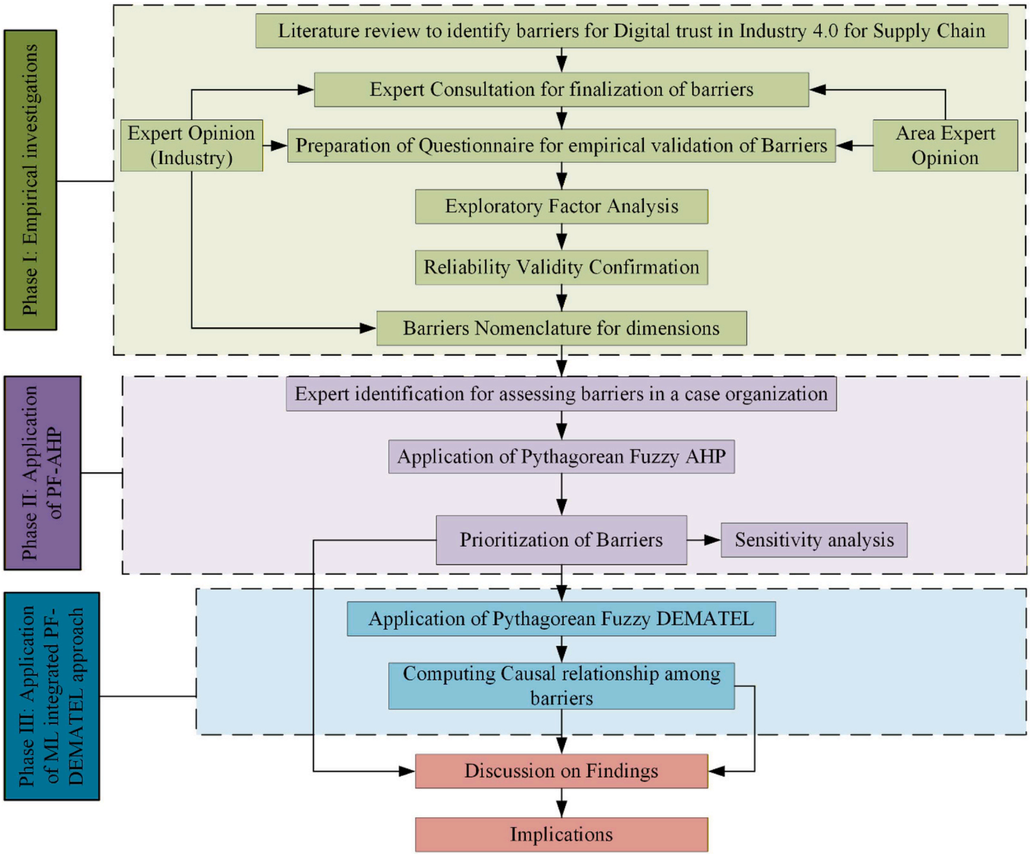

Phase I: Empirical investigations
Literature review to identify barriers for Digital trust in Industry 4.0 for Supply Chain
Expert Consultation for finalization of barriers
Expert Opinion (Industry)
Preparation of Questionnaire for empirical validation of Barriers
Area Expert Opinion
Exploratory Factor Analysis
Reliability Validity Confirmation
Barriers Nomenclature for dimensions
Phase II: Application of PF-AHP
Expert identification for assessing barriers in a case organization
Application of Pythagorean Fuzzy AHP
Prioritization of Barriers
Sensitivity analysis
Phase III: Application of ML integrated PF-DEMATEL approach
Application of Pythagorean Fuzzy DEMATEL
Computing Causal relationship among barriers
Discussion on Findings
Implications

Fig. 2. Research Methodology.

resulting in a total of seventeen barriers, as shown in Table 3.

SC4.0 brings significant benefits of advanced digital technologies to enhance operational efficiency, responsiveness, and decision-making [16]. It enables real-time monitoring and enhances agility by making SCs more adaptable to disruptions caused by external factors such as global crises. It provides better visibility and traceability across different stages, leading to improved demand forecasting and reduced likelihood of stockouts or overproduction [9,44]. Despite the clear benefits, the operation of SC4.0 depends on information sharing, which is hindered by significant barriers to DT to achieve its full potential. One barrier is the perceived risk of cyberattacks and data breaches, which has grown with the increasing reliance on interconnected systems [76]. Additionally, the lack of standardized data security and privacy protocols across global SCs exacerbates these concerns, making it difficult for stakeholders to establish mutual trust [12,16]. The complexity of integrating disparate digital systems from various vendors introduces compatibility issues [77]. Trust in digital platforms is also undermined by the opacity of decision-making processes, where stakeholders may find it difficult to verify and validate algorithmic decisions, raising concerns about fairness and accountability [8,78,79]. Organizational and cultural resistance to change, particularly in industries with low digital literacy, represents a non-technical barrier to DT, slowing the digital transformation to SC4.0 [35,75].

Previous studies have focussed on identifying barriers to the digitalization of SC [16,32,75,78,79] with some studies [35,80–82] specifically focusing on BC implementing barriers while Yadav and Kumar [83], Khan, Haleem [84] analyzed barriers related to BC implementation in the emerging economy's manufacturing SC. However, the discussion and analysis of DT in adopting I4.0 for the SC present a significant gap.

3.2. Significance testing of barriers

This study is part of a broader investigation comprising barriers and enablers of DT in I4.0 for SCs in the emerging economy's manufacturing sector to establish DT and provide suggestions. Conducted between February and June 2024, this study analyzes barriers faced by manufacturing organizations. Professionals from 971 manufacturing firms with experience in SC and digital technologies were contacted via email and LinkedIn, with a brief explanation of the study's objectives. Using a five-point Likert scale, as used in previous studies [75,90], respondents rated the significance of various barriers with a questionnaire presented in Appendix A in the supplementary material. Initially, only 50 responses were received; however, after weekly reminders, 171 valid responses were collected, with 12 discarded due to biases or incomplete data, resulting in a 19.14 % response rate. This response rate is acceptable compared to similar studies [75,91]. The respondents' demographics are shown in Table 4. The mean and descriptive statistics of the identified barriers are detailed in Table 5.

3.3. Exploratory factor analysis

Exploratory Factor Analysis is a statistical method used for data reduction and analysis [92]. The reliability and validity testing of the factor was conducted using statistical software, SPSS, to validate the barriers. As per Table 5, the 'Kaiser-Meyer-Olkin' (KMO) value obtained was 0.873, i.e., higher than the minimum recommended range, i.e., 0.6. Bartlett's sphericity test is also found significant ($p < 0.01$). This suggests that the barriers collected for exploratory factor analysis are relevant. Further, the eigenvalues of discontinuity for the barriers are greater than 1.0, factor loadings are greater than 0.5, which suggests the collected data has convergent validity, and Cronbach's alpha is >0.7 as per the available literature by Nunnally [93]. Table 5 shows the results of the EFA after performing the dimension reduction. Four key dimensional

Table 1
Comparison of related works.

Reference	Problem addressed	Methodology	Technique
Yadav and Singh [72]	Traditional SC issues transparency, traceability, and data security. Factors such as human errors, fraud, delays, and documentation inefficiencies affect sustainability. Blockchain adoption to make a sustainable SC. Framework for critical factors of blockchain success.	Expert survey. Reduction of 39 blockchain variables to key factors. Identify cause–and–effect relationships among 12 key factors.	Principal Component Analysis and Fuzzy DEMATEL.
Shayganmehr, Gupta [73]	Investigated the humanitarian SC struggling with poor coordination, low information quality, and a lack of swift trust. Disaster events like pandemics require collaboration and smooth information exchange for effective relief operations.	Reduction of 29 critical success factors into 5 meaningful constructs: Logistics, Learning, Transparency, Information Quality, and Infrastructure. 3 Iranian Case-study.	Principal Component Analysis with Varimax rotation, Hierarchical Fuzzy Expert System (fuzzification, inference engine, defuzzification).
Lahane and Kant [74]	Circular SC adoption by evaluating multiple operational, economic, technological, and policy barriers. in the emerging economy context (India).	Expert evaluation through linguistic terms. PF sets. Sensitivity analysis for model robustness.	PF-AHP and PF-DEMATEL.
Luthra and Mangla [75]	I4.0 adoption barriers in manufacturing firms for sustainable SC development.	Categorized 18 challenges into 4 major groups. Prioritization of challenges.	Systematic Literature Review (SLR), Exploratory Factor Analysis and Analytic Hierarchy Process.

Table 2
Demographics of area experts.

Expert	Experience (Years)	Area of Expertise
Expert 1	23	Experience in research in SC design and sustainability, leveraging I4.0 technologies to manage SC operations.
Expert 2	10	Experience in research in intelligent manufacturing, with specific interests in Cyber-Physical systems and sustainable, resilient SCs.
Expert 3	8	Experience in research in SC management with a specific focus on SC collaborations

components were extracted: Stakeholder's intent (B1), Organizational (B2), Technical (B3), and Regulatory (B4) based on the experts' input for the nomenclature, which covers around 70.339 percent of the total variance. The nomenclature is discussed below:

Table 3
Barriers to Digital Trust in Supply Chain 4.0 for resiliency.

S. No.	Barrier Name	Description	References
1	Cybersecurity risks	Risks such as breaches, disruptions, and unauthorized access to sensitive data and systems hamper cargo, plant operations, and product specifications, undermining stakeholder confidence and integrity.	[76,77,81, 83–85]
2	Understanding the importance of trust in collaborative action	The real-time trust-based collaboration through data sharing between suppliers and organizations ensures product quality, innovation, and consistency for flexible decision-making and long-term partnerships.	[8,28]
3	Data Ownership, Quality, and Value	Data governance is data quality and value, which depends on the transparency of data ownership, enabling reliable decision-making, operational efficiency, and transparency among stakeholders.	[64,72,86]
4	Risk of information security and privacy	The potential loss of confidential and sensitive SC partners' information and the misuse of the private information for any of the data holder's benefit.	[8,32,82, 87]
5	Lack in the implementation of interactive digital communications	The limited adoption of real-time, transparent tools for seamless SC stakeholder collaboration. reduces transparency, delaying decision-making, and increasing the risks of misinformation.	[35,82,88]
6	Lack of information asymmetry over the shared platforms	Accurate, timely, and high-quality information over shared platforms enhances SC integration and decision-making and fosters DT.	[29,77]
7	Lack of digital culture	The organization's openness to acceptance is based on trust in utilizing digital technologies among its employees, reflecting its beliefs and confidence in utilizing these technologies.	[8,35,75,78, 79]
8	Distributed digital identity of supply chain entities	Verified SC entities (farmer, customer, grocer, warehouse) promote a smooth, secure, transparent operation; a lacuna creates an information gap.	[86]
9	Unwillingness to share information among SC partners	Insecurity for information sharing of the trading partners in the SC due to competitive disadvantage, which provides a de-collaborated environment, resulting in a loss of DT.	[78,79]
10	Lack of real-time information sharing	Lack of confidence in security and privacy makes digital SC participants hesitant to share real-time information, causing congestion and supply delays.	[35,78,84]
11	Interoperability and scalability, and issues	Interoperability is the ability of an information system to connect and exchange information among different SC entities. Scalability is the	[35,78, 80–82,85, 89]

(continued on next page)

Table 3 (*continued*)

S. No.	Barrier Name	Description	References
		ability of the digital system to be expanded without undue loss of performance.	
12	High investment cost of digital technologies	Capital investment in digital infrastructure for digitizing the SC ecosystem with perceived benefits of technology adoption enhances DT, customer retention, and competitive differentiation.	[78,79,84]
13	Unclear organizational objectives	Organizations assessing suppliers with digital technologies often lack data management policies, leading to reduced employee performance and customer loss due to diminished DT.	[78,79]
14	Lack of top management support	Management leadership commitment and support to instill confidence among the SC stakeholders to understand the value of digital technologies for enhanced SC performance	[8,78,79]
15	Low understanding of Industry 4.0 implications	SC participants' low comprehension of the benefits of I4.0 technologies in the SC, which include transparency, accuracy, collaboration, and agility for improved decision-making.	[8,75,78]
16	Lack of digital infrastructure	The absence of a secure and robust digital infrastructure helps to build DT, creating a sense of negligence toward adopting and practising digital technologies by the SC stakeholders.	[8,78,79]
17	User Experience and Usability for Technology	It is characterized by complex interfaces, a lack of transparency, security, and integration challenges, which prevent the user from working on SC4.0.	Expert input

- **Stakeholders' intent (B1):** This component includes five barriers. The Total Variance Explained for this component is 21.925 %. This component consists of the barriers to implementing I4.0 for SC operations due to the stakeholder's perceived intent for enhancing resiliency through technology trust.
- **Organizational (B2):** This dimension includes six barriers with 17.589 % of explained variance. These barriers correspond to the organizational hurdles to implementing digital technologies for SC4.0 to address the disruption through a resilient system.
- **Technological (B3):** This component consists of four barriers, which account for 17.440 % of the total variance. These correspond to the technological factors preventing stakeholders from having DT in I4.0 technologies for the resilient SC.
- **Regulatory (B4):** The regulatory component consists of three factors related to the regulations the stakeholders follow to utilize the stakeholders' data safely and properly and build digitally resilient SC. It accounts for 13.385 % of the Total Variance Explained.

$$\left(\breve{X}_1,\ \ldots,\breve{X}_d\right)=\left\langle\left[\prod_{k=1}^{d}\mu_{k,L}^{w_k},\prod_{k=1}^{d}\mu_{k,U}^{w_k}\right],\left[\left(1-\prod_{k=1}^{d}\left(1-\nu_{k,L}^2\right)^{w_k}\right)^{0.5},\left(1-\prod_{k=1}^{d}\left(1-\nu_{k,U}^2\right)^{w_k}\right)^{0.5}\right]\right\rangle \tag{2}$$

The identified components of the validated barriers are then further utilized to determine the barriers' priorities and their causal interrelationship to develop the framework. The priorities have been evaluated by implementing the PF-AHP to understand the hurdles for DT in SC4.0 to achieve resiliency. The implementation of PF-AHP is discussed in the next sub-sections.

3.4. *Multicriteria analysis in the case organization*

The identified barriers were then tested in a case organization. The case organization is a pioneering Robotics manufacturing organization working to develop cutting-edge solutions for various sectors, including Fire-Fighting Robots, Defence Robots, Humanoid Robots, and Manhole Cleaning Robots. The company is expected to have a turnover of over Rs. 5 billion. It is also ranked among the 100 top competitors in India. The company aims to integrate I4.0 technologies into its SC. However, trust issues in using these technologies are still a concern due to the infrastructural, employee, supplier, and customer concerns over the security of their data on the digital platforms utilized by organizations. The identified barriers were consulted with a questionnaire in Appendix B in the supplementary material, with a panel of five experts with experience in managing manufacturing SCs with I4.0 technologies. The demographic of the Expert panel is shown in Table 6. The PF-AHP is applied to compute the weights of the validated barriers. The top ten of the seventeen barriers are further analyzed to get the causal relationship through PF-DEMATEL [74].

3.5. *Pythagorean fuzzy AHP*

AHP was developed by Prof. Thomas L. Saaty in 1980 [94]. However, the implementation of AHP can generate inconsistencies in decision-making. Different studies [95,96] have employed fuzzy sets. This study employs PF sets to eliminate vagueness and inconsistencies. The following steps implement PF-AHP:

Step 1: The initial linguistic pairwise comparison matrix (P_k) as shown in Eq. (1) is generated for each expert, comparing the criteria i over j where $i,j=1,2,\ldots,m$ and based on the linguistic scale in Table C.1.

$$P_k=\begin{bmatrix}p_{11}^1 & \cdots & p_{1m}^k\\ \vdots & \ddots & \vdots\\ p_{m1}^k & \cdots & p_{mm}^k\end{bmatrix} \tag{1}$$

The linguistic scale inputs are converted into PF numbers based on Table C.1. Therefore $p_{ij}^k=\left\langle\left(\mu_{ijL}^k,\mu_{ijU}^k\right),\left(\nu_{ijL}^k,\nu_{ijU}^k\right)\right\rangle$ represents the numerical transformation of linguistic inputs.

Step 2: All expert inputs are aggregated using Eq. (2).

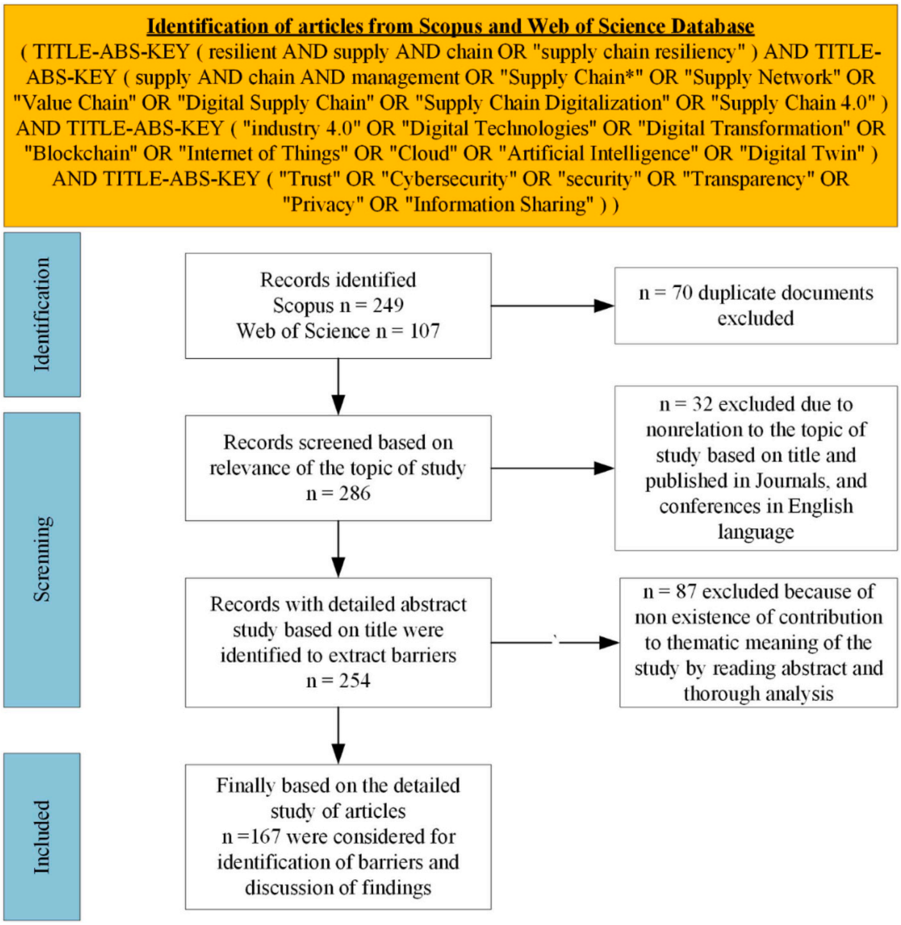

Identification of articles from Scopus and Web of Science Database
(TITLE-ABS-KEY (resilient AND supply AND chain OR "supply chain resiliency") AND TITLE-ABS-KEY (supply AND chain AND management OR "Supply Chain*" OR "Supply Network" OR "Value Chain" OR "Digital Supply Chain" OR "Supply Chain Digitalization" OR "Supply Chain 4.0") AND TITLE-ABS-KEY ("industry 4.0" OR "Digital Technologies" OR "Digital Transformation" OR "Blockchain" OR "Internet of Things" OR "Cloud" OR "Artificial Intelligence" OR "Digital Twin") AND TITLE-ABS-KEY ("Trust" OR "Cybersecurity" OR "security" OR "Transparency" OR "Privacy" OR "Information Sharing"))
Identification
Records identified Scopus n = 249 Web of Science n = 107
n = 70 duplicate documents excluded
Screening
Records screened based on relevance of the topic of study n = 286
n = 32 excluded due to nonrelation to the topic of study based on title and published in Journals, and conferences in English language
Records with detailed abstract study based on title were identified to extract barriers n = 254
n = 87 excluded because of non existence of contribution to thematic meaning of the study by reading abstract and thorough analysis
Included
Finally based on the detailed study of articles n =167 were considered for identification of barriers and discussion of findings

Fig. 3. Search Strategy.

Table 4
Demographics of responding organizations.

S. No.	Professional Demographics	Criteria	Number of respondents	Percentages
1	Type of Manufacturing Industry	Automobile	23	13.45
		Electrical & Electronics	21	12.28
		Healthcare device	34	19.88
		Consulting to manufacturing	65	38.01
		Others	28	16.37
2	Experience in digital technologies	4–10 years	43	25.15
		10–16 years	116	67.84
		16 & above	12	7.02
3	Experience in supply chain	5–10 years	102	59.65
		10–15 years	47	27.49
		15 & above	21	12.28
4	Designation	Top Management	10	5.85
		Senior Level Manager	37	21.64
		Middle-level manager	69	40.35
		IT professional	35	20.47
		Executive	20	11.70
5	Qualifications	Graduate	78	45.61
		Postgraduate	89	52.05
		Doctorate	4	2.34
		Total	171	100

Step 3: The difference matrix $\mathscr{D} = [d_{ij}]_{m\times m}$ where $d_{ij} = (d_{ijL}, d_{ijU})$ is calculated by finding out the difference between the lower (Eq. (2)) and upper values (Eq. (3)) of membership and non-membership functions by using the equations below.

$$d_{ijL} = \mu_{ijL}^2 - \nu_{ijL}^2 \tag{3}$$

$$d_{ijU} = \mu_{ijU}^2 - \nu_{ijU}^2 \tag{4}$$

Step 4: The interval multiplicative matrix $\mathscr{S} = [s_{ij}]_{m\times m}$ where $s_{ij} = (s_{ijL}, s_{ijU})$ is calculated using Eqs. (4) and (5).

$$s_{ijL} = \sqrt{1000^{d_{ijL}}} \tag{5}$$

$$s_{ijU} = \sqrt{1000^{d_{ijU}}} \tag{6}$$

Step 5: The indeterminacy matrix $\mathscr{T} = [\tau_{ij}]_{m\times m}$ is calculated by using Eq. (6) below.

$$\tau_{ij} = 1 - \left(\mu_{ijU}^2 - \mu_{ijL}^2\right) - \left(\nu_{ijU}^2 - \nu_{ijL}^2\right) \tag{7}$$

Step 6: The unnormalized weight matrix $U = [u_{ij}]_{m\times m}$is obtained with Eq. (7).

$$u_{ij} = \tau_{ij}\left(\frac{s_{ijL} + s_{ijU}}{2}\right) \tag{8}$$

Table 5
Exploratory Factor Analysis Results of Barriers.

Dimension	Barriers to digital trust in supply chain 4.0	Mean	Eigenvalues	Variance Explained (Cumulative)	Item Loading
Stakeholder's intent (B1)			6.741	21.925	
SB1	Unwillingness to share information digitally among SC partners	3.41			0.811
SB2	Lack in the implementation of interactive digital communications	3.12			0.838
SB3	Low understanding of Industry 4.0 implications	3.47			0.867
SB4	Understanding the importance of trust in collaborative action	3.59			0.861
SB5	User Experience and Usability for Technology	2.95			0.745
Organizational (B2)			2.357	39.514	
OB1	Unclear organizational objectives	3.69			0.744
0B2	Lack of top management support	3.60			0.682
OB3	Lack of digital culture	3.44			0.712
OB4	Lack of digital infrastructure	3.25			0.707
OB5	High investment cost of digital technologies	3.36			0.698
Technological (B3)			1.605	56.954	
TB1	Cybersecurity risks	3.12			0.849
TB2	Risk of information security and privacy	3.12			0.791
TB3	Data Ownership, Quality, and Value	3.36,			0.801
TB4	Interoperability and scalability and issues	3.26			0.822
Regulatory (B4)			1.254	70.339	
RB1	Lack of information asymmetry over the shared platforms	3.37			0.812
RB2	Lack of real-time information sharing	3.27			0.780
RB3	Distributed digital identity of SC entities	3.24			0.817

KMO: 0.865, Approx. Chi-Square: 1666.649, Cronbach's alpha = 0.897.
Barlett's test of sphericity: df: 136, Sig.0.000.
Extraction method: Principal Component Analysis, Rotation method: Varimax with Kaiser Normalization converged in 5 iterations.

Table 6
Demographics of the Expert panel.

Expert	Designation	Experience (Years)	Industry/Sector	Roles and Responsibilities
1	Chief Executive Officer	30	Manufacturing	Overall Management through AI for Quality Assurance in the SC.
2	Chief Digital Officer	25	Manufacturing	I4.0 Technologies Implementation in SC Operations
3	Head-Technology Risk and Cyber Controls	24	Consulting I4.0 technologies in SC	Handling non-financial risks like Cyber, Information security, Technology Risk, cloud security, Third Party Risk, and Operational Resilience
4	Assistant Manager	18	Manufacturing	Involved in managing the packaging, transportation, and dispatch of products.
5	Director Research	12	Robotics Manufacturing	Oversees and manages the production and distribution of the products to the customer.

Step 7: The weight of each criterion is determined by Eq. (8).

$$w_i = \frac{\sum_{i=1}^{m} w_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}} \tag{9}$$

3.5.1. Pythagorean Fuzzy-AHP implementation for weight computations

In this stage, a panel of five experts is constituted. The experts in the panel were asked to fill the pairwise comparison matrices for the main components and the barriers categorized under them per the linguistic scale in Table C.1 [95,96]. The initial pairwise matrix for the four dimensions by the five Experts is shown in Table C.2, provided in the Appendix C in the supplementary material, and their corresponding fuzzy numbers for Expert 1 are shown in Table C.3. The experts' inputs were combined in a single decision matrix, as shown in Table C.4. The sample calculation for the PF-AHP is shown in Tables C.5-C.7. The final global and local ranking of the DT dimensions and barriers with their weights is shown in Table 7.

3.5.2. Sensitivity analysis

The study has implemented a sensitivity analysis to test the robustness of ranking the barriers. Sensitivity analysis tests different configurations or combinations of weights that influence the prioritization of factors and criteria, and assesses the potential bias of the experts. We have addressed this by systematically changing the expert's weight by interchanging the weights and giving maximum weights to each expert by generating five scenarios (S1-S5), as shown in Table 8.

The weights of the barriers are obtained by interchanging the experts' weights while keeping the other main components constant. This procedure is followed with each of the five scenarios. According to Fig. 4, the weights of the barriers have not varied much from the original configuration, except for the barriers SB1, OB2, and TB1. This shows that the obtained weights are acceptable for further analysis.

3.6. Pythagorean fuzzy DEMATEL

The DEMATEL was first developed by Gabus and Fontela [97] and is a highly effective tool for identifying strengths and visualizing a causal relationship between the components in a complex system with minimum data input as compared to other MCDM techniques, such as interpretive structural modelling (ISM) and total interpretive structural modelling (TISM) [74,98,99]. Recently, studies such as [100] utilized PF-DEMATEL to identify the dependency of criteria for I4.0 sectoral prioritization. Similarly, Giri, Molla [101] utilized PF-DEMATEL for supplier selection in sustainable SC management, while Shafiee, Zare-Mehrjerdi [102] evaluated the perishable product SC risks during the COVID-19 outbreak. The steps of the PF-DEMATEL process are discussed below:

Step 1 Construction of Direct relationship matrix: The initial direct relationship matrix R_k as shown in Table C.9, is developed from k experts' inputs based on the scale in Table C.8.

Table 7
Ranking of barriers.

Main Criteria	Weight	Rank	Sub-Criteria	Local Weight	Local Rank	Global Weight	Global Rank
Stakeholder's intent (B1)	0.130	3	SB1	0.292	2	0.038	9
			SB2	0.344	1	0.045	8
			SB3	0.165	3	0.021	12
			SB4	0.133	4	0.017	13
			SB5	0.066	5	0.009	16
Organizational (B2)	0.398	1	OB1	0.140	3	0.056	7
			OB2	0.483	1	0.192	1
			OB3	0.054	4	0.022	11
			OB4	0.293	2	0.117	4
			OB5	0.029	5	0.012	15
Technological (B3)	0.371	2	TB1	0.384	2	0.143	3
			TB2	0.421	1	0.156	2
			TB3	0.155	3	0.058	6
			TB4	0.039	4	0.015	14
Regulatory (B4)	0.101	4	RB1	0.225	2	0.023	10
			RB2	0.719	1	0.072	5
			RB3	0.056	3	0.006	17

Table 8
Expert's weights scenario for sensitivity analysis.

Expert	Original	S1	S2	S3	S4	S5
E1	0.2	0.42	0.26	0.17	0.1	0.05
E2	0.2	0.05	0.42	0.26	0.17	0.1
E3	0.2	0.1	0.05	0.42	0.26	0.17
E4	0.2	0.17	0.1	0.05	0.42	0.26
E5	0.2	0.26	0.17	0.1	0.05	0.42

Step 2 Conversion of Initial direct relationship matrix: The initial direct relationship matrix R_k is converted into PF numbers with elements $r_{ij}^{k} = \left\langle \left(\mu_{ij}^{k}, \nu_{ij}^{k} \right) \right\rangle$ as shown in Table C.10.
Step 3 Computation of aggregated matrix: The experts' inputs are aggregated into a single aggregated matrix H with elements $h_{ij} = \left\langle \left(\mu_{ij}^{\mathrm{H}}, \nu_{ij}^{\mathrm{H}} \right) \right\rangle$.
Step 4 Computation of Average Crisp Matrix: The average crisp matrix M with elements $m_{ij} = \left(\mu_{ij}^{\mathrm{H}}\right)^2 - \left(\nu_{ij}^{\mathrm{H}}\right)^2$ is obtained.
Step 5 Calculate the normalized average crisp matrix: The normalized average crisp matrix $\mathscr{N}$ is calculated by Eq. (9).

$$\mathscr{N} = \wp \cdot M \tag{10}$$

where,

$$\wp = \frac{1}{\max \sum_{i=1}^{n} m_{ij}} \quad i, j = 1, 2, 3, \ldots, \ n$$

Step 6 Construct the total relationship matrix: The total relation matrix T is calculated by using Eq. (10).

$$T = \mathscr{N}(I - \mathscr{N})^{-1} \tag{11}$$

where I is an identity matrix.
Step 7 Plotting digraph: The digraph is plotted based on the threshold value. The threshold value is calculated to determine internal relations as the average of the Total relationship matrix, excluding partial relations within the matrix. A digraph is plotted for the values exceeding the threshold value only, setting zeros for the values below the threshold. The threshold found in this study was 0.04934. The adjusted total relationship matrix is shown in Table C.17.

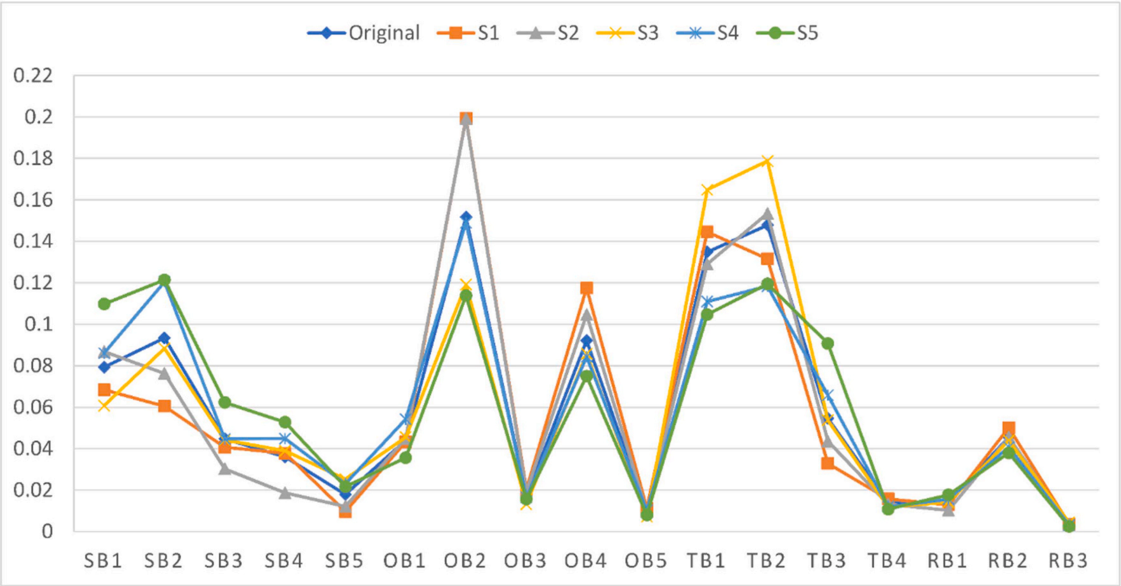


Fig. 4. Barrier weights after sensitivity analysis.

Step 8 Identify the causal relationship: The causal relationship is computed by finding the summation of rows. (R) and column (C) by using Eq. (11) and Eq. (12).

$$R_i = \sum_{j=1}^{n} t_{ij} \tag{12}$$

$$C_j = \sum_{i=1}^{n} t_{ij} \tag{13}$$

Step 9 Calculate the $(R+C)$ *and* $(R-C)$. $(R+C)$ denotes prominence effect and $(R-C)$ denotes the effect strength. The causal relation diagram is drawn by plotting $(R+C)$ on the horizontal axis and $(R-C)$ on the vertical axis. The above matrices can be found in Appendix C in the supplementary material from Table C.11-C.17.

3.6.1. *Application of Pythagorean fuzzy-DEMATEL*

The application of PF-AHP gives the top ten criteria for further evaluation to identify the causal relationship. The inputs were received again from the same experts as shown in Table C.9 of the Appendix C in the supplementary material, the experts were asked to provide an influence relationship between the criteria for applying PF-DEMATEL based on the scale of Table C.8. The inputs received from experts were further processed as per the steps in Section 3.3, and the calculations are presented in Table C.10-C.17. The final causal relationship is shown below in Table 9. Further, the causal diagram and interaction digraph are shown in Fig. 5.

4. Discussions

The assessment of barriers is utilized to develop the framework, which is shown in Fig. 6. The discussion on the results of the application of the methodology is discussed below in the following sections.

4.1. *Discussion of findings*

To develop a digitally trusted SC4.0 toward resiliency. It is imperative to identify and eliminate the critical barriers that prevent stakeholders from sharing their information on digital platforms. Based on the assessment through PF-AHP, the results are shown in Table 7, the order of priority of main dimensions is as follows: Organisational ≻Technological ≻ Stakeholder's intent ≻ Regulatory. Also, per Table 9 and Fig. 5, PF-DEMATEL classifies the barriers under 'Cause' and 'Effect' groups (Fig. 5(a)), with their interrelationship (Fig. 5(b)) shown. The organization's role is pertinent to building DT among the SC stakeholders to meet their obligations. Also, the 'Technological' barrier is the second important hurdle organizations must remove for a digitally trusted SC4.0. so that resiliency can be sustained, which is the result of the study by Agarwal and Seth [103] in the Indian automotive context. This points to the top management's decision-making regarding developing and deploying technologies for SC4.0. Third, an important dimension is the 'Stakeholder's intent' to share information and participate in the SC4.0 operations. The trust in SC4.0 depends on the organization's robust security and ethical responsibilities to maintain the integrity of SC stakeholders' data [104]. Fourth, 'Regulatory' concerns the oversight of the transactions of information and data, making good governance of the management leadership's actions and decisions accountable for SC4.0 sustainability. For instance, trust in BC fosters regulatory oversight with self-governance and coordination [105].

Table 9
Causal relationship of the top ten barriers.

Barriers	R_i	C_i	$R_i + C_i$	$R_i - C_i$	Causal Relationship
OB2	0.388	−0.079	0.309	0.467	Cause
TB2	−0.009	0.461	0.452	−0.470	Effect
TB1	0.150	0.675	0.825	−0.525	Effect
SB2	0.654	1.439	2.093	−0.785	Effect
OB4	1.086	0.491	1.577	0.594	Cause
SB1	1.335	1.387	2.722	−0.052	Effect
TB3	−0.148	0.673	0.525	−0.821	Effect
RB2	0.261	0.406	0.668	−0.145	Effect
RB1	0.554	−0.410	0.144	0.964	Cause
OB1	0.662	−0.110	0.552	0.773	Cause

The study reveals critically important barriers to establishing DT in SC4.0. Globally, 'Lack of top management support' (OB2) is the highest weighted factor. This underscores the integral role top management commitment in developing a digitally secure and resilient SC in the context of emerging economies, which has been emphasized by Luthra and Mangla [75]. This is also positioned within the 'Cause' group and reflects the strategic commitment at the top management level. As noted by Kalaitzi and Tsolakis [106] an organization's responsibility to protect stakeholder data fundamentally shapes trust in technology for visibility-enhancing resilience. It is essential for top executives to proactively lead digital initiatives and embed cybersecurity and data governance as regulatory measures into the core of the organizational strategies. This can be overcome through transparent performance metrics, and cross-functional trainings strengthen managerial cognition of reliability and competence building Cognition-based Trust, ensuring commitment to build DT.

The second highest-ranked barrier is the 'Risk of information security and privacy' (TB2) categorized under the 'Effect' group and most significant under 'Technological'. It emphasized the necessity for reliable, secure, and real-time information sharing among SC partners [106]. Organizations may adopt encrypted communication protocols with periodic vulnerability assessments for maintaining shared data integrity. With regular audits, encryption, access control, and real-time monitoring, reliability can be reinforced, while third-party and stakeholders' certification, regulatory adherence, and credibility. These mechanisms can institutionalize safety that enhances System-based trust.

'Cybersecurity Risks' (TB1) is ranked third globally and second among 'Technological' under the 'Effect' group. Pertaining to resilient digital systems, a majorly cybersecure SC4.0 will mitigate unintended data breaches and protect the stakeholders' data related to supply, operational, and demand risks, as suggested by Pandey, Singh [76] by conducting a case study in the Indian automobile sector. The stakeholders' primary demand for secure environments can be addressed through trust labels, data assurance certifications, and a clearly defined access control system, as has also been suggested by Wu and Zhang [86] in the Chinese coalmine context. The organizations may consider multi-layered defense frameworks and third-party certifications fostering safety, accountability, and oversight in the digital SC4.0 ecosystem. For resilient SC4.0, inadequate addressing of cybersecurity risks gives unauthorized access, makes the data of SC stakeholders vulnerable to cyber assaults, or destroys sensitive data [16]. BC-based security, strict compliance with international security standards, regular penetration testing, continuous monitoring for ensuring safety, accountability, oversight, and ensuring System-based Trust.

The fourth-ranked barrier is the 'Lack of digital infrastructure' (OB4), classified under the 'Cause' group and second within the organizational category. For a resilient I4.0 infrastructure, firms should emphasize the secure integration of software, hardware, and cyber-physical systems instead of focusing on a single technology [107]. Within the organizational level, this integrated perspective supports or impedes DT throughout the digital transformation process, as also outlined by Dixit, Malviya [16] by conducting a case study in the Indian automobile context and Strazzullo [20] by conducting a survey in the Italian Manufacturing industry. Here, the role of organizations becomes crucial to build cognition-based trust among stakeholders by developing scalable technologies leveraging cloud-based platforms, adopting an interoperable system to demonstrate measurable improvements in

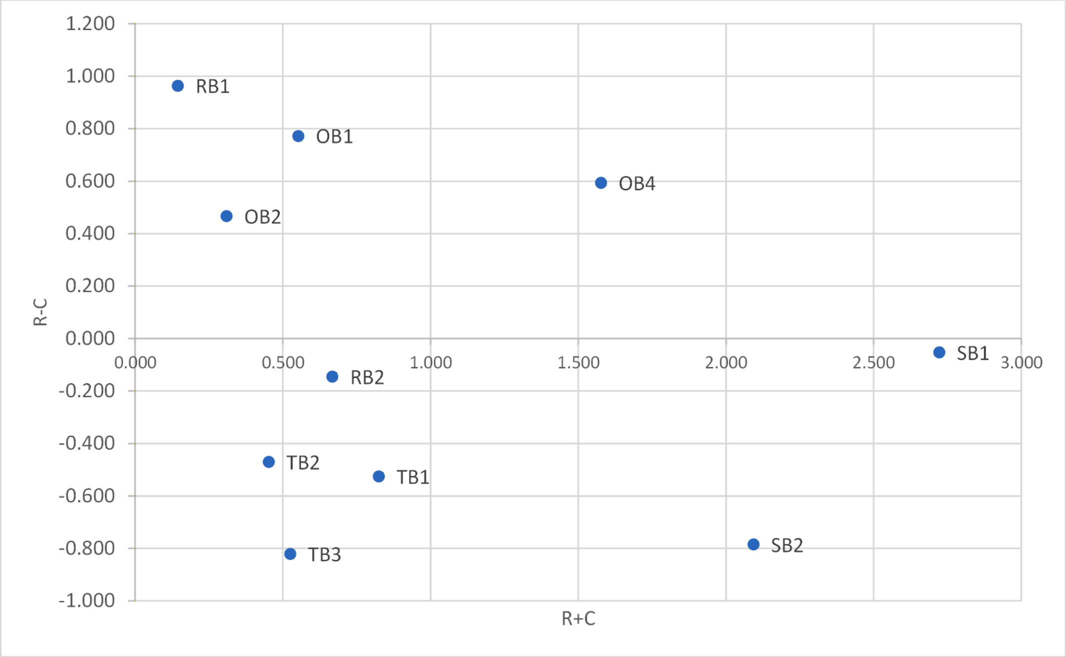


Fig. 5 (a)

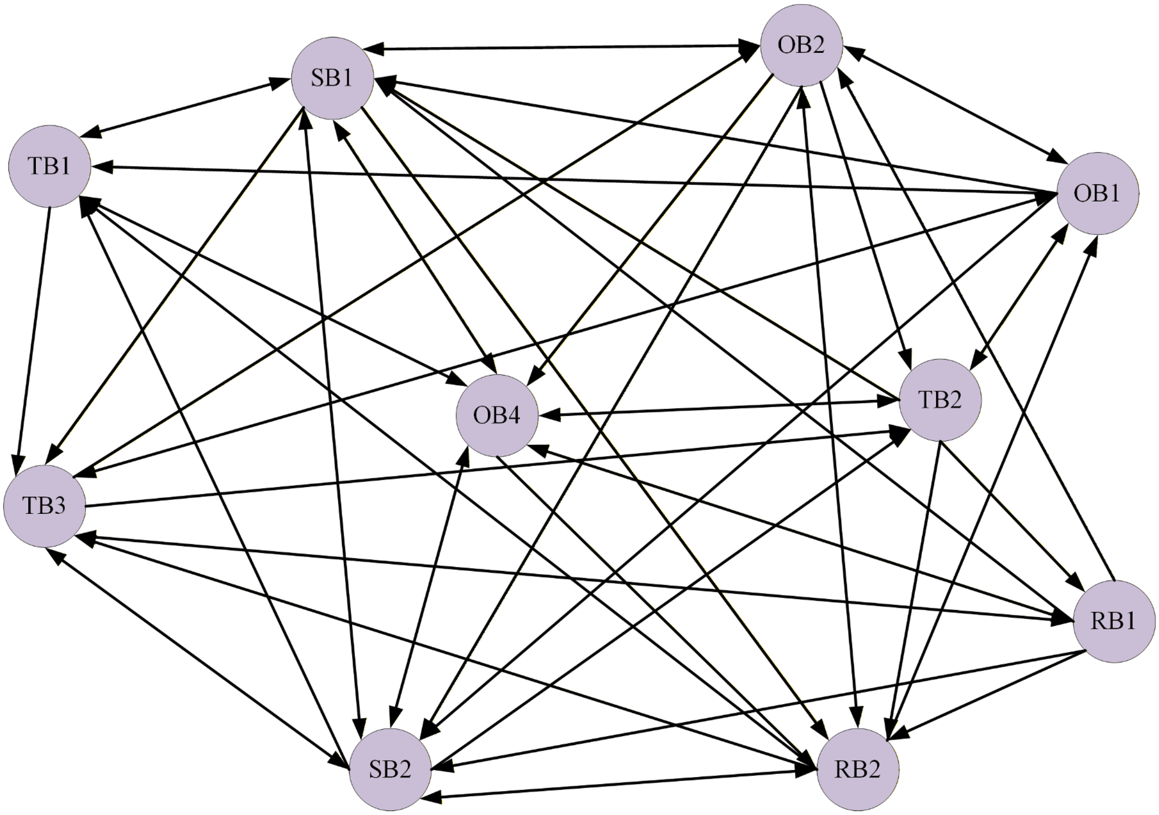


Fig. 5 (b)

Fig. 5. Causal Diagram and Interaction Digraph.

efficiency and transparency through pilot projects.

The fifth important barrier is the ‘Lack of real-time information sharing’ (RB2) placed under the ‘Effect’ group and ranked highest in the regulatory domain. Real-time, bidirectional, accurate data exchange on demand and supply is essential for SC4.0, enabling efficient logistics, inventory management, and financial transactions. This is well observed and pointed out by Wu and Zhang [86] and others Attaran [12], Pandey, Singh [76], Chaouni Benabdellah, Zekhnini [81] underline the role of IoT-based real-time information sharing in the emerging economies. This contributes to improved communication and collaboration quality among SC partners [83]. Firms may implement integrated IoT dashboards and predictive analytics tools to support just-in-time

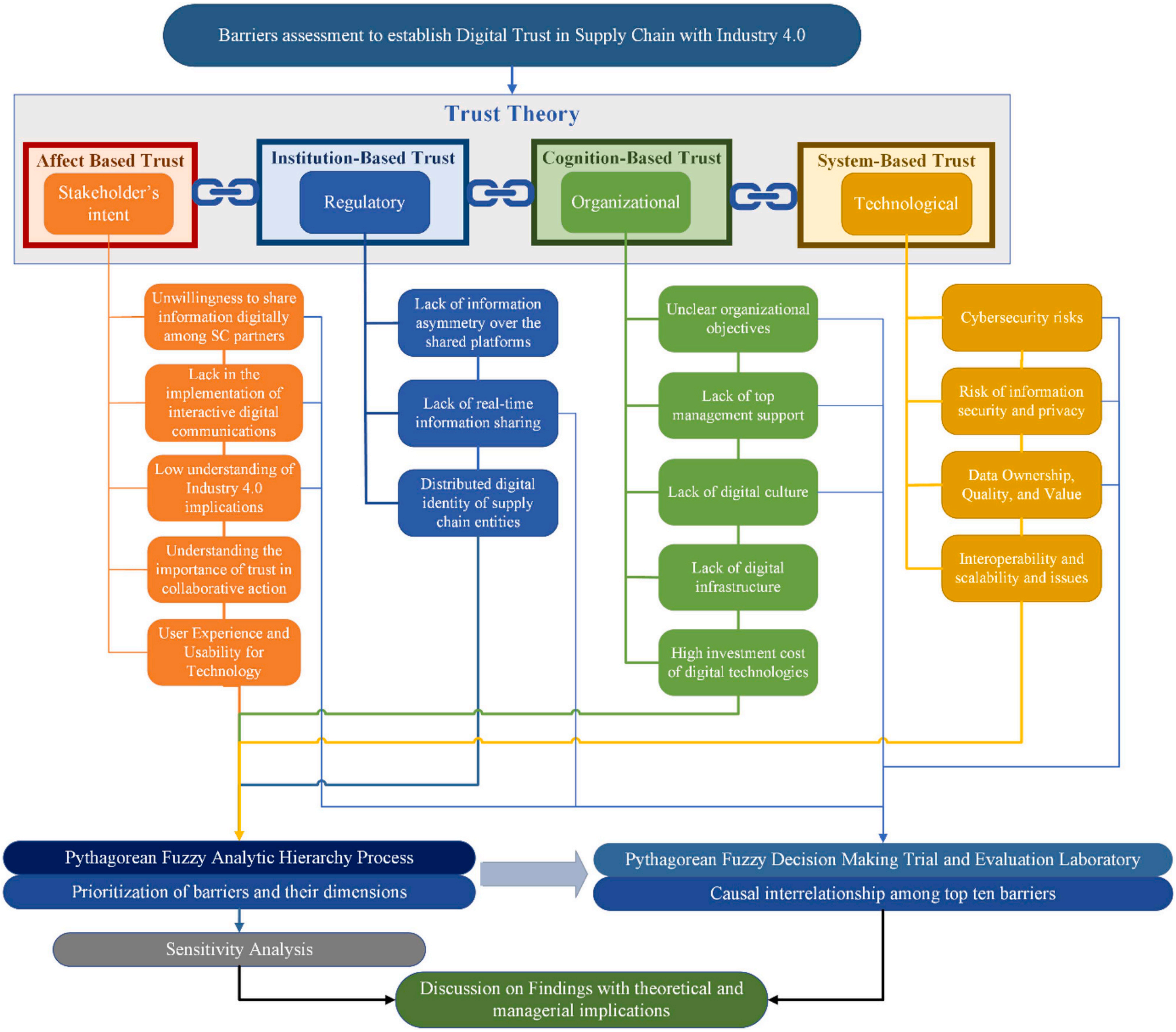

Barriers assessment to establish Digital Trust in Supply Chain with Industry 4.0
Trust Theory
Affect Based Trust
Stakeholder's intent
Institution-Based Trust
Regulatory
Cognition-Based Trust
Organizational
System-Based Trust
Technological
Unwillingness to share information digitally among SC partners
Lack in the implementation of interactive digital communications
Low understanding of Industry 4.0 implications
Understanding the importance of trust in collaborative action
User Experience and Usability for Technology
Lack of information asymmetry over the shared platforms
Lack of real-time information sharing
Distributed digital identity of supply chain entities
Unclear organizational objectives
Lack of top management support
Lack of digital culture
Lack of digital infrastructure
High investment cost of digital technologies
Cybersecurity risks
Risk of information security and privacy
Data Ownership, Quality, and Value
Interoperability and scalability and issues
Pythagorean Fuzzy Analytic Hierarchy Process
Prioritization of barriers and their dimensions
Pythagorean Fuzzy Decision Making Trial and Evaluation Laboratory
Causal interrelationship among top ten barriers
Sensitivity Analysis
Discussion on Findings with theoretical and managerial implications

Fig. 6. Framework for Digital Trust in Supply Chain 4.0 towards Resiliency.

decision-making through improved end-to-end visibility. By instituting standardized data-sharing protocols, ensuring compliance with the regulatory framework for data transparency and accuracy, formalizing accountability and governance, fostering Institution-based trust to move towards DT.

'Data Ownership, Quality, and Value' (TB3) constitutes the sixth important barrier situated in 'Effect' group. In the interconnected SC4.0 environment, weak security practices for suppliers and contractors target cyberattacks on logistics systems and Tier-1 suppliers due to limited visibility [87]. Data ownership is crucial to building DT, as it ensures security, transparency, visibility, compliance, and trusted collaboration in BC-enabled resilient SC [86]. To establish, DT companies can establish data ownership protocols and use smart contracts to develop accountability. Transparently assigning data stewardship roles, with real-time validation mechanisms, and linking data use for measurable business outcomes, reinforces system reliability and thereby instill System-based-Trust among the stakeholders.

The seventh-ranked barrier is 'Unclear organizational objectives' (OB1), falling under the 'Cause' group. Ambiguity in goals hampers cross-functional interactions, which can be diminished by SC collaboration and competence [79]. The organization may prioritize flexibility and strategic objectives for enhancing competitiveness [17], profitability, and quality, and deliver robust operational measures to build stakeholders' confidence. Aligning digital SC initiatives with strategic goals, setting transparent performance metrics, and communicating a unified digital vision can build Cognition-based trust among stakeholders, consistent with the findings of Pandey, Daultani [2] in the Indian context and Jum'a and Bushnaq [27] in the Jordanian manufacturing context.

'Lack in the implementation of interactive digital communications' (SB2) is in eighth position as a barrier under the 'Effect' group. This has appeared due to the low level of understanding of I4.0 implications among participants [75] for digitally trusted collaborative action through data exchange for better resilient SC performance [8]. Organizations with strong leadership commitment can work on digital literacy through targeted training, fostering a culture of transparency and collaboration. Organizations can provide and deploy collaborative platforms, virtual engagement tools, and AI-driven interfaces. Frequently, empathetic interactions among SC partners are crucial for creating continuous relational communication, which nurtures mutual confidence and can establish Affect-based Trust.

The ninth-ranked barrier is the 'Unwillingness to share information digitally among SC partners' (SB1) under the 'Effect' group. The resistance to sharing critical information by SC stakeholders due to security issues hinders partners in interactive digital communication [79]. Transitioning to a digital system in manufacturing often faces resistance to change, obsolescence, and concerns over data privacy. To overcome these challenges, comprehensive training on the functional use of technology can build participants' trust in digital technologies [20]. Cultivating innovative openness through training, transparency, nurtures mutual respect and assurance, thereby strengthening Affect-based trust. These findings align with the suggestions and findings of Dixit, Malviya [16], Chauhan, Singh [34,79], Yadav and Kumar [83] within the emerging economy context.

As for a resilient SC4.0, visibility through information sharing is crucial for businesses to adapt quickly to uncertain disruptions. This underscores the need for the involvement of top management commitment to allocate a budget for a cyber-secure digital infrastructure that can ensure trustworthy interactive digital communication for efficient SC management. Our findings align with the findings of Pandey, Daultani [2], Khan, Haleem [84]. The users' resistance to change to adapt to new digital, organizational, and process transformations often stems from a lack of training or resources. has been highlighted by Caliskan, Eryilmaz [108] in the Turkish context. Similarly, Vietnamese SC, in the early stage of I4.0 adoption, suffers from a gap between the high expected impact of these technologies and the relatively low planned investments [15], suggesting missed opportunities for competitive advantage [109]. Unwillingness to share information restricts digital communication due to the lack of sophisticated cybersecure networks, which questions the integrity of SC [66], if appropriately addressed, it can provide a competitive advantage [81] with SC collaborations, cooperation, interaction, and consistent decision-making [65,89].

'Lack of information asymmetry over the shared platforms' (RB1) is the tenth important barrier and is under 'Cause' group. Participants' concern about the level of organizational and technological capability to prevent the misuse of critical information shared through digital platforms diminishes DT. Information asymmetry creates trust asymmetry, financial cost, and reputation damage. This has also been pointed out by Brookbanks and Parry [29] and Strazzullo [20]. Establishing a unified data taxonomy creates shared benchmarks for fairness, accuracy, and accountability, reinforcing Institutional trust. SC managers should explore regulatory compliance checks on distributed digital identities to enhance transparency, control, and verification in shared digital spaces.

'Lack of digital culture' (OB3) is the eleventh important barrier. A weak digital culture in the organizations can lead to a loss of confidence among the SC stakeholders, leading to ineffective and inefficient SC operations. Here, the importance of human element I4.0 adoption has been linked to workforce management by organizations to foster a digital culture for enhancing flexibility, as suggested by Pandey, Daultani [2]. This attitude will also ensure the trustworthiness of the SC stakeholders in SC4.0 [86]. Incentivizing technology adoption to embed shared values and competencies across the organization demonstrates consistent readiness and capability, which cultivates Cognition-based trust. This has also been suggested as a counter-strategy to secure dedicated support and incentives from top management in the Turkish healthcare SC context [11]. The twelfth and thirteenth barriers, 'Low understanding of Industry 4.0 implications' (SB3) and 'Understanding the importance of trust in collaborative action' (SB4) further signify the knowledge and awareness gap among SC actors for digitally trusted collaborative action through data exchange for better SC performance [8,75]. Education and training through workshops and hands-on training can build relational connections [23], while simultaneously embedding trust and collaborative practices, allowing organizations to build affect-based trust, strengthening DT.

'Interoperability and scalability issues' (TB4) is the fourteenth important barrier. This corresponds to the lack of infrastructure, which in turn leads to interoperability and scalability issues, thereby raising security concerns and creating a lack of data trust (DT) in the infrastructure [81]. Organizations can work on adopting globally recognized standards and participating in the industry-wide standardization initiatives to overcome this barrier. These actions, utilizing modular digital architectures and cloud-based platforms, can promote System-Based Trust, enabling seamless integration and ensuring future growth. 'User Experience and Usability for Technology' (SB5) ranked sixteenth as an important barrier, creating the importance of intuitive and efficient digital platforms, as poor usability can hinder adoption and create resistance. By designing systems with a human-centric approach and integrating regular user feedback under compliance management, organizations can reinforce Affect-based trust that is vital for sustaining digital collaboration across SC networks.

Lastly, the seventeenth barrier, the 'Distributed digital identity of SC entities' (RB3), underscores the growing importance of digital identity protocols in enhancing traceability, authentication, and trust within SC4.0 environments. To address this, regulatory bodies and industrial alliances can collaborate to establish unified frameworks for verification of digital identities that ensure secure, verifiable, and permissioned access across the SC ecosystem, while being vigilant about the stakeholders' data protection regulations to build DT in the technology [20]. Such institutional safeguards formalize credibility and governance, thereby fostering Institution-based trust in SC4.0.

A digitally trusted SC4.0 instigates the requirements of data rights, which require regulatory and legal procedures. A digitally trusted and cyber-resilient platform provides a guarantee of data reliability, quality, security, safety, and credibility for transactions on an automated platform such as BC [17,83,86]. The data collection through real-time information for smarter SCM [44] with trusted BC-IoT devices [110] provides an efficient exchange of information in a trusted environment [65]. The linkage between manufacturing and logistics industry is studied by Li and Wang [64] in the Chinese context and it is highlighted that it is broken by unstable trust, which results in lagging and insecure information sharing, which can result in loss of effective collaboration. This indicates the importance of real-time information sharing to build DT in SC4.0. Its security, reliability, and quality depend on tackling barriers such as cybersecurity risks and a lack of stakeholders' DT, which, if resolved, will not hinder effective collaboration and decision-making.

4.2. Roadmap for the success of DT in SC4.0

Building DT in the SC4.0 environment requires framing digital initiatives as strategic business imperatives rather than focusing solely on technical upgrades. C-executives, i.e., the organization's senior executive, can buy-in only if a clear vision for profitability, risk mitigation, and competitiveness is secured while satisfying customers [44]. Once achieved, it can help the execution of functions in a resilient and trustworthy ecosystem. Within the SC context, this study presents a digital transformational strategy to build DT in SC4.0 in a phased roadmap from vision creation to localized pilots for the SC organizations of the emerging economies to advance their DT journey. To overcome the persistent challenges, the specific implementation roadmap is also provided in a multi-phased manner as shown in Fig. 7 and discussed below.

Phase 1. Establish a Chief Executive Suite-driven mandate

As indicated by the finding, the requirement of organizational commitment, this phase positions executive leadership to ensure board-level sponsorship to overcome barriers such as inadequate commitment, unclear objectives, and limited I4.0 understanding, indicating organizations' focus to act towards (1) Cybersecurity as a strategic Incumbent: Reframe cybersecurity from merely as a cost center to a "cornerstone of trust and motivation" with Chief Executive Officers (CEO). Chief Financial Officers (CFO) and the board are defining it as a whole business responsibility.(2) Long-term vision: Formulating a unified long-term trust strategy to stabilize SC relationships, which translates to growing investment and raised cyber budgets. (3) Cyber-risk quantification to drive investments: Demonstrate the financial strength by quantifying the high cost of breaches and linking cybersecurity risk

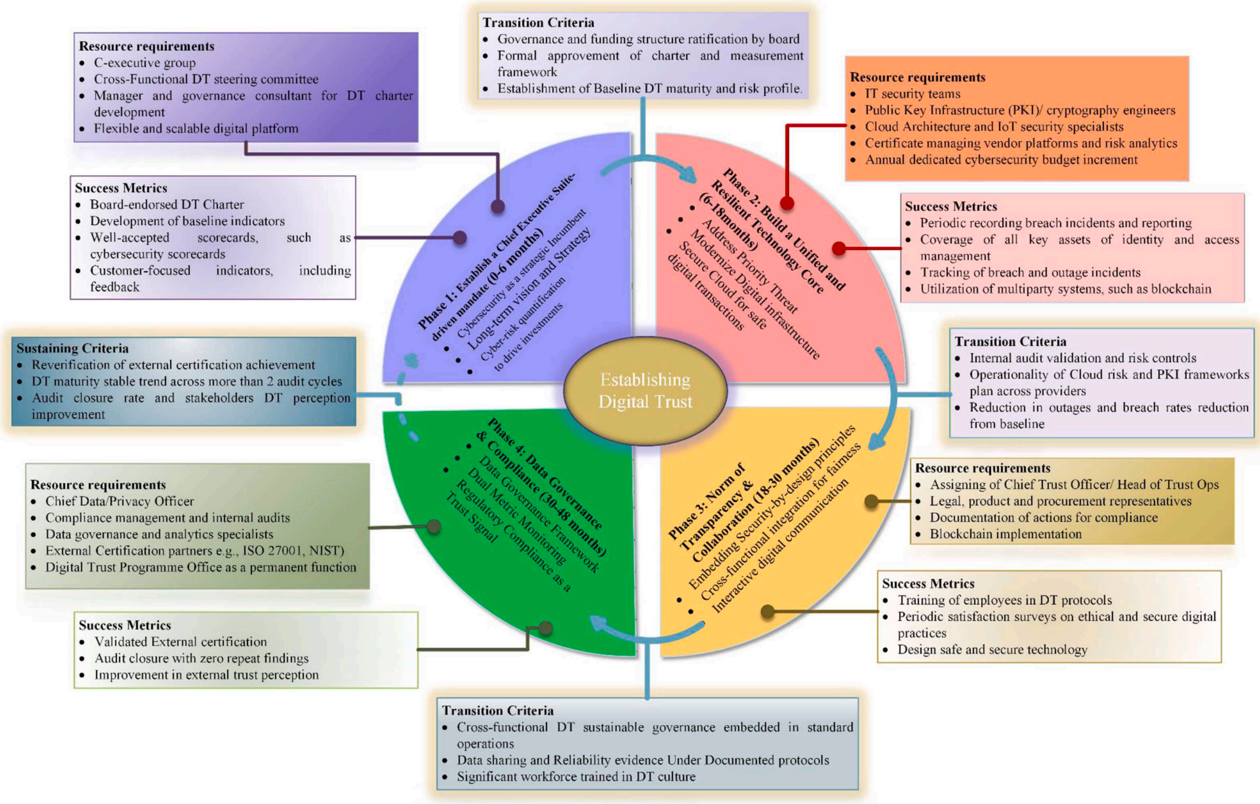

Resource requirements
• C-executive group
• Cross-Functional DT steering committee
• Manager and governance consultant for DT charter development
• Flexible and scalable digital platform
Transition Criteria
• Governance and funding structure ratification by board
• Formal approvement of charter and measurement framework
• Establishment of Baseline DT maturity and risk profile.
Resource requirements
• IT security teams
• Public Key Infrastructure (PKI)/ cryptography engineers
• Cloud Architecture and IoT security specialists
• Certificate managing vendor platforms and risk analytics
• Annual dedicated cybersecurity budget increment
Success Metrics
• Board-endorsed DT Charter
• Development of baseline indicators
• Well-accepted scorecards, such as cybersecurity scorecards
• Customer-focused indicators, including feedback
Success Metrics
• Periodic recording breach incidents and reporting
• Coverage of all key assets of identity and access management
• Tracking of breach and outage incidents
• Utilization of multiparty systems, such as blockchain
Phase 1: Establish a Chief Executive Suite-driven mandate (0-6 months)
• Cybersecurity as a strategic Incumbent
• Long-term vision and Strategy
• Cyber-risk quantification to drive investments
Phase 2: Build a Unified and Resilient Technology Core (6-18months)
• Address Priority Threat
• Modernize Digital infrastructure
• Secure Cloud for safe digital transactions
Sustaining Criteria
• Reverification of external certification achievement
• DT maturity stable trend across more than 2 audit cycles
• Audit closure rate and stakeholders DT perception improvement
Establishing Digital Trust
Transition Criteria
• Internal audit validation and risk controls
• Operationality of Cloud risk and PKI frameworks plan across providers
• Reduction in outages and breach rates reduction from baseline
Phase 4: Data Governance & Compliance (30-48 months)
• Data Governance Framework
• Dual Metric Monitoring
• Regulatory Compliance as a Trust Signal
Phase 3: Norm of Transparency & Collaboration (18-30 months)
• Embedding Security-by-design principles
• Cross-functional integration for fairness
• Interactive digital communication
Resource requirements
• Assigning of Chief Trust Officer/ Head of Trust Ops
• Legal, product and procurement representatives
• Documentation of actions for compliance
• Blockchain implementation
Resource requirements
• Chief Data/Privacy Officer
• Compliance management and internal audits
• Data governance and analytics specialists
• External Certification partners e.g., ISO 27001, NIST)
• Digital Trust Programme Office as a permanent function
Success Metrics
• Training of employees in DT protocols
• Periodic satisfaction surveys on ethical and secure digital practices
• Design safe and secure technology
Success Metrics
• Validated External certification
• Audit closure with zero repeat findings
• Improvement in external trust perception
Transition Criteria
• Cross-functional DT sustainable governance embedded in standard operations
• Data sharing and Reliability evidence Under Documented protocols
• Significant workforce trained in DT culture

Fig. 7. Guidelines for building Digital Trust in Supply Chain 4.0.

associated with digital technologies as a compelling case for investments.

An organization's DT plans need to be executed, but they require auditing and monitoring of I4.0 technologies through a board-endorsed DT Charter with the safety and security of SC stakeholders' data as its prime and long-term goal. Here, the role of effective governance by the organization's C-suite is crucial in enabling security. Support from top leadership can take the form of executing strong governance programs that build DT among SC stakeholders. Leaders can invest more in DT factors and take measures to sustain them by aligning the organization's goals with those of improving skills and training in technology, as well as increasing understanding of DT and its dimensions. The trust measurement can be accomplished by deploying the indicator tools with cross-functional teams working together, which requires the time and effort of the executive teams. These baseline maturity indicators tools can act as a lubricant in maintaining the DT between human-human and human-machine interactions. Execution of data-driven scrutiny under governance and compliance mechanisms through user control requires a flexible and scalable digital platform that does not introduce vulnerabilities to cyberattacks and data breaches.

Organizations can track their progress through various metrics, such as the number of cyberattacks or breaches on their IoT devices and connected devices. The C-suite's responsibility is to monitor for continuous improvement regarding investment allocations, staff learning on DT initiatives, and the development of trusted relationships not only within the firm but also in business-to-business experience. Organizations can develop the DT index by continuously monitoring data governance and implementation practices to ensure security, privacy, safety, and accountability. As the monitoring of compliance management is of great concern, failure may lead to stakeholders losing confidence in the data handling authority. Well-accepted scorecards, such as cybersecurity scorecards within the internal monitoring system. Organizations can monitor delivery DT across SC stakeholders with provenance in digital rights, as well as in other spaces where the stakeholders are connected. Even customer-focused indicators, including feedback on organization data integrity practices, can be used to make them confident in the digital information exchange process. These steps can enable organizations to become capable of making strategic moves and be competitive in the dynamic market environment running on digital platforms.

The above steps lay the foundation for transitioning toward Phase 2, making their cyber space robust and breach-free from cyber-attacks. The investments to develop a strong technology core require the C-suite's formalization of budget implementation through the DT Charter. This can be executed through the approved official communication by the designated C-suite officer, allowing collective involvement of all active teams. Moreover, the organization should consider that, before deploying the DT execution, the mandate must make stakeholders confident through the baseline metrics verification, as DT connects not only with the organization but also with people and technology.

Phase 2. Build a Unified and Resilient Technology Core: Increasing cybercrime, specifically targeted attacks on devices, can cause potential damage to DT due to a lack of awareness and expertise in cyber safety. DT is an inclusive concept that indicates organizations, and their individuals must be understand the policy laid as the mission and objective of the organization with proper governance measures to become resilient to cyberattacks. These can be achieved through the following: (1) Address Priority Threats: The organizations can work on prioritising threats by responding to severe threats in time, as this could help organizations in reducing their liability and enhancing their goodwill among SC stakeholders. (2) Modern digital infrastructure: The organization can install a modern technological setup to map the functioning of the digital identities of all SC stakeholders. Implementing cryptographic keys can help in maintaining the confidentiality of shared data as they run on specific algorithms. The robust infrastructure includes secured hardware and software for uninterrupted communication among all SC stakeholders. (3) Secure Cloud for safe digital transactions: In the modern SC4.0 operation, where data is stored on the cloud that must be secured from attacks, breaches, and unauthorized access as it is being transferred between systems. Organizations may define policies to ensure security under required regulatory compliance for devices that are cloud-oriented and exchange sensitive and crucial data.

DT establishment requires resources such as DT leaders, including technology innovators, who possess expertise in cybersecurity, privacy, and technology ethics, and have the capability to provide verification and assurance for all SC stakeholders. Despite the ubiquitous digital fabric woven into digital devices, traditional digital security remains important. For example, Public Key Infrastructure (PKI) verifies digital identities and data authentication, ensuring integrity in digital transactions. These PKIs being flexible can also be deployed in any number of environments, making them interoperable. Furthermore, DT has the capacity to unify all parties, including stakeholders and regulators, through verified certificate management vendors and the allotment of specific cybersecurity budget.

The organization can build a resilient technology core for SC4.0 by strengthening the critical information protection infrastructure against cyberattacks. This can be accomplished by periodically recording breach incidents and reporting them to the relevant cybersecurity agency. The organization can take actions to protect its users from harmful and malicious digital content. This will build confidence in the organization's stakeholders that the necessary actions regarding the safety of their data are being taken. This can also be further administered by C-executives through feedback, satisfaction surveys, and data flow analysis. Organizations can utilize multiparty systems, such as blockchain, distributed ledger, and tokenization, that mitigate the risks of security, privacy, and control.

The transition to Phase 3 requires the implementation of the actions to make a trusted technology deployment. This may be achieved by the continuous commitment of top management over concerns of data protection of all stakeholders implemented through regular auditing. These activities can build a firm's reputation and reduce data breaches and cyberattack incidents due to confidence of the SC stakeholders in the organization's technology and systems' robustness.

Phase 3. Promote a Norm of Transparency and Secure Collaboration: This phase underscores the need for transparency about the protection policies and actions taken by organizations for the stakeholders' information handling. Also, there must be informed decision-making while displaying crucial information on the SC stakeholders' common dashboard. These practices are critical to limit damage to the enterprise's reputation and build DT by (1) Embedding Security-by-design principles: While designing a technology, the established governance process with multiple supports, such as self-service, dedicated personnel for a particular digital technology. Further, while designing the data processing mechanism individual autonomy through notice and control must be maintained. (2) Empower Cross-functional integration for fairness: To make decision-making process quicker, the cross functional teams such as cybersecurity, data engineers, must work collectively rather in silos in processing data so that outcomes are equitable for all SC stakeholders. (3) Strengthen interactive digital communication: DT is not only a technology mediated solution; however, it involves people, process and technology as a two-way communication. This demands their interactive digital communication in the entire SC4.0 ecosystem.

This phase highlights the need for the assignment of Trust Officers who can execute personnel training on security and privacy through workshops to help them understand how their role affects DT. The lack of training opportunities due to misalignment of the organization's goals can create serious concerns. While the establishment of DT lies in responsible management actions that continuously utilize resources to improve the DT factors and support C-executives consisting of legal representatives, these are the main drivers of DT that ensure a clear understanding of policy and framework implementation across the SC. These actions must also be documented and distributed among all stakeholders, and their compliance must be monitored.

The improvement can be monitored through continuous assessment of the implementation of cybersecurity practices. The stakeholders must be aware and trained to take action and report threats and attacks on devices as a necessity. As the Trust Theory suggests, Trust is a collaborative approach; this suggests that each partner of the SC can contribute to the common goals of a safe and secure SC4.0 ecosystem through their feedback and monitoring system. These actions can help in understanding the importance of I4.0 implications and trust in collaborative actions for the development of digital culture among the SC stakeholders. Further, the C-executive can work on designing the technology that is safe and secure by default. However, only by deciding and acting for the establishment of DT can organizations and board members meet their obligations, which must be synthesized.

As organizations can track technology with security principles and cross-functional teams for DT governance, it enables transparent communication. The information deployed on digital technologies must be assessed periodically to ensure its safety and privacy, and provide reliable evidence to its stakeholders. Organizations can track the status of their DT workshops' deliverables through continuous monitoring, departmental tests, and the generation of report cards as a documented protocol to track improvement in DT culture. These actions can enable the organization to move to Phase 4.

Phase 4. Robust Data Governance and Anticipatory Compliance: In this Phase, organizations assess their data processing capabilities to confirm their auditable discourse. This signals an organization and its leaders' commitment to fulfilling the stakeholders' expectations through collaborative communication on data governance practices. The governance measures relate to the compliance with safety, quality, privacy, and security of digital assets. Even governance over ethics and data integrity is crucial to enable DT in SC4.0, with (1) A comprehensive data governance framework: A framework that covers each and every stakeholder of the SC for significant collaboration. This can become effective when other components, such as risk management, data quality, ownership and stewardship assurances, resilience, and ethics, are interwoven within the organization's DT strategy. (2) A Dual Metric Monitoring System: Auditability of the organization can make a long-lasting impact by drawing comparisons with previous governance measures. These metrics and monitoring systems can help prevent damage due to breaches beforehand. This monitoring can validate data accuracy, authenticity, and reliability through a secure cloud, blockchain with immutable records. (3) Regulatory Compliance as a Trust Signal: A well-structured compliance mechanism and implementation standards can help in verifiable operation, such as certificate status under NIST, ISO 27001 protocols. This enables transparency with additional details, such as where, when, and to whom to report, are provided.

The checking of governance can be conducted through the establishment of a permanent DT programme office, which checks for organization's validation of external certification such as ISO27001, NIST, etc. The office can work on sustaining the data governance mechanism that focuses on the availability of data in the right amount to the concerned stakeholders under a transparent environment. This can be achieved through defined data ownership and stewardship rights that restrict continued access to data. The C-suite can execute the cross-functional teams to make this a priority, as this directly impacts stakeholders.

Assessing the metrics, such as data quality scores, alone is insufficient for effective governance. Organizations can smartly invest in data and regular scrutiny of digital risks to demonstrate better oversight and reduce information asymmetry among SC stakeholders. In addition, collective action can be executed only through a sense of responsibility among stakeholders when firms monitor progress by facilitating documentation for regulatory compliance. This can also assure provenance in the SC inputs to ensure data ownership and transaction records, and build valuable connections through valuable information flow to the right people. An organization's C-executive can run a risk assessment program to monitor participants, as loyal participants may not hesitate to break a relationship that compromises sensitive information.

The implementation of the four-phased framework can be converted into a sustainable execution programme by continuous monitoring. Once the governance, monitoring, and compliance structures are mature enough and validated for autonomous operation. The transition readiness is to be evaluated continuously to sustain. To ensure the readiness and operationality of the system, enforce quality rules, and ensure traceable auditability across all information exchange in the SC4.0. The dual-metric monitoring system consistently generates reliable insights, with year-on-year improvements in both internal and external indicators, thereby enhancing DT perceptions. Certification as a mandatory document and regulatory alignment have been achieved and retained across at least two audit cycles. The decision-making process for DT is institutionalized, such that governance and compliance routines are embedded within organizations as a digital culture, rather than project-driven targets. Meeting these criteria confirms the organization has transitioned from implementing to sustaining DT to build a resilient SC4.0 environment for all the stakeholders.

5. Implications

The implications for researchers, managers, and policymakers are discussed in the next subsections.

5.1. Theoretical implications

The present study contributes to the academic literature on the theoretical advancement of DT in the SC4.0 in the emerging economy context. This study utilised the Trust Theory to analyze the barriers of DT in the digitally mediated SC4.0 context. This study contributes to the theory of DT as well as to the SC4.0 by proposing a roadmap to establish DT and increase understanding of barriers through a framework and Trust theory dimensions. Studies like Strazzullo [20] explored DT within the manufacturing factory and Mubarak and Petraite [55] has explored I4.0 and DT for open innovation, lacking the empirical validation in the manufacturing SC context. This study fills this gap by exploring DT in the horizontal value chain for resiliency through an industry survey in the emerging economy context. This study has investigated seventeen barriers for their validation and discussed the findings based on the Trust Theory. The study conducted a case study to prioritize and find causal relationships of barriers utilising PF-AHP-DEMATEL, which introduces a strong methodological approach to handle uncertainty in decision-making, providing practitioners with actionable insights. These insights may be applicable to other emerging economies undergoing digital transformation to establish DT in their SCs. Further, the study contributes by discussing findings that position DT as a foundational capability to be pursued by organizations to achieve resilience in their operations. As SC is a stakeholder-operated process through collaboration, the intervention of I4.0 technologies enables communication through digital platforms, where the role of DT is inexcusable. The study has also provided a theoretical roadmap for action that organizations may adopt to build DT in their digitally mediated SCs.

5.2. Managerial implications

The present study provides a framework and roadmap for the success of DT in SC4.0. This study has investigated seventeen key barriers to DT for better SC4.0 operations in the merging economy context. Based on this, a few managerial implications can be presented. Industry managers can look upon these barriers to build a cyber-secure and resilient system for all SC stakeholders. The top management can work on revising their policies and extending them to DT establishment policies. Further, organizations can establish officials and managers responsible for monitoring the data handling practices of the organization. The discussion reveals that, to build a digitally resilient SC, DT is not an option to consider for future ventures, but rather it is a present-day need as a foundational capability. These capabilities can be developed among stakeholders in the form of digital culture through workshops/trainings for increasing awareness on disseminating data on digital platforms within a regulatory policy framework, and gathering feedback on the implemented learnings. These actions can influence improved collaboration through digital platforms, as DT is built for technology adoption. The top management's strong commitment towards budgetary allocation for building a cyber-resilient infrastructure and viewing it as a long-term strategic goal. The implications of this can be in the form of reduced risks of security attacks and help in providing transparency, safety, and accountability for stakeholders' information, which are crucial components of DT.

Manufacturing SC managers and digital consultants to manufacturing firms can consider the identified barriers and their dimensions while designing technologies for the firms. As SC4.0 operates on information sharing, the aspect of cybersecurity in the technologies developed remains a challenge for designers. Also, managers can bridge the gap between technology and people by undergoing a process of technology development that involves cross-functional teams trained for DT building. For emerging economies such as India, this study has reported the challenges and insights for DT development that are crucial in the times of digital transformation when the digital systems are in the development stage and are vulnerable to cyberattacks.

5.3. Policy implications

The present study highlights the need for policy frameworks that institutionalize DT as the core pillar in national and industrial SC initiatives. Regulatory bodies need to encourage the standardization of DT metrics in sectors that are adopting digital technologies. As evidenced in existing studies, the absence of data governance norms and interoperable frameworks affects trust building across stakeholders. Government and industry bodies need to develop clear guidelines for transparency and cybersecurity in digital SCs. Policies based on emerging global practices can help address trust-related issues in SC and help to build resilient SCs. Policy makers can pursue I4.0 technologies strategically, which can enhance trust and thereby a resilient SC through real-time visibility and proactive monitoring suggested by Hossain, Talapatra [1] in the Bangladeshi context.

Moreover, policy interventions should focus on capacity building and digital literacy. There is a need for targeted national programs to reskill manufacturing workforces in small and medium enterprises, where digital mistrust often happens from a lack of awareness and technological exposure. Apart from this, the present study also suggests that public-private collaborations are important to promote DT in supplier authentication and data exchange in SC.

This study also highlights the need for context-specific trust policies that reflect regional market realities. For emerging economies, trust in digital platforms is influenced by infrastructure readiness, institutional credibility, and socio-cultural factors. Hence, policy frameworks should not be universally imposed but instead co-created with localized inputs from industry and academia to ensure relevance, scalability, and long-term adoption.

6. Conclusion, limitations, and future research

This paper has investigated the barriers to DT in SC4.0 for resiliency by performing an SLR along with expert consultation. Through the industry-wide survey in the emerging economy context, the barriers were critically analyzed for their significance. The study discusses the importance and implications of DT barriers by extending the Trust Theory to the digital interaction context by prioritizing them and identifying causal relationships through a framework, using an example case. Further, the study presents a roadmap to establish DT for its success in the SC4.0 context for emerging economies. The study is built on the foundation of Trust Theory, which discusses Trust as a multidimensional construct consisting of System-Based Trust, Affect-Based Trust, Institution-Based Trust, and Cognition-Based Trust. All these dimensional constructs have been utilized to discuss the implications of barriers and organizations' actions to develop DT throughout the supply network. This illustrates the applicability of the Trust Theory in the modern digital context of SC, which extends beyond organizational and interpersonal boundaries. Our findings have demonstrated the importance of stakeholders' willingness to accept vulnerability in data-driven SC4.0 operations, as well as the role of organizational compliance and policies regarding data, and the reliability of stakeholders in the organization's resilient cyberspace in the event of a breach or attack. This indicates how the establishment of DT depends on both human (SC stakeholders) and robust technical assurances. Moreover, in an emerging economy like India, challenges to SC digitalization still persist despite numerous government digital initiatives. This study presents DT at the forefront as a precautionary step that organizations may consider in their SC in I4.0 environment to sustain themselves in this digitally evolving, competitive market, which is open to any uncertain events. By applying the conventional Trust Theory in SC4.0 environments which include multi-actor data systems and automated governance architectures intrinsic to SC4.0. the study demonstrates how DT can stabilize digital interaction and reduce perceived risks and strengthen governance in digital SC settings.

The success of SC4.0 greatly depends on the reliability of data shared and trust in the technology. Previous studies have discussed the implications of the digitalization of SC, such as strategies' assessment [16], analysis of barriers to I4.0 and sustainability in SMEs [78], barriers to BC implementation in SC [84], barriers analysis of I4.0 adoption for SC competency and operational performance [34], analysis for a viable circular digital SC with BC technology [81]. The extant literature depicts a continuous lack of evaluation of barriers to DT in SC4.0. This study bridges this gap by substantiating the identified seventeen barriers through a survey in the emerging economies' manufacturing SC context for establishing DT in SC4.0. Therefore, this study contributes to the present literature by providing valuable insights and suggestions to practitioners and managers. The study's contributions are summarized below:

- This research has utilized a mixed-methods approach by utilizing EFA to group barriers into four dimensions, and further case-based evaluation through experts with PF-AHP-DEMATEL.
- The results have underlined the need for top management commitment and support with the development of robust digital infrastructure, ensuring the protection of data of all the SC stakeholders.
- With current digital transformations going on in India, the question of security, reliability, and protection rights of the data of the SC stakeholders is of great concern to the data managers and holders. The findings highlight that DT is essential as the digital SC works on real-time data, which can help managers to prevent disruptions that cause ripple effects.
- It examines the barriers and concludes that a digitally trusted SC can build a resilient SC network that can withstand shocks during disruptions, ensuring business excellence and customer satisfaction.
- This study has identified barriers for their prioritization and relationship identification through a framework. Further, presenting a detailed roadmap for the success of establishing DT in the SC4.0 environment.

The above contributions were generated from the analysis of barriers by conducting a survey within the emerging economy's manufacturing SC landscape. The findings highlight that top management commitment, risk of information security, privacy, cybersecurity risks, lack of digital infrastructure, and lack of real-time information sharing are barriers to be taken care of which can build confidence in stakeholders to share their quality data accurately for building DT, strengthening resilient SC 4.0, and fostering sustained economic growth.

6.1. Limitations and further research

This work has identified seventeen barriers to DT for SC4.0 through literature, expert inputs, and a survey-based approach integrated with expert subjective opinions for evaluating barriers in the emerging economy's manufacturing SC context. To contextualise the findings from the mixed-method analysis involving the industry survey, it is important to acknowledge that the demographics of the industry participants skewed towards respondents from consulting to manufacturing (38 %), which constitutes a limitation of the study. This perspective bias can be attributed to the Government of India's digital initiatives, e.g., Digital India, which offers advantages to companies to leverage digital technologies for improving their SC operations to be competitive in the market and further puts pressure towards digital transitions. Here, organizations that may lack digital competency often look to third-party digital consultancy for their digital execution by hiring a digital consultant to take on the role within the company. As part of our survey process, we have also reached out to experts who provide digital consulting services to manufacturing firms, as they may have a better understanding of digital technologies. However, the response rate suggests that the interest of those experts in participating in the survey skewed the demographics. While the response from other experts, such as department heads and plant managers, may provide more diverse perspectives. Therefore, this presents an opportunity for researchers to consult the experts excluded from this study and produce future research on this niche topic of DT in SC4.0.

The study utilized the PF set theory to judge the identified barriers between AHP and DEMATEL. However, the PF set theory can be integrated with other decision-making techniques such as FUCOM (Full Consistency Method), SWARA (Stepwise Weight Assessment Ratio Analysis), WASPAS (Weighted Aggregated Sum Product Assessment), CoCoSo (Combined Compromise Solution), LBWA (Level Based Weight Assessment**)**, RAFSI (Ranking of Alternatives through Functional mapping of criterion sub-intervals into a single Interval), etc. The lack of qualitative analysis of the identified barriers can be further validated by conducting interviews with industry practitioners. We encourage future research to focus on exploring DT for sustainable buyer-supplier relationships executed through I4.0 technologies. Future studies can focus on the development of analytical and empirical models to analyze DT in different SC contexts. Also, a comparative analysis of the barriers in a specific industry across different geographical settings can be conducted. While this study contributes to the new knowledge of DT in SC4.0, it has some limitations, such as conducting a single case with a panel of five experts, underscoring the need for future studies with more case studies. This study has discussed findings on the barriers through a framework utilising the Trust Theory to generate insights and a practical, indicative roadmap for industry practitioners in emerging economies to establish the DT in SC4.0 environments and move toward resiliency. Like every study, this study also has shortcomings in the form of its basis on a single country and a single case; the findings may not be completely generalizable to all cases. However, the guiding roadmap is an indicative action plan; practical generalizability can be concretized through multi-case analysis in future works. Future studies could extend this study through longitudinal analyses employing structural equation modeling, coupled with validation of the DT measurement instruments, and taking responses from a large sample of diverse industries' SCs, such as pharmaceuticals, within a multi-country context.

Funding

The research work carried out has received no external funding from any source.

CRediT authorship contribution statement

Vaibhav Sharma: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Rajeev Agrawal:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Formal analysis. **Anbesh Jamwal:** Writing – review & editing, Visualization, Validation, Formal analysis. **Vijaya Kumar Manupati:** Writing – review & editing, Supervision, Formal analysis. **Vikas Kumar:** Writing – review & editing, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tbench.2025.100247.

References

[1] M.I. Hossain, et al., From theory to practice: leveraging digital twin technologies and supply chain disruption mitigation strategies for enhanced supply chain resilience with strategic fit in focus, Glob. J. Flex. Syst. Manage. (2024) 1–23.

[2] A.K. Pandey, et al., Analyzing industry 4.0 adoption enablers for supply chain flexibility: impacts on resilience and sustainability, Glob. J. Flex. Syst. Manage. (2024) 1–24.

[3] R.K. Singh, Transforming humanitarian supply chains with digital twin technology: a study on resilience and agility, Int. J. Logist. Manage. (2025).

[4] N.K. Jain, K. Chakraborty, P. Choudhary, Building supply chain resilience through industry 4.0 base technologies: role of supply chain visibility and environmental dynamism, J. Busi. Indust. Market. 39 (8) (2024) 1750–1763.

[5] B. Bhatnagar, V. Dixit, Resilient supply chains: advancing technology integration with pre-and post-disruption technology roadmap, J. Enterp. Inf. Manage. (2025).

[6] A.Z. Piprani, S.A.R. Khan, Z. Yu, Driving success through digital transformation: influence of industry 4.0 on lean, agile, resilient, green supply chain practices, J. Manuf. Tech. Manage. 35 (6) (2024) 1175–1198.

[7] A. Patidar, et al., Supply chain resilience and its key performance indicators: an evaluation under industry 4.0 and sustainability perspective, Manage. Environ. Qual. Int. J. 34 (4) (2023) 962–980.

[8] A. Ghadge, et al., The impact of industry 4.0 implementation on supply chains, J. Manuf. Tech. Manage. 31 (4) (2020) 669–686.

[9] G.F. Frederico, et al., Supply Chain 4.0: concepts, maturity and research agenda, Supp. Chain Manage. Int. J. 25 (2) (2020) 262–282.

[10] J.W. Veile, et al., The transformation of supply chain collaboration and design through industry 4.0, Int. J. Logist. Res. Appl. 27 (6) (2024) 986–1014.

[11] S. Seker, N. Aydin, Analyzing barriers and strategies in digital transformation for resilient SC in healthcare using AHP and MABAC under uncertain environment, J. Enterp. Inf. Manage. (2024) ahead-of-print(ahead-of-print).

[12] M. Attaran, Digital technology enablers and their implications for supply chain management. Supply Chain Forum: An International Journal, Taylor & Francis, 2020.

[13] C.L. Garay-Rondero, et al., Digital supply chain model in Industry 4.0, J. Manuf. Tech. Manage. 31 (5) (2020) 887–933.

[14] M. Ghobakhloo, et al., Industry 4.0 digital transformation and opportunities for supply chain resilience: a comprehensive review and a strategic roadmap, Prod. Plan. Cont. 36 (1) (2025) 61–91.

[15] A. Al Tera, A. Alzubi, K. Iyiola, Supply chain digitalization and performance: a moderated mediation of supply chain visibility and supply chain survivability, Heliyon (2024).

[16] V.K. Dixit, et al., An analysis of the strategies for overcoming digital supply chain implementation barriers, Deci. Anal. J. 10 (2024) 100389.

[17] P.C. Kandarkar, V. Ravi, Investigating the impact of smart manufacturing and interconnected emerging technologies in building smarter supply chains, J. Manuf. Tech. Manage. (2024).

[18] K.F. Cheung, M.G. Bell, J. Bhattacharjya, Cybersecurity in logistics and supply chain management: an overview and future research directions, Transp. Res. Part E Logist. Transp. Rev. 146 (2021) 102217.

[19] B. Hammi, S. Zeadally, J. Nebhen, Security threats, countermeasures, and challenges of digital supply chains, ACM Comput. Surv. 55 (14s) (2023) 1–40.

[20] S. Strazzullo, Fostering digital trust in manufacturing companies: exploring the impact of industry 4.0 technologies, J. Innov. Knowl. 9 (4) (2024) 100621.

[21] O. James, 8 Recent Cyber Attacks on the Manufacturing Industry, 2024 [cited 2025 15/03/2025]; Available from, https://wisdiam.com/publications/recent-cyber-attacks-manufacturing-industry/.

[22] PricewaterhouseCoopers, Manufacturer Cybersecurity and Supply Chain, PwC, 2022, 2022/02/24/; Available from, https://www.pwc.com/us/en/industries/industrial-products/library/cyber-supply-chain.html.

[23] A. Jena, S.K. Patel, Analysis and evaluation of Indian industrial system requirements and barriers affect during implementation of industry 4.0 technologies, Int. J. Adv. Manuf. Tech. 120 (3) (2022) 2109–2133.

[24] R. Gadekar, B. Sarkar, A. Gadekar, Model development for assessing inhibitors impacting industry 4.0 implementation in Indian manufacturing industries: an integrated ISM-Fuzzy MICMAC approach, Int. J. Syst. Assur. Eng. Manage. 15 (2) (2024) 646–671.

[25] N. Borana, T.S. Gaur, V. Yadav, Modeling of barriers to digital transformations in Indian manufacturing small and medium-sized enterprises, J. Sci. Tech. Policy Manage. (2024).

[26] S. Amoujavadi, A. Nemati, Developing sustainability, resiliency, agility, and security criteria for cloud service providers' viability assessment: a comprehensive hierarchical structure, Sustain. Fut. 7 (2024) 100219.

[27] L. Jum'a, M. Bushnaq, Investigating the role of flexibility as a moderator between supply chain integration and firm performance: the case of manufacturing sector, J. Adv. Manage. Res. 21 (2) (2024) 203–227.

[28] A. Rejeb, et al., Potentials of blockchain technologies for supply chain collaboration: a conceptual framework, Int. J. Logist. Manage. 32 (3) (2021) 973–994.

[29] M. Brookbanks, G. Parry, The impact of a blockchain platform on trust in established relationships: a case study of wine supply chains, Supp. Chain Manage. Int. J. 27 (7) (2022) 128–146.

[30] R. Kumar, et al., Prioritising elements of digitalisation for lean and green SME operations: an ISM-MICMAC study in the Indian context, J. Adv. Manage. Res. (2025).

[31] V. Sharma, R. Agrawal, V.K. Manupati, Blockchain technology as an enabler for digital trust in supply chain: evolution, issues and opportunities, Int. J. Syst. Assur. Eng. Manage. 15 (9) (2024) 4183–4209.

[32] R. D'Hauwers, J. Van Der Bank, M. Montakhabi, Trust, transparency and security in the sharing economy: what is the Government's role? Tech. Innov. Manage. Rev. 10 (5) (2020).

[33] P. Pietrzak, J. Takala, Digital Trust–Asystematic Literature Review, 2021.

[34] C. Chauhan, A. Singh, S. Luthra, Barriers to industry 4.0 adoption and its performance implications: an empirical investigation of emerging economy, J. Clean. Prod. 285 (2021) 124809.

[35] S. Kumar, M.K. Barua, Exploring the hyperledger blockchain technology disruption and barriers of blockchain adoption in petroleum supply chain, Resour. Policy 81 (2023) 103366.

[36] M. Hrouga, Towards a new conceptual digital collaborative supply chain model based on industry 4.0 technologies: a conceptual framework, Int. J. Qual. Reliab. Manage. 41 (2) (2023) 628–655.

[37] D. Ivanov, Digital supply chain management and technology to enhance resilience by building and using end-to-end visibility during the COVID-19 pandemic, IEEE Trans. Eng. Manage. (2021).

[38] H. Fatorachian, H. Kazemi, Impact of industry 4.0 on supply chain performance, Prod. Plan. Contr. 32 (1) (2021) 63–81.

[39] Global Risks Report 2024, WEF. 2024. p. 7–9.

[40] A. Patil, et al., Digital twins' Readiness and its Impacts on Supply Chain Transparency and Sustainable Performance, Industrial Management & Data Systems, 2024.

[41] L. Schilling, S. Seuring, Linking the digital and sustainable transformation with supply chain practices, Int. J. Prod. Res. 62 (3) (2024) 949–973.

[42] E. Pessot, et al., Empowering supply chains with industry 4.0 technologies to face megatrends, J. Busi. Logist. 44 (4) (2023) 609–640.

[43] Y. Lv, Y. Shang, Investigation of industry 4.0 technologies mediating effect on the supply chain performance and supply chain management practices, Environ. Sci. Poll. Res. 30 (48) (2023) 106129–106144.

[44] R. Preindl, K. Nikolopoulos, K. Litsiou, Transformation strategies for the supply chain: the impact of industry 4.0 and digital transformation, Supply Chain Forum 21 (1) (2020) 26–34.

[45] D.J. McAllister, Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations, Acad. Manage. J. 38 (1) (1995) 24–59.

[46] C.D. Duong, et al., Blockchain-based food traceability system and pro-environmental consumption: a moderated mediation model of technology anxiety and trust in organic food product, Digital Business 4 (2) (2024) 100095.

[47] K. Yavaprabhas, M. Pournader, S. Seuring, Blockchain and trust in supply chains: a bibliometric analysis and trust transfer perspective, Int. J. Prod. Res. 63 (14) (2025) 5071–5098.

[48] D.M. Rousseau, et al., Not so different after all: a cross-discipline view of trust, Acad. Manage. Rev. 23 (3) (1998) 393–404.

[49] R.C. Mayer, J.H. Davis, F.D. Schoorman, An integrative model of organizational trust, Acad. Manage. Rev. 20 (3) (1995) 709–734.

[50] W.K. Wong, et al., A framework for trust in construction contracting, Int. J. Proj. Manage. 26 (8) (2008) 821–829.

[51] Jason Challender, P. F, Peter McDermott, The theory of trust: concept, components, and characteristics. Building Collaborative Trust in Construction Procurement Strategies, 2019, pp. 37–54.

[52] C. Lin, M. Lin, The determinants of using cloud supply chain adoption, Indust. Manage. Data Syst. 119 (2) (2019) 351–366.

[53] R.M. Morgan, S.D. Hunt, The commitment-trust theory of relationship marketing, J. Mark. 58 (3) (1994) 20–38.

[54] D.H. McKnight, L.L. Cummings, N.L. Chervany, Initial trust formation in new organizational relationships, Acad. Manage. Rev. 23 (3) (1998) 473–490.

[55] M.F. Mubarak, M. Petraite, Industry 4.0 technologies, digital trust and technological orientation: what matters in open innovation? Technol. Forecast. Soc. Change 161 (2020) 120332.

[56] D. Dobrygowski, in: A.M. Assaf Ben-Atar, Amanda Stanhaus (Eds.), Earning Digital Trust: Decision-Making for Trustworthy Technologies, World Economic Forum, 2022.

[57] A.B.A. Daniel Dobrygowski, Augustinus Mohn, Amanda Stanhaus, Earning digital trust: decision-making for trustworthy technologies, World Economic Forum (2022).

[58] D. Treat, How to Build Trust in a New Digital World, Accenture, 2021.

[59] E. Boehm, Digital Trust in a Connected World: Navigating the State of IoT Security, KEYFACTOR, 2023. https://www.keyfactor.com/state-of-iot-security-report-2023/.

[60] ISACA, Digital Trust: A-Modern-Day-Imperative, 2022. https://www.isaca.org/resources/white-papers/digital-trust-a-modern-day-imperative.

[61] S.K. Sivarama Krishnan, Manu Dwivedi, The C-suite playbook: Putting security at the Epicentre of Innovation, PwC India, 2024.

[62] Y. Guo, Digital Trust and the reconstruction of trust in the Digital society: an integrated model based on trust theory and expectation confirmation theory, Digit. Govern. Res. Pract. 3 (4) (2022) 1–19.

[63] S. Han, J.P. Ulhøi, H. Song, Digital trust in supply chain finance: the role of innovative fintech service provision, J. Enterp. Inf. Manage. (2024) ahead-of-print(ahead-of-print).

[64] Q. Li, L. Wang, Research on the information sharing in the linkage between manufacturing and logistics industry based on blockchain, in: Journal of Physics: Conference Series, IOP Publishing, 2021.

[65] Y.M. Pfaff, H. Birkel, E. Hartmann, Supply chain governance in the context of industry 4.0: investigating implications of real-life implementations from a multi-tier perspective, Int. J. Prod. Econ. 260 (2023).

[66] R.K. Ray, F.R. Chowdhury, M.R. Hasan, Blockchain applications in retail cybersecurity: enhancing supply chain integrity, secure transactions, and data protection, J. Busi. Manage. Stud. 6 (1) (2024) 206–214.

[67] D. Ivanov, A. Dolgui, A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0, Prod. Plan. Contr. 32 (9) (2021) 775–788.

[68] A.E. Meafa, et al., Driving resiliency and digitalization in the sourcing process: integration of blockchain technology and smart contracts, Benchmark. Int. J. (2024).

[69] R. Raj, V. Kumar, B. Shah, Big data analytics adaptive prospects in sustainable manufacturing supply chain, Benchmark. Int. J. 31 (9) (2023) 3373–3397.

[70] R. Manzoor, B. Sahay, S.K. Singh, Examining the factors that facilitate or hinder the use of blockchain technology to enhance the resilience of supply chains, IEEE Trans. Eng. Manage. 71 (2024) 10626–10649.

[71] P. Roozkhosh, A. Pooya, R. Agarwal, Blockchain acceptance rate prediction in the resilient supply chain with hybrid system dynamics and machine learning approach, Oper. Manag. Res. (2022) 1–21.

[72] S. Yadav, S.P. Singh, Blockchain critical success factors for sustainable supply chain, Resour. Conserv. Recycl. 152 (2020) 104505.

[73] M. Shayganmehr, et al., Assessing the role of industry 4.0 for enhancing swift trust and coordination in humanitarian supply chain, Ann. Oper. Res. (2021).

[74] S. Lahane, R. Kant, Evaluating the circular supply chain implementation barriers using Pythagorean fuzzy AHP-DEMATEL approach, Clean. Logist. Supp. Chain 2 (2021) 100014.

[75] S. Luthra, S.K. Mangla, Evaluating challenges to industry 4.0 initiatives for supply chain sustainability in emerging economies, Process Saf. Environ. Prot. 117 (2018) 168–179.

[76] S. Pandey, et al., Cyber security risks in globalized supply chains: conceptual framework, J. Glob. Oper. Strat. Sour. 13 (1) (2020) 103–128.

[77] R. Agrawal, et al., Opportunities for disruptive digital technologies to ensure circularity in supply Chain: a critical review of drivers, barriers and challenges, Comput. Ind. Eng. (2023) 109140.

[78] S. Kumar, et al., Barriers to adoption of industry 4.0 and sustainability: a case study with SMEs, Int. J. Comput. Integr. Manuf. 36 (5) (2023) 657–677.

[79] D. T.S, V. Ravi, An ISM-MICMAC approach for analyzing dependencies among barriers of supply chain digitalization, J. Model. Manage. (2022).

[80] A. Mohammed, et al., Blockchain Adoption in Food Supply Chains: A Systematic Literature Review on Enablers, Benefits, and Barriers, IEEE Access, 2023.

[81] A. Chaouni Benabdellah, et al., Blockchain technology for viable circular digital supplychains: an integrated approach for evaluating the implementation barriers, Benchmarking (2023).

[82] O. Bak, A. Braganza, W. Chen, Exploring blockchain implementation challenges in the context of healthcare supply chain (HCSC), Int. J. Prod. Res. (2023) 1–16.

[83] A.K. Yadav, D. Kumar, Blockchain technology and vaccine supply chain: exploration and analysis of the adoption barriers in the Indian context, Int. J. Prod. Econ. 255 (2023) 108716.

[84] S. Khan, et al., Barriers to blockchain technology adoption in supply chains: the case of India, Oper. Manage. Res. (2023).

[85] P. Bottoni, et al., Intelligent smart contracts for innovative supply chain management, Front. Blockchain 3 (2020) 52.

[86] Y. Wu, Y. Zhang, An integrated framework for blockchain-enabled supply chain trust management towards smart manufacturing, Adv. Eng. Inform. 51 (2022) 101522.

[87] C. Colicchia, A. Creazza, D.A. Menachof, Managing cyber and information risks in supply chains: insights from an exploratory analysis, Supp. Chain Manage. Int. J. 24 (2) (2019) 215–240.

[88] V. Naumov, et al., Methodological principles of forming multichannel digital communication in the supply chains, in: E3S Web of Conferences, EDP Sciences, 2020.

[89] F.F. Rad, et al., Industry 4.0 and supply chain performance: a systematic literature review of the benefits, challenges, and critical success factors of 11 core technologies, Indust. Market. Manage. 105 (2022) 268–293.

[90] A. Jamwal, R. Agrawal, M. Sharma, Challenges and opportunities for manufacturing SMEs in adopting industry 4.0 technologies for achieving sustainability: empirical evidence from an emerging economy, Oper. Manage. Res. (2023) 1–26.

[91] V. Jain, P. Ajmera, J.P. Davim, SWOT analysis of industry 4.0 variables using AHP methodology and structural equation modelling, Benchmark. Int. J. 29 (7) (2022) 2147–2176.

[92] J.F. Hair, W.C. Black, B.J. Babin, RE Anderson Multivariate Data Analysis: A Global Perspective, Pearson Prentice Hall, New Jersey, 2010.

[93] J. Nunnally, Psychometric Methods, 1978.

[94] T.L. Saaty, The analytic hierarchy process (AHP), J. Oper. Res. Soc. 41 (11) (1980) 1073–1076.

[95] I. Otay, M. Jaller, A novel pythagorean fuzzy AHP and TOPSIS method for the wind power farm location selection problem, J. Intel. Fuzz. Syst. 39 (5) (2020) 6193–6204.

[96] E. Ilbahar, et al., A novel approach to risk assessment for occupational health and safety using Pythagorean fuzzy AHP & fuzzy inference system, Saf. Sci. 103 (2018) 124–136.

[97] A. Gabus, E. Fontela, World problems, an invitation to further thought within the framework of DEMATEL, 1, Battelle Geneva Research Center, Geneva, Switzerland, 1972, pp. 12–14.

[98] S.L. Si, et al., DEMATEL technique: a systematic review of the state-of-the-art literature on methodologies and applications, Math. Probl. Eng. 2018 (1) (2018) 3696457.

[99] M.H.H. Hemal, F. Parvin, A. Aziz, Analyzing the obstacles to the establishment of sustainable supply chain in the textile industry of Bangladesh, Bench Coun. Trans. Benchm. Stand. Eval. 4 (3) (2024) 100185.

[100] İ. Kaya, et al., An integrated Pythagorean fuzzy-based methodology for sectoral prioritization of industry 4.0 with lean supply chain perspective, Int. J. Comput. Integr. Manuf. (2024) 1–30.

[101] B.C. Giri, M.U. Molla, P. Biswas, Pythagorean fuzzy DEMATEL method for supplier selection in sustainable supply chain management, Expert. Syst. Appl. 193 (2022) 116396.

[102] M. Shafiee, et al., A causality analysis of risks to perishable product supply chain networks during the COVID-19 outbreak era: an extended DEMATEL method under pythagorean fuzzy environment, Transp. Res. Part E Logist. Transp. Rev. 163 (2022) 102759.

[103] N. Agarwal, N. Seth, Analysis of supply chain resilience barriers in Indian automotive company using total interpretive structural modelling, J. Adv. Manage. Res. 18 (5) (2021) 758–781.

[104] G.M. Razak, L.C. Hendry, M. Stevenson, Supply chain traceability: a review of the benefits and its relationship with supply chain resilience, Prod. Plan. Cont. 34 (11) (2023) 1114–1134.

[105] J. Rahman, et al., Regulatory landscape of blockchain assets: analyzing the drivers of NFT and cryptocurrency regulation, Bench Coun. Trans. Benchmark. Stand. Eval. (2025) 100214.

[106] D. Kalaitzi, N. Tsolakis, Supply chain analytics adoption: determinants and impacts on organisational performance and competitive advantage, Int. J. Prod. Econ. 248 (2022) 108466.

[107] K. Huang, et al., The impact of industry 4.0 on supply chain capability and supply chain resilience: a dynamic resource-based view, Int. J. Prod. Econ. 262 (2023) 108913.

[108] A. Caliskan, S. Eryilmaz, Y. Ozturkoglu, Investigating the effects of barriers and challenges on Logistics 4.0 in the era of evolving digital technology, J. Model. Manage. 20 (3) (2025) 949–973.

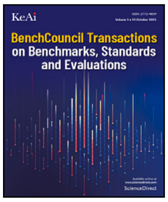
[109] M. Akbari, J.L. Hopkins, Digital technologies as enablers of supply chain sustainability in an emerging economy, Oper. Manage. Res. 15 (3) (2022) 689–710.

[110] W. Viriyasitavat, et al., Building trust of blockchain-based internet-of-thing services using public key infrastructure, Enterp. Inf. Syst. 16 (12) (2022) 2037162.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Full length article

An evaluation framework for measuring prompt wise metrics for large language models in resource-constrained edge

Partha Pratim Ray *, Mohan Pratap Pradhan

Department of Computer Applications, Sikkim University, Gangtok, Sikkim, India

ARTICLE INFO

Keywords:
Large language model
LLM evaluation
LLM metrics
Evaluation framework
Quantized LLMs
Resource-constrained environment

ABSTRACT

Existing challenges in deploying large language models (LLMs) on resource-constrained devices stem from limited CPU throughput, memory capacity, and power budgets. Motivated by the lack of edge-specific evaluation tools, we introduce **LLMEvaluator**, a framework that profiles quantized LLMs — Qwen2.5, Llama3.2, Smollm2, and Granite3 — on a Raspberry Pi 4B using a suite of core and derived metrics. Our contributions include (i) a unified taxonomy that integrates latency, throughput, power variation, memory stability, and thermal behavior; (ii) prompt-wise analyses across ten NLP tasks; and (iii) correlation studies guiding optimizations. Key results show that Qwen2.5 leads in energy efficiency and throughput with a 68.44 MB memory standard deviation; Granite3 excels in memory stability , minimal load overhead, and per-token latency; Smollm2 suffers the highest total duration, longest prompt overhead, and lowest power efficiency; and Llama3.2 balances latency, throughput (8.12 tokens/s), and energy per token with moderate power variability (1.05 W std dev). Correlation analysis reveals that reducing model load time yields the largest improvement in end-to-end latency ($r > 0.9$), and that throughput gains directly translate into energy savings ($r \approx -0.81$). **LLMEvaluator** empowers selection and tuning of LLMs for low-power environments.

1. Introduction

LLMs have become central to a wide range of applications, from complex natural language processing tasks to conversational AI [1–3]. However, deploying these models on hardware with strict resource limits — such as the Raspberry Pi 4B or other CPU-only platforms — poses significant challenges [4,5]. Specifically, edge devices must contend with high computational and memory requirements, variable power draw, and tight latency constraints [6,7]. Conventional evaluation methods, originally developed for data centers and high-performance clusters, often fail to account for the dynamic power fluctuations and memory ceilings characteristic of edge environments [8–10]. As a result, there is an urgent demand for specialized frameworks that generate actionable performance insights under these localized conditions [11,12].

Simultaneously, the rapid adoption of edge computing brings AI workloads closer to data sources, reducing latency and improving privacy [13,14]. Yet this shift also introduces new obstacles for large-scale models [15–17]. Existing benchmarks typically ignore critical edge-specific factors — limited CPU headroom, constrained onboard memory, and the imperative for energy efficiency — leading to suboptimal deployments and scalability issues on low-power devices [18–21]. To address these shortcomings, a robust evaluation framework tailored to edge requirements is essential, ensuring LLMs remain both practical and efficient under resource constraints [22–26].

The aims of this study are threefold such as (i) to design a framework that evaluates LLMs using metrics specific to edge environments, (ii) to measure and quantify the resource consumption and efficiency of LLMs on constrained hardware, and (iii) to introduce novel metrics to provide a deeper, more actionable performance assessment.

To this end, we present LLMEvaluator, a modular framework that unifies power efficiency, memory utilization, and CPU stability metrics to illuminate the trade-offs between computational throughput and resource usage. Validated on quantized versions of Qwen2.5, Llama3.2, Smollm2, and Granite3 running on a Raspberry Pi 4B, LLMEvaluator demonstrates its flexibility across diverse CPU-only platforms. In addition to traditional throughput and accuracy measures, our framework proposes new metrics that capture energy per token and thermal load behavior during inference, thereby bridging the gap between model-centric and system-centric evaluations. This comprehensive approach enables practitioners to optimize model selection, fine-tune hardware configurations, and enhance the reliability of LLM-based services in edge deployments. LLMEvaluator's potential extends across both industrial and research domains [27–30]. For example, in manufacturing automation, it can assess LLM performance for predictive

* Corresponding author.
E-mail address: ppray@cus.ac.in (P.P. Ray).

https://doi.org/10.1016/j.tbench.2025.100249
Received 14 May 2025; Received in revised form 28 November 2025; Accepted 8 December 2025
Available online 11 December 2025

maintenance under strict power and latency constraints, while in healthcare it can validate real-time translation models on portable diagnostic tools without exceeding available resources.

The major contributions of this work can be summarized as follows:

- An extensible framework that unifies model-centric and system-centric assessments for edge deployments by integrating accuracy metrics with CPU/memory utilization and power-draw variability.
- Introduction of various novel metrics to characterize the complex interactions between LLM computation and constrained hardware resources.
- Implementation of evaluations across diverse task categories — from general knowledge to mathematical reasoning — paired with correlation analyses that expose metric interdependencies and guide targeted optimizations.
- Deployment on a Raspberry Pi 4B using quantized Qwen2.5, Llama3.2, Smollm2, and Granite3 models, uncovering distinct trade-offs: Qwen2.5 leads in throughput and energy efficiency, Granite3 excels in memory stability, and Smollm2 exhibits the highest resource consumption.

The remainder of this paper is structured as follows. Section 2 reviews related work and identifies existing gaps. Section 3 details the architecture and implementation of LLMEvaluator framework. Section 4 describes the curated prompt set for evaluating model behavior across NLP tasks. Section 5 presents the LLMEvaluator algorithm. Section 6 defines our taxonomy of core, resource-aware, and derived metrics. Section 7 showcases experimental results, including correlation and prompt-wise analyses. Section 8 presents elaborative discussion of this work. Finally, Section 9 concludes with future research directions, emphasizing heterogeneous platform support and sustainability metrics in resource-constrained edge.

2. Related works

A number of benchmarks and evaluation frameworks have emerged to assess the capabilities of large language models (LLMs), yet few address the stringent requirements of edge deployments. Wang et al. [31] introduced PandaLM, which emphasizes instruction-tuning quality by measuring aspects such as brevity, clarity, and conformity to prompts. Leveraging human-annotated datasets and a trained "judge" LLM, PandaLM aligns model outputs with user expectations while avoiding external API calls. However, its scope is limited to optimizing hyperparameters and instruction-following in general-purpose LLMs, without considering energy consumption or memory constraints characteristic of edge hardware.

In the domain of process mining, Berti et al. [32] developed PM-LLM-Benchmark to evaluate LLM proficiency on specialized mining tasks. Their study demonstrated that smaller, edge-suitable architectures often underperform on these complex workflows, highlighting the absence of lightweight evaluation tools tailored to constrained environments. Similarly, Zhou et al. [33] proposed ElecBench for power dispatch scenarios, combining six high-level metrics (factuality, logicality, stability, security, fairness, expressiveness) and 24 submetrics. While ElecBench provides a comprehensive view of decision-making quality in the energy sector, it presumes access to abundant computational resources and overlooks edge-specific efficiency requirements.

Multimodal benchmarks have also been put forth, though chiefly for data-center settings. Xu et al. [34] presented LVLM-eHub to assess vision–language models on tasks such as visual question answering and hallucination detection, and Sun et al. [35] introduced SciEval for scientific research applications, employing dynamic subsets to prevent data leakage. Both frameworks deliver rigorous quantitative and qualitative analyses but omit measurements of power draw, thermal behavior, and memory footprint. Mobile-Bench by Deng et al. [36] evaluates LLM-powered mobile agents with novel planning metrics, yet does not prioritize computational optimization. Conversely, Yang et al. [37]'s LLMCBench focuses on compression techniques to reduce model size, contributing valuable insights into efficiency gains but stopping short of examining thermal or power stability.

Additional efforts have explored LLMs as evaluators or in multilingual and planning contexts. Chen et al. [38]'s MLLM-as-a-Judge framework addresses hallucination and bias in multimodal evaluations, while Spangher et al. [39]'s Project MPG merges accuracy and cost into aggregate scores. Son et al. [40]'s MM-Eval covers performance across 18 languages, and Valmeekam et al. [41]'s PlanBench examines planning and reasoning capabilities. In recommendation systems, Liu et al. [42] introduced LLMRec for explainability tasks, and Liu et al. [43] extended safety evaluations with MM-SafetyBench. Chu et al. [44]'s PRE uses peer review to reduce evaluator bias, and Yu et al. [45]'s KoLA benchmarks world-knowledge across 19 tasks with a novel self-contrast metric. Although each contributes important evaluation perspectives, none systematically measure the power, memory, and thermal constraints critical to edge computing. Table 1 compares our proposed work with existing literature.

These prior works exhibit increasing sophistication in LLM assessment but consistently neglect edge-specific concerns such as limited CPU throughput, memory ceilings, and energy efficiency. Our LLMEvaluator framework addresses this gap by targeting resource-constrained environments with a lightweight design, support for quantized LLMs, and CPU-optimized inference. We introduce localized metrics —
including power usage variation, thermal load factor, and sustained inference efficiency — to provide a comprehensive, actionable evaluation tailored to low-power edge platforms.

Limitations of Existing Works

Despite the wealth of LLM benchmarking tools, current frameworks share several critical shortcomings when it comes to edge deployment. First, they assume abundant compute and memory resources, making them unsuitable for CPU-only single-board computers with strict power and thermal budgets. Second, they focus almost exclusively on aggregate throughput or quality scores, ignoring transient behaviors in power draw, memory usage spikes, and prompt-specific latency that dominate real-world edge scenarios. Third, existing benchmarks treat inference as a monolithic operation, failing to separate model loading, prompt processing, and token generation—each of which can present a unique bottleneck on constrained hardware. Finally, there is no facility for prompt-wise analysis across diverse task types, so practitioners cannot tailor model selection or optimization strategies to the particular mix of queries their applications will encounter.

Novelty of LLMEvaluator

LLMEvaluator overcomes these gaps by introducing a fine-grained, resource-aware evaluation methodology expressly designed for low-power, CPU-only environments. It unifies a rich taxonomy of core performance metrics (load time, token throughput), system-level telemetry (real-time CPU, RAM, power sampling), and derived efficiency indices (energy per token, thermal load factor) into a prompt-wise analysis pipeline. By decomposing inference into loading, prompt ingestion, and generation phases and correlating each phase's behavior with dynamic resource usage, LLMEvaluator delivers actionable insights that no existing tool provides. Its modular, extensible design supports quantized model evaluation on single-board computers like the Raspberry Pi 4B — and can be readily ported to other edge platforms — empowering practitioners to make informed trade-offs between latency, energy, and stability in production deployments.

3. LLMEvaluator framework

Fig. 1 illustrates the end-to-end architecture of LLMEvaluator, a purpose-built framework that unites a lightweight web front-end, an asynchronous request dispatcher, an on-device inference engine, and

Table 1
Comparison of existing benchmarks and LLMEvaluator

Benchmark	Year	Domain/Task	Core metrics	Edge-aware	Key Findings/Limitations	Edge support
PandaLM [31]	2023	Instruction tuning	Correctness; brevity; clarity; instruction adherence; comprehensiveness; formality	No	Matches 93.8% of GPT-3.5 and 88.3% of GPT-4 on F1, but omits power and memory considerations.	No
PM-LLM-Benchmark [32]	2024	Process mining	Domain-specific accuracy; implementation strategy success	No	Demonstrates tiny models underperform on PM tasks; does not track resource usage.	No
ElecBench [33]	2024	Power dispatch	Factuality; logicality; stability; security; fairness; expressiveness (6); 24 submetrics	No	Evaluates eight LLMs on sector scenarios; assumes high compute; ignores energy/memory limits.	No
LVLM-eHub [34]	2024	Vision–language multimodal	VQA accuracy; object hallucination rate; user-level arena score	No	Benchmarks 13 LVLMs quantitatively and via online arena; no power or thermal profiling.	No
SciEval [35]	2024	Scientific research	Bloom's taxonomy (4 levels); objective vs subjective Q/A; dynamic subsets	No	Prevents data leakage, includes subjective items; lacks resource-use measurement.	No
Mobile-Bench [36]	2024	Mobile agents	Multi-app collaboration tasks; planning complexity tiers; CheckPoint metric	No	Covers 200+ tasks with UI API augmentation; does not address edge compute limits.	No
LLMCBench [37]	2024	Model compression	Compression ratio; accuracy retention; latency	No	Analyzes compression methods across LLMs; overlooks power and thermal dynamics.	No
MLLM-as-a-Judge [38]	2024	Multimodal evaluation	Scoring accuracy; pairwise comparison; batch ranking	No	Introduces judge LLMs for vision–language tasks; not optimized for low-power devices.	No
Project MPG [39]	2024	Aggregate performance	"Goodness" (accuracy); "Fastness" (QPS/cost)	No	Aggregates across benchmarks into two scores; lacks resource-aware granularity.	No
MM-Eval [40]	2024	Meta-evaluation	Multilingual judge reliability across 18 languages	No	Reveals evaluator bias in low-resource languages; does not track edge resource usage.	No
PlanBench [41]	2024	Planning/ reasoning	Domain diversity; plan generation success	No	Tests LLMs on classical planning tasks; ignores memory and energy footprints.	No
LLMRec [42]	2023	Recommendation	Rating prediction; sequential recommendation; explainability	No	Shows moderate accuracy in recommendation tasks; does not measure efficiency.	No
MM-SafetyBench [43]	2025	Safety (MLLMs)	Vulnerability detection; image–text manipulation resilience	No	Exposes security flaws in MLLMs; omits performance and power metrics.	No
PRE [44]	2024	Peer-review evaluation	Reviewer selection accuracy; aggregated rankings	No	Automates peer-review style evaluation; does not account for compute constraints.	No
KoLA [45]	2023	World knowledge	Taxonomy-based knowledge tasks (19); self-contrast score	No	Evaluates breadth of world knowledge; does not include resource-use criteria.	No
This Work	2025	Edge deployment	Latency; tokens/sec; CPU/memory usage; power draw; energy/token; thermal load	Yes	Assesses Qwen2.5, Llama3.2, Smollm2, Granite3 on Raspberry Pi 4B; introduces localized, resource-aware metrics.	Yes

a comprehensive metric scoring module into a single, streamlined pipeline. This design has been carefully crafted for resource-sensitive edge platforms — such as the Raspberry Pi 4B, NVIDIA Jetson Nano, NanoPi, Tinker Board, and similar single-board computers — where every CPU cycle, watt of power, and megabyte of RAM is precious.

At the very top of the stack, end users or client applications send RESTful HTTP POST requests containing prompt payloads, formatted in JSON. These requests are directed at a Flask-based web application configured with a single /api/endpoint. As soon as the first request arrives, Flask triggers a startup routine — implemented via a "startup event" hook — that loads quantized LLM instances (for example, Qwen2, Qwen2.5, Llama3.2, Granite3, and Smollm2) into memory and spins up lightweight monitoring threads. Once initialization is complete, Flask passes control over to Uvicorn, an ASGI-compliant server that sits behind the scenes. leveraging the asgiref bridge, LLMEvaluator ensures that Flask's synchronous WSGI interface can coexist seamlessly with Uvicorn's fully asynchronous event loop. In practical terms, this means dozens of simultaneous inference requests can be handled without blocking, even on a single-core CPU.

When a prompt is routed to the inference engine, the local model hub — essentially a directory of preloaded, quantized model binaries — selects the appropriate LLM based on either a prompt-category mapping or developer-specified configuration. The chosen model then generates text on-device, invoking efficient quantized matrix-multiplication kernels and sparse expert-routing logic (in the case of Granite3) as needed. Concurrently, a dedicated psutil monitoring thread samples system-level telemetry at configurable intervals (e.g. every 50–200 ms). These samples capture instantaneous CPU utilization, RAM consumption, and an estimated power draw derived from a calibrated CPU-power curve.

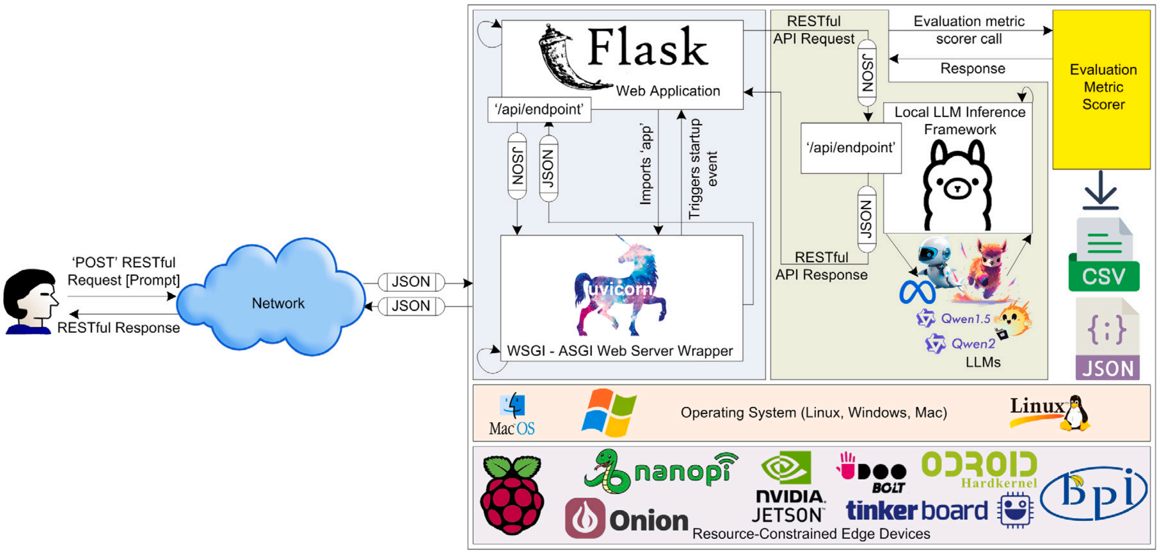


Fig. 1. LLMEvaluator framework.

This sampling logic directly attaches into the inference loop, LLMEvaluator avoids the need for external measurement hardware or disruptive kernel modules.

Once generation concludes, the framework collates the raw textual output and the time-series of resource samples into a single batch, which is then handed off to the Evaluation Metric Scorer. This component processes the timestamped logs and `psutil` readings to compute a suite of core performance metrics—total end-to-end latency, tokens-per-second throughput, and average power consumption. It then derives a richer set of edge-aware indices: power efficiency (tokens/W), memory efficiency (tokens/MB), CPU stability index (1 − normalized CPU usage variance), thermal load factor (average CPU load divided by average power draw), and memory-to-power ratio (average RAM divided by average power). In addition, peak, minimum, and standard-deviation values for CPU, memory, and power are recorded to surface transient spikes or volatility. To ensure easy downstream consumption, all computed metrics are serialized both to CSV (ideal for spreadsheet analyses) and JSON (suitable for dashboards or automated reporting pipelines).

Under the hood, all inter-module HTTP calls — whether between Flask and the inference service, or between the scorer and external APIs — are driven by the `requests` library, which provides robust retry mechanisms, connection pooling, and fine-grained timeout controls. This means LLMEvaluator can be readily extended: developers might plug in new REST endpoints to serve bespoke data, implement custom metric calculators for domain-specific KPIs, or swap in alternative model back ends without touching the core codebase. The framework's clear separation of concerns — web layer, inference layer, and scoring layer — ensures that optimizations or feature additions in one area have minimal impact on the others.

In real-world trials, LLMEvaluator has been stress-tested on a Raspberry Pi 4B using a 1 billion-parameter quantized LLM, sustaining up to 12 tokens/s at under 2 W of average draw for Qwen2.5. Further scalability tests on Jetson Nano and Rockchip boards confirmed the system's ability to handle concurrent inference workloads with virtually no extra overhead. Thus a lightweight web front-end, precise resource instrumentation, and a rich taxonomy of metrics, LLMEvaluator stands as a comprehensive, extensible solution for rigorously assessing LLM performance in environments where efficiency is not just beneficial—it is essential.

4. Prompts used

In order to probe the capabilities of various localized LLMs under constrained conditions, our framework employs ten carefully selected prompt categories, each designed to mirror a distinct real-world application in natural language processing whereby spanning tasks from factual retrieval to creative composition, this suite of prompts ensures a comprehensive evaluation of both core language understanding and the specialized skills required for edge deployment.

The first category, general knowledge, includes straightforward fact-recall questions (e.g. "What is the capital of France?"). These prompts validate the model's ability to access and return accurate encyclopedic information, serving as a baseline for entity recognition and retrieval accuracy. Mathematical queries (such as "What is the square root of 256?") extend this by examining numerical reasoning and exactitude, ensuring that the model can perform basic computations without external tooling.

Summarization prompts ask the model to condense longer passages — "Summarize the following text: The industrial revolution was a period of major industrialization. . . " — testing its capacity to distill salient points while preserving coherence and logical flow. In parallel, sentiment analysis tasks like "Classify the sentiment: 'I absolutely loved the new restaurant!'" evaluate the model's grasp of subjective nuance and its ability to categorize emotional tone accurately.

To assess generative versatility, we introduce creative writing tasks ("Write a short poem about the beauty of nature.") which require the LLM to produce original, expressive content, and conversational scenarios ("Pretend to be a travel assistant. Suggest some attractions in Paris for a family vacation.") that probe dialogue management, context retention, and user engagement. These categories reveal how models craft fluent, contextually relevant text across diverse registers.

Domain-specific prompts further explore practical utility. Coding assistance requests, such as "Write a Python function to calculate the factorial of a number.", gauge the model's knowledge of programming constructs, debugging strategies, and syntactic accuracy. Edge device suitability queries — for example, "Explain the benefits of Raspberry Pi in IoT applications." — test the LLM's proficiency in technical explanations tailored to resource-constrained environments, highlighting its ability to offer implementable guidance on hardware selection and deployment.

Translation tasks ("Translate to French: 'Good morning, how are you?'") measure multilingual competence and syntactic fidelity, ensuring that the LLM can perform accurate cross-lingual mapping without sacrificing grammar or idiomatic expression. Finally, text completion exercises ("Complete this sentence: The quick brown fox jumps

Table 2
Prompts used in the LLMEvaluator framework.

Prompt Category	Prompt ID	Example Prompt	Selection Criteria	Role in the Study Assessment
General Knowledge	1	*"What is the capital of France?"*	Evaluates the model's ability to provide accurate factual information.	Tests fundamental NLP capabilities like entity recognition and retrieval of general knowledge.
Summarization	2	*"Summarize the following text: The industrial revolution was a period of major industrialization..."*	Represents tasks requiring compression and synthesis of information.	Assesses text understanding, coherence, and brevity in summarizing large inputs.
Creative Writing	3	*"Write a short poem about the beauty of nature."*	Explores generative capabilities and creativity.	Examines the model's ability to produce human-like, imaginative, and contextually relevant content.
Sentiment Analysis	4	*"Classify the sentiment: 'I absolutely loved the new restaurant!'"*	Targets classification and emotional comprehension in text.	Measures the model's precision in understanding and categorizing sentiments.
Text Completion	5	*"Complete this sentence: The quick brown fox jumps over..."*	Tests the ability to continue incomplete textual inputs logically.	Evaluates contextual understanding and language modeling for seamless text generation.
Translation	6	*"Translate to French: 'Good morning, how are you?'"*	Examines multilingual understanding and grammar proficiency.	Assesses the model's effectiveness in cross-lingual tasks and syntactic accuracy.
Coding Assistance	7	*"Write a Python function to calculate the factorial of a number."*	Represents domain-specific assistance in programming.	Tests practical utility in writing, debugging, and optimizing code snippets.
Edge Device Suitability	8	*"Explain the benefits of Raspberry Pi in IoT applications."*	Focuses on IoT-related technical queries relevant to resource-constrained environments.	Evaluates domain-specific expertise and model applicability to edge scenarios.
Mathematical Queries	9	*"What is the square root of 256?"*	Targets computational reasoning and precision.	Tests the model's capability to handle arithmetic and mathematical problem-solving.
Conversational	10	*"Pretend to be a travel assistant. Suggest some attractions in Paris for a family vacation."*	Simulates real-world interaction scenarios.	Measures adaptability, relevance, and user engagement in dialogue-based tasks.

over...") focus on next-token prediction in incomplete contexts, providing insight into the model's internal language modeling and its seamless continuation of prompts.

By integrating these ten categories, LLMEvaluator captures a holistic view of model performance across factual retrieval, reasoning, synthesis, creative generation, and domain-specific expertise. Each prompt category contributes unique diagnostic information: core metrics such as response latency and tokens per second uncover throughput characteristics, while resource-aware measurements (CPU usage, memory footprint, and power draw) reveal the operational cost associated with each task type. Derived indicators — power efficiency (tokens/W), memory efficiency (tokens/MB), and the thermal load factor — further illuminate the trade-offs between computational demands and resource consumption for each prompt category.

This diversified prompt suite ensures that model evaluations reflect the full spectrum of challenges encountered in edge computing scenarios. Table 2 details the example prompts, selection criteria, and the specific role each category plays in the overall assessment strategy, enabling researchers to pinpoint strengths, identify weaknesses, and guide targeted optimizations for deploying LLMs on devices with stringent resource constraints.

In our semantic audit of the ten prompt categories, we identified seven core clusters — Factual Retrieval, Text Transformation, Free-Form Generation, Classification/Sentiment, Cross-Lingual Mapping, Code Synthesis, and Technical Explanation — that collectively span the major axes of language model capability. *Factual Retrieval* (Prompts 1 and 9) demands concise, single-fact responses, while *Text Transformation* (2 and 5) reshapes existing inputs via summarization or completion. *Free-Form Generation* (3 and 10) exercises the model's creative and conversational fluency. Unique standalone tasks include sentiment classification (4), translation (6), programming assistance (7), and domain-specific explanation (8).

Overlap emerges at the boundaries: summarization and completion both manipulate text structure, and poetic versus dialogic outputs share generative mechanics. However, gaps remain—there is no intent-detection prompt to classify tone or politeness, no structured-data interpretation challenge (e.g., analyzing a CSV snippet), and no robustness test for unanswerable queries (e.g., "What is the square root of −1 in real numbers?") (see Table 3).

Our qualitative semantic similarity matrix quantifies these relationships, marking high similarity within clusters (H), moderate overlap at cluster interfaces (M), and low similarity across distinct tasks (L) (see Table 4). This landscape confirms broad linguistic coverage while pinpointing areas for extension.

5. LLMEvaluator framework algorithm

LLMEvaluator adopts a structured algorithm to deliver end-to-end benchmarking of localized LLMs within edge environments, where computational power, memory, and energy are severely limited. The framework takes as input a collection of user-defined prompts $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$, a quantized LLM instance $\mathcal{M}$, and a resource monitoring component R that samples system metrics at regular intervals Δt. Its goal is to produce a comprehensive set of evaluation results $\mathcal{E} = \{e_1, e_2, \dots, e_n\}$, each e_i synthesizing raw performance logs with derived efficiency indicators.

At the core of this methodology lies the metric computation function $\mathcal{F}$, which encapsulates the conversion of prompt, response, and

Table 3
Semantic audit of prompt set.

Prompt IDs	Semantic Cluster	Representative Prompts	Notes on Overlap/Uniqueness
1, 9	Factual Retrieval	“What is the capital of France?” (1) “What is the square root of 256?” (9)	Both ask for precise, single-fact answers.
2, 5	Text Transformation	“Summarize …” (2) “Complete this sentence …” (5)	Both reshape or extend existing text; differ in direction (condense vs. continue).
3, 10	Free-form Generation	“Write a short poem …” (3) “Pretend to be a travel assistant …” (10)	Both require creative, open-ended output.
4	Classification/Sentiment	“Classify the sentiment …” (4)	Pure classification—unique among the set.
6	Cross-lingual Mapping	“Translate to French …” (6)	Translation is its own semantic axis.
7	Code Synthesis/Assistance	“Write a Python function …” (7)	Domain-specific (programming), no direct overlap.
8	Technical Explanation (Edge)	“Explain the benefits of Raspberry Pi …” (8)	Domain knowledge + explanation—unique blend of semantics.

resource data into a unified evaluation record. Formally:

$$\mathcal{F}(p_i, r_i, m_i) \longrightarrow e_i, \tag{1}$$

where $r_i = \mathcal{M}(p_i)$ represents the text generated by the model, and $m_i = R(p_i, \Delta t)$ denotes the time-series measurements (CPU usage, memory footprint, inferred power draw) collected during inference. Each e_i comprises:

- core LLM metrics: total inference latency, tokens per second, token count;
- edge-aware resource metrics: average and peak CPU utilization, average and peak RAM usage, power consumption statistics (mean, variance, peaks);
- derived efficiency indices: energy per token (E/T), power efficiency index (tokens/W), CPU stability index, thermal load factor, and memory-to-power ratio.

To assemble the full evaluation suite, LLMEvaluator aggregates per-prompt metrics as follows:

$$\mathcal{E} \;=\; \bigcup_{i=1}^{n} \mathcal{F}\big(p_i,\; r_i,\; R(p_i, \Delta t)\big). \tag{2}$$

This union of results enables cross-prompt comparisons, revealing patterns of trade-offs between speed, resource consumption, and stability under varying task complexities.

We can try uniting model-centric measures (e.g. response latency, throughput) with system-centric observations (e.g. CPU cycles, memory pressure, power draw), the algorithm delivers a holistic profile of an LLM’s edge performance. The inclusion of derived metrics such as energy per token and power efficiency underscores the emphasis on energy-aware deployment, critical for battery-powered or thermally constrained hardware. Algorithm 1 outlines the operational steps in detail:

Inputs

- $\mathcal{P} = \{p_1, \dots, p_n\}$: a suite of test prompts spanning general knowledge, summarization, coding assistance, and more.
- $\mathcal{M}$: the localized, quantized LLM under evaluation (e.g. Qwen2.5, Llama3.2, Smollm2, Granite3).
- $R(p_i, \Delta t)$: the monitoring routine that samples CPU, memory, and power at each Δt interval during inference of p_i.

Outputs

- $\mathcal{E} = \{e_1, \dots, e_n\}$: the collection of evaluation records for each prompt.

Algorithm 1 LLMEvaluator Framework

```
Inputs: Prompt set $\mathcal{P}$, model $\mathcal{M}$, resource monitor $R$, sampling interval $\Delta t$
Output: Evaluation set $\mathcal{E}$
Initialize $\mathcal{E} \leftarrow \emptyset$
Activate resource monitor $R$
for all $p_i \in \mathcal{P}$ do
    $r_i \leftarrow \mathcal{M}(p_i)$    // Obtain model response
    $m_i \leftarrow R(p_i, \Delta t)$    // Record CPU, memory, power metrics
    $e_i \leftarrow \mathcal{F}(p_i, r_i, m_i)$    // Compute combined metrics
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{e_i\}$
end for
Deactivate resource monitor $R$
Persist $\mathcal{E}$ as CSV and JSON files
return $\mathcal{E}$
```

Table 4
Qualitative prompt-pair semantic similarity matrix.

Prompt ID	1	2	3	4	5	6	7	8	9	10
1	–	L	L	L	L	M	L	L	H	L
2	L	–	M	L	H	L	L	L	L	M
3	L	M	–	L	M	L	L	L	L	H
4	L	L	L	–	L	L	L	L	L	L
5	L	H	M	L	–	L	L	L	L	M
6	M	L	L	L	L	–	L	L	M	L
7	L	L	L	L	L	L	–	L	L	L
8	L	L	L	L	L	L	L	–	L	L
9	H	L	L	L	L	M	L	L	–	L
10	L	M	H	L	M	L	L	L	L	–

- Each e_i includes:
  1. Core performance metrics: inference time, token throughput.
  2. Resource usage metrics: CPU/RAM averages and peaks, power consumption stats.
  3. Derived efficiency indices: energy per token, power efficiency, stability, thermal and memory-to-power ratios.

Overall, this algorithm provides a repeatable, extensible procedure for measuring how LLMs perform when deployed on edge devices, guiding both model selection and system-level optimization for energy- and resource-sensitive applications.

Metric computation

At the heart of LLMEvaluator lies the computation function $\mathcal{F}$, which transforms each prompt, its resulting output, and the corresponding system measurements into a unified evaluation record. Formally, for a given prompt p_i, model response r_i, and resource trace m_i, the computation is defined by

$$\mathcal{F}(p_i, r_i, m_i) \longrightarrow e_i, \tag{3}$$

where:

- p_i denotes the ith input prompt drawn from the test suite.
- r_i represents the sequence of tokens generated by the LLM $\mathcal{M}$.
- m_i comprises the time-series of system metrics (CPU utilization, memory usage, power draw) sampled during inference.

The output e_i bundles three categories of metrics:

(1) *Core performance metrics*: including total inference time td_i, token throughput tps_i, and token count ec_i.
(2) *Edge resource usage metrics*: such as average and peak CPU percentages (acup_i, pcu_i), average and peak RAM usage (aru_i, pru_i), and power statistics (ap_i, pp_i, mp_i, psd_i).
(3) *Derived efficiency indices*: for example, energy per token ept_i, power efficiency index pei_i, thermal load factor tlf_i, and memory-to-power ratio mtpr_i.

To obtain a global view of the model's behavior over the entire prompt set $\mathcal{P}$, LLMEvaluator aggregates these individual records via

$$\mathcal{F}(p_i, r_i, m_i) \longrightarrow e_i, \tag{4}$$

$$\mathcal{E} = \bigcup_{i=1}^{n} \mathcal{F}\big(p_i,\ r_i,\ R(p_i, \Delta t)\big). \tag{5}$$

This union of e_i entries yields the complete metric set $\mathcal{E}$, which encapsulates performance, stability, and energy characteristics across all test scenario whereby examining $\mathcal{E}$, practitioners can identify patterns of resource bottlenecks, energy hotspots, and throughput limitations, guiding both model refinement and hardware configuration for edge deployments.

6. Proposed LLM evaluation metrics

To capture the multifaceted behavior of localized LLMs operating on edge hardware, we define a rich taxonomy of evaluation metrics grouped into three categories: (i) core LLM evaluation metric as inspired from Ollama inference framework, (ii) edge-aware resource usage metrics, and (iii) derived efficiency metrics (Fig. 2). Each metric is precisely formulated to illuminate distinct aspects of inference performance, system load, and energy utilization, thereby enabling a holistic assessment.

At the heart of LLMEvaluator lies a suite of core metrics that quantify the raw computational performance of a language model, abstracting away system-level noise and focusing squarely on the efficiency of the model's own inference pipeline. Total duration measures the end-to-end latency — from the moment weights begin loading to the final token emission — expressed in nanoseconds. This captures both initialization and generation overheads in a single, comprehensive timespan. To disentangle these constituents, load duration isolates the cost of deserializing model parameters, building tokenizer vocabularies, and allocating memory buffers, while prompt evaluation duration quantifies the time spent parsing, tokenizing, and embedding the input text before any token is generated. Once the prompt is fully ingested, the evaluation duration strictly tracks the forward-pass execution responsible for generating new tokens. Finally, evaluation count provides the absolute number of tokens produced under a fixed stopping criterion, and tokens per second (throughput) normalizes that count by the generation interval. Together, these six metrics form a rigorous baseline—a compute-centric "fingerprint" of each LLM's intrinsic speed and verbosity. They allow researchers to directly compare how different quantization schemes, sparsity patterns, or architectural variants affect latency and throughput on identical hardware, unencumbered by fluctuations in CPU scheduling or power management.

While core metrics reveal a model's computational profile, deploying LLMs on constrained edge devices demands an understanding of their real-world hardware footprint. LLMEvaluator's edge-resource-aware metrics accomplish this by instrumenting system telemetry throughout inference. Average and peak CPU utilization record the sustained and maximum processor loads, highlighting potential contention with background tasks or thermal throttling triggers. Average and peak RAM usage reveal a model's working memory requirements and the magnitude of transient allocation spikes—critical for avoiding out-of-memory failures on platforms with limited DRAM. Power consumption is characterized by average, peak, and minimum wattage, augmented by power standard deviation, which measures draw variability, and power usage variation index, a dimensionless ratio of fluctuation to mean power. Similarly, memory standard deviation and memory variation index signal instability in RAM footprint. Derived directly from these measurements are compound indicators such as the thermal load factor — the ratio of CPU load to power draw — which approximates heat generation per watt, and the time-weighted power factor, which normalizes average power by total inference time to capture sustained energy demands. Metrics like evaluation memory efficiency (tokens per second per megabyte) and average CPU-to-power ratio (percent CPU per watt) quantify how effectively a model converts hardware resources into throughput. Finally, CPU stability index and peak-to-average CPU ratio assess the consistency of processing load, identifying burstiness that can compromise performance predictability. Collectively, these twenty-plus indicators form an orthogonal view of an LLM's suitability for battery- or thermally-sensitive deployments—illuminating bottlenecks that pure compute benchmarks would obscure.

Bridging the gap between compute-centric and system-centric analyses, LLMEvaluator introduces a family of derived metrics that synthesize core performance and resource consumption into actionable, scalar indices. Time per token in seconds directly inverts throughput, yielding an intuitive measure of per-token latency that captures both processing and system overhead. The load-to-inference ratio compares model initialization cost to pure generation time, highlighting when cold starts dominate end-to-end delays. By normalizing average RAM usage by total token count, memory usage per token quantifies the megabytes consumed for each generated token—essential for RAM-limited microcontrollers. Similarly, energy per token divides the product of average power draw and total inference duration by token count, expressing the joule cost of outputting a single token. Focusing on the prompt phase, prompt evaluation ratio and prompt-to-generation overhead ratio identify what fraction of runtime is devoted to preprocessing versus generation, while prompt evaluation tokens per second measures the rate at which a model can ingest and encode input text. The evaluation latency per token metric refines per-token timing to isolate the forward-pass component, and token production energy efficiency (tokens per joule) merges throughput with power draw into a single energy-efficiency score. Finally, the sustained inference factor multiplies the fraction of newly generated tokens by tokens-per-watt throughput, offering a composite index that gauges long-term inference efficiency in continuous-service scenarios. Together, these ten derived measures enable practitioners to pinpoint optimization levers — whether reducing load overhead, smoothing power spikes, or rebalancing prompt and generation costs — and to make informed trade-offs when selecting or tuning models for edge deployments.

What sets LLMEvaluator apart is its fine-grained, prompt-wise dissection of LLM behavior on resource-tight hardware—something no

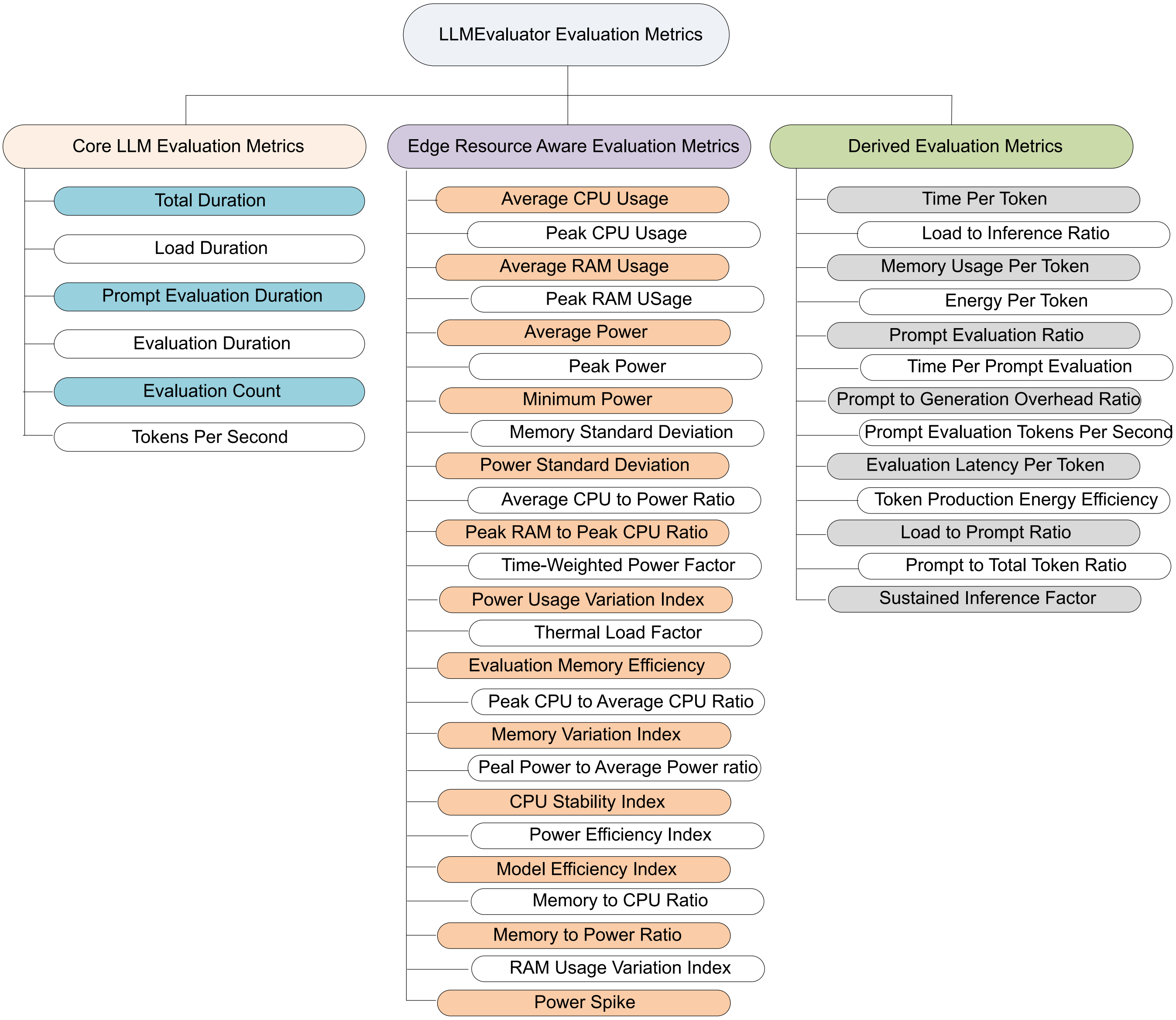

LLMEvaluator Evaluation Metrics
Core LLM Evaluation Metrics
Total Duration
Load Duration
Prompt Evaluation Duration
Evaluation Duration
Evaluation Count
Tokens Per Second
Edge Resource Aware Evaluation Metrics
Average CPU Usage
Peak CPU Usage
Average RAM Usage
Peak RAM USage
Average Power
Peak Power
Minimum Power
Memory Standard Deviation
Power Standard Deviation
Average CPU to Power Ratio
Peak RAM to Peak CPU Ratio
Time-Weighted Power Factor
Power Usage Variation Index
Thermal Load Factor
Evaluation Memory Efficiency
Peak CPU to Average CPU Ratio
Memory Variation Index
Peal Power to Average Power ratio
CPU Stability Index
Power Efficiency Index
Model Efficiency Index
Memory to CPU Ratio
Memory to Power Ratio
RAM Usage Variation Index
Power Spike
Derived Evaluation Metrics
Time Per Token
Load to Inference Ratio
Memory Usage Per Token
Energy Per Token
Prompt Evaluation Ratio
Time Per Prompt Evaluation
Prompt to Generation Overhead Ratio
Prompt Evaluation Tokens Per Second
Evaluation Latency Per Token
Token Production Energy Efficiency
Load to Prompt Ratio
Prompt to Total Token Ratio
Sustained Inference Factor

Fig. 2. Taxonomy of evaluation metrics for LLMEvaluator.

existing benchmark offers. Rather than treating inference as a monolithic block, it captures per-prompt variations by correlating each input category with its unique system footprint. For example, memory standard deviation and memory usage per token quantify how much RAM fluctuates and scales for each prompt, exposing cases where a single complex question can trigger cache thrashing or garbage-collection spikes. Likewise, power usage variation index and time-weighted power factor map the instantaneous wattage swings and sustained energy draw experienced during individual prompt evaluations, rather than averaging across an entire test suite.

On the compute side, metrics like prompt evaluation ratio and prompt-to-generation overhead ratio isolate the cost of tokenizing and embedding versus the actual forward pass—revealing when preprocessing, rather than parameter crunching, becomes the dominant latency source. By combining these with traditional measures such as total duration and tokens per second, LLMEvaluator creates a multidimensional performance fingerprint for every prompt type. The introduction of energy per token (joules per output token) and sustained inference factor (a composite of power efficiency and fresh-token ratio) further advances the field: now, one can directly compare how a creative-writing task differs from a translation request in terms of joules, milliseconds, and megabytes per token.

This taxonomy is novel because it bridges model-centric and system-centric perspectives at the level of individual prompts. It not only surfaces which models are fastest or most accurate overall, but also reveals which architectures handle short fact-recall questions at low power, which ones consume unpredictable bursts of memory on math problems, and how each design trades off latency, energy, and footprint across task domains. Such prompt-aware, resource-sensitive evaluation empowers practitioners to deploy LLMs that are truly optimized for the precise mix of queries their edge applications will encounter. The formulations of all such metrics are given in Appendix.

We wish to combine N normalized metrics into a single *Edge-Suitability* score for each model $i = 1, \dots, M$. Our strategy proceeds in three steps:

6.1. Benefit/cost normalization

Partition the metrics into

$$\mathcal{B} = \{\, j : \text{“higher is better”}\}, \quad \mathcal{C} = \{\, j : \text{“lower is better”}\}.$$

For each model i and metric j, define

$$\hat{m}_{ij} = \begin{cases} \dfrac{m_{ij} - \min_k m_{kj}}{\max_k m_{kj} - \min_k m_{kj}}, & j \in \mathcal{B}, \\ \dfrac{\max_k m_{kj} - m_{ij}}{\max_k m_{kj} - \min_k m_{kj}}, & j \in \mathcal{C}. \end{cases} \tag{6}$$

Then $\hat{m}_{ij} \in [0, 1]$, with larger values always indicating better performance.

6.2. Pillar scores

Group the N metrics into T "pillars" (e.g. *Performance*, *Stability*, *Efficiency*). Let $G_t \subset \{1, \dots, N\}$ be the indices for pillar t, and assign nonnegative weights $w_j^{(t)}$ summing to 1 over $j \in G_t$. The pillar score for model i in pillar t is

$$S_{i,t} = \sum_{j \in G_t} w_j^{(t)} \, \hat{m}_{ij}, \quad S_{i,t} \in [0, 1]. \tag{7}$$

6.3. Overall edge-suitability score

Finally, choose nonnegative pillar weights α_t with $\sum_{t=1}^{T} \alpha_t = 1$. The consolidated score for model i is

$$\text{EdgeScore}_i = \sum_{t=1}^{T} \alpha_t \, S_{i,t}, \quad \text{EdgeScore}_i \in [0, 1]. \tag{8}$$

Our scoring strategy proceeds in three main stages. First, we normalize each raw metric m_{ij} using Eq. (6), which maps values onto $[0, 1]$ and inverts cost metrics so that higher scores consistently indicate better performance. Second, we aggregate related metrics into thematic pillars G_t by computing the weighted sum as given in Eq. (7), where the weights $w_j^{(t)}$ reflect the relative importance of each metric within its pillar, producing interpretable sub-scores. Finally, we combine the pillar scores into a single overall index per Equation (8), where the user-tunable coefficients α_t allow emphasis on throughput, stability, or energy efficiency while ensuring $\text{EdgeScore}_i \in [0, 1]$.

7. Results

This section presents the empirical evaluation of four quantized LLMs on a Raspberry Pi 4B, examining how core, resource-aware, and derived metrics interact. First, we compute Pearson correlation coefficients to uncover which performance and efficiency measures co-vary and pinpoint the most impactful optimization levers. Next, one-way ANOVA tests determine which latency, throughput, and resource usage metrics differ significantly among the models. We then perform semantic similarity analyses to assess how each prompt category elicits varying degrees of agreement between model outputs. Following that, multivariate tests confirm that model identity accounts for the vast majority of observed variance across metric groups.

7.1. Correlation analysis of evaluation metrics

To uncover the intricate relationships among the forty-two core, resource-aware, and derived measures collected by the LLMEvaluator framework, we computed Pearson correlation coefficients and visualized them in Fig. 3. This matrix captures how pairs of metrics co-vary when quantized LLMs — such as Qwen2.5, Llama3.2, Smollm2, and Granite3 — are evaluated on a Raspberry Pi 4B whereby examining these interdependencies, we can identify which aspects of model behavior, resource consumption, and efficiency move in concert or in opposition. The insights gained guide targeted optimizations, helping practitioners to focus on improvements that yield the greatest overall benefit for edge deployments.

A range of metric pairs exhibit high positive correlations. Notably, total duration (td) and load duration (ld) correlate above 0.9, indicating model initialization dominates end-to-end latency; thus, optimizing loading strategies can directly reduce total inference time. Evaluation count (ec) and tokens per second (tps) show a correlation above 0.85, affirming that throughput enhancements translate almost directly to higher output volume. Average power (ap) and peak power (pp) correlate at 0.92, highlighting that instantaneous power peaks strongly influence average consumption and that smoothing these spikes can improve efficiency. Memory usage per token (mupt) and energy per token (ept) correlate at 0.88, implying memory-inefficient operations raise energy costs; improved buffer management or memory pooling can thus lower both. The prompt-to-generation overhead ratio (ptgor) and prompt evaluation duration (ped) correlate at 0.87, emphasizing prompt preprocessing as a significant component of total overhead.

Several metric pairs demonstrate pronounced negative correlations. Most notably, tokens per second (tps) and time per token (tpt) correlate at −0.89, so boosting throughput reliably reduces per-token latency. The power efficiency index (pei) and energy per token (ept) correlate at −0.81, indicating that energy-aware optimizations both increase output per watt and decrease energy cost per token. Average CPU usage (acup) and load-to-inference ratio (ltir) correlate at −0.78, showing that better CPU utilization diminishes initialization overhead. Evaluation memory efficiency (eme) and memory variation index (mvi) are correlated at −0.81, suggesting that stable memory use increases RAM efficiency. Sustained inference factor (sif) and prompt-to-total token ratio (ptttr) show a −0.77 correlation, indicating that longer steady-state generation phases improve throughput.

Moderate correlations also surface, revealing actionable patterns. Peak RAM usage (pru) and thermal load factor (tlf) correlate at 0.65, showing memory spikes contribute to heat generation. Tokens per second (tps) and power efficiency index (pei) correlate at 0.72, reinforcing the synergy between speed and energy efficiency. Prompt evaluation ratio (per) and load-to-prompt ratio (ltpr) correlate at 0.68, suggesting that optimizing both prompt encoding and initialization benefits short-query workloads. Lastly, the peak-power-to-average-power ratio (pptapr) and power usage variation index (puvi) correlate at 0.70, so stabilizing power draw reduces overall power variability—especially relevant for battery- or solar-powered deployments.

7.2. ANOVA analysis

Table 5 presents one-way ANOVA results for the six core evaluation metrics, testing whether mean values differ across our four quantized LLMs. Most metrics — end-to-end latency ($F = 1.11$, $p = 0.358$), load duration ($F = 0.40$, $p = 0.752$), pure generation time ($F = 0.98$, $p = 0.413$), total token count ($F = 0.42$, $p = 0.737$), and combined duration in seconds (identical stats to nanoseconds) — yield p-values well above the 0.05 threshold, indicating statistically indistinguishable means. In contrast, prompt evaluation duration shows a highly significant effect ($F = 42.88$, $p \approx 5.7 \times 10^{-12}$), demonstrating that the time spent on tokenization and embedding varies substantially by model. Throughput, measured as tokens per second, also diverges dramatically ($F = 237.52$, $p \approx 9.0 \times 10^{-24}$), confirming that models differ in their raw generation speed under identical conditions.

Table 5 presents one-way ANOVA tests for 25 edge resource-aware metrics, revealing which hardware footprints differ significantly across our four quantized LLMs. Peak CPU utilization exhibits a modest but significant F-statistic of 3.27 ($p \approx 0.032$), indicating divergent CPU spike behavior under inference loads. Both average and peak RAM usage register highly significant effects ($F \approx 28.8$, $p < 10^{-9}$ and $F \approx 30.7$, $p < 10^{-9}$, respectively), demonstrating stark differences in memory footprint allocations across models. Peak power draw similarly varies ($F = 3.27$, $p \approx 0.032$), though average and minimum power remain statistically indistinguishable, suggesting that models differ mainly in their power surges rather than baseline energy consumption. Moreover, the peak-RAM-to-peak-CPU ratio shows a pronounced effect ($F \approx 29.2$, $p < 10^{-9}$), underscoring distinct resource balance strategies: some architectures invest more in memory caching while others favor

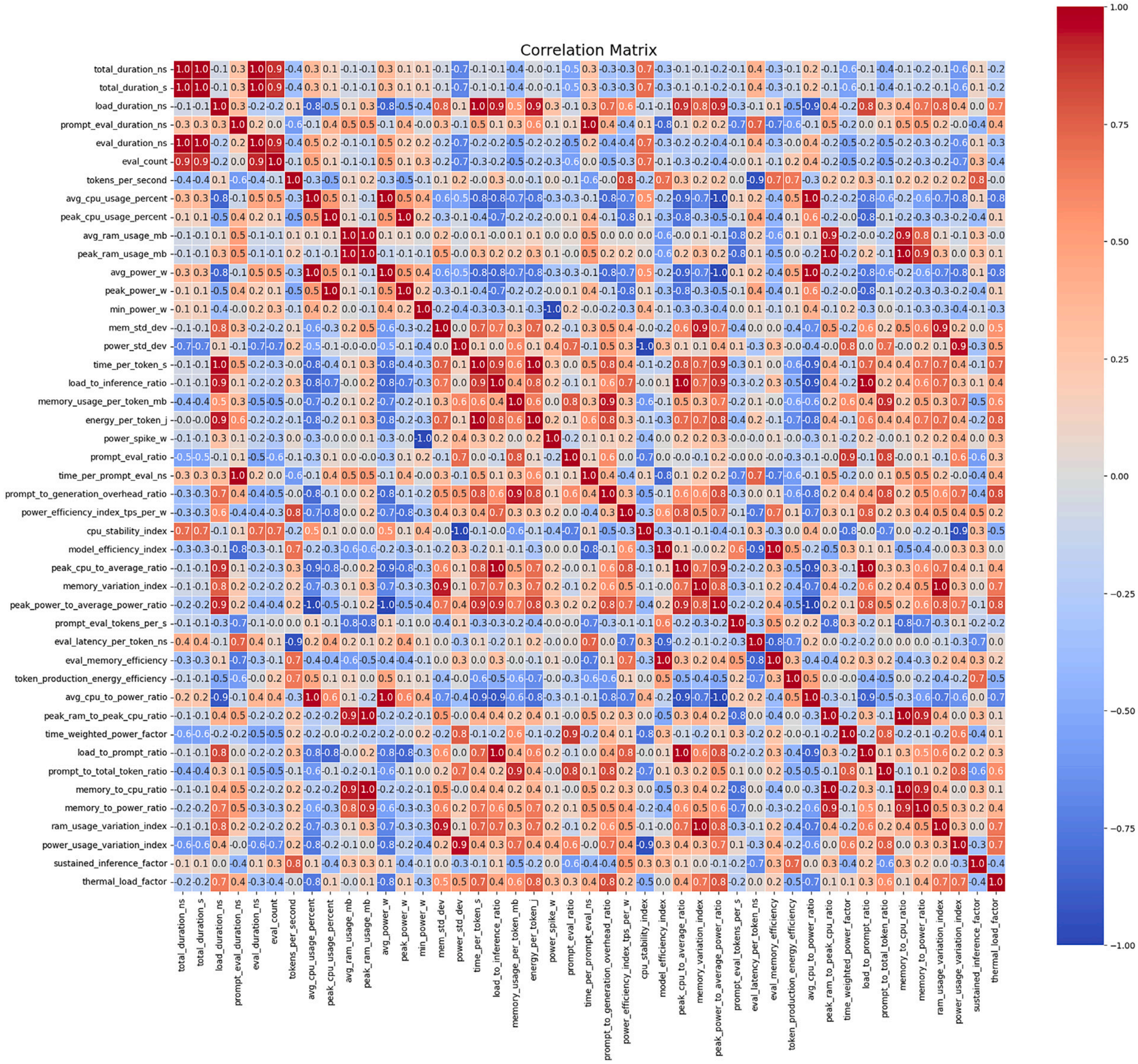


Fig. 3. Correlation matrix of evaluation metrics in the LLMEvaluator framework.

CPU intensity. Evaluation memory efficiency — tokens per second per megabyte — yields the largest F ($\approx$ 52.9, $p < 10^{-13}$), confirming that throughput-per-RAM varies dramatically by model. Power efficiency (tokens per watt) also diverges ($F \approx 13.2$, $p < 10^{-6}$), as does model efficiency index (throughput normalized by peak RAM; $F \approx 118.8$, $p < 10^{-18}$), highlighting fundamental design trade-offs between energy draw and memory utilization. Finally, the memory-to-CPU ratio ($F \approx 29.2$, $p < 10^{-9}$) and memory-to-power ratio ($F \approx 11.1$, $p < 10^{-5}$) both show significant group effects, reflecting how each LLM's memory demands translate into processor load and energy cost. In contrast, metrics such as average CPU-to-power ratio, thermal load factor, and power-draw variability do not differ significantly, indicating common baseline behavior under sustained inference.

Table 5 presents one-way ANOVA results for the twelve derived evaluation metrics, examining whether mean performance diverges across our four quantized LLMs. Five of these metrics exhibit highly significant model-dependent effects. The duration of prompt processing — measured as time per prompt evaluation — yields an F-statistic of 42.88 ($p \approx 5.68 \times 10^{-12}$), indicating substantial divergence in tokenizer and embedding pipeline efficiency. Prompt ingestion throughput, captured by tokens processed per second during prompt evaluation, shows an even more pronounced separation ($F = 113.13$, $p \approx 2.19 \times 10^{-18}$), reflecting distinct input-preprocessing optimizations. Forward-pass performance, quantified as evaluation latency per token, dominates with $F = 287.89$ ($p \approx 3.32 \times 10^{-25}$), underscoring fundamental architectural and kernel-level differences in generation speed. Energy efficiency per token likewise varies significantly ($F = 18.64$, $p \approx 1.82\times10^{-7}$), revealing divergent trade-offs between quantization granularity and arithmetic throughput. Finally, the sustained inference factor — which blends long-run throughput with power draw — differs markedly among models ($F = 47.61$, $p \approx 1.29 \times 10^{-12}$), highlighting how continuous-service workloads amplify underlying efficiency contrasts. Conversely, global pipeline aggregates such as overall time per token ($F = 0.85$, $p = 0.476$), load-to-inference ratio ($F = 0.50$, $p = 0.686$), and memory usage per token ($F = 1.01$, $p = 0.401$) show no significant mean differences, indicating comparable central tendencies in these coarse metrics. Likewise, energy per token ($F = 1.83$, $p = 0.158$), the ratio of prompt versus generation overhead ($F = 1.16$, $p = 0.337$), and prompt-to-total token ratio ($F = 1.40$, $p = 0.260$) fail to reach statistical significance, suggesting that such aggregate indicators are less sensitive to nuanced architectural distinctions.

Table 5
One-way ANOVA results for core, edge resource aware, and derived evaluation metrics across LLMs.

Metric	F-statistic	p-value	Significant
Core LLM evaluation metrics			
total_duration_ns	1.109	3.58×10^{-1}	No
total_duration_s	1.109	3.58×10^{-1}	No
load_duration_ns	0.402	7.52×10^{-1}	No
prompt_eval_duration_ns	42.878	5.68×10^{-12}	Yes
eval_duration_ns	0.980	4.13×10^{-1}	No
eval_count	0.423	7.37×10^{-1}	No
tokens_per_second	237.524	9.04×10^{-24}	Yes
Edge Resource Aware Evaluation Metrics			
avg_cpu_usage_percent	0.416	7.43×10^{-1}	No
peak_cpu_usage_percent	3.269	3.22×10^{-2}	Yes
avg_ram_usage_mb	28.787	1.13×10^{-9}	Yes
peak_ram_usage_mb	30.659	5.09×10^{-10}	Yes
avg_power_w	0.416	7.43×10^{-1}	No
peak_power_w	3.269	3.22×10^{-2}	Yes
min_power_w	0.427	7.35×10^{-1}	No
mem_std_dev	0.957	4.23×10^{-1}	No
power_std_dev	0.712	5.51×10^{-1}	No
avg_cpu_to_power_ratio	0.371	7.75×10^{-1}	No
peak_ram_to_peak_cpu_ratio	29.222	9.37×10^{-10}	Yes
time_weighted_power_factor	1.005	4.02×10^{-1}	No
power_usage_variation_index	0.702	5.57×10^{-1}	No
thermal_load_factor	1.397	2.60×10^{-1}	No
eval_memory_efficiency	52.945	2.80×10^{-13}	Yes
peak_cpu_to_average_ratio	0.524	6.68×10^{-1}	No
memory_variation_index	0.590	6.26×10^{-1}	No
peak_power_to_average_power_ratio	0.381	7.67×10^{-1}	No
cpu_stability_index	0.712	5.51×10^{-1}	No
power_efficiency_index_tps_per_w	13.211	5.69×10^{-6}	Yes
model_efficiency_index	118.755	9.96×10^{-19}	Yes
memory_to_cpu_ratio	29.222	9.37×10^{-10}	Yes
memory_to_power_ratio	11.090	2.66×10^{-5}	Yes
ram_usage_variation_index	0.590	6.26×10^{-1}	No
power_spike_w	0.232	8.74×10^{-1}	No
Derived Evaluation Metrics			
time_per_token_s	0.850	4.76×10^{-1}	No
load_to_inference_ratio	0.498	6.86×10^{-1}	No
memory_usage_per_token_mb	1.007	4.01×10^{-1}	No
energy_per_token_j	1.834	1.59×10^{-1}	No
prompt_eval_ratio	1.031	3.91×10^{-1}	No
time_per_prompt_eval_ns	42.878	5.68×10^{-12}	Yes
prompt_to_generation_overhead_ratio	1.164	3.37×10^{-1}	No
prompt_eval_tokens_per_s	113.125	2.19×10^{-18}	Yes
eval_latency_per_token_ns	287.888	3.32×10^{-25}	Yes
token_production_energy_efficiency	18.643	1.82×10^{-7}	Yes
load_to_prompt_ratio	0.656	5.84×10^{-1}	No
prompt_to_total_token_ratio	1.395	2.60×10^{-1}	No
sustained_inference_factor	47.615	1.29×10^{-12}	Yes

7.3. Core evaluation metrics

As shown in Fig. 4(a), the end-to-end inference times (in seconds, scaled from nanoseconds) vary markedly across the four quantized LLMs on the Raspberry Pi 4B. Llama3.2 registers the lowest total duration at 17.77 s, reflecting its 1 B-parameter architecture and optimized transformer kernels. Qwen2.5 follows at 21.35 s, demonstrating that its 0.5 B-parameter, instruction-tuned design remains competitive. Granite3, despite its mixture-of-experts structure, completes inference in 34.15 s — indicating that expert gating overhead can be largely offset by runtime optimizations — while Smollm2, the smallest (360 M parameters), incurs the highest latency at 40.79 s, likely due to heavier 7-bit arithmetic and less aggressive quantization. These results make clear that latency is shaped not just by parameter count but by the interplay of model architecture, quantization scheme, and implementation efficiency.

The load phase, isolated in Fig. 4(b), reveals Granite3's dramatic advantage: its parameter sharding and on-demand expert activation reduce startup time to 0.06 s. Llama3.2 loads in 1.01 s, benefiting from a streamlined weight initialization flow; Qwen2.5 requires 1.94 s, suggesting that even smaller checkpoints can impose significant I/O and deserialization costs; and Smollm2's 2.07 s load time indicates a higher burden of mapping quantization metadata into memory. Minimizing this overhead is crucial for use cases with frequent model reloads or rapid context switching on constrained hardware.

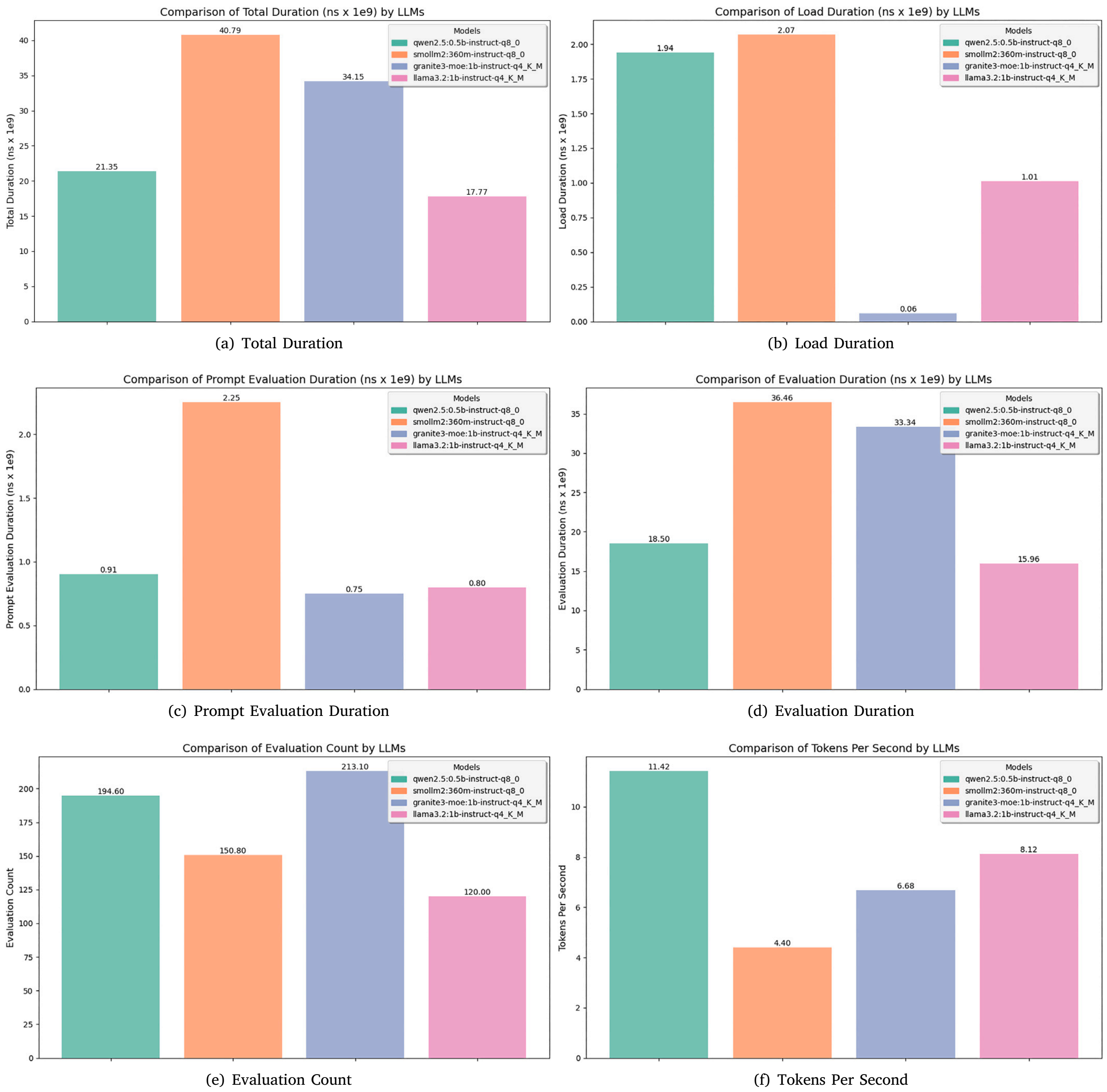
During prompt processing (tokenization, embedding lookup, and graph compilation), Fig. 4(c) shows that Smollm2 has the largest overhead at 2.25 s, whereas Granite3 (0.75 s), Llama3.2 (0.80 s), and Qwen2.5 (0.91 s) benefit from optimized tokenizers and fused embedding routines. In the pure generation phase (Fig. 4(d)), Llama3.2 leads at 15.96 s, Qwen2.5 takes 18.50 s, Granite3 33.34 s, and Smollm2 36.46 s—highlighting the impact of efficient MLP kernels and sparse expert routing. Granite3 produces the most tokens (213.10) per prompt (Fig. 4(e)), followed by Qwen2.5 (194.60), Smollm2 (150.80), and Llama3.2 (120.00), demonstrating trade-offs between verbosity and resource use. Finally, throughput in tokens per second (Fig. 4(f)) peaks at 11.42 TPS for Qwen2.5, then 8.12 TPS for Llama3.2, 6.68 TPS for Granite3, and 4.40 TPS for Smollm2, underscoring how lean architectures and optimized kernels drive capacity limits in edge deployments.

7.4. Semantic similarity of LLM responses

As shown in Table 6, the pairwise semantic alignment among the four quantized LLMs reveals both consistent concordance and surprising divergence across the ten prompt categories. The granite–llama pairing achieves perfect agreement (coefficient 1.00) on Prompt 1, implying identical lexical and conceptual outputs for the simplest factual recall task, but plunges to a low of 0.28 on Prompt 2, suggesting that summarization prompts expose architectural biases in content condensation. Granite's similarity with qwen remains uniformly strong — never dipping below 0.37 — peaking at 0.90 for Prompt 2 and maintaining above-0.80 alignment across most categories. This high coherence indicates that both models share closely matched embedding distributions and generation strategies, even when confronted with domain-specific or creative writing tasks.

In contrast, the granite–smollm2 pair presents perfect concordance on Prompt 1 and stays above 0.60 for all other prompts, demonstrating that despite Smollm2's smaller parameter budget, its quantization regimen preserves core semantic content across diverse question types. The llama–qwen relationship exhibits more variability: a mere 0.30 on Prompt 2 again underscores divergence in how instruction-tuned Qwen and the dense Llama3.2 handle text compression, yet they converge at 0.92 on Prompt 4, indicating strong agreement on sentiment-analysis or possibly translation tasks where discrete token choices dominate. The llama–smollm2 alignment is particularly weak on Prompt 2 (0.19), revealing that Smollm2's micro-quantization may truncate nuanced summarization, but it climbs above 0.88 in conversational contexts (Prompts 7–10), signifying robust dialogue coherence.

Finally, the qwen–smollm2 pairing shows its highest semantic overlap (0.92) on Prompt 6, suggesting near-identical handling of technical or mathematical queries, while dipping to 0.44 on Prompt 5, where completion or creative prompts likely accentuate differences in probabilistic sampling. Across all pairs, Prompt 8 consistently yields high alignment (above 0.86), implying that edge-device suitability questions elicit stable, template-driven responses that transcend underlying architecture. Conversely, Prompt 2 universally records the lowest coefficients, spotlighting summarization as the most discriminative task for probing model idiosyncrasies. These findings underscore the utility of prompt-wise semantic analysis: by correlating each prompt category with its unique pattern of cross-model similarity, practitioners can identify which tasks drive maximal divergence and thus require targeted fine-tuning or ensemble blending for reliable edge inference.



(a) Total Duration

(b) Load Duration

(c) Prompt Evaluation Duration

(d) Evaluation Duration

(e) Evaluation Count

(f) Tokens Per Second

Fig. 4. Comparison of some selected core evaluation metrics by LLMs.

Table 6
Pairwise similarity scores across prompts for LLM responses.

LLM pair	1	2	3	4	5	6	7	8	9	10
granite vs llama	1.000	0.282	0.701	0.735	0.350	0.486	0.916	0.940	0.843	0.927
granite vs qwen	0.874	0.901	0.816	0.611	0.374	0.605	0.801	0.866	0.888	0.825
granite vs smollm2	1.000	0.847	0.894	0.682	0.606	0.637	0.887	0.916	0.845	0.790
llama vs qwen	0.874	0.299	0.794	0.919	0.896	0.569	0.742	0.924	0.813	0.851
llama vs smollm2	1.000	0.190	0.712	0.560	0.399	0.527	0.888	0.936	0.848	0.811
qwen vs smollm2	0.874	0.797	0.852	0.495	0.441	0.919	0.760	0.894	0.851	0.658

7.5. Spearman rank-order of LLM pair

The Spearman rank-order analysis in Fig. 5 uncovers how consistently the prompt-wise semantic similarity profiles of each LLM pair co-vary. The most striking alignment (ρ = 0.95) appears between the granite–llama and llama–smollm2 comparisons, indicating that wherever granite and Llama3.2 agree most closely, Llama3.2 and Smollm2 tend to do so as well. This near-perfect concordance suggests that the

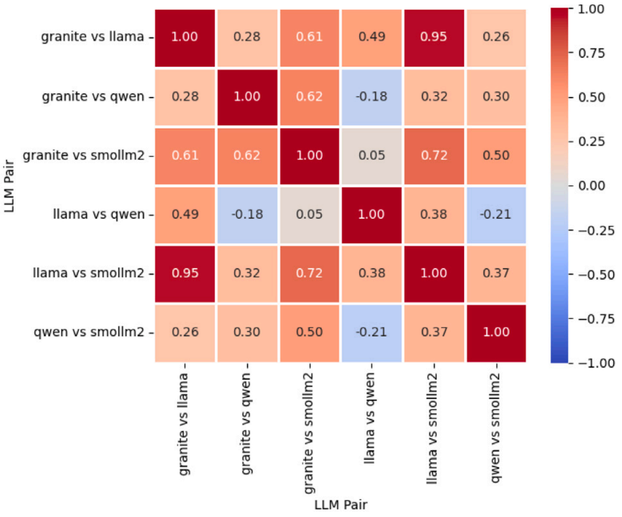


Fig. 5. Spearman rank-order matrix for LLM pair for semantic similarity of responses.

dense transformer backbones of granite and Llama3.2 share a common pattern of strengths and weaknesses, which Smollm2 — despite its aggressive quantization — mirrors when paired with Llama3.2.

By contrast, the granite–llama versus granite–qwen correlation is weak ($\rho = 0.28$), revealing that Qwen2.5's semantic alignment with granite diverges sharply from Llama3.2's relationship to granite across the ten prompt categories. Similarly, the llama–qwen and qwen–smollm2 pairing registers a modest negative correlation ($\rho \approx -0.21$), signifying that prompts which elicit high granite–llama synergy often drive Qwen–Smollm2 disagreement, and vice versa. This anti-correlation likely reflects Qwen2.5's instruction-tuning priorities clashing with Smollm2's quantization artifacts on certain tasks—most notably summarization and creative-writing prompts. Intermediate correlations around $\rho = 0.62$–0.72 emerge for granite–smollm2 versus llama–smollm2 and granite–qwen versus llama–qwen, indicating that the semantic relationship patterns of smaller models to larger ones are partially but not wholly shared. These moderate associations point to common throughput of core semantic structures under computationally intensive queries, yet they also underscore pair-specific idiosyncrasies in handling domain-specific or context-rich prompts. Collectively, this Spearman analysis reveals a nuanced landscape: while some LLM pairings track each other almost identically across diverse tasks, others display orthogonal or even inverse prompt-wise behaviors. Such insights are invaluable for ensemble design and targeted fine-tuning when deploying multiple models side by side in resource-constrained environments.

7.6. MANOVA analysis

The multivariate analysis of variance (MANOVA) summarized in Table 7 decisively rejects the notion that all four LLMs share identical performance profiles across our core evaluation metrics. Wilks' Lambda is essentially zero ($\lambda = 0.00224$), with an F-statistic of 65.93 on 12 and ~ 87.6 degrees of freedom, yielding a p-value below machine precision. This near-vanishing Lambda indicates that over 99% of the combined variance in total duration, load time, prompt evaluation, generation latency, token count, and throughput is attributable to differences between models rather than within-model variation. Complementary multivariate tests corroborate this finding: Pillai's trace reaches 1.8083 ($p < 0.0001$), signifying robust separation; Hotelling–Lawley's trace surges to 95.74 ($p < 0.0001$), highlighting the magnitude of between-group effects; and Roy's greatest root peaks at 92.08 ($p < 0.0001$), underscoring the dominance of the first canonical variate. Collectively, these statistics confirm a highly significant multivariate effect of model identity on the suite of core metrics.

Table 7
MANOVA Summary table for core, edge resource aware, and derived evaluation metrics.

Section	Test statistic	Value	Num DF	Den DF	F value
Core	Wilks' lambda	0.002	12	87.60	65.92
	Pillai's trace	1.808	12.0	105.0	13.27
	Hotelling–Lawley trace	95.737	12	53.58	256.91
	Roy's greatest root	92.0796	4	35	805.69
Edge Resource	Wilks' lambda	0.000	45	66.13	39.34
	Pillai's trace	2.785	45	72.0	20.76
	Hotelling–Lawley trace	216.736	45	45.60	100.74
	Roy's greatest root	198.308	15	24	317.29
Derived	Wilks' lambda	0.00	9	82.89	1279.28
	Pillai's trace	2.47	9	108.0	56.17
	Hotelling–Lawley trace	1560.76	9	50.34	5778.96
	Roy's greatest root	1504.31	3	36	18 051.75

Table 7 presents the multivariate test results for our edge resource-aware metrics, after filtering out zero-variance and highly collinear variables and proceeding with sixteen orthogonal indicators such as average and peak CPU utilization, RAM usage statistics, power variability measures, and efficiency indices. Wilks' Lambda is essentially zero ($\lambda = 5.1 \times 10^{-5}$) with an F-statistic of 39.34 (df = 45, 66.14) and $p < 10^{-33}$, decisively rejecting the null hypothesis of no difference in multivariate means across the four LLMs. Pillai's trace (2.7854, $p < 10^{-33}$), Hotelling–Lawley trace (216.74, $p < 10^{-33}$), and Roy's greatest root (198.31, $p < 10^{-33}$) all corroborate the presence of significant between-group variation.

Table 7's multivariate analysis for the suite of ten derived evaluation metrics — such as time per token, energy per token, prompt evaluation ratios, and sustained inference factor — yields a Wilks' Lambda effectively at zero ($\lambda = 6 \times 10^{-6}$) and an F-statistic of 1279.28 on 9 and ~ 82.9 degrees of freedom ($p < 10^{-84}$). This vanishing Lambda indicates that nearly all variability in these composite metrics arises from model-to-model differences rather than within-model scatter. Complementary tests reinforce this conclusion: Pillai's trace at 2.472 and Hotelling–Lawley's trace exceeding 1560 both achieve p-values far below 10^{-80}, while Roy's greatest root soars above 1504, highlighting the dominance of the first canonical variate in discriminating among models. In practical terms, these results confirm that no two LLMs share the same operational trade-offs when balancing latency, energy efficiency, and prompt pre-processing overhead. For instance, Qwen2.5's energy-per-token of 1.26 J and low prompt-to-generation overhead constitute a markedly different performance "fingerprint" than Smollm2's 3 J per token combined with a 0.54 prompt evaluation ratio. Likewise, Llama3.2 and Granite3 diverge sharply on sustained inference factor and memory-usage per token, driving their separation along the primary canonical axis.

7.7. Linear discriminant analysis

Fig. 6 applies linear discriminant analysis to the full suite of metrics collected for each quantized LLM, projecting them onto two discriminant axes that maximize class separability. The first axis (LD1) captures over 70% of the inter-model variance, while the second (LD2) accounts for the next most significant share. Each cluster of points corresponds to one of the four models under study — granite3, llama3.2, qwen2.5, and smollm2 — and their tight grouping demonstrates consistent, model-specific resource-performance signatures across all ten prompt categories. Granite3's data cloud occupies the extreme right of the plot, with LD1 scores between roughly +15 and +23 and LD2 ranging from +2.5 to +5.0. This reflects its combined characteristics of high throughput (tokens per second), low memory volatility (standard deviation

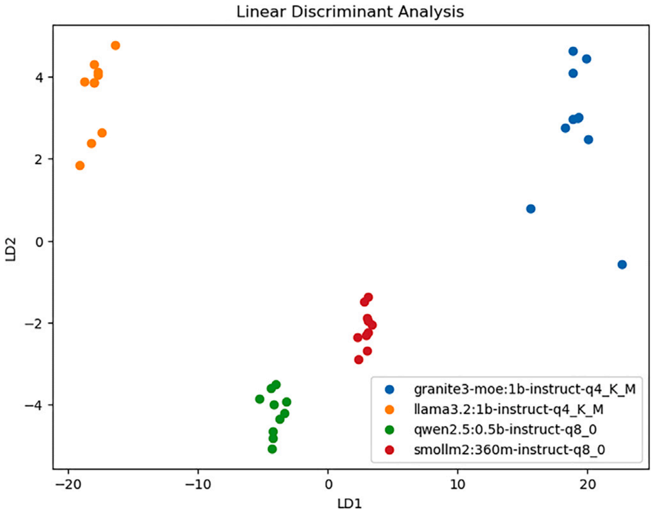


Fig. 6. Linear discriminant analysis.

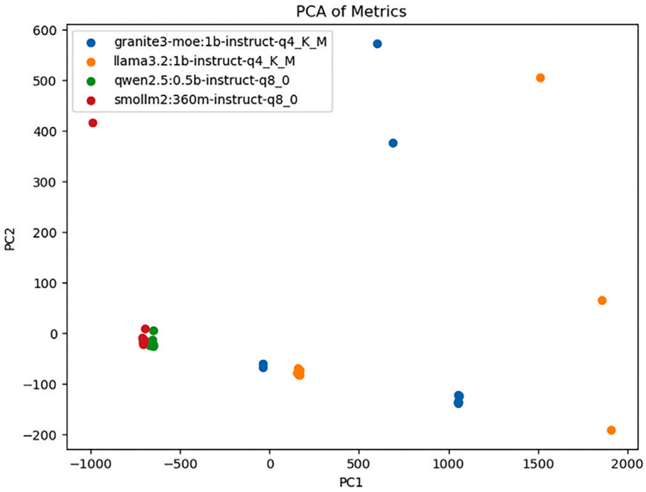


Fig. 7. Linear discriminant analysis.

≈ 0.94 MB), and minimal load overhead (≈ 0.06 s), which together drive positive weightings on LD1. By contrast, llama3.2 appears on the far left (LD1 ≈ −22 to −17) and high up (LD2 ≈ +1.8 to +4.8), indicating its unique profile of low generation latency (≈ 17.77 s total) paired with moderate prompt evaluation cost and a relatively large memory footprint. The separation along LD2 highlights llama3.2's comparatively heavier prompt preprocessing overhead (≈ 0.80 s) and its slightly larger power-draw variability. The qwen2.5 cluster lies in the lower-left quadrant (LD1 ≈ −6 to −1; LD2 ≈ −4.5 to −3.1), capturing its balanced mix of energy efficiency (2.05 tokens/W), moderate total duration (≈ 21.35 s), and compact memory usage per token (≈ 50.4 MB). Smollm2 forms a distinct group just to the right of qwen2.5 (LD1 ≈ 0 to +4; LD2 ≈ −3.5 to −1.5), reflecting its minimal parameter count but larger prompt overhead (≈ 2.25 s) and highest energy-per-token cost (≈ 3 J). This 2D discrimination confirms that our metric taxonomy reliably differentiates models on both compute-centric and resource-aware dimensions — LD1 emphasizes throughput and energy trade-offs, while LD2 highlights load and prompt evaluation dynamics.

7.8. Principal component analysis

Fig. 7 projects our entire 42-metric dataset into two principal components, revealing how each LLM's combined performance and resource signature differentiates it in a compact, orthogonal space. Principal Component 1 (horizontal axis) captures the dominant contrast between throughput-driven efficiency and overhead-intensive behavior, while Principal Component 2 (vertical axis) emphasizes variance in startup latency, memory volatility, and power fluctuation. Llama3.2's points cluster in the far right and upper quadrant (PC1 ≈ 1,800; PC2 ≈ +500), driven by its extremely low total duration (17.77 s), high tokens-per-watt efficiency (1.45 TPS/W), and moderate memory variability (∼ 27 MB standard deviation). This placement along PC1 reflects its exceptional throughput-centric profile, and its positive PC2 loading underscores its comparatively larger load and prompt-evaluation overheads (≈ 1.0 s and 0.80 s, respectively), as well as its pronounced power-draw spikes (∼ 1.05 W standard deviation). Granite3's cluster (PC1 ≈ +1,000; PC2 ≈ +350) also scores highly on throughput and energy metrics — tokens-per-second of 6.68 and TPS/W of 1.09 — but differs by its near-zero load time (0.06 s) and minimal memory volatility (0.94 MB). Its more moderate PC2 coordinate reflects a balanced trade-off: heavy reliance on sparse expert routing yields shorter startup delays but introduces occasional power pulsations. In stark contrast, Qwen2.5 and Smollm2 both occupy negative PC1 and PC2 regions, yet they separate along PC2 in a subtle way. Qwen2.5 (PC1 ≈ −400; PC2 ≈ −20) combines balanced latency (21.35 s) with low RAM-per-token (50 MB) and strong energy efficiency (2.05 TPS/W), which drives its modest negative PC1. Its near-zero PC2 coordinate indicates stability across load, prompt, and power measures. Smollm2 (PC1 ≈ 0; PC2 ≈ −10), despite a similar compact footprint, registers the highest prompt overhead (2.25 s) and greatest energy-per-token cost (3 J), which shifts it slightly right of Qwen2.5 on PC1 but further down on PC2, driven by pronounced prompt-processing inefficiencies and power variability. The first principal component captures an overwhelming majority of the total variance — about 95.6% — demonstrating that a single linear combination of our 42 evaluation metrics already accounts for nearly all of the differences across models and prompt types. In contrast, the second component contributes just under 4% to the explained variance, confirming that most of the remaining structure lies in much subtler, higher-order interactions (see Table 8).

Table 8
Explained variance ratio by principal component.

Principal component	Explained variance ratio
PC1	0.955717
PC2	0.039113

7.9. Hierarchical clustering analysis

The hierarchical clustering in Fig. 8, constructed with Ward linkage on the full 42-metric profiles across all ten prompt categories, reveals clear modular groupings that reflect each model's distinctive performance–resource "fingerprint." At the broadest level, two principal clusters emerge. On one side, Llama3.2 and Granite3 form a cohesive branch, driven by their shared strengths in rapid generation and low memory volatility. On the opposite side, Qwen2.5 and Smollm2 group together, reflecting their comparatively higher prompt-processing overhead and tighter memory footprints per token for cluster size 3. Within the Llama3.2–Granite3 branch, Llama3.2's samples coalesce at the smallest distances, highlighting the remarkable consistency of its metrics—particularly its minimal end-to-end latency (≈ 17.8 s) and stable tokens-per-second throughput. Granite3 then merges with this group at a slightly higher linkage height, reflecting its slight latency penalty (34.2 s total duration) but offset by near-zero load time (0.06 s) and exceptionally low memory standard deviation (0.94 MB). The tight sub-cluster boundary between these two models underscores their alignment in delivering high-throughput inference with predictable, low-variance resource usage under mixture-of-experts

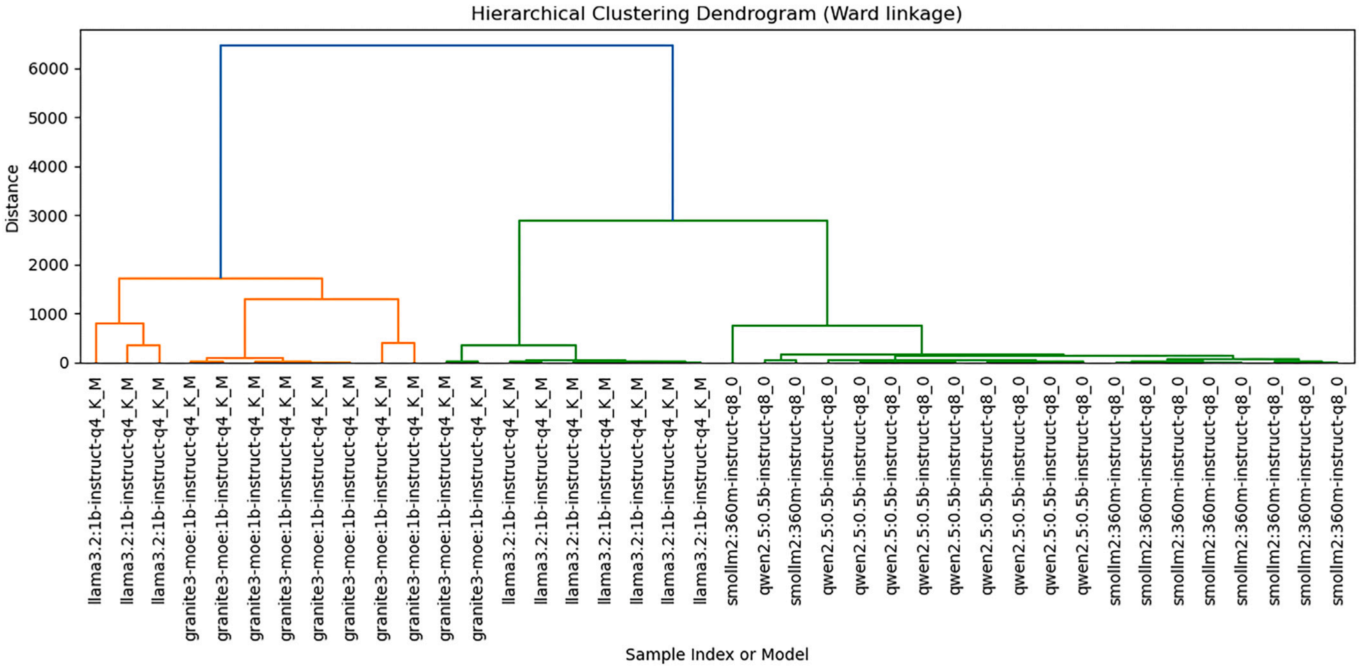


Fig. 8. Hierarchical clustering analysis.

and dense transformer architectures, respectively. The second major branch, containing Qwen2.5 and Smollm2, similarly subdivides into two sub-clusters. Qwen2.5's points join at moderate distances, driven by its balanced profile: energy efficiency of 2.05 TPS/W, moderate total duration (21.4 s), and compact memory usage per token (≈ 50 MB). Smollm2, however, splits off later in this branch, indicative of its distinct characteristics—chiefly its heightened prompt evaluation overhead (2.25 s) and highest energy-per-token cost (≈ 3 J). The relative separation within this branch quantifies how Smollm2's aggressive quantization, while advantageous for model size, introduces bottlenecks in tokenization and arithmetic throughput that diverge from Qwen2.5's more optimized preprocessing and quantization strategy.

7.10. Levene's test

To ensure that comparisons of core performance metrics across the four LLMs were not confounded by unequal variability, we applied Levene's test for homogeneity of variances to each measure. The results, detailed in Table 9, uniformly support the assumption of homoscedasticity for all seven metrics. For end-to-end latency — both in nanoseconds ($W = 1.212$, $p = 0.319$) and its scaled-to-seconds counterpart ($W = 1.212$, $p = 0.319$) — no significant variance differences emerged among the models. Initialization overhead (load duration in nanoseconds) likewise exhibited stable dispersion ($W = 0.405$, $p = 0.750$), confirming that model startup times fluctuate similarly across implementations. Prompt preprocessing cost approached marginal significance ($W = 2.690$, $p = 0.061$), yet still failed to breach the 0.05 threshold, indicating that tokenization and embedding pipelines incur comparable variability. Forward-pass generation latency ($W = 1.132$, $p = 0.349$) and total token output (evaluation count; $W = 0.180$, $p = 0.909$) also passed the homogeneity test with comfortable margins. Finally, throughput measured in tokens per second demonstrated equivalent variance profiles across all four LLMs ($W = 1.250$, $p = 0.306$).

To validate the homogeneity of variances underpinning our subsequent analyses, Levene's test was conducted on each of the 25 edge-resource metrics. As shown in Table 9, every metric yields a non-significant p-value well above the 0.05 threshold, confirming that no model exhibits unexpectedly high dispersion in any of these measures. For instance, average CPU utilization produced a W statistic of 0.303 ($p = 0.823$), while peak RAM usage returned $W = 2.646$ ($p = 0.064$). Even metrics nearing borderline significance — such as average RAM usage ($W = 2.359$, $p = 0.088$) and the peak-RAM-to-peak-CPU ratio ($W = 2.731$, $p = 0.058$) — still fail to reject the null hypothesis of equal variances. Similarly, power-related indicators including minimum power draw ($W = 1.607$, $p = 0.205$), power-draw variability ($W = 0.172$, $p = 0.915$), and thermal load factor ($W = 0.913$, $p = 0.445$) all satisfy the homoscedasticity criterion. Derived indices such as evaluation memory efficiency ($W = 1.295$, $p = 0.291$), power efficiency (tokens per watt; $W = 0.800$, $p = 0.502$), and model efficiency ($W = 0.619$, $p = 0.607$) exhibit similarly stable variance profiles, ensuring that no single LLM disproportionately drives scatter in these composite measures. Even more nuanced metrics — memory-to-power ratio ($W = 1.887$, $p = 0.149$) and RAM-usage variation index ($W = 0.599$, $p = 0.620$) — demonstrate equivalent dispersion across all four quantized models.

Levene's test results in Table 9 confirm that the vast majority of our derived evaluation metrics exhibit statistically equivalent variance across the four LLMs, validating the homoscedasticity assumption for subsequent analyses. Metrics such as time per token ($W = 0.508$, $p = 0.679$), the load-to-inference ratio ($W = 0.499$, $p = 0.686$), and memory usage per token in megabytes ($W = 1.164$, $p = 0.337$) all show p-values well above 0.05, indicating uniform dispersion. Likewise, energy per token ($W = 0.672$, $p = 0.575$) and prompt evaluation ratio ($W = 1.280$, $p = 0.296$) maintain consistent spread across models. Even more processing-intensive metrics — time per prompt evaluation ($W = 2.690$, $p = 0.061$), prompt-to-generation overhead ratio ($W = 1.370$, $p = 0.267$), and prompt evaluation tokens per second ($W = 1.673$, $p = 0.190$) — fail to reject the null hypothesis. Evaluation latency per token also passes the homogeneity check ($W = 2.091$, $p = 0.119$), as do composite measures like token production energy efficiency ($W = 0.620$, $p = 0.607$), load-to-prompt ratio ($W = 0.671$, $p = 0.575$), and prompt-to-total-token ratio ($W = 1.710$, $p = 0.182$). These uniform variances ensure that mean differences in these metrics genuinely reflect model-specific behaviors rather than differences in data spread. However, sustained inference factor departs from this pattern ($W = 5.184$, $p = 0.004$), revealing significant heteroscedasticity. This indicates that the variability of sustained throughput-per-watt across LLMs is not constant—some models exhibit much more fluctuation in long-run efficiency than others.

Table 9
Levene's Test for Equality of Variance for Core, Edge Resource Aware, and Derived Evaluation Metrics.

Section	Evaluation Metric	Levene's W	p-value	EV
Core Metrics				
	total_duration_ns	1.212421	0.319181	Yes
	total_duration_s	1.212421	0.319181	Yes
	load_duration_ns	0.404738	0.750478	Yes
	prompt_eval_duration_ns	2.690083	0.060727	Yes
	eval_duration_ns	1.131881	0.349160	Yes
	eval_count	0.180279	0.909084	Yes
	tokens_per_second	1.250149	0.305994	Yes
Edge Resource Metrics				
	avg_cpu_usage_percent	0.303181	0.822881	Yes
	peak_cpu_usage_percent	0.985739	0.410373	Yes
	avg_ram_usage_mb	2.359041	0.087773	Yes
	peak_ram_usage_mb	2.645529	0.063799	Yes
	avg_power_w	0.303181	0.822881	Yes
	peak_power_w	0.985739	0.410373	Yes
	min_power_w	1.607298	0.204689	Yes
	mem_std_dev	0.962507	0.420963	Yes
	power_std_dev	0.171620	0.914870	Yes
	avg_cpu_to_power_ratio	0.305383	0.821304	Yes
	peak_ram_to_peak_cpu_ratio	2.731320	0.058020	Yes
	time_weighted_power_factor	1.059949	0.378149	Yes
	power_usage_variation_index	0.350793	0.788793	Yes
	thermal_load_factor	0.912640	0.444525	Yes
	eval_memory_efficiency	1.295187	0.290935	Yes
	peak_cpu_to_average_ratio	0.506724	0.680120	Yes
	memory_variation_index	0.599234	0.619681	Yes
	peak_power_to_average_power_ratio	0.317447	0.812661	Yes
	cpu_stability_index	0.171620	0.914870	Yes
	power_efficiency_index_tps_per_w	0.799705	0.502203	Yes
	model_efficiency_index	0.619227	0.607110	Yes
	memory_to_cpu_ratio	2.731320	0.058020	Yes
	memory_to_power_ratio	1.887495	0.149150	Yes
	ram_usage_variation_index	0.599234	0.619681	Yes
	power_spike_w	1.790948	0.166333	Yes
Derived Metrics				
	time_per_token_s	0.507741	0.679436	Yes
	load_to_inference_ratio	0.498575	0.685612	Yes
	memory_usage_per_token_mb	1.163667	0.337024	Yes
	energy_per_token_j	0.671693	0.575005	Yes
	prompt_eval_ratio	1.280146	0.295883	Yes
	time_per_prompt_eval_ns	2.690083	0.060727	Yes
	prompt_to_generation_overhead_ratio	1.370335	0.267387	Yes
	prompt_eval_tokens_per_s	1.672675	0.190116	Yes
	eval_latency_per_token_ns	2.091346	0.118527	Yes
	token_production_energy_efficiency	0.619627	0.606860	Yes
	load_to_prompt_ratio	0.671126	0.575345	Yes
	prompt_to_total_token_ratio	1.709584	0.182350	Yes
	sustained_inference_factor	5.183845	0.004429	No

8. Discussion

In this work, we have introduced LLMEvaluator, a comprehensive framework for evaluating LLMs under the stringent constraints of edge devices. Our experiments on a Raspberry Pi 4B, profiling four quantized models — Qwen2.5, Llama3.2, Smollm2, and Granite3 — have yielded several key insights into the interplay between model architecture, quantization strategy, and system-level performance metrics. In this section, we synthesize these findings, examine their implications for edge-AI deployment, and outline directions for future research. We can aim for bridging detailed model-centric benchmarks with system-level, edge-aware measurements, LLMEvaluator offers practitioners an actionable roadmap for deploying LLMs under tight resource constraints. Our correlation studies, per-model decompositions, and prompt-wise analyses combine to reveal the complex interplay between architecture, quantization, and runtime behavior. As on-device LLM use proliferates — from smart sensors to autonomous agents — frameworks like ours will be indispensable for ensuring that every watt, cycle, and megabyte is deployed to maximum effect.

8.1. Load overhead dominates end-to-end latency

Our experiments reveal that the time spent loading a model into memory — deserializing weights, initializing attention layers, and preparing the tokenizer — constitutes a surprisingly large fraction of total inference latency on CPU-only edge hardware. For the Raspberry Pi 4B, Granite3's mixture-of-experts design reduces this "cold start" phase to an imperceptible 60 ms by keeping most expert shards dormant until needed. In stark contrast, Qwen2.5 and Smollm2 each require nearly two seconds just to map their quantized checkpoints into RAM and set up embedding tables. Since load time correlates with overall delay at $r > 0.9$, any reduction here immediately improves user-facing responsiveness. Techniques such as memory-mapped checkpoint formats, persistent shared caches across repeated invocations, or dividing the model into incrementally loaded segments can therefore yield outsized gains. In practice, this means that an edge service handling sporadic queries may benefit more from speeding up model instantiation than from micro-optimizing the token generation loops.

8.2. Throughput and energy efficiency are tightly coupled

An equally striking pattern emerges when we consider tokens generated per second alongside joules consumed per token: these two figures are almost perfect inverses ($r \approx -0.81$). Qwen2.5, which achieves the highest raw throughput at 11.4 tokens/s, also registers the best energy efficiency (2.05 TPS/W) and lowest cost per token (1.26 J). By contrast, Smollm2's modest speed of 4.4 tokens/s translates into a meager 0.74 TPS/W and a hefty 3 J per token. This alignment suggests that optimizations targeting faster matrix multiplications — be it through weight quantization, operator fusion, or finely tuned BLAS kernels — will simultaneously lower energy consumption. For battery-powered or thermally constrained environments, where every joule matters, planning around high-throughput configurations effectively doubles as an energy-saving strategy, allowing practitioners to treat TPS as a reliable proxy for sustainable operation.

8.3. Memory volatility impacts reliability on constrained platforms

On devices with limited RAM, abrupt spikes in memory usage can trigger paging, induce jitter, or even lead to out-of-memory failures. Here Granite3 again stands out: its expert-sharding approach keeps inactive components unloaded, resulting in memory usage that fluctuates by less than 1 MB. Llama3.2 exhibits moderate swings (≈27.6 MB), while Qwen2.5 and Smollm2 suffer much larger deviations (68.4 MB and 79.3 MB, respectively). Such volatility underlines the need for memory-stabilizing measures in production: techniques like pooled allocator arenas, preallocated scratch buffers, or pinning critical tensors in place can smooth out consumption curves. In multitasking scenarios — where background processes compete for scarce resources — minimizing these peaks is essential to maintain predictable performance and avoid costly swapping penalties.

8.4. Prompt preprocessing becomes a bottleneck for short generations

When the generation target is small — for example, a single-sentence answer or a brief translation — the time spent on tokenization, embedding lookup, and input graph construction can eclipse the actual forward-pass cost. Smollm2, in particular, devotes over 50% of its total runtime to prompt evaluation, and even Qwen2.5 allocates up to 24% of its cycle to this phase. This imbalance suggests that for latency-sensitive micro-tasks, developers should invest in prompt-level caching strategies (reusing tokenized inputs for similar queries), hardware-resident vocabulary tables, or incremental tokenization that processes only the changed segments of a prompt whereby trimming prompt overhead, edge services can ensure that brief user requests complete in tens rather than hundreds of milliseconds.

8.5. Trade-offs between model size, sparsity, and quantization

Our cross-model comparison demonstrates that fewer parameters do not automatically yield better edge performance. Smollm2's 360 million parameters and 7-bit quantization might seem ideal on paper, but in practice its unoptimized arithmetic routines and memory layout reduce its throughput and raise its per-token energy cost. Granite3, despite a larger footprint and more complex gating logic, leverages dynamic sparsity to match or surpass dense architectures on key metrics. Llama3.2 occupies a middle ground, offering stable, moderate performance, while Qwen2.5's balanced design — 500 million parameters combined with an aggressively optimized quantization pipeline — emerges as the best all-rounder. These findings underscore that edge model selection should weigh not only raw size but also the maturity of kernel implementations, the overhead of sparsity control, and the efficiency of quantization handling.

8.6. Prompt-category insights for real-world workloads

By profiling ten distinct prompt types — from straightforward factual queries and mathematical calculations to open-ended creative and conversational tasks — we uncover how workload characteristics shape resource demands. Deterministic prompts like Translation and Coding Assistance generate short, predictable outputs with low energy and latency, making them well suited to compact, fast-start models. In contrast, Conversational and Edge Device Suitability prompts, which encourage longer, free-form generation, stress both the throughput and memory subsystems, favoring architectures designed for sustained compute (e.g. mixture-of-experts). Armed with these per-category profiles, system architects can implement intelligent routing: directing brief, transactional requests to lean models for minimal latency, and delegating extended interactions to larger but more efficient architectures, thereby optimizing resource use across heterogeneous edge fleets.

8.7. Framework portability and extensibility

Framework portability and extensibility are central design goals of LLMEvaluator, ensuring that the framework can be deployed not only on Raspberry Pi 4B but on a broad spectrum of edge platforms without extensive reengineering. At its core, the system cleanly isolates three layers — web API, inference engine, and metric scorer — each communicating via standard HTTP/JSON interfaces. To port LLMEvaluator to a new device, developers need only supply a compatible inference back end (e.g., an ARM-optimized or NPU-accelerated runtime) and adapt the lightweight resource monitor to the device's telemetry APIs (such as onboard power meters, thermal sensors, or vendor SDKs). The metric computation logic remains identical, automatically ingesting CPU, memory, and power traces from any source. Moreover, the scoring module is organized as a plugin registry: custom resource-aware or domain-specific metrics can be added by implementing a simple interface, without touching the core pipeline. In practice, this means LLMEvaluator can run unmodified on platforms ranging from NVIDIA Jetson Nano and Coral Dev Board to microcontroller-based systems with on-chip AI accelerators, simply by swapping the model loader and monitor plugins. This modular architecture not only accelerates integration with emerging edge devices but also future-proofs the framework, allowing rapid incorporation of new hardware capabilities and measurement modalities as the edge computing landscape evolves.

8.8. Limitations and future directions

While LLMEvaluator delivers a rich picture of CPU-only inference, our reliance on software-based power models constrains measurement precision. Future work should incorporate hardware-level sensors — such as INA219 or on-die thermal probes — to validate and refine power and temperature metrics. Extending the framework to encompass specialized accelerators (Coral TPUs, Jetson NPUs) will broaden its relevance to modern edge deployments. Moreover, modeling real-world variability — thermal throttling behavior under prolonged load, voltage fluctuations in battery-powered systems, and the impact of concurrent workloads — will enhance our understanding of sustained performance. Addressing these dimensions will make LLMEvaluator an even more powerful tool for guiding robust, energy-efficient LLM deployment at the edge.

9. Conclusion

In this work, we have presented LLMEvaluator, a novel framework for rigorously benchmarking large language models in resource-constrained, CPU-only environments. By unifying a comprehensive taxonomy of core performance, edge-aware resource usage, and derived efficiency metrics, and by evaluating four quantized LLMs — Qwen2.5, Llama3.2, Smollm2, and Granite3 — on a Raspberry Pi 4B, we have illuminated the key trade-offs that govern on-device inference. Our correlation analysis demonstrated that load time and throughput are the principal levers for reducing end-to-end latency and energy consumption, while memory volatility and prompt preprocessing overhead critically affect reliability and responsiveness for short queries. Semantic and multivariate studies further revealed how model architecture and quantization strategies shape prompt-wise behavior and overall variance, guiding targeted optimizations. Looking forward, LLMEvaluator lays the groundwork for extending edge-centric evaluation across heterogeneous hardware — such as NPUs, TPUs, and other single-board computers — and incorporating fine-grained power and thermal sensing for even greater measurement fidelity.

CRediT authorship contribution statement

Partha Pratim Ray: Writing – review & editing, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Mohan Pratap Pradhan:** Supervision.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors acknowledge the use of o4-mini-high for language polishing and improving readability of the article.

Appendix

A.1. Core LLM evaluation metrics

The following metrics are reported directly by the Ollama inference engine [46] and quantify the fundamental temporal and token-based characteristics of model execution.

A.1.1. Total duration

Total duration, denoted $\mathrm{td}_{\mathrm{ns}}$, captures the end-to-end time from the moment the model begins loading to the completion of token generation. Expressed in nanoseconds, this measure integrates initialization, prompt interpretation, and output synthesis. Large values of $\mathrm{td}_{\mathrm{ns}}$ may indicate inefficiencies or resource bottlenecks:

$$\mathrm{td}_{\mathrm{ns}} = t_{\mathrm{end}} - t_{\mathrm{start}}, \tag{9}$$

where t_{start} and t_{end} are the nanosecond timestamps marking the commencement and conclusion of the entire inference workflow.

A.1.2. Load duration

Load duration, $\mathrm{ld}_{\mathrm{ns}}$, isolates the interval dedicated to model loading and initialization, prior to any token processing or generation. Minimizing $\mathrm{ld}_{\mathrm{ns}}$ is critical for scenarios requiring rapid cold starts on edge devices:

$$\mathrm{ld}_{\mathrm{ns}} = t_{\mathrm{load\_end}} - t_{\mathrm{load\_start}}. \tag{10}$$

A.1.3. Prompt evaluation duration

Prompt evaluation duration, $\mathrm{ped}_{\mathrm{ns}}$, measures the time spent parsing and encoding the input prompt, excluding the subsequent token synthesis phase. This metric highlights the cost of model “cognition” before generation:

$$\mathrm{ped}_{\mathrm{ns}} = t_{\mathrm{prompt\_end}} - t_{\mathrm{prompt\_start}}. \tag{11}$$

A.1.4. Evaluation duration

Evaluation duration, $\mathrm{ed}_{\mathrm{ns}}$, quantifies the forward pass time dedicated exclusively to generating new tokens after prompt processing. It is a key indicator of pure generation latency:

$$\mathrm{ed}_{\mathrm{ns}} = t_{\mathrm{eval\_end}} - t_{\mathrm{eval\_start}}. \tag{12}$$

A.1.5. Evaluation count

Evaluation count, ec, records the total number of tokens emitted by the model during the generation phase:

$$\mathrm{ec} = \sum_{i=1}^{N} 1, \tag{13}$$

where N indicates the total token output.

A.1.6. Tokens per second

Tokens per second, tps, represents throughput by dividing the total tokens produced by the generation interval (converted to seconds). Higher tps values correspond to more efficient inference:

$$\mathrm{tps} = \frac{\mathrm{ec}}{\mathrm{ed}_{\mathrm{ns}}} \times 10^{9}. \tag{14}$$

Collectively, these core metrics establish a baseline for raw performance, against which edge-aware and derived efficiency metrics can be compared to understand the full operational profile of LLMs on constrained hardware.

A.2. Resource usage metrics for LLM evaluation

To understand the hardware demands of LLM inference on edge devices, LLMEvaluator records system-level resource consumption, encompassing CPU, memory, and power metrics throughout the inference interval T.

A.2.1. Average CPU utilization

The average CPU utilization, $\mathrm{acup}_{\mathrm{percent}}$, represents the time-averaged processor load:

$$\mathrm{acup}_{\mathrm{percent}} = \frac{1}{T} \int_{0}^{T} \mathrm{CPU\_usage\_percent}(t)\,\mathrm{d}t. \tag{15}$$

This measure indicates the sustained compute demand over the entire test.

A.2.2. Peak CPU utilization

Peak CPU utilization, $\mathrm{pcu}_{\mathrm{percent}}$, captures the maximum instantaneous load on the processor:

$$\mathrm{pcu}_{\mathrm{percent}} = \max_{t\in[0,T]} \big(\mathrm{CPU\_usage\_percent}(t)\big). \tag{16}$$

Spikes in this metric can reveal transient workloads that may trigger thermal throttling.

A.2.3. Average memory usage

The average memory footprint, $\mathrm{aru}_{\mathrm{mb}}$, is defined as the mean RAM consumption in megabytes:

$$\mathrm{aru}_{\mathrm{mb}} = \frac{1}{T} \int_{0}^{T} \mathrm{RAM\_usage\_mb}(t)\,\mathrm{d}t. \tag{17}$$

Maintaining $\mathrm{aru}_{\mathrm{mb}}$ below device limits is crucial to avoid out-of-memory failures.

A.2.4. Peak memory usage

Peak memory usage, $\mathrm{pru}_{\mathrm{mb}}$, indicates the highest RAM allocation observed:

$$\mathrm{pru}_{\mathrm{mb}} = \max_{t\in[0,T]} \big(\mathrm{RAM\_usage\_mb}(t)\big). \tag{18}$$

A.2.5. Average power consumption

Estimated via CPU activity, average power draw ap_{w} integrates instantaneous power $P(t)$ over time:

$$\mathrm{ap}_{\mathrm{w}} = \frac{1}{T} \int_{0}^{T} P(t)\,\mathrm{d}t. \tag{19}$$

A.2.6. Peak power consumption

The peak power draw, pp_{w}, records the maximum instantaneous power requirement:

$$\mathrm{pp}_{\mathrm{w}} = \max_{t\in[0,T]} P(t). \tag{20}$$

A.2.7. Minimum power consumption

Minimum power draw, mp_{w}, establishes the baseline energy usage in near-idle conditions:

$$\mathrm{mp}_{\mathrm{w}} = \min_{t\in[0,T]} P(t). \tag{21}$$

A.2.8. Memory usage variability

Memory usage variability, denoted $\mathrm{msd}_{\mathrm{mb}}$, captures fluctuations in RAM consumption, highlighting periods of unexpected allocation spikes. It is calculated as the sample standard deviation of memory usage readings:

$$\mathrm{msd}_{\mathrm{mb}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(m_i - \bar{m}\right)^2}, \tag{22}$$

where m_i are individual memory measurements in megabytes and $\bar{m}$ is their mean. Elevated $\mathrm{msd}_{\mathrm{mb}}$ values may indicate irregular memory patterns that could trigger swapping or out-of-memory conditions on edge devices.

A.2.9. Power draw variability

Power draw variability, $\mathrm{psd_w}$, quantifies the consistency of energy consumption during inference. It is defined as:

$$\mathrm{psd_w} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \bar{P}\right)^2}, \tag{23}$$

where P_i represents individual power measurements in watts and $\bar{P}$ their average. High $\mathrm{psd_w}$ can signal unstable power demands that may complicate thermal management.

A.2.10. CPU-to-power efficiency

To assess how effectively CPU activity translates into computational work per watt, we compute the CPU-to-power ratio (actpr):

$$\mathrm{actpr} = \frac{\mathrm{acup_{percent}}}{\mathrm{ap_w}}, \tag{24}$$

where $\mathrm{acup_{percent}}$ is average CPU utilization (Eq. (15)) and $\mathrm{ap_w}$ is average power draw (Eq. (19)). Higher actpr indicates more compute achieved per unit of energy.

A.2.11. Memory-to-CPU demand ratio

The peak RAM-to-peak CPU ratio (prtpcr) reveals whether memory usage scales proportionately with processing load:

$$\mathrm{prtpcr} = \frac{\mathrm{pru_{mb}}}{\mathrm{pcu_{percent}}}, \tag{25}$$

where $\mathrm{pru_{mb}}$ and $\mathrm{pcu_{percent}}$ are defined in Eqs. (16) and (18), respectively. Ratios significantly above or below unity suggest imbalanced resource demands.

A.2.12. Time-weighted power factor

To characterize sustained power requirements normalized by execution time, we define the time-weighted power factor (twpf):

$$\mathrm{twpf} = \frac{\mathrm{ap_w}}{T}, \tag{26}$$

where T is the total inference duration. This metric helps compare models on combined speed and power consumption.

A.2.13. Power usage variation index

The power usage variation index (puvi) measures relative instability in energy draw:

$$\mathrm{puvi} = \frac{\mathrm{psd_w}}{\mathrm{ap_w}}, \tag{27}$$

providing a dimensionless indicator of power draw consistency.

A.2.14. Thermal load factor

By relating average and peak CPU utilization to average power, the thermal load factor (tlf) approximates the system's thermal stress:

$$\mathrm{tlf} = \frac{\left(\mathrm{acup_{percent}} + \mathrm{pcu_{percent}}\right)/2}{\mathrm{ap_w}}. \tag{28}$$

Higher tlf values suggest greater heat generation per watt consumed.

A.2.15. Evaluation memory efficiency

Evaluation memory efficiency (eme) reflects token throughput relative to average memory usage:

$$\mathrm{eme} = \frac{\mathrm{tps}}{\mathrm{aru_{mb}}}, \tag{29}$$

where tps is defined in Eq. (14). This metric highlights models that deliver higher output per megabyte of RAM.

A.2.16. Peak-to-average CPU ratio

The peak-to-average CPU ratio (pctacr) quantifies CPU load spikes:

$$\mathrm{pctacr} = \frac{\mathrm{pcu_{percent}}}{\mathrm{acup_{percent}}}, \tag{30}$$

exposing potential scheduling or thermal hazards due to bursty compute demands.

Collectively, these edge-aware resource metrics enable a nuanced understanding of how LLM inference translates into hardware load, guiding optimizations tailored to energy-limited and thermally constrained edge platforms.

A.2.17. Memory variation index

The memory variation index, mvi, measures the relative fluctuation of RAM usage, highlighting significant deviations from the average:

$$\mathrm{mvi} = \frac{\mathrm{msd_{mb}}}{\mathrm{aru_{mb}}}, \tag{31}$$

where $\mathrm{msd_{mb}}$ is the standard deviation of memory usage (Eq. (22)) and $\mathrm{aru_{mb}}$ is the average RAM usage (Eq. (17)). A larger mvi implies more pronounced memory oscillations that could impact inference stability.

A.2.18. Peak-to-average power ratio

The peak-to-average power ratio, pptapr, quantifies the extremity of power draw relative to its mean level:

$$\mathrm{pptapr} = \frac{\mathrm{pp_w}}{\mathrm{ap_w}}, \tag{32}$$

with $\mathrm{pp_w}$ denoting peak power (Eq. (20)) and $\mathrm{ap_w}$ the average power (Eq. (19)). High pptapr values indicate power spikes that may trigger thermal or voltage regulation events.

A.2.19. CPU stability index

To evaluate the steadiness of processor utilization, the CPU stability index, csi, is defined as:

$$\mathrm{csi} = 1 - \frac{\sigma_{\mathrm{CPU}}}{\max\left(100, \mathrm{acup_{percent}}\right)}, \tag{33}$$

where σ_{CPU} is the standard deviation of CPU usage readings and $\mathrm{acup_{percent}}$ is the average CPU utilization (Eq. (15)). Values closer to one denote more uniform CPU loads.

A.2.20. Power efficiency index

The power efficiency index, pei, captures how effectively tokens are generated per watt of power:

$$\mathrm{pei} = \frac{\mathrm{tps}}{\mathrm{ap_w}}, \tag{34}$$

where tps is tokens per second (Eq. (14)). A greater pei reflects lower energy cost per token.

A.2.21. Model efficiency index

Model efficiency, mei, normalizes token throughput by peak memory usage:

$$\mathrm{mei} = \frac{\mathrm{tps}}{\mathrm{pru_{mb}}}, \tag{35}$$

with $\mathrm{pru_{mb}}$ defined in Eq. (18). Higher mei indicates more tokens generated per megabyte at peak demand.

A.2.22. Memory-to-CPU ratio

The memory-to-CPU ratio, mtcr, examines whether memory demands grow in tandem with processor load:

$$\mathrm{mtcr} = \frac{\mathrm{pru_{mb}}}{\mathrm{pcu_{percent}}}, \tag{36}$$

where $\mathrm{pcu_{percent}}$ comes from Eq. (16). Imbalanced values can indicate suboptimal resource utilization.

A.2.23. Memory-to-power ratio

To assess the influence of memory usage on energy draw, the memory-to-power ratio, mtpr, is defined as:

$$\mathrm{mtpr} = \frac{\mathrm{aru_{mb}}}{\mathrm{ap_w}}, \tag{37}$$

linking average RAM consumption to mean power usage.

A.2.24. RAM usage variation index

The RAM usage variation index, ruvi, normalizes the memory standard deviation by the average RAM usage:

$$\mathrm{ruvi} = \frac{\mathrm{msd_{mb}}}{\mathrm{aru_{mb}}}, \tag{38}$$

mirroring the definition of mvi but emphasizing variability per unit memory.

A.2.25. Power spike

Finally, the power spike metric, ps, captures the absolute difference between peak and minimum power:

$$\mathrm{ps} = \mathrm{pp_w} - \mathrm{mp_w}, \tag{39}$$

where $\mathrm{mp_w}$ is from Eq. (21). This value quantifies the energy burst magnitude during inference.

A.3. Derived LLM evaluation metrics

Building on the core and resource-aware measurements, derived metrics synthesize these values to reveal deeper aspects of an LLM's operational profile, including its responsiveness, overhead distribution, and per-token resource footprint.

A.3.1. Time per token

The time per token, tpt, quantifies the average latency incurred for generating each token. Shorter tpt values reflect more rapid token production, as expressed by

$$\mathrm{tpt} = \frac{\mathrm{total\_duration\_s}}{\mathrm{ec}}, \tag{40}$$

where total_duration_s is the overall inference time in seconds and ec is the total token count (Eq. (13)).

A.3.2. Load-to-inference ratio

To understand the relative cost of initialization versus generation, the load-to-inference ratio, ltir, is defined as

$$\mathrm{ltir} = \frac{\mathrm{ld_{ns}}}{\mathrm{ed_{ns}}}, \tag{41}$$

where $\mathrm{ld_{ns}}$ and $\mathrm{ed_{ns}}$ are the load and evaluation durations (Eqs. (10) and (12)). Larger values indicate that model startup contributes disproportionately to total latency.

A.3.3. Memory usage per token

The memory usage per token, mupt, assists in assessing memory scalability by normalizing average RAM consumption against throughput:

$$\mathrm{mupt} = \frac{\mathrm{aru_{mb}}}{\mathrm{ec}}, \tag{42}$$

where $\mathrm{aru_{mb}}$ is the average RAM usage (Eq. (17)). Lower mupt values suggest more memory-efficient token production.

A.3.4. Energy per token

Energy per token, ept, captures the joule cost of producing a single token by combining average power draw with inference duration:

$$\mathrm{ept} = \frac{\mathrm{ap_w} \times T_{\mathrm{local}}}{\mathrm{ec}}, \tag{43}$$

where $\mathrm{ap_w}$ denotes the average power (Eq. (19)) and T_{local} is the inference time in seconds. Minimizing ept is essential for energy-constrained deployments.

A.3.5. Prompt evaluation ratio

The prompt evaluation ratio, per, measures the fraction of total time consumed by prompt processing:

$$\mathrm{per} = \frac{\mathrm{ped_{ns}}}{\mathrm{td_{ns}}}, \tag{44}$$

where $\mathrm{ped_{ns}}$ is the prompt evaluation duration (Eq. (11)) and $\mathrm{td_{ns}}$ is the total duration (Eq. (9)). Values approaching unity indicate heavy preprocessing overhead relative to token generation.

A.3.6. Time per prompt evaluation

The time per prompt evaluation, denoted tppe, directly captures the overhead incurred during prompt parsing and preparation. This metric equals the prompt evaluation duration:

$$\mathrm{tppe} = \mathrm{ped_{ns}}, \tag{45}$$

where $\mathrm{ped_{ns}}$ is defined in Eq. (11). A lower tppe suggests more efficient prompt handling, which is especially important when multiple or iterative prompts are processed.

A.3.7. Prompt-to-generation overhead ratio

To compare the relative expense of prompt processing versus token generation, the prompt-to-generation overhead ratio, ptgor, is introduced:

$$\mathrm{ptgor} = \frac{\mathrm{ped_{ns}}}{\mathrm{ed_{ns}}}, \tag{46}$$

with $\mathrm{ed_{ns}}$ from Eq. (12). Higher values indicate that a larger portion of total execution time is consumed by prompt evaluation.

A.3.8. Prompt evaluation tokens per second

The prompt evaluation tokens per second, petps, reflects how many tokens the model can process during the prompt interpretation stage:

$$\mathrm{petps} = \frac{\mathrm{pec}}{\mathrm{ped_{ns}}/10^9}, \tag{47}$$

where pec is the count of tokens reread or reprocessed in the prompt. A higher petps denotes more rapid prompt throughput.

A.3.9. Evaluation latency per token

Evaluation latency per token, elpt, measures the average time the model requires to generate each token during inference:

$$\mathrm{elpt} = \frac{\mathrm{ed_{ns}}}{\mathrm{ec}}, \tag{48}$$

where ec is the total token count (Eq. (13)). Smaller elpt values indicate reduced per-token latency.

A.3.10. Token production energy efficiency

The token production energy efficiency, tpee, quantifies the number of tokens generated per joule of energy consumed:

$$\mathrm{tpee} = \frac{\mathrm{ec}}{\mathrm{ap_w} \times T_{\mathrm{local}}}, \tag{49}$$

with $\mathrm{ap_w}$ as in Eq. (19) and T_{local} the inference time in seconds. Higher tpee indicates superior energy efficiency.

A.3.11. Load-to-prompt ratio

The load-to-prompt ratio, ltpr, contrasts model initialization time against prompt processing:

$$\mathrm{ltpr} = \frac{\mathrm{ld_{ns}}}{\mathrm{ped_{ns}}}, \tag{50}$$

where $\mathrm{ld_{ns}}$ is given in Eq. (10). Values much greater than one suggest that startup overhead dominates prompt handling.

A.3.12. Prompt-to-total token ratio

The prompt-to-total token ratio, ptttr, indicates the fraction of tokens originating from the prompt versus those newly generated:

$$\mathrm{ptttr} = \frac{\mathrm{pec}}{\mathrm{ec}}, \tag{51}$$

where pec and ec are defined as above. A lower ratio implies that the model produces more novel tokens relative to prompt tokens.

A.3.13. Sustained inference factor

To capture steady-state operational efficiency, the sustained inference factor, sif, combines the proportion of generative output with energy-aware throughput:

$$\mathrm{sif} = \left(\frac{\mathrm{ec}}{\mathrm{pec} + \mathrm{ec}}\right) \times \left(\frac{\mathrm{tps}}{\mathrm{ap}_{\mathrm{W}}}\right), \tag{52}$$

where tps is tokens per second (Eq. (14)). Higher sif values reflect a favorable balance between fresh token generation and power-efficient operation.

References

[1] Z. Li, X. Wu, H. Du, H. Nghiem, G. Shi, Benchmark evaluations, applications, and challenges of large vision language models: A survey, 2025, arXiv preprint arXiv:2501.02189.

[2] V. Veeramachaneni, Large language models: A comprehensive survey on architectures, applications, and challenges, Adv. Innov. Comput. Program. Lang. 7 (1) (2025) 20–39.

[3] M. Shao, A. Basit, R. Karri, M. Shafique, Survey of different large language model architectures: Trends, benchmarks, and challenges, IEEE Access (2024).

[4] J. Li, W. Lu, H. Fei, M. Luo, M. Dai, M. Xia, Y. Jin, Z. Gan, D. Qi, C. Fu, Y. Tai, A survey on benchmarks of multimodal large language models, 2024, arXiv preprint arXiv:2408.08632.

[5] Z. Li, X. Wu, H. Du, H. Nghiem, G. Shi, Benchmark evaluations, applications, and challenges of large vision language models: A survey, 2025, arXiv preprint arXiv:2501.02189.

[6] C. Xu, S. Guan, D. Greene, M. Kechadi, Benchmark data contamination of large language models: A survey, 2024, arXiv preprint arXiv:2406.04244.

[7] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, A survey of large language models, 2023, arXiv preprint arXiv:2303.18223.

[8] Z. Zheng, K. Ning, Y. Wang, J. Zhang, D. Zheng, M. Ye, J. Chen, A survey of large language models for code: Evolution, benchmarking, and future trends, 2023, arXiv preprint arXiv:2311.10372.

[9] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, Y. Shan, SEED-bench: Benchmarking multimodal large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13299–13308.

[10] H. Zhou, F. Liu, B. Gu, X. Zou, J. Huang, J. Wu, Y. Li, S.S. Chen, P. Zhou, J. Liu, Y. Hua, A survey of large language models in medicine: Progress, application, and challenge, 2023, arXiv preprint arXiv:2311.05112.

[11] Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, J. Zhang, Benchmarking foundation models with language-model-as-an-examiner, Adv. Neural Inf. Process. Syst. 36 (2024).

[12] Z. Feng, W. Ma, W. Yu, L. Huang, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications, 2023, arXiv preprint arXiv:2311.05876.

[13] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, Holistic evaluation of language models, 2022, arXiv preprint arXiv:2211.09110.

[14] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, Pixiu: A large language model, instruction data and evaluation benchmark for finance, 2023, arXiv preprint arXiv:2306.05443.

[15] J. Li, X. Cheng, W.X. Zhao, J.Y. Nie, J.R. Wen, Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023, arXiv preprint arXiv:2305.11747.

[16] L. Xu, A. Li, L. Zhu, H. Xue, C. Zhu, K. Zhao, H. He, X. Zhang, Q. Kang, Z. Lan, Superclue: A comprehensive chinese large language model benchmark, 2023, arXiv preprint arXiv:2307.15020.

[17] T. Shen, S. Li, Q. Tu, D. Xiong, Roleeval: A bilingual role evaluation benchmark for large language models, 2023, arXiv preprint arXiv:2312.16132.

[18] X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W.L. Tam, X. Zhang, Alignbench: Benchmarking chinese alignment of large language models, 2023, arXiv preprint arXiv:2311.18743.

[19] Y. Zhao, J. Zhang, I. Chern, S. Gao, P. Liu, J. He, Felm: Benchmarking factuality evaluation of large language models, Adv. Neural Inf. Process. Syst. 36 (2024).

[20] Q. Huang, J. Vora, P. Liang, J. Leskovec, Benchmarking large language models as ai research agents, in: NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023.

[21] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T.B. Hashimoto, Benchmarking large language models for news summarization, Trans. Assoc. Comput. Linguist. 12 (2024) 39–57.

[22] R. Xu, Z. Wang, R.Z. Fan, P. Liu, Benchmarking benchmark leakage in large language models, 2024, arXiv preprint arXiv:2404.18824.

[23] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance, Adv. Neural Inf. Process. Syst. 36 (2024).

[24] Y. Chen, H. Wang, S. Yan, S. Liu, Y. Li, Y. Zhao, Y. Xiao, Emotionqueen: A benchmark for evaluating empathy of large language models, 2024, arXiv preprint arXiv:2409.13359.

[25] Z. Wang, CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models, in: Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing, SIGHAN-10, 2024, pp. 143–151.

[26] Z. Qiu, J. Li, S. Huang, X. Jiao, W. Zhong, I. King, Clongeval: A chinese benchmark for evaluating long-context large language models, 2024, arXiv preprint arXiv:2403.03514.

[27] A.C. Doris, D. Grandi, R. Tomich, M.F. Alam, M. Ataei, H. Cheong, F. Ahmed, Designqa: A multimodal benchmark for evaluating large language models? understanding of engineering documentation, J. Comput. Inf. Sci. Eng. 25 (2) (2025).

[28] X. Zhang, C. Li, Y. Zong, Z. Ying, L. He, X. Qiu, Evaluating the performance of large language models on gaokao benchmark, 2023, arXiv preprint arXiv:2305.12474.

[29] I. Jahan, M.T.R. Laskar, C. Peng, J.X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks, Comput. Biol. Med. 171 (2024) 108189.

[30] S. Wang, P. Wang, T. Zhou, Y. Dong, Z. Tan, J. Li, Ceb: Compositional evaluation benchmark for fairness in large language models, 2024, arXiv preprint arXiv:2407.02408.

[31] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization, 2023, arXiv preprint arXiv:2306.05087.

[32] A. Berti, H. Kourani, W.M. van der Aalst, PM-LLM-benchmark: Evaluating large language models on process mining tasks, 2024, arXiv preprint arXiv:2407.13244.

[33] X. Zhou, H. Zhao, Y. Cheng, Y. Cao, G. Liang, G. Liu, W. Liu, Y. Xu, J. Zhao, Elecbench: a power dispatch evaluation benchmark for large language models, 2024, arXiv preprint arXiv:2407.05365.

[34] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, P. Luo, Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[35] L. Sun, Y. Han, Z. Zhao, D. Ma, Z. Shen, B. Chen, L. Chen, K. Yu, Scieval: A multi-level large language model evaluation benchmark for scientific research, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, 2024, pp. 19053–19061.

[36] S. Deng, W. Xu, H. Sun, W. Liu, T. Tan, J. Liu, A. Li, J. Luan, B. Wang, R. Yan, S. Shang, Mobile-bench: An evaluation benchmark for llm-based mobile agents, 2024, arXiv preprint arXiv:2407.00993.

[37] G. Yang, C. He, J. Guo, J. Wu, Y. Ding, A. Liu, H. Qin, P. Ji, X. Liu, Llmcbench: Benchmarking large language model compression for efficient deployment, 2024, arXiv preprint arXiv:2410.21352.

[38] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, L. Sun, Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024, arXiv preprint arXiv:2402.04788.

[39] L. Spangher, T. Li, W.F. Arnold, N. Masiewicki, X. Dotiwalla, R. Parusmathi, P. Grabowski, E. Ie, D. Gruhl, Project MPG: towards a generalized performance benchmark for LLM capabilities, 2024, arXiv preprint arXiv:2410.22368.

[40] G. Son, D. Yoon, J. Suk, J. Aula-Blasco, M. Aslan, V.T. Kim, S.B. Islam, J. Prats-Cristia, L. Tormo-Banuelos, S. Kim, MM-eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models, 2024, arXiv preprint arXiv:2410.17578.

[41] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan, S. Kambhampati, Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change, Adv. Neural Inf. Process. Syst. 36 (2024).

[42] J. Liu, C. Liu, P. Zhou, Q. Ye, D. Chong, K. Zhou, Y. Xie, Y. Cao, S. Wang, C. You, P.S. Yu, Llmrec: Benchmarking large language models on recommendation task, 2023, arXiv preprint arXiv:2308.12241.

[43] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, Y. Qiao, Mm-safetybench: A benchmark for safety evaluation of multimodal large language models, in: European Conference on Computer Vision, Springer, Cham, 2025, pp. 386–403.

[44] Z. Chu, Q. Ai, Y. Tu, H. Li, Y. Liu, Pre: A peer review based large language model evaluator, 2024, arXiv preprint arXiv:2401.15641.

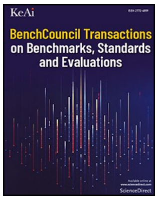
[45] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li, C. Li, Kola: Carefully benchmarking world knowledge of large language models, 2023, arXiv preprint arXiv:2306.09296.

[46] Ollama, https://github.com/ollama/ollama/blob/main/docs/api.md. (Accessed 11 January 2025).

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Full Length Article

“We don’t plagiarise, we parrot”: Cognitive load and ethical perceptions in higher education written assessment

Ibnatul Jalilah Yusof [*], Zakiah Mohamad Ashari, Lukman Hakim Ismail, Mira Panadi

Universiti Teknologi Malaysia, Malaysia

ARTICLE INFO

Keywords:
Cognitive load
Ethical perceptions
Parroting behaviour
Plagiarisms

ABSTRACT

Generative artificial intelligence has reshaped written assessment in higher education and sharpened concerns about “parroting,” the undisclosed use of AI-generated text with minimal cognitive engagement. This study examines the cognitive and ethical mechanisms underlying parroting among undergraduates in one Malaysian research university. Drawing on Cognitive Load Theory and Dual-System Theory, parroting is conceptualised across three dimensions: intrinsic load, extraneous load, and ethical rationalisation. Survey responses from 211 students were analysed using Rasch measurement to evaluate item reliability, construct separation, differential item functioning (DIF) across academic fields and item hierarchies. Results indicate that items function equivalently for engineering, non-engineering, and science students, supporting the instrument’s fairness and stability. Overall, findings show that parroting is most strongly driven by extraneous pressures such as vague instructions and heavy workload, followed by intrinsic challenges related to writing confidence and conceptual understanding. Ethical rationalisation is endorsed least frequently but becomes more salient when institutional guidance on AI use is unclear. The study offers implications for pedagogy and policy, underscoring the need for explicit AI-use guidelines, improved task design, and learning environments that promote ethically responsible engagement with generative technologies.

Background

The advent of generative artificial intelligence (GenAI) tools such as ChatGPT has profoundly transformed perspectives on written assessments in higher education [1,2]. These tools offer students instant access to grammatically polished, coherent text with minimal cognitive effort. While this may appear to support productivity, a more insidious challenge has emerged: the increasing prevalence of cognitively superficial writing, whereby students submit polished work that lacks evidence of internalisation, reflection, or intellectual authorship [3–5].

This phenomenon aligns with what scholars refer to as *parroting*; the act of reproducing content verbatim or with minimal alterations, devoid of meaningful comprehension or engagement [6,7]. Traditionally, parroting has been associated with rote learning and low-level memorisation [8,9], but in the context of AI-generated content, it signals a more complex and ethically ambiguous behaviour. The student’s work may appear fluent and structured, yet lacks the deeper cognitive investment expected in higher education writing tasks.

Why do students do it? Why do they submit assignments that are not the product of their own thinking, but rather stitched together from AI-generated text? This study assumes that such behaviour is not simply a matter of laziness or rebellion, but a reflection of deeper psychological and situational mechanisms.

Firstly, consider the existing definitions of plagiarism. Whether framed in legal, institutional, or scholarly terms, plagiarism is typically defined as the use of another human’s intellectual work without proper attribution [10,11]. However, AI-generated text does not originate from a person, and current plagiarism frameworks are often ill-equipped to address the complexities introduced by AI-assisted writing [12,13]. While the Committee on Publication Ethics [14] maintains that AI cannot be listed as an author due to its lack of agency and accountability, many students may continue to use such tools in ways that obscure their own contribution.

This grey area has led to a dangerous misconception; that parroting AI-generated content is acceptable because *‘it is not technically plagiarism’*. Furthermore, as institutions race to integrate AI literacy into curricula, few have developed clear policies addressing the boundaries between AI support and AI substitution [1]. In the absence of clear

Peer review under the responsibility of The International Open Benchmark Council.

[*] Corresponding author.
E-mail address: ijalilah@utm.my (I.J. Yusof).

https://doi.org/10.1016/j.tbench.2025.100254
Received 25 August 2025; Received in revised form 6 December 2025; Accepted 10 December 2025
Available online 12 December 2025

ethical guidelines, students may normalise parroting as a legitimate academic strategy [1,4,15], especially under cognitive strain, time pressure, or a desire for efficiency.

Secondly, a more complex reason lies in the psychological interplay between Dual-System Theory (DST) and Cognitive Load Theory (CLT). Students may not always make deliberate ethical judgments; rather, they respond impulsively under pressure. For example, some may think, *"ChatGPT writes better than I ever could, why should I bother?"* is a reflection of System 1 thinking, the fast, intuitive, and automatic mode described in DST [16]. Others may feel cognitively overwhelmed by the task itself and conclude, *"This is too hard, AI can do it faster and better than me"* is a coping response aligned with CLT [17,18].

In both cases, students opt out of the mental effort required for authentic academic work, either because the task exceeds their working memory capacity especially in academic writing tasks requiring synthesis, originality, and technical precision, [19–21] or because an easier alternative presents itself [3–5]. This convergence of automatic decision-making and cognitive strain may explain why students rationalise GenAI use without engaging in meaningful writing processes.

Problem statement

Despite growing attention to the academic implications of GenAI, the specific phenomenon of parroting remains under-theorized and insufficiently measured, particularly within Malaysia higher education [22]. As such, parroting exists in a grey zone technically legal, ethically ambiguous, and psychologically convenient. Yet, little is known about how students themselves rationalise or justify this behaviour, and under what conditions they are more likely to engage in it.

Therefore, the present study focuses on two key areas:

(i) the cognitive (intrinsic and extraneous load) mechanisms that lead students to rely on AI-generated content in their academic writing,
(ii) their ethical perceptions regarding the acceptability of such practices.

Literature review

Parroting is not a new term in educational discourse. It draws from the observable behaviour of parrots mimicking sounds without true comprehension. While animal cognition research continues to debate the cognitive capacities of parrots, particularly in the case of *Alex the African Grey* who demonstrated the ability to respond meaningfully to prompts [23], this study adopts a more conservative interpretation. Here, parroting refers to surface-level mimicry that lacks cognitive depth, shaped more by repetition and performance than by meaningful understanding.

In classroom contexts, parroting has traditionally been associated with rote memorisation, where students reproduce information with little or no conceptual engagement [8,9]. Such practices may yield correct answers but rarely reflect deep learning or analytical thought. The concern becomes even more pressing when parroting migrates into written assessments, where the stakes are higher, and the appearance of originality can mask a lack of authentic intellectual contribution. It takes on a more deliberate and strategic form.

Glatt [7] examined this phenomenon in students' written work and identified three interrelated drivers:

(i) Insecurity about writing ability, which leads learners to appropriate existing text rather than risk producing original prose; and
(ii) Time pressure, which encourages shortcuts that prioritize rapid completion over thoughtful composition; and
(iii) Low motivation, prompting students to avoid the effortful stages of planning, drafting, and revision.

Glatt's [7] conceptualization of parroting reflects a traditional academic context, where students, despite resorting to copying, would still engage with the source material by reading and minimally modifying it. Parroting was thus a low-effort strategy, but not entirely void of cognitive involvement. In contrast, the rise of GenAI introduces a more detached form of mimicry, one that allows students to bypass even the most basic interaction with the text. Building on this, the present study conceptualizes parroting not merely as mindless copying but as a form of cognitive outsourcing, where students offload the demanding aspects of intellectual work to an external system. This shift invites deeper inquiry into how such behaviour is shaped not only by academic pressure but also by the cognitive architecture of decision making and the ethical reasoning (or absence thereof) that justifies it.

According to DST [16], human thinking operates through two modes which are System 1, which is fast, automatic, and intuitive, and System 2, which is slow, effortful, and reflective. In academic settings, writing typically demands System 2 processing such as analysing ideas, structuring arguments, and evaluating sources. However, when confronted with cognitively overwhelming tasks or vague assignment instructions, students often default to shortcuts such as copy-and-paste (System 1) [3,7,24]. This reliance on intuition over analysis is not a sign of laziness but a systematic response to cognitive strain, particularly when combined with high intrinsic or extraneous cognitive load [18].

Here, CLT complements DST. Writing tasks that involve abstract or unfamiliar content increase intrinsic load, while unclear instructions add extraneous load. Both can easily exceed students' working memory capacity. Instead of engaging in laborious thought, students may rely on GenAI tools as cognitive crutches [25] inserting or modifying text with minimal intellectual contribution. The ethical dimension deepens this issue. Students may engage in what Bandura [26] calls moral disengagement, a psychological process in which individuals justify unethical behaviour through diffusion of responsibility, minimization of harm, or ambiguity in norms [27,28]. If there are no explicit university policies about GenAI citation [1,4,15,29], or if peers do it without consequence [30,31]. In doing so, they enter a grey zone where academic mimicry masquerades as competence.

This tendency was operationalised in the present study, where parroting behaviours were measured across three dimensions:

(i) Intrinsic Load-Induced Parroting: When students perceive the task as beyond their own ability (e.g., abstract topic, unfamiliar concepts, requirement for synthesis), students may experience intrinsic cognitive overload. Believing that AI will produce better results than they could, they bypass deep thinking (System 2) and uncritically use AI-generated content.
(ii) Extraneous Load-Induced Parroting: When the assignment is poorly structured, vague, or confusing, students may encounter extraneous cognitive load. Rather than seeking clarification or attempting to navigate unclear instructions, they impulsively rely on AI as a shortcut, again engaging System 1 thinking.
(iii) Ethically Rationalised Parroting: Even when aware of academic norms, students may morally disengage, justifying their use of AI based on peer behaviour, lack of institutional guidelines, or the perception that AI-generated text is not 'real plagiarism'. This rationalisation supports System 1 decisions that minimize effort and suppress ethical reflection.

Methodology

Research design

This study employed a quantitative, cross-sectional survey design to investigate parroting behaviour defined as the undisclosed use of generative AI content with minimal cognitive engagement among education students. Drawing on Cognitive Load Theory and Dual-System Theory, parroting was conceptualized across three dimensions:

intrinsic load–induced, extraneous load–induced, and ethically rationalised behaviours. A Rasch measurement approach was adopted to examine the psychometric properties of the instrument and to establish endorsement hierarchies within each dimension.

Sample

Table 1 presents the distribution of respondents selected through purposive sampling. A total of 211 participants from one research university in Malaysia were included in this study. The largest proportion came from Non-Engineering/Social Sciences fields (n = 87, 41.2%), representing disciplines such as Education, Business, Psychology, and Human Resource. This was followed by respondents from Engineering (n = 65, 30.8%), which included Electrical, Electronic, Biomedical, and Mechanical Engineering. The remaining 28.0% (n = 59) consisted of participants from Science and Technology, such as Building Surveying, Architecture, Science, and Mathematics.

In Rasch measurement, sample adequacy is determined not only by the total number of respondents but also by the distribution of responses across rating scale categories. Linacre [32,33] suggested that at least 10 responses per rating scale category are recommended. Linacre [34] also notes that a sample of 80–100 well-targeted respondents provides stable item calibrations within approximately ±1 logit at a 95% confidence level. In the present study, all rating categories met this criterion, supporting the adequacy of N = 211 for generating stable Rasch estimates for descriptive purposes.

Nevertheless, although the sample size and response distribution were sufficient for Rasch analysis, the use of purposive sampling from a single university limits the generalisability of the findings. The results primarily reflect the assessment practices and GenAI-related behaviours of undergraduates within this specific institutional context and may not fully represent students from other universities or educational settings. Accordingly, the findings should be interpreted as context-specific and exploratory rather than universally generalisable.

Instrument

The instrument used in this study was a structured questionnaire developed to investigate students' parroting behaviour in AI-assisted academic writing, along with their general awareness of AI tools.

The questionnaire consisted of two sections (Table 2). Section A captured demographic information such as, awareness of generative AI, knowledge of access, and knowledge of use. These items were measured using a four-point Likert agreement scale (Strongly Agree = 4 to Strongly Disagree = 1). Section B measured parroting behaviour across three constructs: Intrinsic-Induced (10 items), Extraneous-Induced (10 items), and Ethically Rationalised (10 items). Responses were rated on a five-point Likert frequency scale ranging from Always (5) to Never (1).

Table 3 shows item reliabilities across all subscales were acceptable (>0.70), indicating consistent measurement of items within each construct. Item separation values also suggest adequate to strong differentiation among items (>2.0). For the Intrinsic-Induced construct, the item reliability was 0.98 with an item separation of 6.71, while the Extraneous-Induced construct recorded an item reliability of 0.98 and item separation of 6.66, both reflecting strong item stability. The Ethically Rationalised construct demonstrated an item reliability of 0.92 with an item separation of 3.42, indicating meaningful variation in item difficulty. Taken together, these indices meet recommended thresholds for Rasch measurement and support the suitability of the instrument for descriptive use in this study [35,36]

Table 1
Respondent distribution.

Field		
	N	%
Engineering	65	30.8%
Non-Engineering / Social Sciences	87	41.2%
Science and Technology	59	28.0%

Table 2
Items tabulation.

Section	Measure	No of Item	Likert-Type Scale
A	Demographic Awareness of Generative AI Know How to Access Know How to Use	3	4 Point Likert (Agreement) Strongly Agree (4) Agree (3) Disagree (2) Strongly Disagree (1)
B	**Parroting Behaviour**		5 Point Likert Scale (Frequency) Always (5) Often (4) Sometimes (3) Rarely (2) Never (1)
	Intrinsic Load-Induced	10	
	Extraneous Load-Induced	10	
	Ethically Rationalised	10	

Table 3
Item and person reliability and separation for parroting behaviour subscales.

Parroting Behaviour	Item		Person	
	Reliability	Separation	Reliability	Separation
Intrinsic Induced	.98	6.71	.86	2.43
Extraneous Induced	.98	6.66	.85	2.41
Ethically Rationalised	.92	3.42	.80	2.01

Group behaviour across constructs

Group-based item functioning analysis was conducted to examine whether items performed equivalently across different respondent groups: Engineering, Non-Engineering/Social Sciences, and Science and Technology. This analysis evaluates the extent to which item difficulty estimates remain stable when comparing students from distinct academic fields, thereby identifying any potential measurement bias. The analysis ensures that observed differences in item responses reflect true variations in the underlying construct rather than systematic advantages or disadvantages for particular groups [37]

Table 4 summarises the T-scores, size indices, and standard errors for each construct across the three groups. For the Intrinsic-Induced construct, T-scores were 1112 (Engineering), 1857 (Non-engineering), and 1074 (Science and Technology); for Extraneous-Induced, the corresponding T-scores were 1260, 2043, and 1178; and for Ethically Rationalised, 1017, 1659, and 876. The size statistics index for all constructs and groups was consistently 0.00, and standard errors remained small (0.04–0.05) indicate no evidence of differential item functioning, meaning that the items behaved equivalently across academic fields. Such consistency is a strong indicator of the instrument's validity and fairness. It shows that the items are robust across disciplinary contexts and that the constructs are generalisable beyond a single group. For a behavioural measure that spans cognitive load responses and ethical rationalisation processes, this invariance is particularly important, as it confirms that patterns observed in the hierarchy are not discipline-specific but represent broader student tendencies.

Data collection

Data were collected online via a google form. The introductory section stated the purpose of the study, ensured anonymity, and reiterated participants' right to refuse or withdraw. No identifying information was collected.

Table 4
Group-based item functioning analysis.

PERSON class/group specification is: DIF=@FIELD ITEM class/group specification is: DPF=$S1W1

Construct	Engineering			Non-Engineering			Science and Technology		
	T. SCORE	SIZE	S.E	T. SCORE	SIZE	S.E	T. SCORE	SIZE	S.E
Intrinsic-Induced	1112	0.00	0.05	1857	0.00	0.04	1074	0.00	0.05
Extraneous-Induced	1260	0.00	0.05	2043	0.00	0.04	1178	0.00	0.05
Ethically Rationalised	1017	0.00	0.04	1659	0.00	0.04	876	0.00	0.05

Results

Awareness, access, and usage of AI-based writing tools

Table 5 presents the distribution of responses regarding participants' awareness and use of AI-based writing tools. A large majority of students reported high awareness of generative AI tools (e.g., ChatGPT) for academic writing, with 77.3% strongly agreeing and 19.0% agreeing. Only a small proportion (3.3%) disagreed, while 0.5% strongly disagreed. Similarly, most students indicated they know how to access AI-based writing tools, with 68.7% strongly agreeing and 26.1% agreeing. A small percentage (4.7%) disagreed and (0.5%) strongly disagreed, suggesting some gaps in accessibility knowledge. In terms of usage skills, 71.6% strongly agreed and 20.9% agreed that they know how to use AI-based writing tools. Only 5.2% disagreed, while 2.4% strongly disagreed.

Person–item mean distribution across constructs

The person–item distribution across constructs provides an overview of how respondents' endorsement levels align with the difficulty of items within each subscale. By comparing the mean person measures against the item mean set at 0.00 logits, this analysis illustrates how well each construct targets the sample and highlights relative tendencies in endorsing intrinsic-induced, extraneous-induced, and ethical-induced parroting behaviours.

Fig. 1 illustrates the distribution of person mean measures across the three constructs relative to the item mean (0.00 logits). The Extraneous-Induced Parroting construct was positioned slightly above the item mean (+0.18 logits), indicating that students, on average, displayed a comparatively higher tendency to rely on AI when encountering unclear or poorly structured assignments. In contrast, both Intrinsic-Induced Parroting (−0.14 logits) and Ethically Rationalised Parroting (−0.38 logits) fell below the item mean.

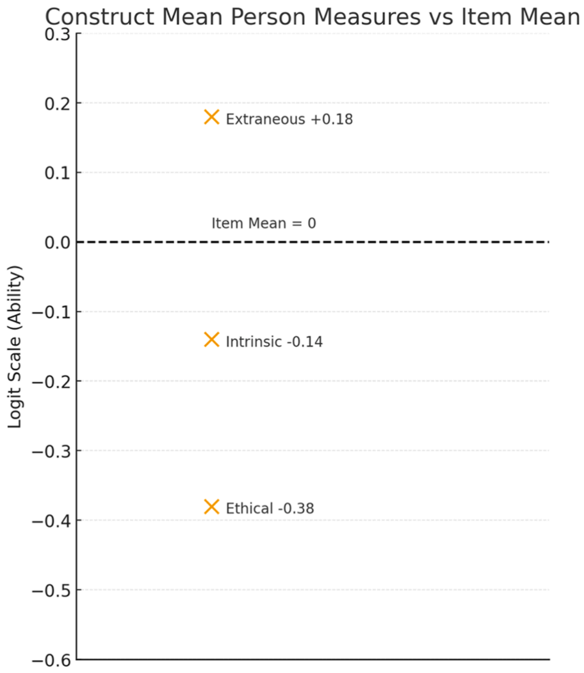


Fig. 1. Comparative mean person–item distribution.

Item hierarchy (intrinsic load induced)

Fig. 2 presents the Rasch item map for intrinsic-induced parroting behaviours, displaying item difficulty along the logit scale. Using the mean item measure of 0.00 logits and the standard deviation of 0.58 logits as reference points, items located below –0.58 logits represent behaviours that are easier for students to endorse. Items falling within one standard deviation of the mean (–0.58 to +0.58 logits) reflect *moderately endorsed* behaviours, aligning with the overall difficulty level of the construct. Items positioned above +0.58 logits exceed one standard deviation from the mean and represent behaviours that are *harder to endorse*, requiring a stronger inclination toward intrinsic-induced parroting before respondents agree with them. This pattern illustrates how the distribution of item difficulties differentiates the relative ease or challenge of endorsing each behaviour within the construct.

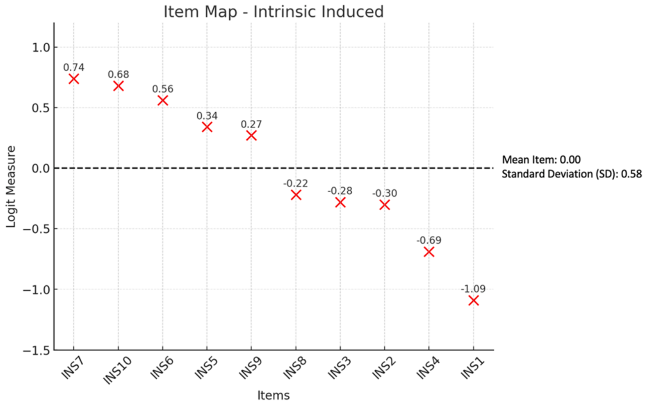


Fig. 2. Item hierarchy (intrinsic load-induced).

Table 5
Agreement levels on awareness, access, and usage of AI-based writing tools.

Item	Agreement (%)			
	1 (Strongly Disagree)	2 (Disagree)	3 (Agree)	4 (Strongly Agree)
I am aware that AI tools such as (e.g. Chatgpt) can assist with academic writing.	1 (0.5%)	7 (3.3%)	40 (19.0%)	163 (77.3%)
I know how to access AI-based writing tools.	1 (0.5%)	10 (4.7%)	55 (26.1%)	145 (68.7%)
I know how to use AI-based writing tools.	5 (2.4%)	11 (5.2%)	44 (20.9%)	151 (71.6%)

Easier endorsement

The items most easily endorsed include INS1 (“I use AI-generated text directly when the assignment topic feels too complex”, –1.09 logits) and INS4 (“I paraphrase AI output minimally when I cannot make sense of the full question”, –0.69 logits). Their positions well below the -0.58 indicate that, within the intrinsic-induced construct, students are comparatively more willing to admit relying on AI-generated text when they feel overwhelmed by task demands. In practical terms, these behaviours require less “activation” of intrinsic-induced parroting tendencies for respondents to agree with them, suggesting that using AI as a direct shortcut in challenging tasks is a relatively common strategy in this sample.

Moderate endorsement

Several items fall within this range. For instance, INS2 (“I insert AI text into my assignment when the task involves too many unfamiliar concepts,” –0.30 logits) and INS3 (“I rely on AI content without changes when I do not know how to begin writing,” –0.28 logits) indicate reliance on AI when students feel uncertain about how to initiate their work or structure their ideas.

Meanwhile item INS8 (“I use AI content that sounds professional to cover up my confusion with the topic,” –0.22 logits), item INS9 (“When the task is too hard, I submit AI-generated writing to make it look like I understand,” 0.27 logits) and INS5 (“I believe AI-generated content is better than mine, so I use it without making significant changes,” 0.34 logits) reflect a more sophisticated form of reliance, where students use AI not only to complete tasks but also to enhance the perceived quality or intellectual depth of their work.

Lastly, INS6 (“I skip outlining my ideas when I use AI to help with writing difficult topics,” 0.56 logits) illustrates how intrinsic cognitive strain can interrupt students’ ability to plan and structure their ideas. Collectively, these items reflect behaviours that students engage in situationally, often when facing cognitive uncertainty, feeling underprepared, or perceiving AI as a tool that compensates for gaps in their understanding or ability.

Less likely endorsement

Items positioned above one standard deviation from the mean (> +0.58 logits) represent behaviours that are less likely to be endorsed, reflecting actions that only students with a stronger tendency toward intrinsic-induced parroting admit to engaging in. Two items fall into this category. INS10 (“I use AI tools to make my assignments appear thoughtful even when I do not fully understand them,” 0.68 logits) captures a more deliberate form of reliance on AI, where students strategically enhance the perceived depth of their work despite limited comprehension. Similarly, INS7 (“I do not revise AI-generated responses because the assignment feels mentally exhausting,” 0.74 logits) reflects a high level of dependence driven by mental fatigue or cognitive overload, where students bypass revision entirely. Their positions above +0.58 logits indicate that these behaviours require a comparatively stronger inclination toward intrinsic-induced parroting before students are willing to endorse them. In other words, although students may frequently rely on AI for support in challenging tasks, only a smaller subset acknowledges engaging in these more intensive or effort-avoiding behaviours.

Item hierarchy (extraneous load induced)

Fig. 3 presents the Rasch item map for extraneous-induced parroting behaviours, displaying the relative difficulty of each item along the logit scale. Using the mean item measure of 0.00 logits and the standard deviation of 0.58 logits as reference points, items located below –0.58 logits represent behaviours that are easier for students to endorse, typically reflecting actions taken in response to external pressures such as limited time or heavy workload. Items falling within one standard deviation of the mean (–0.58 to +0.58 logits) reflect *moderately endorsed* behaviours, indicating tendencies that students acknowledge in certain situations but not as consistently as the easier items. Items positioned above +0.58 logits exceed one standard deviation from the mean and therefore represent *harder-to-endorse* behaviours, requiring a stronger inclination toward extraneous-induced parroting before respondents agree with them. This distribution demonstrates how the item hierarchy differentiates the relative ease or challenge of endorsing each behaviour within the extraneous-induced construct.

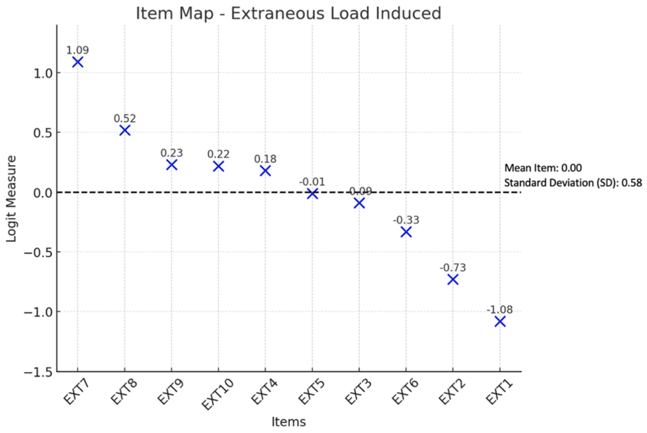


Fig. 3. Item hierarchy (extraneous load-induced).

Easier endorsement

Items falling more than one standard deviation below the mean (logits < –0.58) reflect behaviours that students most readily admitted when experiencing extraneous cognitive load. Two items; EXT1 (–1.08 logits) and EXT2 (–0.73 logits) were positioned in this category. EXT1 (“I use AI content in my work when the instructions are hard to follow.”) and EXT2 (“I use AI-generated paragraphs as-is when I cannot understand the assignment requirements”) indicate a strong tendency to rely on AI when assignment expectations are unclear or difficult to interpret. Their comparatively low logit values suggest that these behaviours represent common coping mechanisms, with students turning to AI to navigate ambiguity or incomprehensible instructions. These behaviours require minimal extraneous-induced prompting, making them the most easily endorsed within the construct.

Moderate endorsement

Items located within one standard deviation of the mean (–0.58 to +0.58 logits) represent behaviours that students acknowledge under certain conditions but do not endorse as consistently as the easier items. This group includes EXT6 (“When writing requirement is high, I use AI to generate content for assignment”, –0.33 logits), EXT3 (“I use AI content as if it were my own when I have too many assignments”, –0.09 logits), EXT5 (“When a deadline is approaching, I focus on changing just a few words from the AI output”, –0.01 logits), EXT4 (“I slightly reword AI-generated content and submit it when the assignment guidelines are unclear”, 0.18 logits), EXT10 (“If I run out of time, I rely on AI to produce work that seems well-structured and original. “, 0.22 logits), and EXT9 (“I use AI to complete writing quickly when I do not understand how to structure the assignment”, 0.23 logits). These behaviours capture situational reliance on AI driven by external pressures such as high workload, impending deadlines, vague requirements, or difficulty structuring assignments.

Meanwhile, EXT8 (“I avoid thinking critically when AI already provides a full answer under time pressure") sits at the upper boundary of this range, reflecting a behavioural shift toward cognitive offloading. Overall, items in this band illustrate behaviours that are neither frequent defaults nor rare occurrences, but rather responses enacted when external task demands intensify.

Less likely endorsement

Items positioned above one standard deviation from the mean (logits > +0.58) reflect behaviours that students are least willing to endorse. Only one item, EXT7 (1.09 logits), fell into this category. EXT7 ("When assignment instructions are vague, I allow AI to do most of the writing without my input") represents a more substantial level of reliance on AI, indicating near-total delegation of writing responsibilities under extraneous load. Its high logit value suggests that this behaviour requires the strongest extraneous-induced prompting and is endorsed only by a smaller subset of respondents. This pattern indicates that while students often rely on AI to navigate confusing or demanding tasks, fully surrendering authorship to AI without contributing their own ideas remains a comparatively uncommon practice.

Item hierarchy (ethical rationalised)

Fig. 4 illustrates the Rasch item–person map for ethical rationalisation–induced parroting behaviours, with item difficulty plotted along the logit scale. Using the mean item measure of 0.00 logits and the standard deviation of 0.29 logits as interpretive thresholds, items below –0.29 logits are classified as *easier to endorse*, items between –0.29 and +0.29 logits as *moderately endorsed*, and items exceeding +0.29 logits as *less likely to be endorsed*. This framework provides a clear differentiation of behavioural tendencies related to ethical rationalisation in the context of AI-assisted academic work.

Easier endorsement (negative logits)

Items falling below –0.29 logits represent ethical-induced behaviours that students are more inclined to endorse. Three items were located in this range: ETC8 (–0.32 logits), ETC4 (–0.38 logits), and ETC2 (–0.39 logits). These items reflect behaviours such as assuming AI use is acceptable when no explicit rules are provided ("If there are no clear rules about AI usage, I assume it is acceptable to use it freely," ETC8, –0.32 logits), withholding disclosure when institutional or course policies are unclear ("If there are no clear policies, I do not mention AI assistance in my work," ETC4, –0.38 logits), and avoiding citation when assignment guidelines do not address AI usage ("I avoid mentioning AI use when assignment instructions do not specify citation rules," ETC2, –0.39 logits). Their comparatively low logit values indicate that students readily rationalise nondisclosure of AI use when external guidance is uncertain, suggesting that ambiguity around rules and expectations strongly facilitates ethical rationalisation and non-transparent practices.

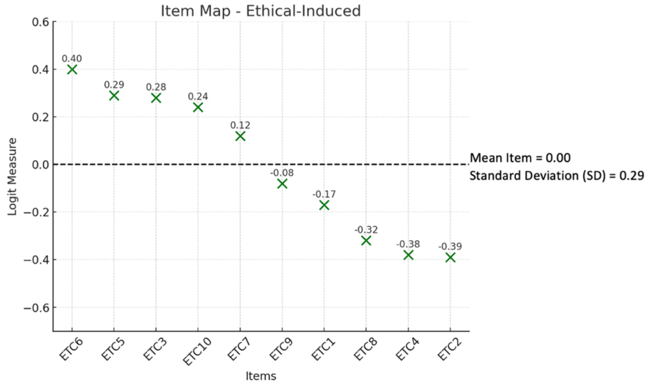


Fig. 4. Item hierarchy (ethical rationalised).

Moderate endorsement

Items positioned between –0.29 and +0.29 logits represent ethical-induced behaviours that students endorse under certain circumstances but not as consistently as the easier items. Six items were in this range: ETC1, ETC9, ETC7, ETC10, ETC3, and ETC5. These items reflect behaviours such as choosing not to cite AI-generated content when it appears generic or obvious ("I do not cite AI tools when the content seems generic or obvious," ETC1, –0.17 logits) and normalising AI use because peers do the same ("I believe it is fine to use AI to complete difficult tasks, especially when others do the same," ETC9, –0.08 logits).

Students also report feeling comfortable relying on AI due to perceived low detection risk ("I know it is hard to detect AI use, so I feel safe using it without making changes," ETC7, 0.12 logits) and submitting AI-generated work without acknowledgment ("I submit AI-generated work without acknowledgment, assuming that my lecturer will not notice," ETC10, 0.24 logits).

Additional behaviours in this category include uncertainty about proper citation ("I submit content from AI tools without citing them because I am unsure how to do it properly," ETC3, 0.28 logits) and a belief that AI-generated material does not constitute plagiarism ("I do not consider it plagiarism if I use content generated by AI tools in my assignments," ETC5, 0.29 logits). Together, these items illustrate ethically ambiguous practices that emerge when students weigh convenience, perceived norms, and uncertainty about proper academic procedures.

Harder endorsement

Only one item exceeded the +0.29-logit threshold, indicating a behaviour that students are least willing to endorse. ETC6 reflects a more explicit rationalisation of nondisclosure, capturing the belief that AI-generated content does not require citation because it lacks a human author ("I believe AI-generated content does not need to be cited because it has no human author," ETC6, 0.40). Its comparatively higher logit value indicates that students generally resist endorsing this stance, suggesting that outright dismissal of citation norms represents a more extreme form of ethical rationalisation within this construct.

Discussion

Extraneous-load induced parroting

The Rasch analysis highlights a clear pattern across the three dimensions of parroting behaviour. The extraneous dimension shows the highest average endorsement, meaning that students are most likely to use AI-generated text when they face external pressures such as tight deadlines, unclear instructions, or heavy workloads. This finding is consistent with cognitive load theory (CLT) which explains that poorly designed tasks can overload students' working memory and push them toward surface-level strategies [17,38,39]. In this context, AI-related parroting should not be viewed simply as disengagement or academic dishonesty; rather, it emerges as a compensatory strategy students adopt when instructional design unintentionally overwhelms their working memory [3,7,24]. When assignment clarity is low or deadlines converge, students may perceive AI tools not as shortcuts, but as necessary scaffolds to manage competing academic demands.

A closer look at the item difficulty hierarchy reveals a clear pattern in how students respond to these pressures. The most easily endorsed behaviours indicate that when assignment requirements are confusing or instructions lack specificity, students resort to AI as a compensatory strategy. This is consistent with CLT which emphasizes that ambiguous or poorly structured tasks create unnecessary processing demands unrelated to the actual learning goal [17,18,25]. When students report that they submit AI content “as-is” under such conditions, it underscores how extraneous load can redirect effort away from meaningful learning toward superficial task completion. Rather than investing effort in interpreting ambiguous expectations, students, consistent with Dual-Process Theory (DPT) bypass deeper cognitive processing (System 2) and instead rely on the automatic, heuristic responses (System 1) by outsourcing their work to AI tools [16,24]

A second set of behaviours reflects reliance on AI in response to situational pressures such as approaching deadlines, multiple competing assignments, demanding writing requirements, or difficulty structuring ideas. Students report using AI to accelerate writing, to navigate dense or vague guidelines, or to produce text that appears coherent when they cannot organise their thoughts under pressure. Unlike the more automatic reliance observed in the easily endorsed behaviours, these patterns emerge when external task demands escalate and students feel unable to allocate sufficient cognitive resources to meet expectations. This aligns with prior research showing that when extraneous load interacts with heavy workload, learners adopt “satisficing” strategies meeting minimum requirements with minimal effort [40,41]. Consequently, AI is used strategically to maintain productivity, reduce the time needed to generate or revise content, or meet performance standards with minimal cognitive investment. Here, the issue is not lack of capability, but rather misalignment between the demands of the task and the cognitive resources available to students.

A more concerning behavioural tendency appears when students describe relinquishing critical thinking to AI under time pressure or allowing AI to produce the majority of the writing when instructions are vague. These behaviours signal a transition from targeted assistance to broader cognitive offloading [42–44], where AI is used not only to support writing but also to replace central components of the writing process. Although fewer students endorse these behaviours, their presence indicates that extraneous load can create conditions where learners begin to withdraw from meaningful cognitive participation. Such reliance suggests that when instructional demands exceed students’ perceived capacity for engagement, AI becomes a surrogate writer rather than a supplemental tool.

Taken collectively, the extraneous-induced construct highlights how avoidable instructional factors play a central role in prompting parroting behaviours [18,39]. When assignment guidelines are unclear, expectations are ambiguous, or workload feels unmanageable, students shift toward AI-driven shortcuts to manage the external pressures placed upon them. Importantly, reliance on AI in these situations is not necessarily a reflection of limited ability or motivation, but rather a response to the cognitive disruption caused by poorly structured or insufficiently supported tasks.

These findings underscore a key pedagogical implication: reducing extraneous cognitive load can substantially diminish the need for students to adopt AI-based coping strategies [17]. For instance, when lecturers provide explicit instructions, structured assignment scaffolds, and timely formative feedback [45,46], students experience fewer cognitive pressures that push them toward superficial shortcuts. By minimising unnecessary cognitive strain and clarifying the pathways to task completion, educators can foster deeper engagement and create learning environments that support authentic academic effort rather than reliance on parroting behaviours.

Intrinsic-load induced parroting

Academic writing involves complex cognitive skills such as synthesizing ideas and using disciplinary language. Intrinsic load–induced parroting arises when students perceive the task as exceeding their own ability. When topics feel abstract, concepts are unfamiliar, or the assignment requires synthesis they are not confident performing, students experience intrinsic cognitive overload. In these moments, they tend to believe that AI can produce better work than they could on their own [47,48], which leads them to bypass deeper reasoning and rely uncritically on AI-generated content as a way to cope with the demands of the task.

The item hierarchy results show that the most readily acknowledged behaviours occur at the initial stages of task engagement, where students face uncertainty about how to interpret questions or how to begin formulating ideas. In such moments, students often turn to AI-generated text as a way to reduce the cognitive friction associated with understanding complex topics. Rather than struggling through conceptual ambiguity, they use AI as an immediate scaffold, allowing them to bypass the mental effort required for early comprehension and idea generation [49–51]. This pattern aligns with cognitive load theory, which posits that learners experiencing high intrinsic load are more likely to seek external aids that alleviate the burden on working memory.

As writing tasks become more conceptually demanding, reliance on AI evolves into more nuanced forms. Students report using AI to fill gaps in understanding, articulate unfamiliar concepts, or supply language that feels more polished than what they believe they can produce independently. These behaviours are situationally driven, often triggered when students lack the confidence or capacity to engage fully with the assignment [49,52]. When students doubt their ability to synthesise disciplinary ideas or express them with sufficient clarity, AI becomes a compensatory mechanism that stabilises their performance. This is consistent with scholarship showing that low writing self-efficacy increases the use of external supports and shortcuts [51,53], particularly when students face tasks requiring specialised vocabulary or higher-order reasoning [54,55].

The least acknowledged behaviours involve more profound forms of cognitive disengagement. These include relying on AI to produce work that conveys understanding the student does not possess, or skipping revision entirely because the task feels mentally exhausting. These patterns signal a shift from targeted support to near-total cognitive offloading, where students relinquish ownership of the writing process [25,56]. Although these behaviours are less frequently endorsed, their presence is there: they illustrate how overwhelming intrinsic load can push some students toward avoidance rather than assistance [56,57], resulting in limited interaction with the cognitive processes that underlie learning.

Overall, the intrinsic-induced construct reveals a continuum of behaviours shaped by cognitive strain, writing insecurity, and perceived deficits in conceptual understanding. Students tend to rely on AI when they cannot interpret a task or when the complexity of disciplinary content overwhelms their working memory, and this reliance becomes more sophisticated as cognitive challenges deepen. For some, AI serves as pragmatic support to initiate writing; for others, it becomes a tool to enhance the appearance of competence or, in a smaller subset, a means of withdrawing from cognitive effort altogether. This pattern is consistent with earlier work showing that writing insecurity is a major driver of plagiarism-like behaviours [6,7,51], highlighting that academic integrity policies alone are insufficient. Students who lack confidence in their writing or conceptual abilities are more likely to view parroting as their only viable option. These findings underscore the need for targeted scaffolding, and instructional interventions that build students’ confidence in engaging with complex ideas [51,52]. Strengthening these capacities may reduce dependence on AI-driven shortcuts and promote more sustained and meaningful cognitive participation in academic writing.

Ethical rationalised parroting

Ethical rationalisation–induced parroting refers to behaviours in which students justify the nondisclosure or improper use of AI by appealing to ambiguity, perceived norms, or personal interpretations of academic rules [58]. Rather than being driven by cognitive overload, these behaviours arise when students reinterpret ethical boundaries in ways that make AI reliance seem acceptable, excusable, or harmless [4, 15]. This construct captures the psychological mechanisms through which students normalise nondisclosure and minimise perceived wrongdoing in the context of generative AI.

Students most readily acknowledge behaviours that involve using ambiguity as a justification for nondisclosure. When institutional policies do not explicitly address AI use, or when assignment guidelines fail to clarify citation expectations, students report feeling comfortable omitting acknowledgment of AI-generated content. This pattern suggests that uncertainty around rules creates a permissive space where students interpret silence as implicit approval. Ethical boundaries become negotiable when learners believe responsibility lies not in adhering to academic norms but in the institution's ability to articulate them. Such reasoning reflects classic moral disengagement processes, in which individuals shift accountability away from themselves by claiming ambiguous standards or external omissions [27,42].

This finding raises an important concern: if students believe using AI without proper engagement is "normal," the issue is less about individual dishonesty and more about how academic norms are communicated [49]. Research suggests that students are more likely to act ethically when institutions clearly teach why integrity matters [3,13] and connect it to learning and personal growth [59,60]. Without helping students reflect on the ethical side of authorship, any rules or deterrents risk being ignored or followed only superficially.

The item hierarchy finding illustrates a clear gradient in how students rationalise the undisclosed use of AI tools, revealing important perspectives into the ethical dimension of emerging academic practices. At the easier end of the continuum, students' willingness to bypass disclosure is closely tied to contextual ambiguity. When institutional guidelines are vague or non-existent, or when detection appears unlikely, students perceive little ethical risk in using AI content without attribution [13,61]. Such patterns indicate that policy opacity and weak enforcement function as critical enabling conditions for non-disclosure.

A second set of behaviours highlights more situational forms of ethical rationalisation. Students describe choosing not to cite AI when its output appears generic, when they feel unsure about proper citation procedures, or when they perceive AI use as a common practice among peers. They also report a willingness to rely on AI because detection feels unlikely, which diminishes the perceived consequences of nondisclosure. These behaviours indicate that ethical decision-making is influenced by convenience, social norms, and perceptions of risk rather than by principled adherence to integrity [13,62]. In these cases, students do not reject academic values outright; instead, they interpret them flexibly in ways that reduce personal effort, limit vulnerability to punishment, or align with what they perceive others are doing.

In contrast, students are less willing to explicitly deny established plagiarism norms or challenge the principle that AI-generated content requires acknowledgement. The most difficult rationalisations involve outright claims that AI-assisted writing is not plagiarism or that it does not require citation because it lacks a human author. The relative reluctance to endorse such views highlights that most students still recognise a baseline of academic integrity [15,62,63] even when those contributions come from non-human sources. The boundary between acceptable and unacceptable behaviour is therefore not entirely eroded; instead, it becomes selectively expanded in areas where students feel uncertain, unsupported, or shielded from accountability.

In summary, the ethical rationalisation construct reveals how students navigate the moral grey areas surrounding generative AI. Their justifications emerge not from malicious intent but from ambiguity, perceived norms, and uncertainty about proper academic conduct [10, 61]. This highlights a crucial implication: strengthening academic integrity in the age of AI requires more than punitive policies [1,3,61]. Clear guidance on citation, explicit communication about acceptable and unacceptable uses of AI, and transparent expectations for academic honesty are essential to limiting the space in which students morally rationalise nondisclosure. When institutions articulate expectations clearly and model ethical use, students are less likely to reinterpret boundaries in self-serving ways and more likely to engage with AI in ways that align with academic values.

Implications

The findings of this study offer important implications for theory, pedagogy, and institutional policy in higher education. The Rasch-derived hierarchies show that parroting behaviours emerge through a structured progression shaped by cognitive load, intuitive reasoning shortcuts, and ethical rationalisation. This progression demonstrates that inappropriate reliance on AI is rarely the result of intentional misconduct alone; instead, it reflects the interaction of intrinsic cognitive strain, environmental pressures, and students' efforts to morally justify their actions. Understanding these layered influences is crucial for designing interventions that meaningfully support students rather than merely penalising them.

From a theoretical perspective, the results highlight the need to understand AI misuse as a cognitive and motivational phenomenon rather than a purely moral one. Intrinsic load explains why students rely on AI when they feel unable to understand or synthesise complex ideas. Extraneous load accounts for patterns that emerge when assignment expectations are unclear or when the volume of academic work overwhelms students' ability to think through tasks carefully. Dual mode reasoning helps explain why students resort to rapid, low-effort responses when under strain, allowing AI to fill the space where reflective thought would ordinarily occur. Ethical rationalisation explains how students maintain a sense of personal integrity while engaging in questionable practices by reframing rules, shifting responsibility, or interpreting ambiguity as permission. Together, these perspectives suggest that AI misuse is not a single behaviour but a predictable progression shaped by context, cognition, and self-perception.

There are also important implications for academic integrity practices. When expectations surrounding AI are vague or inconsistent, students fill the gaps with their own interpretations, often in ways that justify nondisclosure. Policies should therefore be explicit, accessible, and aligned across courses and programmes. Ethical guidelines must be communicated clearly enough to eliminate uncertainty that students might use as justification for misleading practices. However, clarity alone is insufficient. Institutions must cultivate a culture in which responsible use of AI is consistently modelled and in which discussions of ethical reasoning accompany the development of academic skills.

At the policy level, the findings argue for a shift from reactive enforcement to proactive capacity building. Approaches that rely solely on detection or punishment overlook the cognitive and contextual pressures that lead students to rely on AI in the first place. Institutions should focus on creating learning environments that minimise avoidable sources of cognitive strain, strengthen students' sense of academic competence, and promote ethical self-awareness. Policies that position integrity as a shared responsibility rather than an individual obligation are more likely to reduce the conditions that give rise to rationalisation and misuse.

Overall, the implications of this work point toward the need for coordinated changes in teaching, institutional communication, and student support. Addressing intrinsic challenges, reducing extraneous pressures, guiding ethical reasoning, and clarifying expectations can collectively reduce the appeal of AI-driven shortcuts and foster more sustained and authentic learning.

Recommendations

While the present study offers descriptive information into students' intrinsic, extraneous, and ethical rationalisations for AI misuse, several limitations suggest important directions for future work. These recommendations aim to strengthen both the theoretical and practical implications of subsequent studies.

Firstly, this research was conducted within a single institution, which limits the generalisability of the Rasch-based findings on differential functioning and item hierarchy. Although the analysis demonstrated stable item behaviour across academic fields within the university, it remains unclear whether the same invariance would hold in institutions with different curricular structures, assessment cultures, or student profiles. Future studies should therefore replicate the Rasch differential analysis and item hierarchy across multiple institutions to determine whether the progression of intrinsic, extraneous, and ethical rationalisation behaviours reflects a broader pattern or is influenced by local academic contexts. Such cross-institutional validation would strengthen confidence in the construct stability of the instrument and provide a more comprehensive understanding of how students in diverse environments engage with generative AI.

Secondly, the current study relied exclusively on questionnaire data, which, although appropriate for establishing item hierarchy through Rasch analysis, limits the depth with which students' underlying reasoning can be understood. The purpose of the instrument was to measure behavioural tendencies and their structured progression across intrinsic, extraneous, and ethical dimensions, but the quantitative approach cannot fully capture the nuances of how students interpret task difficulty, negotiate cognitive pressures, or construct ethical justifications in real time. Future research would benefit from incorporating qualitative methods such as interviews, think-aloud protocols, or analysis of student writing processes to contextualise the hierarchical patterns identified in the Rasch model. Such methodological triangulation would provide richer insight into why certain behaviours are more foundational, how students transition from one behavioural level to another, and how cognitive and ethical factors interact during actual engagement with academic tasks.

Thirdly, the study focused solely on student behaviours and did not capture the perspectives of lecturers or the institutional policies that shape students' interpretations of acceptable AI use. In practice, students' ethical rationalisations and responses to cognitive load are strongly influenced by how instructors articulate expectations, structure assignments, and model responsible use of AI. Likewise, institutional policies often vary in clarity, enforcement, and alignment across departments, creating inconsistencies that students may use to justify nondisclosure. Future research should therefore examine lecturers' beliefs, teaching practices, and assessment designs, as well as the institutional policy landscape that frames AI usage. Understanding how these structural and pedagogical factors interact with student behaviours would allow researchers to identify systemic contributors to AI misuse and design interventions that operate at classroom and institutional levels rather than at the level of student behaviour alone.

These recommendations suggest a research agenda that moves beyond descriptive measurement toward deeper, contextually grounded understanding of AI-related academic behaviours. Future studies should expand across institutions to test the stability of the item hierarchy, combine quantitative measurement with qualitative inquiry to illuminate the cognitive and ethical mechanisms underlying each behavioural level, and include lecturer perspectives to capture the broader pedagogical ecosystem in which students make decisions about AI use. Institutional policy research should also be prioritised to identify how clarity, consistency, and cultural norms influence students' ethical interpretations. By integrating these directions, future scholarship can develop a more comprehensive model of AI engagement in higher education, one that accounts for cognitive constraints, instructional design, social norms, and institutional contexts. This integrated approach will support the development of interventions that not only discourage misuse but also strengthen authentic learning, ethical reasoning, and student agency in an AI-mediated academic environment.

Funding

This work was supported by the Universiti Teknologi Malaysia under the Potential Academic Staff Research University Grant (Q.J130000.2753.03K65).

Declaration

During the preparation of this manuscript, the author(s) used ChatGPT to assist with language refinement and grammar checking. After using this tool, the author(s) carefully reviewed and edited the content as necessary and assume full responsibility for the final version of the publication.

CRediT authorship contribution statement

Ibnatul Jalilah Yusof: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Zakiah Mohamad Ashari:** Validation, Conceptualization. **Lukman Hakim Ismail:** Resources, Project administration, Funding acquisition. **Mira Panadi:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.R. King, A conversation on artificial intelligence, chatbots, and plagiarism in higher education, Cell Mol. Bioeng. 16 (1) (2023) 1–2, https://doi.org/10.1007/s12195-022-00754-8.

[2] C. Wang, S.J. Aguilar, J.S. Bankard, E. Bui, B. Nye, Writing with AI: what college students learned from utilizing ChatGPT for a writing assignment, Educ. Sci. 14 (9) (2024) 976, https://doi.org/10.3390/educsci14090976 (Basel).

[3] J. Crawford, M. Cowling, K.A. Allen, Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI), J. Univ. Teach. Learn. Pract. 20 (3) (2023) 02, https://doi.org/10.53761/1.20.3.02.

[4] L.A. Gammoh, ChatGPT in academia: Exploring university students' risks, misuses, and challenges in Jordan, J. Furth. High. Educ. 48 (6) (2024) 608–624, https://doi.org/10.1080/0309877X.2024.2378298.

[5] Y.Y. Lo, D. Fung, X. Qiu, Assessing content knowledge through L2: mediating role of language of testing on students' performance, J. Multiling. Multicult. Dev. 44 (10) (2023) 1013–1028, https://doi.org/10.1080/01434632.2020.1854274.

[6] B.T. Gallant, M. Picciotto, G. Bozinovic, E. Tour, Plagiarism or not? Investigation of Turnitin®-detected similarity hits in biology laboratory reports, Biochem. Mol. Biol. Educ. 47 (4) (2019) 370–379, https://doi.org/10.1002/bmb.21236.

[7] Glatt, B. S. (1987). The cognitive consequences of parroting. (T-30189) [Doctoral dissertation, The University of Chicago]. ProQuest Dissertations & Theses Global.

[8] C.B. Paulston, M.N. Bruder, Teaching English as a second language, Tech. Proced. (1976). Retrieved from, https://files.eric.ed.gov/fulltext/ED153499.pdf.

[9] J. Tepperman, T. Stanley, K. Hacioglu, B. Pellom, Testing suprasegmental English through parroting, in: Proceedings of the Speech Prosody, Chicago, IL, USA, 2010, pp. 11–14. https://www.isca-archive.org/speechprosody_2010/tepperman10_speechprosody.pdf.

[10] A.B. Cyphert, Generative AI, plagiarism, and copyright infringement in legal documents, Minn. JL Sci. Tech. 25 (49) (2023). https://scholarship.law.umn.edu/mjlst/vol25/iss2/9.

[11] L. Stearns, Copy wrong: Plagiarism, process, property, and the law, Calif. Law Rev. 80 (1992) 513. https://www.jstor.org/stable/pdf/3480772.pdf.

[12] B.L. Frye, Should using an AI text generator to produce academic writing be plagiarism? Fordham Intell. Prop. Media Ent. LJ 33 (2022) 946. https://ssrn.com/abstract=4292283.

[13] A. Kleebayoon, V. Wiwanitkit, Artificial intelligence, chatbots, plagiarism and basic honesty: comment, Cell Mol. Bioeng. 16 (2) (2023) 173–174, https://doi.org/10.1007/s12195-023-00759-x.

[14] COPE Council (2023). COPE position - Authorship and AI - English. 10.24318/cCVRZBms.

[15] A.M. Jarrah, Y. Wardat, P. Fidalgo, Using ChatGPT in academic writing is (not) a form of plagiarism: what does the literature say, Online J. Commun. Media Technol. 13 (4) (2023) e202346, https://doi.org/10.30935/ojcmt/13572.

[16] D. Kahneman. Thinking, Fast and Slow, Farrar, Straus and Giroux, 2011. https://psycnet.apa.org/record/2011-26535-000.

[17] O. Chen, F. Paas, J. Sweller, A cognitive load theory approach to defining and measuring task complexity through element interactivity, Educ. Psychol. Rev. 35 (2) (2023) 63, https://doi.org/10.1007/s10648-023-09782-w.

[18] J. Sweller, Cognitive load theory, learning difficulty, and instructional design, Learn. Instr. 4 (4) (1994) 295–312, https://doi.org/10.1016/0959-4752(94)90003-5.

[19] S. Gupta, A. Jaiswal, A. Paramasivam, J. Kotecha, Academic writing challenges and supports: perspectives of international doctoral students and their supervisors, in: Frontiers in Education, 7, Frontiers Media SA, 2022 891534, https://doi.org/10.3389/feduc.2022.891534.

[20] L.P. Patac, A.V Patac Jr, Using ChatGPT for academic support: managing cognitive load and enhancing learning efficiency–A phenomenological approach, Soc. Sci. Humanit. Open 11 (2025) 101301, https://doi.org/10.1016/j.ssaho.2025.101301.

[21] M.F. Teng, M. Yue, Metacognitive writing strategies, critical thinking skills, and academic writing performance: a structural equation modeling approach, Metacogn. Learn. 18 (1) (2023) 237–260, https://doi.org/10.1007/s11409-022-09328-5.

[22] S. Mat Yusoff, A. Mohamad Marzaini, L. Hao, et al., Understanding the role of AI in Malaysian higher education curricula: an analysis of student perceptions, Discov. Comput. 28 (2025) 62, https://doi.org/10.1007/s10791-025-09567-5.

[23] I.M. Pepperberg, Cognition and communication in an African grey parrot (Psittacus erithacus): studies on a Nonhuman Nonprimate, nonmammalian subject. Language and Communication, Psychology Press, 2013, pp. 221–248.

[24] B.O. Omer, Why do students still plagiarize? Perceptions of EFL and non-EFL students on plagiarism, J. Univ. Hum. Dev. 8 (2) (2022) 54–60, https://doi.org/10.21928/juhd.v8n2y2022.pp54-60.

[25] L. Yan, V. Pammer-Schindler, C. Mills, A. Nguyen, D. Gašević, Beyond efficiency: empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning, Br. J. Educ. Technol. 56 (2025) 1675–1685, https://doi.org/10.1111/bjet.70000.

[26] A. Bandura, Social cognitive theory of self-regulation, Organ. Behav. Hum. Decis. Process 50 (2) (1991) 248–287, https://doi.org/10.1016/0749-5978(91)90022-L.

[27] A. Abbas, A. Fatima, A. Arrona-Palacios, et al., Research ethics dilemma in higher education: impact of internet access, ethical controls, and teaching factors on student plagiarism, Educ. Inf. Technol. 26 (2021) 6109–6121, https://doi.org/10.1007/s10639-021-10595-z (Dordr).

[28] R.A. Vasquez, The ethical decision-making gap in student ethics: examining how university students approach ethical dilemmas, Int. J. Ethics Educ. 7 (1) (2022) 53–71, https://doi.org/10.1007/s40889-021-00133-3.

[29] D. Ayton, C. Hillman, K. Hatzikiriakidis, T. Tsindos, S. Sadasivan, S. Maloney, D. Illic, Why do students plagiarise? Informing higher education teaching and learning policy and practice, Stud. High. Educ. 47 (9) (2021) 1921–1934, https://doi.org/10.1080/03075079.2021.1985103.

[30] J.Á. De Lima, Á. Sousa, A. Medeiros, B. Misturada, C. Novo, Understanding undergraduate plagiarism in the context of students' academic experience, J. Acad. Ethics 20 (2) (2022) 147–168, https://doi.org/10.1007/s10805-021-09396-3.

[31] B. Mujtaba, Clarifying ethical dilemmas in sharpening students' artificial intelligence proficiency: dispelling myths about using AI tools in higher education, Bus. Ethics Leadersh. 8 (2) (2024) 107–127, https://doi.org/10.61093/bel.8(2).107-127.2024.

[32] J.M. Linacre, Understanding Rasch measurement: estimation methods for Rasch measures, J. Outcome Meas. 3 (4) (1999) 382–405. https://www.researchgate.net/profile/Alfred-Stenner/publication/12729624_Mapping_variables/links/577c092908ae355e74f169aa/Mapping-variables.pdf#page=92.

[33] J.M. Linacre, Optimizing rating scale category effectiveness, J. Appl. Meas. 3 (1) (2002) 85–106. https://www.researchgate.net/profile/John-Linacre/publication/11372384_Understanding_Rasch_measurement_Optimizing_rating_scale_category_effectiveness/links/65548bccb1398a779d8f59be/Understanding-Rasch-measurement-Optimizing-rating-scale-category-effectiveness.pdf.

[34] J.M. Linacre, Sample size and item calibration (or person measure) stability, Rasch Meas. Trans. 7 (4) (1994) 328. Retrieved from, http://www.rasch.org/rmt/rmt74m.htm.

[35] S. Dabaghi, F. Esmaielzadeh, C. Rohani, Application of Rasch analysis for development and psychometric properties of adolescents' quality of life instruments: a systematic review, Adolesc. Health Med. Ther. 11 (2020) 173–197, https://doi.org/10.2147/AHMT.S265413.

[36] A. Tennant, P.G. Conaghan, The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 57 (2007) 1358–1362, https://doi.org/10.1002/art.23108.

[37] Jr.G. Engelhard, Using item response theory and model—data fit to conceptualize differential item and person functioning for students with disabilities, Educ. Psychol. Meas. 69 (4) (2009) 585–602, https://doi.org/10.1177/0013164408323240.

[38] R. Mulenga, H. Shilongo, Academic integrity in higher education: understanding and addressing plagiarism, Acta Pedagogia Asiana 3 (1) (2024) 30–43, https://doi.org/10.53623/apga.v3i1.337.

[39] A. Skulmowski, K.M. Xu, Understanding cognitive load in digital and online learning: a new perspective on extraneous cognitive load, Educ. Psychol. Rev. 34 (1) (2022) 171–196, https://doi.org/10.1007/s10648-021-09624-7.

[40] D.R. Blazek, J.T. Siegel, Preventing satisficing: a narrative review, Int. J. Soc. Res. Methodol. 27 (6) (2024) 635–648, https://doi.org/10.1080/13645579.2023.2239086.

[41] T.C. Cheng, Y. Zhang, Y.H. Chou, V. Koshy, T.W. Li, K. Karahalios, H. Sundaram, Organize, then vote: Exploring cognitive load in quadratic survey interfaces, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–35, https://doi.org/10.1145/3706598.3714193.

[42] P. Ayala-Enríquez, J. Guerrero-Dib, Moral disengagement leading to social acceptance of academic misconduct: a predictor of behavior. Handbook of Academic Integrity, Springer, 2023, pp. 1–24.

[43] C.K.Y. Chan, Understanding AI guilt: The development, pilot-testing, and validation of an instrument for students, Educ. Inf. Technol. (2025) 1–20, https://doi.org/10.1007/s10639-025-13629-y (Dordr).

[44] A.H. Perry, D.A. Rettinger, J.M. Stephens, E.M. Anderman, M.L. McTernan, H. Tatum, T. Bertram Gallant, From institutional climate to moral attitudes: examining theoretical models of academic misconduct, Ethics Behav. (2025) 1–18, https://doi.org/10.1080/10508422.2025.2514577.

[45] D.J. Nicol, D. Macfarlane-Dick, Formative assessment and self-regulated learning: a model and seven principles of good feedback practice, Stud. High. Educ. 31 (2) (2006) 199–218, https://doi.org/10.1080/03075070600572090.

[46] S. Quinn, Using "how to…" videos in feedforward practices to support the development of academic writing, J. Empower. Teach. Excell. 5 (3) (2022) 42–54. https://digitalcommons.usu.edu/jete/vol5/iss3/6.

[47] A. Morrison, Meta-writing: AI and writing, Compos. Stud. 51 (1) (2023). https://rws511.pbworks.com/w/file/fetch/154174671/Morrison%20Meta%20Writing.pdf.

[48] W. Simon, Distinguishing between student and AI-generated writing: a critical reflection for teachers, Metaphor (3) (2023) 16–23. https://search.informit.org/doi/10.3316/informit.274843605275892.

[49] R.W. Black, B. Tomlinson, University students describe how they adopt AI for writing and research in a general education course, Sci. Rep. 15 (1) (2025) 8799, https://doi.org/10.1038/s41598-025-92937-2.

[50] S. Koudsia, M. Kirchner, Reducing cognitive overload for students in higher education: a course design case study, J. High. Educ. Theory Pract. (10) (2024) 24, https://doi.org/10.33423/jhetp.v24i10.7382.

[51] J. Rodríguez-Ruiz, I. Marín-López, R. Espejo-Siles, Is artificial intelligence use related to self-control, self-esteem and self-efficacy among university students? Educ. Inf. Technol. 30 (2) (2025) 2507–2524, https://doi.org/10.1007/s10639-024-12906-6 (Dordr).

[52] A.S. Nelson, P.V. Santamaría, J.S. Javens, Students' perceptions of generative AI use in academic writing, in: Proceedings of the Conference Proceedings. Innovation in Language Learning 2024, 2024, https://doi.org/10.3390/educsci15050611.

[53] E. Shmeleva, T. Semenova, Academic dishonesty among college students: academic motivation vs. contextual factors, Вопросы образования 3 (2019) 101–129. https://cyberleninka.ru/article/n/academic-dishonesty-among-college-students-academic-motivation-vs-contextual-factors/viewer.

[54] E. Bensalem, R. Harizi, A. Boujlida, Exploring undergraduate students' usage and perceptions of AI writing tools, Glob. J. Foreign Lang. Teach. 14 (2) (2024) 53–65, https://doi.org/10.18844/gjflt.v14i2.9344.

[55] S. Bongsu, W.N.M Nik Mohamed, W.N Wan Mohamed, Determinants of students' intention to use AI-powered writing tools in academic writing, Int. J. Res. Innov. Soc. Sci. (IJRISS) 9 (03) (2025) 5616–5630, https://doi.org/10.47772/IJRISS.2025.903SEDU0411.

[56] Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.H., Beresnitzky, A. V., Maes, P. (2025). Your brain on ChatGPT: accumulation of cognitive debt when using an ai assistant for essay writing task. arXivLabs 10.48550/arXiv.2506.08872.

[57] K. Blackwell-Starnes, I prefer my own writing": Engaging first-year writers' agency with generative AI, Thresholds Educ. 48 (1) (2025) 25–39. https://files.eric.ed.gov/fulltext/EJ1468038.pdf.

[58] P.V. Sysoyev, Ethics and AI-plagiarism in an academic environment: students' understanding of compliance with author's ethics and the problem of plagiarism in the process of interaction with generative artificial intelligence, Vyss. Obraz. v Ross. Higher Educ. Russ. 33 (2) (2024) 31–53, https://doi.org/10.31992/0869-3617-2024-33-2-31-53.

[59] S. Atlas, ChatGPT for Higher Education And Professional Development: A Guide To Conversational AI, University of Rhode Island, 2023. Retrieved from, https://digitalcommons.uri.edu/cba_facpubs/548.

[60] S. Grassini, Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings, Educ. Sci. 13 (7) (2023) 692, https://doi.org/10.3390/educsci13070692 (Basel).

[61] J.M. Higgs, A. Stornaiuolo, Being human in the age of generative AI: young people's ethical concerns about writing and living with machines, Read. Res. Q. 59 (4) (2024) 632–650, https://doi.org/10.1002/rrq.552.

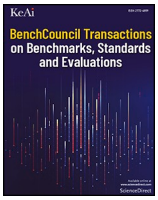
[62] Graf, A. E. (2024). "You Do You": Digital-Native College Students' Perceptions of Cheating and Plagiarism (Order No. 30812633). Available from ProQuest Dissertations & Theses Global. (2895144273). https://vpn.utm.my/dissertations-theses/you-do-digital-native-college-students/docview/2895144273/se-2.

[63] P. Shukla, S. Naik, I. Obi, J. Backus, N. Rasche, P. Parsons, Rethinking citation of AI sources in student–AI collaboration within HCI design education, in: Proceedings of the 7th Annual Symposium on HCI Education, 2025, pp. 1–8, https://doi.org/10.1145/3742901.3742909.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Full Length Article

Benchmark-based prioritizing sustainable consumption and production practices for achieving SDG 12 in India: A multi-criteria decision-making approach

Neha Gupta [a,*], Srikant Gupta [b]

[a] *Amity School of Business, Amity University Uttar Pradesh, Noida, India*
[b] *Jaipuria Institute of Management, Jaipur, India*

ARTICLE INFO

Keywords:
Sustainable development goal
Sustainable consumption and production
Multi-criteria decision making
Grey DEMATEL
India

ABSTRACT

This study prioritizes sustainable consumption and production (SCP) practices to advance SDG 12 in India by employing a hybrid Grey Delphi–Grey DEMATEL framework. Twelve SCP practices identified through a comprehensive literature review were assessed by ten sustainability experts, with Grey Delphi confirming their relevance and Grey DEMATEL mapping the causal structure and influence dynamics within the system. The results show that circular economy practices, multi-stakeholder partnerships, and life cycle assessment function as core driving practices that exert substantial influence on the broader SCP landscape, while sustainable supply chain management, consumption education, urban planning, and green procurement appear as dependent practices shaped by these drivers. By integrating expert judgment and uncertainty-aware analytical techniques, the study provides a structured and replicable decision-support approach that assists policymakers, industry stakeholders, and practitioners in prioritizing impactful SCP interventions tailored to India's socio-economic context, thereby supporting more effective progress toward sustainable development.

1. Introduction

Sustainable Consumption and Production (SCP) has evolved as a highly critical concentration area for sustainable development, particularly in those economies that are increasing rapidly, such as India's. The United Nations has Sustainable Development Goal 12 (SDG 12), which specifies the proper "sustainable consumption and production patterns" [1]. To the Indian nation facing resource depletion, environmental degradation, and socioeconomic challenges, SCP is critical in balancing economic growth with environmental and social sustainability [2]. Especially in India, the second most populous country in the world, with a growing middle class whose consumption and production systems have ramifications for resource use and impacts on the environment at a global level [3], there has been an increasingly focused application and realization of best practices of SCP for achieving the SDG 12 for India while making its contribution to broader sustainable development goals.

SCP is highly relevant for the sustainable development of India. According to Goyal et al. [4], traditional production and consumption systems have been identified as the main causes of environmental degradation and the rapid depletion of natural resources in developing economies such as India. A strategic transition toward sustainability can help achieve several goals, including resource efficiency, waste reduction, pollution mitigation, and social equity [5]. Approaches to SCP can, in turn, open newer avenues towards innovating, increasing competition, and opening new job opportunities in emerging green sectors [6]. India indeed has opened up ambitious development and economic growth, but she needs to put SCP principles into practice accordingly within major sectors and domains of consumption.

Even though SCPs have been considered important, their implementation in India faces tremendous obstacles and challenges. Several studies reported a lack of awareness, financial constraints, technological limitations, regulatory gaps, and aversion to new changes as some of the obstacles among producers and consumers [7,8]. They are also compounded by the complexity and interrelation of the consumption and the production systems, with consequent difficulty in prioritization and practical application of interventions [9]. Against these, there is a pressing need for research studies to be conducted that could

Peer review under the responsibility of The International Open Benchmark Council.

* Corresponding author.

E-mail address: ngupta10@amity.edu (N. Gupta).

https://doi.org/10.1016/j.tbench.2025.100253
Received 22 August 2025; Received in revised form 28 November 2025; Accepted 10 December 2025
Available online 11 December 2025

significantly assist in identifying, assessing, and prioritizing those SCP practices that would be most pertinent and impactful for the Indian scenario. In lieu of this, the current research attempts to fill this bridging gap by giving the SCPs the utmost importance in fulfilling SDG 12 in India using a systematic methodology. The research will shed light on valuable literature and employ sound analytical methods to provide inputs for policymakers, businesses, and other stakeholders to promote SCPs in India. This will introduce a more comprehensive understanding of what practices may need utmost importance and how the adoption hurdles can be intercepted and minimized. The end intention of this research is to add to efforts put in place by India to achieve SDG 12 and achieve its overall success in sustainable development.

2. Literature review

2.1. Current state and challenges of sustainable consumption and production

SCP is a focal point of SDG 12 to "ensure sustainable consumption and production patterns" [1]. SCP practices, when implemented in India, are confronted with multiple challenges. Sharma et al. [2] noted organizational barriers as SCP's foremost hurdle for deploying digital technologies within the supply chain for SCP. Comparable to Goyal et al. [4], it can be noted from the paper that government-linked, management-linked, and finance-linked barriers constitute the core of SCP's hurdles for adoption within Indian manufacturing. These are representative indicators of multiple dimensions of bringing SCP to India, and a multiple-faceted approach is needed to work through several hurdles to be implemented. Luthra et al. [7] elaborated more on these barriers when they specified 15 particular hiccups of SCP adoption within the supply chain for which government support and policies were the most pivotal. Challenges are hardly Indian-specific; Liu et al. [10] found identical barriers even within China, where weakly enforced environmental laws and poor environmental education were some of the key issues. This suggests that developing countries have a unifying challenge of establishing SCP practices and thus need shared knowledge and collaboration.

A lack of awareness and inadequate policy frameworks characterize the current state of SCP in India. According to Mensah et al. [11], the SDG 12 targets are insufficient for monitoring sustainable food consumption, requiring enhanced policy indicators. Abbas et al. [12], in the Indian context, have focused on the importance of eco-labeling and green advertising for achieving SDG 12. The strategies influence consumer perceptions and attitudes toward environmentally friendly products. However, price sensitivity remains a significant challenge that needs to be addressed to promote sustainable consumption patterns in India. It risks what Gladkova calls the "miscommunication of harms," meaning that when communicating accomplishments of such a project, harm through environmental degradation and damage to people's livelihoods goes unrecognized [13]; this can be applicable when considering the implementation process for SDG 12 regarding India. Historically, economic growth has generally been rapid but in tension with sustainability. Stevens, in 2010, said that SCP approaches needed to be more holistic, as sustainable consumption initiatives need to be reconciled with policies that would enhance the private sector's sustainability of production.

The absence of sustainable production practices remains a challenge in several Indian sectors in reaching SDG 12. Shaikh et al. [3] maintain that responsible consumption and production will be achieved when the agri-food system addresses its national cropland limits. Luthra et al. [7] identified 15 barriers to SCP adoption in the Indian manufacturing sector, including government support and essential policies. These studies also highlight the need for a sector-specific approach to help India ensure sustainable production methods. Chiou et al. [14] presented bi-level programming models for sustainable intercity passenger transport systems that could potentially be used in the booming transportation sector of India. Bocean [15] noted that using digital technologies to make the supply chain more sustainable is an area through which the production practices of India's industries could be improved.

The country's rapid urbanization and economic development have created further complexities in implementing the SCP practices in India. Summerhayes et al. [16] emphasized that food diversity and urban environments should be accessible to create sustainable consumption. This requirement in an Indian context for urban planning calls for it to provide for consumption with sustainable consumption patterns. In addition, Carlsen [17] found waste generation to be a pertinent issue in achieving SDG 12 for India's growing urban centers. Han et al. [18] studied the balance between urbanization, economy, and eco-environment in Chinese cities, and their findings can be informative for India's urban development process. Skare et al. [19] identified important synergies and trade-offs between several SDGs, which warrants an integrated approach to urban development to consider the simultaneous implementation of multiple SDGs in India.

Nevertheless, despite the mentioned difficulties, there is still an opportunity for India to make a step ahead toward the achievement of SDG 12. A virtual community platform for promoting responsible consumption and production has been suggested by Whitaker and Pawar [20] that can be adapted to suit the Indian scenario. In addition, Cosentino et al. [21] discussed the fast-growing bio-based construction materials' potential for accelerating the achievement of SDG 12. There is a promising route in sustainable production in the rapidly growing Indian construction sector. Bianchet et al. [22] outlined the impacts of the COVID-19 pandemic on responsible production and consumption, opening possible opportunities for India to "build back better" during the post-pandemic recovery. Sebestyén and Abonyi [23] developed a comprehensive SDG performance measurement tool that could be applied to monitor India's performance in achieving SDG 12 and identify areas for improvement.

2.2. Strategies and priorities for achieving SDG 12

Meeting the goal of SDG 12 would require a holistic approach to India's production and consumption aspects. Bengtsson et al. [5] argued that this may not be just an issue of efficiency but an overall consumption volume with concomitant social and institutional change. In the Indian scenario, this would call for policies that not only improve efficiency in production methods but also affect patterns of consumption. Leal Filho et al. [6] brought to our attention the function of design thinking in achieving sustainable development goals and how this may be applicable for innovative SCP strategies tailored to the challenges facing India. Geels et al. [9] had a "reconfiguration" position focusing on the transition in socio-technical systems and the daily practices embedded with it; such a concept can allow India to understand systemic shifts in consumption and production practices. Roy and Singh [24] identified five principal themes of the business focus in the literature of SPC, thereby providing a structured approach to address the implementation issues that India may face at both strategic and operational levels. Thus, by giving importance to such sustainable production practices, India may achieve the goals of SDG 12. Mangla et al. [8] followed the fuzzy analytical hierarchy process to evaluate barriers toward SCP trends, and similarly, it can be applied to finding those areas of focus for the Indian case. Trummer et al. [25] came forward with measures regarding mineral raw materials' consumption by bringing it down further to fulfill sustainability strategies, which the Indian manufacturing sector must follow. Further, Sharma et al. [26] indicated an opportunity for Industry 4.0 technologies to reach SDG 12 by requiring a need to invest in digital transformation for its industries within India. The challenges of SPCs analyzed through a PEST-AHP methodology were suggested with a proposed framework by Goyal et al. [27] for rank evaluation across Indian sectors of barriers. Kunskaja et al. [28] analyzed the implementation of innovative energy technologies and their alignment with SDG 12, giving insight into India's contributions to the energy sector for sustainable production practices.

The other dimension of achieving SDG 12 in India should be promoting sustainable consumption. According to Vallet-Bellmunt et al. (2023), who measured the contribution of the food retailers' sector to SDG 12, it called for increased involvement in disclosing their sustainability performance. Similar actions can be adopted to make Indian consumers aware of the importance of promoting responsible consumption. Jastrzębska [29] put forth the best practices to achieve SDG 12 in cities, which may be tailored to India's cities and towns to help realize sustainable consumption. Vergragt et al. [30] called for more research on sustainable production, consumption, and livelihoods at the global and regional levels and recommended that India invest in such research to inform context-specific SCP strategies [31]. proposed the use of consumer footprint as an indicator to monitor SDG 12, which can be adopted to assess and direct India's progress on sustainable consumption.

Some sector-specific challenges need to be addressed for India to achieve SDG 12. Liu et al. [10] identified barriers related to sustainable food consumption and production from a circular economy perspective that are relevant to India's agricultural and food processing sectors. Olabi et al. [32] discussed the potential of green ammonia to achieve SDG 12, which has some implications for India's chemical and fertilizer industries. These sector-specific approaches will help India focus on SCP practices in key sectors. Raman et al. [33] have identified key research contributions and policy insights for SDG 12, which calls for a multi-faceted approach to e-waste management and sustainable practices in India. Castellano et al. [34] conducted an efficiency analysis to achieve SDG 12 by providing India with guidelines about the performance of different sectors of activities with potential areas for improvement.

In addition, more robust monitoring and assessment frameworks will be needed to track India's progress on SDG 12. According to Confraria et al. [35], an analysis of countries' research priorities about the SDGs shows that their priorities need to be aligned with the national challenges of the SDGs. For India, that means investing in related SDG 12 research. Dinçer et al. [36] provide an integrated decision-making

Table 1
Identified practices.

Practice	Description	Explanation for Non-Domain Readers (Simplified & Accessible)	Refs.
Implement circular economy practices (P1)	Focuses on reducing waste, reusing resources, and recycling materials in production processes to minimize environmental impact and maximize resource efficiency.	A circular economy works like a loop where products and materials are continually reused instead of thrown away. This reduces pollution, saves resources, and cuts costs for industries. Think of it as shifting from a *"use–throw–replace"* model to a *"use–reuse–repair–recycle"* system.	Shaikh et al. [3]; Leal Filho et al. [40]; Liu et al. [10]
Adopt green marketing and eco-labeling (P2)	Promotes environmentally friendly products through transparent communication and labeling, enabling consumers to make informed, sustainable choices.	Green marketing helps consumers identify products that are better for the environment. Eco-labels act as a quick guide—similar to nutrition labels—helping buyers choose products that use fewer chemicals, less energy, or are responsibly sourced.	Abbas et al. [12]; Marcos et al. [37]
Enhance regulatory frameworks and policies (P3)	Develop and enforce stricter environmental regulations and policies to guide businesses towards sustainable practices and responsible resource use.	These are government rules that require companies to reduce pollution, use cleaner technologies, or follow sustainability standards. Clear and strong policies push industries to act responsibly and prevent environmental damage.	Stevens [41]; Trummer et al. [25]; Liu et al. [10]
Invest in sustainable technologies (P4)	Implement innovative technologies like Industry 4.0 and digital solutions to improve resource efficiency, reduce waste, and optimize production processes.	Sustainable technologies include automation, IoT sensors, AI, and energy-efficient machinery. These tools help companies monitor resource use, reduce wastage, and make smarter decisions—similar to how smart meters reduce home electricity bills.	Sharma et al. [26]; Kunskaja et al. [28]; Sharma et al. [2]
Promote sustainable supply chain management (P5)	Integrate sustainability criteria throughout the supply chain, from sourcing to distribution, to ensure responsible production practices.	This means ensuring that raw materials are responsibly sourced, transportation is energy-efficient, and suppliers follow environmental standards. A sustainable supply chain improves transparency and reduces the overall footprint of products—from factory to consumer.	Mangla et al. [8]; Luthra et al. [7]; Liu et al. [10]
Encourage sustainable consumption education (P6)	Raise awareness and educate consumers about sustainable consumption patterns, empowering them to make environmentally conscious decisions.	This involves awareness campaigns, school programs, and digital tools that help people understand how their daily choices—energy use, food waste, product selection—impact the environment. It empowers consumers to make greener decisions.	Cuesta et al. [1]; Marcos et al. [37]; Liu et al. [10]
Develop sustainable product design (P7)	Create products with longer lifespans, easier repair and recycling capabilities, and reduced environmental impact throughout their lifecycle.	Sustainable design means making products that last longer, can be repaired easily, and use materials that can be recycled. For example, modular electronics or biodegradable packaging reduce long-term waste.	Leal Filho et al. [40]; Marcos et al. [37]
Implement sustainable food systems (P8)	Reduce food waste, promote sustainable agriculture, and optimize food distribution to ensure responsible consumption and production in the food sector.	Sustainable food systems make farming more efficient and less harmful while ensuring food reaches people without waste. Examples include precision farming, composting, and cold-chain systems that prevent spoilage.	Summerhayes et al. [16]; Mensah et al. [11]; Sharma et al. [2]
Foster multi-stakeholder partnerships (P9)	Collaborate across sectors and industries to share knowledge, resources, and best practices for sustainable production and consumption.	Sustainability challenges cannot be solved by one group alone. Partnerships between businesses, governments, researchers, and communities help combine expertise, reduce duplication, and scale solutions faster.	de Visser-Amundson [42].; Opoku et al. [43]; Mangla et al. [8]
Adopt life cycle assessment approaches (P10)	Use comprehensive environmental impact assessments to guide product development and production process decision-making.	LCA looks at everything—from extracting raw materials to manufacturing, using the product, and disposing of it. It helps identify where the biggest environmental impacts occur, guiding smarter design and policy decisions.	Sala & Castellani [31]; Cordella et al. [44]
Promote sustainable urban planning (P11)	Design cities and infrastructure to support sustainable consumption patterns and efficient resource use.	Sustainable urban planning includes green buildings, efficient public transport, waste-management systems, and energy-saving infrastructure—creating healthier, low-carbon cities.	Summerhayes et al. [16]; Han et al. [18]
Implement sustainable procurement practices (P12)	Integrate sustainability criteria into purchasing decisions for both public and private sectors to drive market demand for sustainable products.	When governments and businesses choose suppliers based on environmental and social standards, it pushes the entire market toward greener products. This creates demand for sustainable goods and services.	Opoku et al. [43]; Luthra et al. [7]

approach for SDG disclosures that applies to evaluating SCP practices in India that require priority. These could help India identify areas it needs to focus on most and monitor progress toward realizing SDG 12. Marcos et al. [37] reported opportunities and overlooked issues in the implementation of SDG 12, thus allowing India to develop a more holistic way of approaching SCP. Pandey and Asif [38] measured the energy and environmental sustainability in South Asia vis-à-vis the context of SDGs and brought regional background to India’s efforts towards SDG 12. Rweyendela [39] introduced the approach toward industrial ecology principles through environmental impact assessment, one approach India might have while integrating SCP in its regulation Table 1.

Many research studies have been conducted to identify numerous sustainable consumption and production practices. However, relatively few studies focus on the Indian context using the MCDM technique. Very few studies study the barriers concerning sustainable consumption and production, as identified by Goyal et al. [4] and Luthra et al. [7]. Very few research focuses on positive practices, especially for prioritization in India. This gap prevents Indian policymakers and decision-makers from efficiently allocating resources and energy toward the most impactful practices in achieving SDG 12 in the most suitable socio-economic and environmental contexts. A systematic approach for prioritizing appropriate sustainable consumption and production would provide valuable insights for implementational strategies and policy formation in India.

3. Methodology

This study employs a systematic methodology to identify and prioritize sustainable consumption and production (SCP) practices to achieve SDG 12 in India. The methodology integrates the Grey Delphi and Grey DEMATEL approaches to address the challenges associated with human judgment, incomplete information, and the interrelationships among practices Table 2.

3.1. Grey Delphi approach

The Grey Delphi method combines the principles of the traditional Delphi technique with the grey set theory to refine and achieve consensus on SCP practices. The method was proposed by Dalkey and Helmer in 1963. The steps involved are as follows:

Step 1: Identification of SCP practices

An extensive literature review was conducted to identify 12 key SCP practices relevant to SDG 12 in India. Sources included journal articles, policy documents, and industry reports. The identified practices formed the basis of a questionnaire distributed to experts.

Step 2: Expert panel selection

A panel of 10 experts was selected, including policymakers, sustainability professionals, academicians, and industry practitioners with diverse expertise in sustainable development and SCP practices Table 3.

Step 3: Data collection using a linguistic scale

Experts assessed the importance of each SCP practice using a linguistic scale (see Table 4), which was subsequently converted into grey numbers.

Step 4: Creating the grey numbers

The collected responses changed into the corresponding grey numbers based on Table 4. These grey numbers are then used as the basis for successive analyses. Assume that the expert panel is comprised of k members for evaluation, then the assessment of the barrier $\otimes H_i$ is determined as follows (Bhattacharyya, 2015):

Table 3
Expert panel composition.

Expert No.	Role	Field of Expertise	Experience (Years)	Highest Qualification
1	Policymaker	Regulatory Frameworks and Policies	15	Master’s in Public Administration
2	Sustainability Professional	Circular Economy Practices	12	MBA in Sustainability
3	Academic Researcher	Sustainable Product Design	10	Ph.D. in Sustainable Design
4	Industry Practitioner	Sustainable Technologies	8	Bachelor’s in Engineering
5	Policymaker	Urban Planning	20	Master’s in Urban Development
6	Sustainability Professional	Supply Chain Management	15	MBA in Supply Chain
7	Academic Researcher	Sustainable Consumption Education	18	Ph.D. in Environmental Education
8	Industry Practitioner	Green Marketing and Eco-labeling	14	Master’s in Business
9	Policymaker	Multi-Stakeholder Partnerships	12	Ph.D. in Political Science
10	Sustainability Professional	Life Cycle Assessment	10	Master’s in Environmental Science

Table 2
Justification for using proposed method.

Method	Ability to Handle Uncertainty	Captures Interdependence Between Criteria	Identifies Cause-Effect Relationships	Expert Judgment Requirements	Suitability for SCP Context
AHP	Low - uses crisp pairwise judgments; sensitive to inconsistency	Limited - assumes hierarchical, one-way relationships	No	High cognitive load due to many comparisons; requires consistency	Moderate - works for simple, stable systems but not ideal for dynamic SCP interactions
TOPSIS	Low - distance- based calculation assumes precise data.	None - does not capture interactions	No	Moderate; requires rating alternatives on criteria	Low- Moderate - useful for final ranking but cannot reveal influence structure among SCP practices
ANP	Low - relies on crisp judgments	High - allows feedback and interdependence	No	Very high; complex pairwise comparisons across networks	Moderate - captures complexity but too demanding for data-scarce SCP contexts
Fuzzy DEMATEL	Moderate- High - handles imprecision via fuzzy numbers	High -DEMATEL models interrelationships	Yes- identifies cause/ effect through D-R	Requires defining membership functions. adding subjectivity	High - suitable for uncertainty-heavy systems but depends heavily on quality of membership functions
Grey Delphi-Grey DEMATEL (proposed)	Very High -grey intervals naturally represent incomplete or inconsistent expert data without membership functions	High - DEMATEL structure models mutual influence	Yes-identifies structural drivers using D-R and prominence values	Low-Moderate - linguistic inputs converted to grey intervals reduce cognitive burden	Very High - ideal for SCP where data gaps, expert disagreement, and policy uncertainty are common

Table 4
Linguistic scale.

Linguistics Term	Grey Number
No influence (NI)	[0, 0]
Very low influence (VLI)	[0, 1]
Low influence (LI)	[1, 2]
High influence (HI)	[2, 3]
Very high influence (VHI)	[3, 4]

$$\otimes H_i = \frac{\left(\otimes H_i^1 + \otimes H_i^2 + \ldots + \otimes H_i^h + \ldots + \otimes H_i^k\right)}{k} \tag{1}$$

where $\otimes H_i$ is the total evaluation of barrier importance and $\otimes H_i^h$ denotes that h^{th} expert's evaluation of barrier i from the adopted problem.

Step 5: Whitening of the grey numbers

Considers the $\tilde{\otimes}$ as a whitenisation value of the general interval grey number $\otimes\ H_i = [\underline{H},\ \overline{H}] = [H' \epsilon\ H\ |\underline{H} \leq\ H' \leq\ \overline{H}]$. When the $\tilde{\otimes}$ has unknown distribution, the whitenisation could be done through Eq. (2) (Liu and Forrest, 2010; Liu et al., 2012):

$$\tilde{\otimes} = \beta.\underline{H} + (1 - \beta).\overline{H},\ \ \beta = [0, 1] \tag{2}$$

If the β coefficient is 0.5, $\tilde{\otimes}$ is known as equal weight mean whitenisation, which is a commonly used value fora (Liu and Forrest, 2010).

Step 6: Setting a threshold and refining practices

A predefined threshold value (e.g., $\lambda = 3.5$) was used to filter and finalize the significant SCP practices. If $\tilde{\otimes}$ is equal to or greater than λ, the practice is chosen; otherwise, it is rejected.

3.2. Grey DEMATEL approach

The Grey DEMATEL method was employed to explore the interrelationships among the selected SCP practices and to classify them into cause-and-effect groups. The steps include:

Step 1: Constructing the initial direct-relation matrix

To determine the influence of one practice on another, a pairwise comparison matrix ($d = \ \{d_i | i = 1, 2, \cdots, n\}$) was established using expert inputs and a linguistic scale on five-point grey points (Table 1).

Step 2: Develop the grey direct-relation matrix

Linguistic evaluations were translated into grey numbers to construct the initial grey direct-relation matrix, with each expert's input contributing to the aggregated matrix. With responses from K experts, the resulting K grey direct relationship matrices, denoted as $Y^1,\ Y^2,\ Y^3,\ \ldots,\ Y^k$, are derived. The expression for the representation of the grey matrix depicting direct relations is given by Eq. (3) as follows:

$$Y^k = \begin{bmatrix} 0 & \otimes y_{12k} & \otimes y_{13k} & \cdots & \otimes y_{1(n-1)k} & \otimes y_{1nk} \\ \otimes y_{21k} & 0 & \otimes y_{23k} & \cdots & \otimes y_{2(n-1)k} & \otimes y_{2nk} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \otimes y_{(n-1)1k} & \otimes y_{(n-1)2k} & \otimes y_{(n-1)3k} & \cdots & 0 & \otimes y_{(n-1)nk} \\ \otimes y_{n1k} & \otimes y_{n2k} & \otimes y_{n3k} & \cdots & \otimes y_{n(n-1)k} & 0 \end{bmatrix} \tag{3}$$

The element of $[Y]$, $\otimes\ y_{ijk} = \left(\underline{\otimes} y_{ijk},\ \overline{\otimes} y_{ijk}\right)$ shows the influence of practice 'i' on practice 'j' by the kth expert. The $\underline{\otimes} y_{ijk}$ represents the lower and $\overline{\otimes} y_{ijk}$ the upper limit of grey values.

Step 3: Develop the total grey relation matrix

The total grey relation matrix is obtained by combining all individual grey direct-relation matrices using Eq. (4).

$$\otimes Y = \sum_{i=0}^{K} \left(\frac{\sum \underline{\otimes} y_{ijk}}{K},\ \frac{\sum \overline{\otimes} y_{ijk}}{K} \right) \tag{4}$$

Step 4: Express the normalised grey direct-relation matrix

The grey relation matrix is normalized into the grey direct-relation matrix N using Eqs. (5)–(7).

$$\otimes s = \left[\underline{s},\ \overline{s}\right] = \frac{1}{\underset{0 \leq i \leq n}{max} \sum_{j=0}^{n} \otimes\ Y_{ij}}\ \ i.\ j = 1, 2, 3, \ldots, n \tag{5}$$

$$N = \otimes s.\ X \tag{6}$$

$$\otimes n_{ij} = \left[\underline{s}.\ \ \otimes\ y_{ijk},\ \ \overline{s}.\ \ \otimes\ y_{ijk}\right] \tag{7}$$

Step 5: Compute the total relation matrix

The total relation matrix T is derived from the normalized grey direct-relation matrix using Eq. (8).

$$\otimes T = \otimes N.(\otimes I - \otimes N)^{-1} \tag{8}$$

where "$\otimes\ I$" is the grey identity matrix.

Step 6: Calculate the causal parameters

To determine the causal parameter Eqs. (9) and (10) are used.

$$\otimes r_i = \sum_{j=1}^{n} t_{ij}\ \forall_i \tag{9}$$

$$\otimes c_j = \sum_{i=1}^{n} t_{ij}\ \forall_j \tag{10}$$

The practice "i" influence is shown by $\otimes\ r_i$, which implies the total influence of practices, and the $\otimes\ c_j$ signifies the influence received by practice "j" from the other practices.

Step 7: Computation of the prominence and net effect

The prominence ($\otimes\ P_i$) and net effect ($\otimes\ E_i$) score of the practices are determined using expressions (11) and (12):

$$\otimes P_i = \otimes r_i + \otimes c_j | i = j \tag{11}$$

$$\otimes E_i = \otimes r_i - \otimes c_j | i = j \tag{12}$$

Based on the prominence ($\otimes\ P_i$) and net effect ($\otimes\ E_i$) scores, the causal relationship graph is constructed. A positive value of $\otimes\ E_i$ indicates the net effect (cause) of practice on the system, while a negative value represents the net influence of the system on the practice.

Step 8: Constructing the causal diagram

A causal diagram was developed to visually represent the interdependencies among SCP practices, highlighting their roles in achieving SDG 12.

4. Analysis

4.1. Grey Delphi analysis

The literature review identified 12 SCP practices for achieving SDG 12 in India. The expert panel evaluated these practices. Table 5 summarizes the grey and crisp values derived through the Grey Delphi process.

Practices with crisp values exceeding the threshold of 3.5 were considered significant. All 12 practices were accepted for further analysis. Fig. 1 displays the crisp values for the 12 practices. Practices above the red threshold line (3.5) are significant for achieving SDG 12.

The Grey Delphi analysis in Table 5 successfully validated and prioritized all twelve initially identified SCP practices pertinent to achieving SDG 12 in India. Through expert evaluation using a linguistic scale converted into grey numbers and subsequent crisp value calculation, every practice achieved a crisp score above the threshold value of 3.5, indicating strong consensus on their significance. Practices such as implementing circular economy practices (P1), investing in sustainable technologies (P4), and promoting sustainable urban planning (P11) emerged with particularly high crisp values of 4.5 and 4.6, reflecting the expert panel's strong agreement on their critical role in advancing sustainability goals. The consistency of the high scores across all practices

Table 5
Grey and crisp values for SCP practices.

Practice	Grey Value Range	Crisp Value	Decision
Implement circular economy practices (P1)	[4.2, 4.8]	4.5	Accepted
Adopt green marketing and eco-labeling (P2)	[4.0, 4.6]	4.3	Accepted
Enhance regulatory frameworks and policies (P3)	[3.8, 4.6]	4.2	Accepted
Invest in sustainable technologies (P4)	[4.3, 4.9]	4.6	Accepted
Promote sustainable supply chain management (P5)	[4.1, 4.7]	4.4	Accepted
Encourage sustainable consumption education (P6)	[3.7, 4.3]	4.0	Accepted
Develop sustainable product design (P7)	[3.9, 4.5]	4.1	Accepted
Implement sustainable food systems (P8)	[4.0, 4.6]	4.3	Accepted
Foster multi-stakeholder partnerships (P9)	[4.2, 4.8]	4.5	Accepted
Adopt life cycle assessment approaches (P10)	[3.8, 4.6]	4.2	Accepted
Promote sustainable urban planning (P11)	[4.3, 4.9]	4.6	Accepted
Implement sustainable procurement practices (P12)	[4.1, 4.7]	4.4	Accepted

suggests a broad recognition that an integrated and multi-dimensional approach is required to effectively address India's SCP challenges. Notably, the selection outcome ensures that key domains such as regulatory frameworks, supply chain management, sustainable product design, and life cycle assessment — are given due attention, aligning technical, policy, and behavioral interventions. By leveraging the Grey Delphi method, the study minimized biases and enhanced reliability in expert judgment, thereby strengthening the foundation for subsequent causal analysis using Grey DEMATEL. The outcomes of Table 4 thus provide a validated, context-specific framework for guiding policymakers, businesses, and stakeholders in prioritizing and implementing the most impactful SCP practices in India's unique socio-economic and environmental landscape.

4.2. Grey DEMATEL analysis

Using the refined list of 12 SCP practices, the Grey DEMATEL approach analyzed interrelationships and categorized practices into cause-and-effect groups. Table 6 presents the total grey relation between the practices, which is normalized using Eqs. (5)–(7) (see Table 7). Then, using Eq. (8), the total relation matrix has been computed (see Table 8). Table 9 shows the practices' prominence (β) and net effect (γ) scores and Fig. 1 depicts the causal diagram.

Based on the results presented in Table 9 obtained through the Grey DEMATEL analysis, the study successfully classified the 12 SCP practices into cause-and-effect groups while highlighting their relative importance and interrelations. Practices such as "Implement circular economy practices" (P1), "Adopt green marketing and eco-labeling" (P2), "Enhance regulatory frameworks and policies" (P3), "Invest in sustainable technologies" (P4), "Develop sustainable product design" (P7), "Implement sustainable food systems" (P8), "Foster multi-stakeholder partnerships" (P9), and "Adopt life cycle assessment approaches" (P10) were identified as cause group practices, given their positive net effect (γ) values. Among these, P1, P9, and P10 recorded the highest net effect scores, implying that they are primary drivers influencing other SCP practices within the system. In contrast, practices like "Promote sustainable supply chain management" (P5), "Encourage sustainable consumption education" (P6), "Promote sustainable urban planning" (P11), and "Implement sustainable procurement practices" (P12) were categorized as effect group practices, indicating that they are more likely to be influenced by the cause group actions. The prominence (β) values revealed that "Implement circular economy practices" (P1) and "Foster multi-stakeholder partnerships" (P9) held the most significant total influence within the network, underlining their pivotal roles in promoting SCP outcomes. The Grey DEMATEL analysis provided a clear hierarchical structure of the practices, offering critical insights for policymakers to focus on reinforcing high-causal-effect practices to drive systemic improvements toward achieving SDG 12 in India. This nuanced understanding of cause-and-effect relationships among SCP practices ensures more strategic prioritization and resource allocation in policy and implementation efforts. Fig. 2 illustrates SCP practices as nodes linked by directed edges of varying strength, representing their influence. "Cause" practices are placed on the left-hand side, influencing changes in the right-hand side "effect" practices.

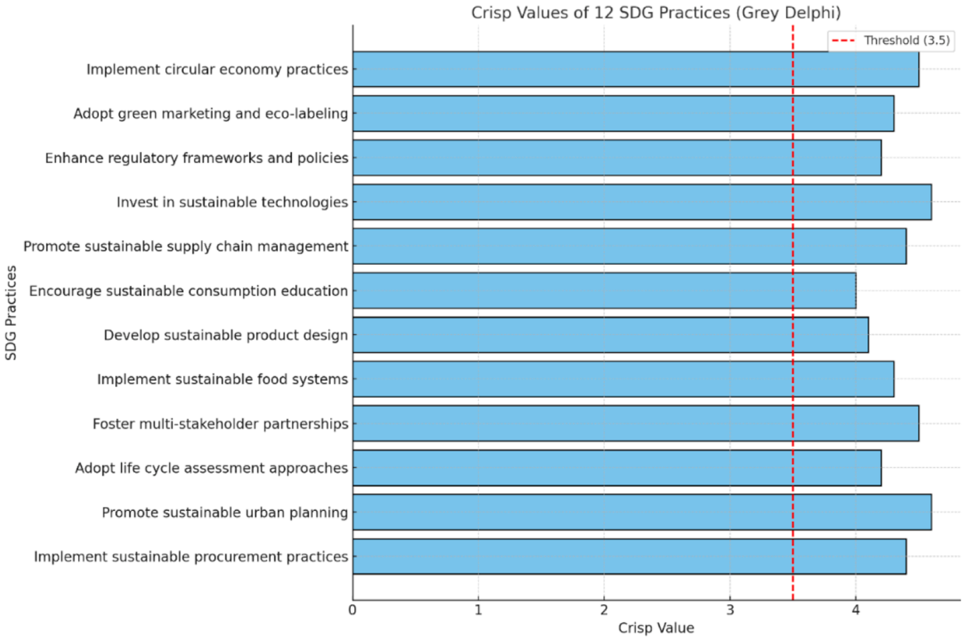


Fig. 1. Finalization of SDG practices based on threshold value.

Table 6
Total grey relation matrix.

SCP	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1	0.00	3.75	3.73	3.75	4.90	4.88	3.73	3.70	3.75	3.75	4.93	3.73
P2	3.60	0.00	3.58	3.60	4.75	4.73	3.58	3.55	3.60	3.60	4.78	3.58
P3	3.48	3.48	0.00	3.48	4.63	4.60	3.45	3.43	3.48	3.48	4.65	3.45
P4	3.40	3.40	3.38	0.00	4.55	4.53	3.38	3.35	3.40	3.40	4.58	3.38
P5	1.80	1.80	1.78	1.80	0.00	2.93	1.78	1.75	1.80	1.80	2.98	1.78
P6	1.88	1.88	1.85	1.88	3.03	0.00	1.85	1.83	1.88	1.88	3.05	1.85
P7	3.53	3.53	3.50	3.53	4.68	4.65	0.00	3.48	3.53	3.53	4.70	3.50
P8	3.40	3.40	3.38	3.40	4.55	4.53	3.38	0.00	3.40	3.40	4.58	3.38
P9	3.70	3.70	3.68	3.70	4.85	4.83	3.68	3.65	0.00	3.70	4.88	3.68
P10	3.65	3.65	3.63	3.65	4.80	4.78	3.63	3.60	3.65	0.00	4.83	3.63
P11	1.73	1.73	1.70	1.73	2.88	2.85	1.70	1.68	1.73	1.73	0.00	1.70
P12	3.28	3.28	3.25	3.28	4.43	4.40	3.25	3.23	3.28	3.28	4.45	0.00

Table 7
Normalized grey relation matrix.

SCP	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1	0.00	0.76	0.76	0.76	0.99	0.99	0.76	0.75	0.76	0.76	1.00	0.76
P2	0.73	0.00	0.73	0.73	0.96	0.96	0.73	0.72	0.73	0.73	0.97	0.73
P3	0.71	0.71	0.00	0.71	0.94	0.93	0.70	0.70	0.71	0.71	0.94	0.70
P4	0.69	0.69	0.69	0.00	0.92	0.92	0.69	0.68	0.69	0.69	0.93	0.69
P5	0.37	0.37	0.36	0.37	0.00	0.59	0.36	0.36	0.37	0.37	0.60	0.36
P6	0.38	0.38	0.38	0.38	0.61	0.00	0.38	0.37	0.38	0.38	0.62	0.38
P7	0.72	0.72	0.71	0.72	0.95	0.94	0.00	0.71	0.72	0.72	0.95	0.71
P8	0.69	0.69	0.69	0.69	0.92	0.92	0.69	0.00	0.69	0.69	0.93	0.69
P9	0.75	0.75	0.75	0.75	0.98	0.98	0.75	0.74	0.00	0.75	0.99	0.75
P10	0.74	0.74	0.74	0.74	0.97	0.97	0.74	0.73	0.74	0.00	0.98	0.74
P11	0.35	0.35	0.35	0.35	0.58	0.58	0.35	0.34	0.35	0.35	0.00	0.35
P12	0.66	0.66	0.66	0.66	0.90	0.89	0.66	0.65	0.66	0.66	0.90	0.00

Table 8
Total relation matrix.

SCP	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1	-0.49	-0.06	-0.06	-0.06	-0.12	-0.12	-0.06	-0.06	-0.06	-0.06	-0.12	-0.06
P2	-0.06	-0.48	-0.06	-0.06	-0.12	-0.11	-0.06	-0.06	-0.06	-0.06	-0.12	-0.06
P3	-0.06	-0.06	-0.47	-0.06	-0.11	-0.11	-0.06	-0.06	-0.06	-0.06	-0.11	-0.06
P4	-0.06	-0.06	-0.06	-0.47	-0.11	-0.11	-0.06	-0.06	-0.06	-0.06	-0.11	-0.06
P5	-0.04	-0.04	-0.04	-0.04	-0.40	-0.02	-0.04	-0.04	-0.04	-0.04	-0.02	-0.04
P6	-0.04	-0.04	-0.04	-0.04	-0.03	-0.40	-0.04	-0.05	-0.04	-0.04	-0.02	-0.05
P7	-0.06	-0.06	-0.06	-0.06	-0.11	-0.11	-0.48	-0.06	-0.06	-0.06	-0.12	-0.06
P8	-0.06	-0.06	-0.06	-0.06	-0.11	-0.11	-0.06	-0.47	-0.06	-0.06	-0.11	-0.06
P9	-0.06	-0.06	-0.06	-0.06	-0.12	-0.12	-0.06	-0.06	-0.49	-0.06	-0.12	-0.06
P10	-0.06	-0.06	-0.06	-0.06	-0.12	-0.12	-0.06	-0.06	-0.06	-0.49	-0.12	-0.06
P11	-0.04	-0.04	-0.04	-0.04	-0.02	-0.02	-0.04	-0.04	-0.04	-0.04	-0.39	-0.04
P12	-0.06	-0.06	-0.06	-0.06	-0.10	-0.10	-0.06	-0.06	-0.06	-0.06	-0.10	-0.46

5. Discussion

The Grey DEMATEL analysis of SCP practices for realizing SDG 12 in India brought rich insights into the causal relationship among the twelve SCP practices. Categorizing practices into cause-and-effect groups gives strategic support to ordering implementation priorities. Findings show that "Implement circular economy practices" (P1) with the highest prominence (β=7.5) and net effect (γ=3.6) scores are the most impactful practice, indicating its consequential status as a catalyst for sustainable development. This is consistent with Shaikh et al. [3], who noted that responsible consumption and production were possible through systemic means subject to resource constraints. The prominence of the circular economy's focus on minimizing waste, reusing resources, and recycling materials directly addresses the issues of resource depletion in India's booming economy. Likewise, "Foster multi-stakeholder partnerships" (P9), with a prominence score of 7.4 and net effect score of 3.5, was another key driver, reinforcing de Visser-Amundson's [42] position that inter-sector collaboration is imperative for resolving complex issues of sustainability. The strong causal power of the "Adopt life cycle assessment approaches" (P10), with a prominence score of 7.3, reaffirms Sala and Castellani's [31] finding that holistic environmental impacts are a prerequisite for tracking progress towards SDG 12. These three practices constitute the core of a systemic approach to SCP in India, as they have multiple domain impacts with synergetic effects and cascades of positive impacts throughout the system. The causal map showcases how these practices directly affect effect group practices with a cascade of positive effects within the entire SCP system. This hierarchical perspective allows policymakers to plan for transmitting systemic gains from these key practices while avoiding singular attention to individual practices.

The effect practices, such as "Promote sustainable supply chain management" (P5), "Encourage sustainable consumption education" (P6), "Promote sustainable urban planning" (P11), and "Implement sustainable procurement practices" (P12), are key areas with high susceptibility towards influence from cause group practices. While categorized as effect practices, they have high prominence scores, suggesting they are the most important parts of the overall SCP framework. Sustainable supply chain management (P5), with a prominence value of 5.9,

Table 9
Prominence (β) and net effect (γ) scores of SCP practices.

Practice	Prominence (β)	Net Effect (γ)	Role
Implement circular economy practices (P1)	7.5	3.6	Cause
Adopt green marketing and eco-labeling (P2)	7.2	3.3	Cause
Enhance regulatory frameworks and policies (P3)	6.9	3.1	Cause
Invest in sustainable technologies (P4)	6.8	2.9	Cause
Promote sustainable supply chain management (P5)	5.9	-2.6	Effect
Encourage sustainable consumption education (P6)	6.0	-2.4	Effect
Develop sustainable product design (P7)	7.0	3.2	Cause
Implement sustainable food systems (P8)	6.7	3.0	Cause
Foster multi-stakeholder partnerships (P9)	7.4	3.5	Cause
Adopt life cycle assessment approaches (P10)	7.3	3.4	Cause
Promote sustainable urban planning (P11)	5.8	-2.8	Effect
Implement sustainable procurement practices (P12)	6.2	-2.5	Effect

endorses Mangla et al.'s [8] finding that supply chains are key areas of intervention for sustainable consumption and production. Its negative net effect value of -2.6 implies that the practice derives a massive benefit from improvements in the application of circular economy thinking and regulatory frameworks. Similarly, sustainable consumption education (P6) with a prominence value of 6.0 aligns with Cuesta et al.'s [1] focus on awareness of alternative consumption patterns. Interrelation analysis infers that educational initiatives are more effective with accompanying green marketing and multi-stakeholder partnerships. Sustainable urban planning (P11), although with the lowest prominence value of 5.8, is key to creating environments conducive to sustainable consumption, as Summerhayes et al. [16] noted. It is evident from the analysis that initiatives in urban areas need to be linked with circular economy thinking and life cycle analyses to achieve maximum effects. Sustainable procurement practices (P12) with a prominence value of 6.2 align with Opoku et al.'s [43] work regarding procurement in construction to achieve SDG 12. It is apparent from the causal diagram that having strong cause group practices helps to build a conducive ecosystem for successfully pursuing these effect group practices. This helps identify a more effective resource utilization by understanding that enhancing effect practices depends on strong cause practices.

Prioritization of SCP practices using Grey DEMATEL provides insightful recommendations for sector-specific action plans for Indian implementation. For the country's manufacturing sector, a major contributor to Indian economic growth, the high priority for establishing circular economy practices (P1) and employing sustainable technologies (P4) implies that these should be high-priority areas. This resonates with Sharma et al.'s [26] discovery that Industry 4.0 technologies provide strategic avenues for achieving SDG 12, especially with the support of institutional pressure. For the agricultural sector, a backbone of India's economy and society, as well as the food sector, the most benefits can be derived from establishing sustainable food systems (P8) and applying life cycle assessment methods (P10). This resonates with Mensah et al.'s [11] discovery of conceptual issues with tracking changes in sustainable consumption of foods and Liu et al.'s [10] research indicating barriers to sustainable food systems from a circular economy viewpoint. For India's burgeoning construction sector, with its rapid growth, the sector can benefit from a focus on sustainable product design (P7) together with regulatory frameworks (P3), as per Cosentino et al.'s [21] research using bio-based construction materials as drivers of the SDGs. For India's growing service sector, incorporating green marketing and eco-labeling (P2) presents a major practice for advancement, complementing Abbas et al.'s [12] research using eco-labeling and green advertising to achieve SDG 12. The role of the public sector is underscored through the significance of improving regulatory frameworks and regulations (P3), as indicated by Trummer et al.'s [25] discovery of measurement sets for achieving SDG target 12.2. These sector-based implications offer a focused method of implementing SCP practices for India's widespread economic base while recognizing that varying sectors may need different focal points while the comprehensive goal of achieving overall SDG 12 is pursued.

Prioritization of SCP practices within India poses significant challenges that necessitate focused policy interventions. The resourcefulness scores of circular economy practices (P1) and multi-stakeholder partnerships (P9) are juxtaposed with the literature-established existing barriers. Goyal et al. [4] highlighted organizational, government, and financial barriers as significant challenges to SCP adoption of Indian manufacturing, while Luthra et al. [7] underscored the importance of government support and policies. This implies that improvements to

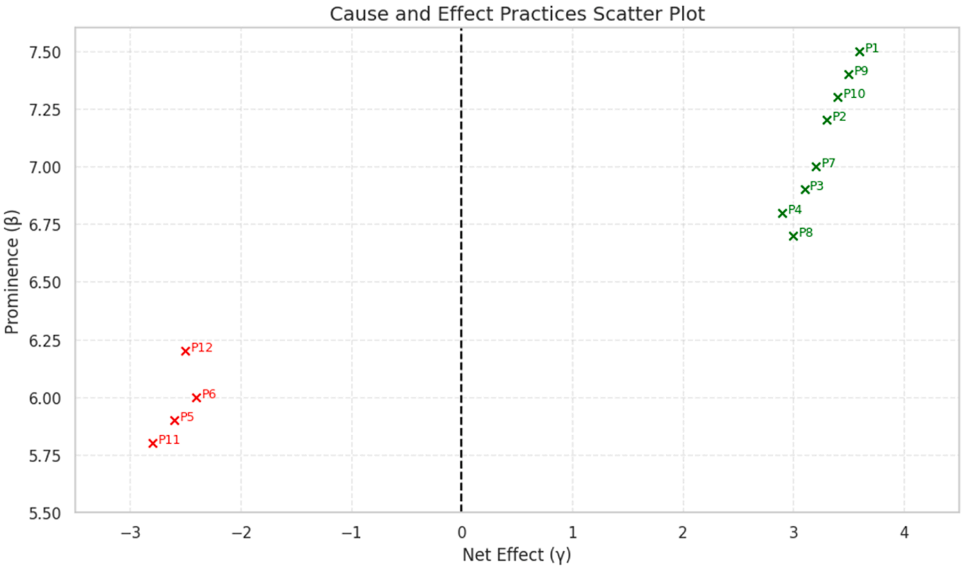


Fig. 2. Causal relationship diagram.

policy frameworks must take precedence to embed high-impact cause group practices. Policies must address technological and financial constraints noted by Mangla et al. [8] to enable the effective adoption of a circular economy. The strength of regulatory frameworks (P3) is that they are significant direct influencers of sustainable supply chain management (P5) and procurement practices (P12), which implies a ripple effect within the system. Awareness, as well as behavioral barriers noted by Liu et al. [10], can be overcome by strategic support for sustainable consumption education (P6) through synchronization with green marketing campaigns (P2). The causal diagram graphically captures interactions of practices, establishing a template for sequencing of policies. The relative position of sustainable urban planning (P11) as a resultant practice with the least prominence implies that it is dependent for effectiveness primarily on strong support from other practices, much as noted by Han et al.'s [18] research seeking municipal-level sustainable development policies. This interconnection necessitates a holistic policy approach rather than individual interventions. Findings validate Bengtsson et al.'s [5] proposition that change in consumption and production patterns requires transcendence of efficiency to confront absolute levels of resource utilization, institutions, and social practices. For India, that implies creating comprehensive packages of policies that set about addressing multiple practices simultaneously while factoring their causal interactions, offering financial rewards for business models with a circular orientation, introducing regulatory frameworks designed to foster innovative technological development for sustainability, as well as educational initiatives promoting changes towards patterns of consumption that are sustainable.

6. Managerial and theoretical implications

This research provides substantial practical implications for Indian policymakers, business leaders, and sustainability practitioners. First, the cause-effect grouping of SCP practices gives decision-makers a distinct strategic path to resource allocation. By focusing on high-prominence cause practices such as the adoption of circular economy approaches (P1), developing multi-stakeholder partnerships (P9), and embracing life cycle approaches (P10), managers can capitalize on cascading effects of their investment in other practices, with a maximum return on investment. For business leaders, the research results recommend giving precedence to incorporating circular economy approaches within business models over stand-alone initiatives for sustainability, as these changes have systemic benefits. The high prominence of eco-labeling and green marketing (P2) implies that open communication of efforts towards sustainability is a strategic market differentiator for operators of businesses in the rapidly environmentalized Indian market. For supply chain managers, positioning sustainable supply chain management (P5) as a resulting practice implies its success depends on the prior development of regulatory frameworks and the adoption of circular economy approaches, making it easier for practitioners to plan their change efforts sequentially. Public sector managers can apply these research results to craft packages of policies addressing multiple practices at a time while accounting for their causal interconnections. Industry-specific insights provide opportunities for industry groups to work towards creating focused guidelines for implementation specific to their unique business contexts. Altogether, such a framework of priorities allows managers to progress beyond ad hoc initiatives for sustainability towards systemic, strategy-based approaches that understand the interlinked nature of SCP practices towards consumption and, ultimately, the efficient execution of the goal of SDG 12 for India.

This research contributes several significant advances to the theory of sustainability and methodology. It develops a theoretical understanding of SCP practices for sustainable consumption and production through a hierarchical framework for identifying the causality of SCP practices, advancing beyond conventional approaches that isolate sustainability practices as independent entities. It expands SCP literature by illustrating how applying Grey theory can successfully manage the ambiguity and subjectivity of sustainability scores, especially for developing economy contexts where data may be incomplete, vague, or even conflicting. Merging the applications of Grey Delphi with Grey DEMATEL methodology constitutes a methodological advancement that improves the quality of multi-criteria decision-making within research on sustainability, providing a model for applying similar studies elsewhere and for other SDGs. It contributes to the theory of systems within sustainability through empirical validation of the interdependent nature of SCP practices, affirming Bengtsson et al.'s [5] position that consumption and transformation of production systems require changes along multiple dimensions. Cause-effect classification contributes to transition theory by identifying focal points within complex socio-technical systems through which interventions can speed up transitions toward sustainability. Additionally, the research addresses the theoretical gap between global sustainability goals and context-specific implementation strategies by providing a structured framework tailored to India's unique socio-economic landscape. The prominence-net effect matrix introduced in this study offers a novel theoretical construct for conceptualizing sustainability practices' relative importance and influence, extending existing sustainability assessment frameworks. Overall, this research advances the theoretical understanding of operationalizing abstract SDG targets into concrete, prioritized action plans while accounting for the complex interactions among diverse sustainability practices in emerging economy contexts.

7. Conclusion and future scope

This study has successfully prioritized sustainable consumption and production practices for achieving SDG 12 in India using an integrated Grey Delphi and Grey DEMATEL approach, providing valuable insights for policymakers and practitioners. The findings revealed that implementing circular economy practices, fostering multi-stakeholder partnerships, and adopting life cycle assessment approaches are primary drivers with the highest prominence and influence across the system. In contrast, sustainable supply chain management, consumption education, urban planning, and procurement practices function as effect group practices that benefit from improvements in driver practices. This causal framework enables the strategic allocation of resources by focusing on high-leverage intervention points within India's complex socio-economic landscape. Future research could extend this work by incorporating quantitative measures of implementation costs and benefits, developing sector-specific implementation roadmaps, examining temporal dimensions of practice implementation, and exploring synergies between SDG 12 and other Sustainable Development Goals in the Indian context. Additionally, comparative studies across different developing economies could identify transferable lessons and context-specific challenges, while action research involving key stakeholders could translate these theoretical insights into practical implementation strategies tailored to India's diverse regional contexts. Future researchers can also discuss how integrating Markov chains or dynamic MCDM extensions can capture evolving policy effectiveness and temporal spillovers.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Neha Gupta: Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Srikant Gupta:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Example 1 — Grey Delphi (worked example)
Using the linguistic scale given in Table 3.
Step 1 — Expert responses for a practice.
Suppose k = 5 experts evaluated Practice P_x and gave these linguistic ratings:
Expert 1: VHI → [3, 4]
Expert 2: HI → [2, 3]
Expert 3: VHI → [3, 4]
Expert 4: LI → [1, 2]
Expert 5: HI → [2, 3]
Step 2 — Aggregate the grey interval (component-wise sum and average).
Sum of lower bounds = 3 + 2 + 3 + 1 + 2 = 11
Sum of upper bounds = 4 + 3 + 4 + 2 + 3 = 16
Average (divide by k = 5):
Lower average = 11 / 5 = 2.2
Upper average = 16 / 5 = 3.2
So, the aggregated grey interval for P_x is:

$$\widetilde{H}_{P_x} = [2.2, 3.2]$$

Step 3 — Whitenisation (crisp value).
Using equal-weight mean whitenisation (β = 0.5):
$\widetilde{H}^* = \beta \cdot H_{\text{upper}} + (1-\beta) \cdot H_{\text{lower}} = 0.5 \times 3.2 + 0.5 \times 2.2 = \frac{3.2+2.2}{2} = 2.7$
So, the crisp (whitened) value for P_x is 2.7.
Step 4 — Threshold check.
If the acceptance threshold is λ = 3.5 (as in the manuscript), then:
2.7 < 3.5 → **P_x would be excluded** (not accepted) after Delphi.
Example 2 — Grey DEMATEL (worked example with 3 practices)
This example demonstrates the DEMATEL arithmetic after grey → whitened (crisp) conversion. For clarity we show a 3-practice system {A, B, C}.
Step 1 — Suppose after grey conversion & whitenisation you obtain the following **crisp direct-relation matrix** X(rows = influencer, columns = influenced):

$$X = \begin{bmatrix} 0 & 2.50 & 1.50 \\ 1.00 & 0 & 2.00 \\ 0.50 & 1.00 & 0 \end{bmatrix}$$

(Each non-diagonal entry is the whitened score representing how strongly row-practice influences column-practice, aggregated across experts.)
Step 2 — Normalize the direct-relation matrix.
Compute row sums of X:

- Row A sum = 0 + 2.50 + 1.50 = 4.00
- Row B sum = 1.00 + 0 + 2.00 = 3.00
- Row C sum = 0.50 + 1.00 + 0 = 1.50

Let $s = 1/\max(\text{row sums}) = 1/4.00 = 0.25$.
Normalized matrix $N = s \cdot X$:

$$N = 0.25 \times \begin{bmatrix} 0 & 2.50 & 1.50 \\ 1.00 & 0 & 2.00 \\ 0.50 & 1.00 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.625 & 0.375 \\ 0.25 & 0 & 0.50 \\ 0.125 & 0.25 & 0 \end{bmatrix}$$

Step 3 — Compute total relation matrix T.
Use the formula $T = N(I-N)^{-1}$. Computing T (matrix arithmetic shown here as results):

$$T \approx \begin{bmatrix} 0.4359 & 1.1795 & 1.1282 \\ 0.5128 & 0.5641 & 0.9744 \\ 0.3077 & 0.5385 & 0.3846 \end{bmatrix}$$

Step 4 — Compute row sums r_i and column sums c_j (total outward and inward influences):

- r = row sums of T:
  - $r_A = 0.4359 + 1.1795 + 1.1282 = 2.7436$
  - $r_B = 0.5128 + 0.5641 + 0.9744 = 2.0513$

  - ○ $r_C = 0.3077 + 0.5385 + 0.3846 = 1.2308$
- c = column sums of T:
  - ○ $c_A = 0.4359 + 0.5128 + 0.3077 = 1.2564$
  - ○ $c_B = 1.1795 + 0.5641 + 0.5385 = 2.2821$
  - ○ $c_C = 1.1282 + 0.9744 + 0.3846 = 2.4872$

Step 5 — Prominence (P) and Net effect (E):

- Prominence $P_i = r_i + c_i$:
  - ○ $P_A = 2.7436 + 1.2564 = 4.0000$
  - ○ $P_B = 2.0513 + 2.2821 = 4.3333$
  - ○ $P_C = 1.2308 + 2.4872 = 3.7179$
- Net effect $E_i = r_i - c_i$:
  - ○ $E_A = 2.7436 - 1.2564 = +1.4872$
  - ○ $E_B = 2.0513 - 2.2821 = -0.2308$
  - ○ $E_C = 1.2308 - 2.4872 = -1.2564$

Step 6 — Interpretation and classification.

- If $E_i > 0$ then practice i is a net cause (driver).
- If $E_i < 0$ then practice i is a net effect (receiver).

From the results:

- A: $E_A = +1.4872$ → Cause / driver
- B: $E_B = -0.2308$ → Effect / receiver
- C: $E_C = -1.2564$ → Effect / receiver

Prominence values show overall involvement; B has the highest prominence here (4.3333), meaning it is highly connected, even though its net effect is slightly negative (i.e., it receives slightly more influence than it gives).

References

[1] L. Cuesta, C. López-Fernández, E. Paños, J.R. Ruiz-Gallardo, Teachers' attitudes towards SDG-12: responsible consumption and production. Development and validation of a measurement scale, Int. Res. Geogr. Environ. Educ. (2024) 1–23.

[2] M. Sharma, S. Joshi, K. Govindan, Overcoming barriers to implement digital technologies to achieve sustainable production and consumption in the food sector: a circular economy perspective, Sustain. Prod. Consum. 39 (2023) 203–215.

[3] M.A. Shaikh, M. Hadjikakou, O. Geyik, B.A. Bryan, Assessing global agri-food system exceedance of national cropland limits for linking responsible consumption and production under SDG 12, Ecol. Econ. 215 (2024) 107993.

[4] S. Goyal, D. Garg, S. Luthra, Sustainable production and consumption: analysing barriers and solutions for maintaining green tomorrow by using fuzzy-AHP–fuzzy-TOPSIS hybrid framework, Environ. Dev. Sustain. 23 (2021) 16934–16980.

[5] M. Bengtsson, E. Alfredsson, M. Cohen, S. Lorek, P. Schroeder, Transforming systems of consumption and production for achieving the sustainable development goals: moving beyond efficiency, Sustain. Sci. 13 (2018) 1533–1547.

[6] W. Leal Filho, I. Schmidberger, A. Sharifi, V.R. Vargas, I.S. Rampasso, T. Dibbern, V. Kozlova, Design thinking for sustainable development: a bibliometric analysis and case study research, J. Clean. Prod. 455 (2024) 142285.

[7] S. Luthra, S.K. Mangla, L. Xu, A. Diabat, Using AHP to evaluate barriers in adopting sustainable consumption and production initiatives in a supply chain, Int. J. Prod. Econ. 181 (2016) 342–349.

[8] S.K. Mangla, K. Govindan, S. Luthra, Prioritizing the barriers to achieve sustainable consumption and production trends in supply chains using fuzzy analytical hierarchy process, J. Clean. Prod. 151 (2017) 509–525.

[9] F.W. Geels, A. McMeekin, J. Mylan, D. Southerton, A critical appraisal of sustainable consumption and production research: the reformist, revolutionary and reconfiguration positions, Glob. Environ. Change 34 (2015) 1–12.

[10] Y. Liu, L.C. Wood, V.G. Venkatesh, A. Zhang, M. Farooque, Barriers to sustainable food consumption and production in China: a fuzzy DEMATEL analysis from a circular economy perspective, Sustain. Prod. Consum. 28 (2021) 1114–1129.

[11] K. Mensah, C. Wieck, B. Rudloff, Sustainable food consumption and sustainable development goal 12: conceptual challenges for monitoring and implementation, Sustain. Dev. 32 (1) (2024) 1109–1119.

[12] S. Abbas, H. Munir, Y. Ahmad, Integrating eco-labeling and green advertising in achieving sustainable development goal 12, Bus. Strategy Dev. 7 (2) (2024) e378.

[13] E. Gladkova, Miscommunication of harms? A critique of SDG 12: responsible consumption and production implementation in the food sector in Northern Ireland, Palgrave Handb. Int. Commun. Sustain. Dev. (2021) 305–323.

[14] Y.C. Chiou, L.W. Lan, K.L. Chang, Sustainable consumption, production and infrastructure construction for operating and planning intercity passenger transport systems, J. Clean. Prod. 40 (2013) 13–21.

[15] C.G. Bocean, A longitudinal analysis of the impact of digital technologies on sustainable food production and consumption in the European union, Foods 13 (8) (2024) 1281.

[16] L. Summerhayes, D. Baker, K. Vella, Food diversity and accessibility enabled urban environments for sustainable food consumption: a case study of Brisbane, Australia, Humanit. Soc. Sci. Commun. 11 (1) (2024) 1–14.

[17] L. Carlsen, Responsible consumption and production in the European union. A partial order analysis of Eurostat SDG 12 data, Green Finance 3 (1) (2021) 28–45.

[18] Z. Han, S. Jiao, X. Zhang, F. Xie, J. Ran, R. Jin, S. Xu, Seeking sustainable development policies at the municipal level based on the triad of city, economy and environment: evidence from Hunan province, China, J. Environ. Manag. 290 (2021) 112554.

[19] M. Skare, B. Gavurova, M. Rigelsky, The relationship between the selected sectoral dimensions and sustainable consumption and production within the sustainable development goal 12, Environ. Sci. Pollut. Res. (2023) 1–18.

[20] M. Whitaker, P. Pawar, Commodity ecology: a virtual community platform for promoting responsible consumption and production to achieve SDG# 12, in: Proceedings of the 2020 IEEE Green Technologies Conference (GreenTech), IEEE, 2020, pp. 59–61.

[21] L. Cosentino, J. Fernandes, R. Mateus, Fast-growing bio-based construction materials as an approach to accelerate united nations sustainable development goals, Appl. Sci. 14 (11) (2024) 4850.

[22] Bianchet, R.T., Provin, A.P., Beattie, V.I., & de Andrade Guerra, J.B.S.O. (2021). COVID-19 and sustainable development goal 12: What are the impacts of the pandemic on responsible production and consumption?. COVID-19: Environmental Sustainability and Sustainable Development Goals, 35-71.

[23] V. Sebestyén, J. Abonyi, Data-driven comparative analysis of national adaptation pathways for sustainable development goals, J. Clean. Prod. 319 (2021) 128657.

[24] V. Roy, S. Singh, Mapping the business focus in sustainable production and consumption literature: review and research framework, J. Clean. Prod. 150 (2017) 224–236.

[25] P. Trummer, G. Ammerer, M. Scherz, Sustainable consumption and production in the extraction and processing of raw materials—measures sets for achieving SDG target 12.2, Sustainability 14 (17) (2022) 10971.

[26] M. Sharma, P. Singh, K. Tsagarakis, Strategic pathways to achieve sustainable development goal 12 through industry 4.0: moderating role of institutional pressure, Bus. Strategy. Environ. (2024).

[27] S. Goyal, D. Garg, S. Luthra, An analysis of sustainable production and consumption challenges: using PEST-AHP approach, Int. J. Logist. Syst. Manag. 37 (3) (2020) 407–426.

[28] S. Kunskaja, J.F. Bauer, A. Budzyński, I.C. Jitea, A research analysis: the implementation of innovative energy technologies and their alignment with SDG 12, East. Eur. J. Enterp. Technol. 5 (13) (2023) 6–25.

[29] E. Jastrzębska, Implementation of sustainable development goal 12 in cities: best practices, Studia Ecol. Bioethicae 20 (3) (2022) 13–24.

[30] P. Vergragt, L. Akenji, P. Dewick, Sustainable production, consumption, and livelihoods: global and regional research perspectives, J. Clean. Prod. 63 (2014) 1–12.

[31] S. Sala, V. Castellani, The consumer footprint: monitoring sustainable development goal 12 with process-based life cycle assessment, J. Clean. Prod. 240 (2019) 118050.

[32] A.G. Olabi, M.A. Abdelkareem, M. Al-Murisi, N. Shehata, A.H. Alami, A. Radwan, E.T. Sayed, Recent progress in green ammonia: production, applications, assessment; barriers, and its role in achieving the sustainable development goals, Energy Convers. Manag. 277 (2023) 116594.

[33] R. Raman, H.H. Lathabai, P. Nedungadi, Sustainable development goal 12 and its synergies with other SDGs: identification of key research contributions and policy insights, Discov. Sustain. 5 (1) (2024) 150.

[34] R. Castellano, G. De Bernardo, G. Punzo, Sustainable well-being and sustainable consumption and production: an efficiency analysis of sustainable development goal 12, Sustainability 16 (17) (2024) 7535.

[35] H. Confraria, T. Ciarli, E. Noyons, Countries' research priorities in relation to the sustainable development goals, Res. Policy 53 (3) (2024) 104950.

[36] H. Dinçer, A. El-Assadi, M. Saad, S. Yüksel, Influential mapping of SDG disclosures based on innovation and knowledge using an integrated decision-making approach, J. Innov. Knowl. 9 (1) (2024) 100466.

[37] A. Marcos, P. Hartmann, J.M. Barrutia, Toward the implementation of SDG12 to ensure sustainable consumption and production patterns: opportunities and neglected issues. Handbook of Sustainability Science in the Future: Policies, Technologies and Education by 2050, Springer International Publishing, Cham, 2023, pp. 1353–1376.

[38] A. Pandey, M. Asif, Assessment of energy and environmental sustainability in South Asia in the perspective of the sustainable development goals, Renew. Sustain. Energy Rev. 165 (2022) 112492.

[39] A.G. Rweyendela, Getting closer to SDG12: incorporating industrial ecology principles into project EIA, J. Environ. Plan. Manag. 65 (6) (2022) 953–974.

[40] W. Leal Filho, J. Barbir, P.G. Özuyar, E. Nunez, J.M. Diaz-Sarachaga, B. Guillaume, T.F. Ng, Assessing provisions and requirements for the sustainable production of plastics: towards achieving SDG 12 from the consumers' perspective, Sustainability 14 (24) (2022) 16542.

[41] C. Stevens, Linking sustainable consumption and production: the government role, in: Natural Resources Forum, 34, Blackwell Publishing Ltd, Oxford, UK, 2010, pp. 16–23.

[42] A. de Visser-Amundson, A multi-stakeholder partnership to fight food waste in the hospitality industry: a contribution to the united nations sustainable development goals 12 and 17, J. Sustain. Tour. 30 (10) (2022) 2448–2475.

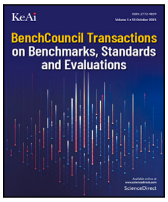
[43] A. Opoku, J. Deng, A. Elmualim, S. Ekung, A.A. Hussien, S.B. Abdalla, Sustainable procurement in construction and the realisation of the sustainable development goal (SDG) 12, J. Clean. Prod. 376 (2022) 134294.

[44] M. Cordella, R. Horn, S.H. Hong, M. Bianchi, M. Isasa, R. Harmens, H. Pihkola, Addressing sustainable development goals in life cycle sustainable assessment: Synergies, challenges and needs, J. Clean. Prod. 415 (2023) 137719.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Full length article

Data-driven financial fraud detection using hybrid artificial and quantum intelligence

Md. Sobuj Mia [a], Sujit Roy [b,*,1], Md Amimul Ihsan [c], Sadek Hossain [a], Md. Khabir Uddin Ahamed [b,2]

[a] Department of Computer Science and Engineering, Jamalpur Science and Technology University, Jamalpur, Bangladesh
[b] Department of CSE, Jamalpur Science and Technology University, Jamalpur, Bangladesh
[c] Department of Electrical and Electronic Engineering, Jamalpur Science and Technology University, Jamalpur, Bangladesh

ARTICLE INFO

Dataset link: http://mlg.ulb.ac.be, https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

Keywords:
Credit card fraud detection
Class imbalance
SMOTE
VQC
Ensemble learning

ABSTRACT

The unauthorized use of a cardholder's financial data, resulting in significant losses to individuals and companies, is known as credit card fraud. The increasing frequency and complexity of such fraud in the digital era highlight the absolutely vital need for reliable and accurate detection systems. Under the specific challenge of extreme class imbalance, this work investigates the credit card fraud identification performance of several Machine Learning (ML), Deep Learning (DL) and Quantum Machine Learning (VQC) algorithms. The study uses a commonly used dataset consisting of 284,807 anonymized credit card transactions, of which only 492 (0.17%) are fraudulent. To solve the class imbalance, we produced synthetic samples of the minority class utilizing the SMOTE, thus raising model sensitivity. Moreover, we enhanced model performance by means of hyperparameter tuning applied with Grid Search, Random Search, and Keras Tuner. Combining deep learning-based feature extraction with ensemble learning approaches, together with effective data balancing and hyperparameter tuning, yields, according to the results, a very accurate and dependable credit card fraud detection system. The hybrid model that includes AutoEncoder for feature extraction, Bagging (Random Forest), and Boosting (XGBoost) was the best, with 100% accuracy. This shows that this integrated technique is better than others. This approach provides a sensible analysis for building robust, real-time fraud detection systems for practical financial applications.

1. Introduction

The ease of buying and conducting digital transactions has significantly improved as credit card use has spread in modern society. Still, this convenience comes with more risk for fraud. Credit card fraud seriously affects individuals as well as financial institutions since it causes financial losses and erodes confidence in electronic payment systems. The strategies used by fraudsters change as digital commerce grows; thus, advanced detection systems must be developed to effectively combat these risks. [1–3].

The ability of a fraud detection system to consistently detect both known and new kinds of fraud determines its effectiveness mostly. This potential thus depends much on the accessibility of extensive, high-caliber datasets. Recent developments in deep learning enable researchers to examine large-scale transaction data and create adaptive models competent in real-time risk detection. These models minimize financial losses and improve system dependability by being more sensitive to complex patterns and instantaneous anomaly detection [4].

Due to their limited adaptability and high false positives, conventional fraud detection methods, mostly dependent on rule-based algorithms and human supervision, are growing increasingly inadequate. On the other hand, highly valuable are complex machine learning techniques that examine transaction trends and identify anomalies. These techniques reduce the possibility of mistakenly identifying legitimate transactions [5] and increase detection accuracy. The design and implementation of an advanced credit card fraud detection system combining modern machine learning techniques with real-time transaction monitoring forms the main focus of this thesis. Using Random

[*] Corresponding author.
E-mail addresses: s20111101@jstu.ac.bd (M.S. Mia), sujit@jstu.ac.bd (S. Roy), amimul@jstu.ac.bd (M.A. Ihsan), s20111132@jstu.ac.bd (S. Hossain), khabir.cse@jstu.ac.bd (M.K.U. Ahamed).
[1] Assistant Professor.
[2] Lecturer.

https://doi.org/10.1016/j.tbench.2025.100252
Received 2 November 2025; Received in revised form 8 December 2025; Accepted 10 December 2025
Available online 13 December 2025

Forest and XGBoost, among other supervised learning, anomaly detection, and ensemble modeling techniques, the proposed architecture distinguishes between legitimate and dubious transactions. Reducing financial risk, raising detection accuracy, and restoring user confidence in digital payment systems is the aim here. The study emphasizes the need to use intelligent and adaptive models that can develop with the dynamic strategies of cybercrime. Using integrated explainable artificial intelligence (XAI) techniques, transparency and interpretability in decision-making are improved, enabling stakeholders to understand and trust the outputs of the detection system [6].

The main contributions of this paper are as follows:

- This work assessed 12 Machine Learning, Deep Learning and Quantum Machine Learning models to find the best model for credit card fraud. We developed a voting classifier consisting of the three most efficient models to utilize their strengths and improve overall accuracy, thereby reducing the limitations of each model.
- Hybrid Model Integration: Developed a robust fraud detection system by combining Autoencoders for feature extraction with Random Forest (Bagging) and XGBoost (Boosting) to enhance accuracy and generalization.
- A significant aspect of this research is employing a Variational Quantum Circuit (VQC) model to investigate the potential enhancements of quantum computing in fraud detection. The VQC employs quantum entanglement along with superposition to transfer data into elevated-dimensional Hilbert spaces. This makes feature representation and classification more accurate than traditional approaches.
- Feature Extraction with Autoencoders: Used deep learning-based Autoencoders to reduce dimensionality and capture meaningful, non-linear patterns in transaction data.
- Addressed class imbalance with the application of the Synthetic Minority Over-sampling Technique (SMOTE), enhancing the model's sensitivity to fraudulent transactions.
- Hyperparameter Tuning: Optimized model performance through thorough hyperparameter tuning, resulting in higher precision, recall, and efficiency in fraud detection.

This all-encompassing approach improves fraud detection abilities and links research creativity with useful applications. The following arrangement of this thesis: the next parts are Review of the present literature in Section 2; definition of the recommended technique in Section 3; a comparative analysis and discussion in Section 4; closing of the thesis and future recommendations in Section 5 (see Table 1).

Table 1
Nomenclature.

Terms	Abbreviation
CCFD	Credit Card Fraud Detection
ECC	European Credit Card
MFA	Multi-Factor Authentication
PCI DSS	Payment Card Industry Data Security Standard
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
GMV	Gross Merchandise Value
XAI	Explainable Artificial Intelligence
CNN	Convolutional Neural Network
LGBM	Light Gradient Boosting
MLP	Multilayer Perceptron
RUS	Random Under-Sampling
GBT	Gradient Boosted Trees
PCA	Principal Component Analysis
KNN	K-Nearest Neighbors
AUPR C	Area Under the Precision-Recall Curve
ReLU	Rectified Linear Unit
XGBoost	Extreme Gradient Boosting
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
QML	Quantum Machine Learning
VQC	Variational Quantum Circuit
QNNs	Quantum Neural Networks
QSVMs	Quantum Support Vector Machines

2. Literature review

Based on our connected topic, we will present a few literature reviews in this area. Additionally, we will go over a few development approaches for machine learning-based CCF prediction.

The study [7] on credit card fraud detection employs neural networks and SMOTE to improve model efficacy, but faces limitations in real-time detection and comparison with advanced systems, necessitating future research.

The research [8] examines the evolution of machine learning models for credit card fraud detection, focusing on Explainable Artificial Intelligence (XAI) for improved openness. It uses techniques like gradient boosting, logistic regression, decision trees, and neural networks. Future research should explore ensemble tactics, advanced feature extraction methods, and larger datasets. [9] employs four machine learning techniques to detect credit card fraud in transactions: Random Forest, Naïve Bayes, Decision Tree, and Support Vector Machine. Despite a 99.96% accuracy rate, the system struggles with real-time data processing and unequal distributions. Future research should focus on developing fraud techniques and advanced algorithms. The authors [10] explore deep learning's potential in fraud detection, evaluating various systems like CNN, RNN, LSTM, GRU, ensemble, and ensemble, along with ML models like logistic regression, decision tree, SVM, ANN, and KNN. They propose improving interpretability, creating hybrid models, and optimizing deep learning architectures for effective fraud detection in practical applications.

The study [11] explores machine learning methodologies for credit card theft, including SVM, Random Forests, Decision Trees, Naive Bayes, and K-Nearest Neighbors. While promising in stationary environments, they struggle with imbalanced datasets and real-time fraud detection. Improvements include explainable artificial intelligence. The authors [12]propose an ensemble-based fraud detection method using under-sampling and SMOTE, addressing class imbalance through various classifiers. This method improves accuracy, even in dataset and classifier selection issues. Future developments should focus on real-time detection systems, dynamic sampling methods, and hybrid architectures. The study [13] proposes a deep learning method for addressing class imbalance in credit card fraud databases, using a Multi-Layer Perceptron (MLP) meta-classifier and LSTM and GRU base classifiers. However, the approach lacks research on resampling methods, datasets, and model interpretability issues. From the study [14] evaluates machine learning models, including support vector machines, artificial neural networks, and random forests. For detecting fraudulent credit card activity. Despite the complexity and need for training resources, the research emphasizes the need for advanced techniques and real-time detection improvements. [15] evaluates machine learning techniques like Random Forest, Adaboost, Support Vector Machines, logistic regression, artificial neural networks, and k-nearest neighbors, but suggests hybrid models, deep learning applications, dataset variation, and behavior-based analytics for improved fraudulent trend identification. Again, [16] examines machine learning approaches for transaction classification using an imbalanced credit card fraud dataset. They found that the adoption of SMOTE improved model correctness for oversampling and feature selection. However, the paper lacks comparisons of algorithms like KNN and outlier identification systems and does not provide a thorough analysis of deep learning techniques.

One of the studies [17] evaluates deep learning techniques like Autoencoder, CNN, and LSTM for credit card fraud detection using hyperparameter tuning and data balancing approaches. However, the study's limitations include limited generalizability and overfitting risk.

Table 2
Existing contribution & research gap within the key technologies.

Authors	Algorithms utilized	Key contributions	Identified research gaps
[7]	Neural Network (NN), SMOTE	Integrated a NN with SMOTE to improve accuracy for fraudulent transactions in imbalanced datasets. & ML.	Lacked real-time deployment validation and offered limited comparative analysis with advanced models.
[8]	Gradient Boosting, Logistic Regression, Decision Trees, Neural Networks (with SMOTE)	Emphasized the use of SMOTE and model performance evaluation through AUC and ROC; highlighted the importance of Explainable AI (XAI).	No actual implementation of XAI; lacked analysis across multiple datasets and advanced feature engineering.
[9]	Random Forest, Naïve Bayes, Decision Tree, SVM	Demonstrated high accuracy (up to 99.96%) using Random Forest for credit card fraud classification.	Faced limitations in processing real-time data and handling class imbalance more effectively.
[10]	CNN, RNN, LSTM, GRU, Easy Ensemble, Balanced Bagging	Reviewed deep learning models, showing their potential in capturing complex fraud patterns compared to ML.	Limited by dataset quality, lack of interpretability, and shallow analysis of training challenges.
[11]	KNN, Naïve Bayes, SVM, Random Forest, Decision Trees	Focused on traditional ML models for detecting fraudulent activity and minimizing financial loss.	Lacked integration of deep learning and feature importance analysis; real-time deployment was not considered.
[12]	Ensemble: Bagging, Boosting, SVM, KNN, RF (with SMOTE/undersampling)	Proposed an ensemble framework addressing class imbalance through hybrid sampling methods.	Provided minimal insight into deep learning, data access limitations, and classifier selection complexities.
[13]	LSTM, GRU (Stacked) with MLP meta-classifier, SMOTE-ENN	Introduced a deep learning stacking ensemble architecture with SMOTE-ENN for better fraud detection performance.	No exploration of alternative sampling techniques or dataset diversity; limited attention to model explainability.
[14]	ANN, SVM, Random Forest	Evaluated performance across models using accuracy and false positive rate; ANN showed variable outcomes.	Required larger datasets, improved real-time detection, and consideration of evolving fraud tactics.
[15]	RF, AdaBoost, SVM, Logistic Regression, ANN, KNN	Compared models based on precision, recall, and F1-score for fraud detection.	The study did not evaluate data diversity or NN advancements, and lacked behavior-based analytics.
[16]	RF, Naïve Bayes, MLP (with SMOTE + Feature Selection)	Demonstrated improved detection using SMOTE and feature selection for imbalanced datasets.	Did not compare with outlier detection or deep learning models; lacked depth in algorithm benchmarking.
[17]	Autoencoder, CNN, LSTM (with SMOTE, ADASYN, RUS)	Presented robust deep learning models for fraud detection, supported by empirical results.	Risk of overfitting and limited dataset generalizability; future work should emphasize ensemble methods.
[18]	RT, RF, DT, DS, GBT (with feature aggregation)	Highlighted the importance of optimization and hybridization for fraud detection.	Limited geographic scope; lacked hybrid model exploration and real-time performance testing.
[19]	Transformer with RF, SVM (baselines)	Applied advanced Transformer architectures to address data sparsity in fraud detection.	Did not explore loss function optimization, additional data sources, or hybrid configurations with other models.
[20]	QNNs, QSVMs, XGBoost, Random Forest	Led the way in comparing Quantum ML models for fraud detection, proving that QNNs can work well even when the data is not balanced.	Limited by simulation on classical hardware, which are not real quantum processors, it lacks real-time validation and incurs extra processing costs.
[21]	QFDNN (Variational Quantum Feature Deep Neural Network), Classical DNN	Proposed a hybrid quantum–classical model that uses fewer qubits and quantum gates. This change makes it more practical for upcoming quantum devices.	Validation was done on quantum simulators, not physical hardware. There is no testing on real-time data streams or a wider range of dataset variability.
[22]	Hybrid Quantum LSTM (HQ-LSTM), Classical LSTM	A new hybrid model combines quantum circuits inside of an LSTM framework that improves capturing complex sequential patterns within transaction data that can be used to detect fraud.	Its complexity is a challenge for existing NISQ-era hardware. Practicality for deployment for real-time inference and robustness on larger, noisier datasets is unproven.

Future developments could include ensemble models, varied datasets, and real-time systems with improved scalability and precision. [18] presents a comprehensive fraud detection system among several models combining Random Trees, Random Forests, Decision Trees, Decision Stumps, and Gradient Boosting Trees. Still, depending just on one geographic dataset reduces the global relevance of the research.. Hybrid models are suggested for better fraud detection, including different datasets, real-time processing, and risk-model adaptability. [19] uses advanced transformer architectures for credit card fraud detection, highlighting the lack of focus on imbalance loss functions and empirical assessment. It suggests exploring unexplored areas like integrating other machine learning models with transformers, tailoring transformers for different fraud types, and using alternative data sources. From the study [20] Quantum Machine Learning (QML) architectures for detecting credit card fraud. They focus on hybrid quantum–classical models like QNNs and QSVMs. QML has great potential, but it faces challenges with processing overhead and scaling. This indicates that we need to understand more about quantum computers in the future.

The study [21] presents QFDNN, a hybrid quantum–classical model aimed at financial tasks such as fraud detection and loan prediction. The authors emphasize its key innovation: better resource efficiency than other Quantum Machine Learning methods. This solves a major issue of high computational demand seen in earlier quantum models. Although the QFDNN shows promising results on benchmark datasets, most research is done on simulators. This highlights the need for testing on real quantum hardware. We should also look at its use with real-time, streaming financial data. Moreover, the research [22] proposes a new Hybrid Quantum Long Short-Term Memory (HQ-LSTM) framework developed for classifying fraudulent activity. The setup integrates quantum circuits with essential features of standard LSTM models. The framework lifts model performance when differentiating complex temporal patterns from transaction sequences. The researchers showcase that performance levels are improved with the HQ-LSTM over a standard classical LSTM, especially when used in sophisticated sequential fraudulent behavior detection. The greater complexity of the framework is a challenge for existing noisy quantum computing devices. It is presumptuous to apply it under real-time analysis without further research.

Research on credit card fraud detection shows a trend towards machine learning and deep learning approaches, with classifiers and ensemble techniques showing performance despite real-time detection challenges. Future developments should focus on hybrid models, sophisticated feature engineering, and improved computational approaches.

Recent work on credit-card fraud detection shows that machine-learning models. From the state-of-art works presented in [8] &[9] are nearing near-perfect performance, but gaps remain. Random Forest classifiers achieved 99.8% accuracy, with precision and recall in the high 0.99 range. Tree-based ensembles [8] &[11] effectively handle extreme class-imbalance in the public European dataset. Simple algorithms can rival complex pipelines when data is pre-processed. Some previous studies add useful historical context. Paper [14,16] showed that, even before the recent surge in deep-learning interest, boosting and bagging methods regularly delivered AUCs around 0.99 on the same dataset. Their findings reinforce the message that, for tabular transaction data, sophisticated feature engineering and resampling often outweigh model novelty. Six studies on a 2013 European dataset under-sample or over-sample the majority or minority classes, resulting in inaccurate results and potential information loss or synthetic-data bias. [19].'s intelligent system improves discriminative power but faces challenges in scalability and real-time scoring. Table 2 describes the summary of recent studies.

Taken together, the literature suggests that future CCFD research should: (i) evaluate on fresher, multi-regional datasets, (ii) couple high-performing ensemble models with explainable-AI add-ons to satisfy regulatory transparency, and (iii) benchmark inference latency alongside accuracy.

3. Methodology

This chapter delineates the methodological approach utilized in this research to effectively distinguish between genuine and fraudulent transactions. Fig. 1 depicts the successive methodologies utilized in the study, offering a visual representation of the entire workflow. Before assessing the various steps of the suggested technique, it is essential to perform a thorough assessment of the dataset used in this study, as it forms the foundation for all subsequent operations.

Upon thorough evaluation and selection of models based on their performance, we identified the optimal one for categorization. This approach guarantees that predictive models are reliable and applicable in real-world scenarios.

3.1. Dataset description

The dataset utilized in this study is sourced from Kaggle and results from a collaboration between Worldline and the Machine Learning Group (MLG) of the Université Libre de Bruxelles (ULB). Accessible via http://mlg.ulb.ac.be, the group specializes in advanced research in large data mining and fraud detection; hence, this dataset is especially suitable for tasks driven by machine learning-based anomaly detection. Dataset Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/dataThe dataset consists of anonymized credit card transaction data from September 2013, primarily focusing on purchases made by European cardholders. Most input features have undergone Principal Component Analysis (PCA) processing to ensure data privacy and ethical standards. The dataset is largely numerical, simplifying preparation and facilitating smooth integration with machine learning techniques. However, two crucial elements, “Time” and “Amount”, are absent from PCA’s transformation. The “Class” column records the intended output for classification jobs, with 0 representing regular transactions and 1 representing fraudulent ones. This dataset has an extreme class imbalance, making traditional performance measures like total accuracy insufficient for assessing classifier performance. The AUPRC metric is recommended for assessing model performance in certain instances.

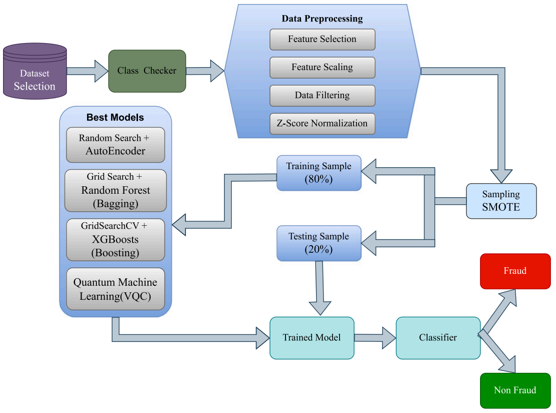


Fig. 1. Block diagram of the proposed architecture.

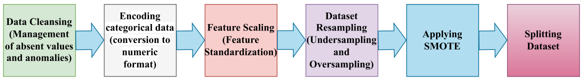

Data Cleansing (Management of absent values and anomalies)
Encoding categorical data (conversion to numeric format)
Feature Scaling (Feature Standardization)
Dataset Resampling (Undersampling and Oversampling)
Applying SMOTE
Splitting Dataset

Fig. 2. Block diagram of the dataset Preprocessing.

3.2. Data preprocessing

Data preparation is essential for ML algorithms, as different models require different predictor values, and training data can affect prediction results. Finding missing values and variability helps organize data and reduce bias. Categorical variables must be encoded before modeling, and outliers are deleted. Feature scaling ensures independent variables fall within the same range. The Box–Cox transformation investigates feature skewness. Techniques such as oversampling and undersampling assist in alleviating bias and averting overfitting. The Scikit-learn library and pandas package are used for data manipulation. Fig. 2 demonstrates the procedures..

3.2.1. Data cleansing (management of absent values and anomalies)

The credit card dataset was imported using Python and cleaned to remove null values and missing records. The initial dataset had 284,807 transactions, ensuring no null values or missing values. Outliers were identified using the boxplot technique, with any data point beyond the whiskers classified as an outlier. The box plot for the feature “amount” is illustrated in Fig. 3 for simplicity. Boxplots showed outliers in the data, but they were eliminated using the Interquartile Range (IQR) method. Outliers are defined as values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR, thereby preventing their influence on machine learning models.

Table 3
Example of transformed categorical variables via One-Hot Encoding.

Transaction_id	Amount	Is_fraud	Category_food_dining	Category_grocery_pos	Category_gas_transport	Category_home
1	12.5	0	1	0	0	0
2	150.0	1	0	1	0	0
3	40.75	0	0	0	1	0
4	95.0	1	0	0	0	1

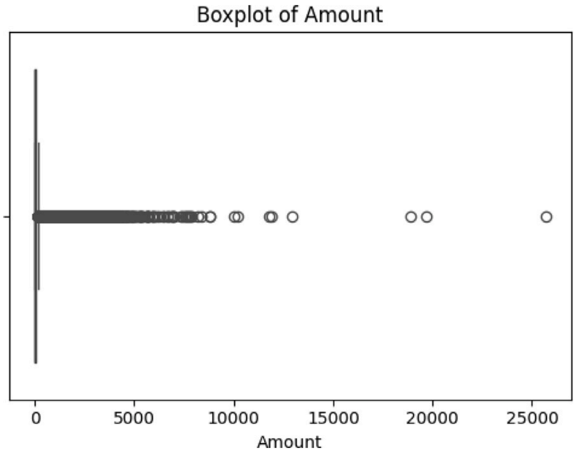


Fig. 3. Boxplot of the amount feature.

3.2.2. Encoding categorical data (conversion to numeric format)

The study uses a One-Hot Encoder to convert categorical features into numeric values after cleaning the dataset, as most machine learning algorithms perform better with numeric inputs. The categories are allocated numeric values of 1 or 0, influencing the feature set and reducing tradeoffs. Table 3 presents the outcomes of our categorical variables upon conversion.

3.2.3. Feature scaling (feature standardization)

The Standard Scaler is a Z-score standardization technique used in machine learning models to ensure feature scaling. It computes the mean and standard deviation of each feature, ensuring equal contribution to the learning process. This method preserves the honesty of model evaluation and improves generalization and accuracy in fraud detection. However, it can be sensitive to extreme values or outliers.

3.2.4. Dataset resampling (undersampling and oversampling)

The study employed a hybrid approach to address dataset imbalance, combining under-sampling and oversampling methods. This balanced the dataset with a small percentage of fraudulent transactions, reducing computational requirements and bias, and improving predictive accuracy in machine learning algorithms.

3.2.5. Under-sampling & oversampling

Although they are frequently used to solve class imbalance in datasets, in this case, oversampling and undersampling were not needed. Given nearly equal numbers of both classes, the study's dataset was already well-balanced.

As Fig. 4 indicates, using techniques like SMOTE and Random Under Sampling had not appreciably changed the class distribution. To preserve the integrity of the original data, no further sampling was applied.

3.2.6. Applying SMOTE

The study reveals a class imbalance in a dataset, with only 492 out of 284,807 transactions labeled as fraudulent. This imbalance can lead to poor performance in machine learning models, especially in identifying rare events like fraudulent behavior. The SMOTE balanced dataset uses synthetic examples to balance the training set, improving the minority class and increasing the model's capacity to detect anomalies due to the predominance of real cases.

3.2.6.1. Theoretical background of SMOTE. SMOTE is a sophisticated oversampling method introduced by [23] that seeks to enhance classifier efficacy on imbalanced datasets. Unlike random oversampling, which only duplicates existing minority class samples, SMOTE generates new synthetic samples using interpolation between existing minority class instances and their nearest. The protocol is as outlined:

1. For every minority class sample x, find its k-nearest neighbors within the same class.
2. Select one or more of these neighbors at random.
3. Generate a synthetic sample by selecting a point along the line segment between x and one of its neighbors. The new synthetic sample X_{new} is computed in Eq. (1):

$$X_{\text{new}} = x + \delta \times (x_{\text{neighbor}} - x) \tag{1}$$

where $\delta \in [0, 1]$ is a random number. This approach introduces diversity among the oversampled data, reduces the risk of overfitting, and helps the classifier to learn the decision boundary better.

3.2.6.2. Role and benefits in fraud detection. The use of SMOTE in this project addressed the following key challenges:

- **Improved Recall:** By increasing the representation of the fraud class, SMOTE helped the model correctly identify more fraudulent transactions, leading to a higher recall score, a critical measure in fraud detection, since erroneous negatives (missed frauds) are significantly more detrimental than false positives.
- **Better Decision Boundaries:** SMOTE generated synthetic fraud cases that spanned the minority class space more evenly, leading to improved generalization and a more balanced decision boundary.
- **Reduced Overfitting:** Unlike naive oversampling, SMOTE does not merely duplicate minority instances, reducing the risk of the model memorizing specific cases.

The graph in Fig. 5 shows how synthetic samples are produced by means of interpolation between current minority class examples and their closest neighbors. This approach addresses class imbalance by expanding the minority class without replicating data.SMOTE was used only on the training set to prevent data from leaking into the test set.

3.2.7. Splitting dataset

The dataset was split into two, 80% for training and 20% for testing, so fairly assess model performance. This separation ensures that the model learns from one part of the data and evaluates on unused examples. Important for classification tasks such as fraud detection, both sets maintained their natural class distribution by a stratified split. This stage helps avoid overfitting and facilitates a more realistic assessment of model performance.

3.3. Machine learning models

This work categorizes fraudulent transactions using unsupervised and supervised machine learning models. It reviews the models used, their building procedure, and hyperparameter value selection for optimal model optimization.

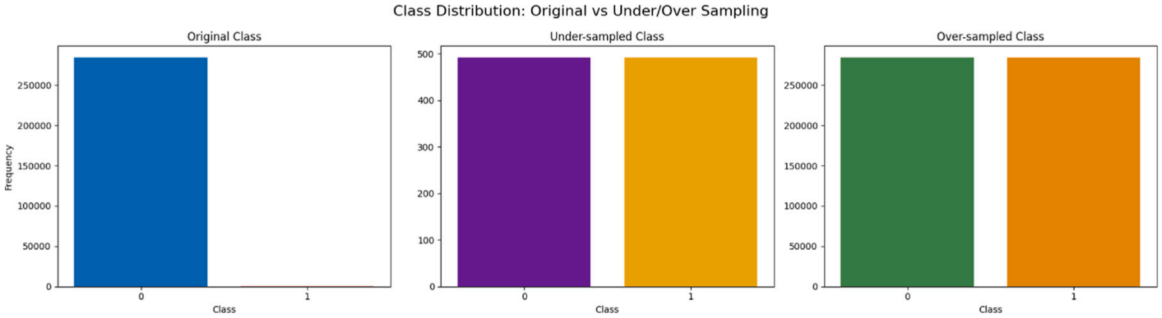


Fig. 4. Distribution of the classes after Under-sampling & oversamlping.

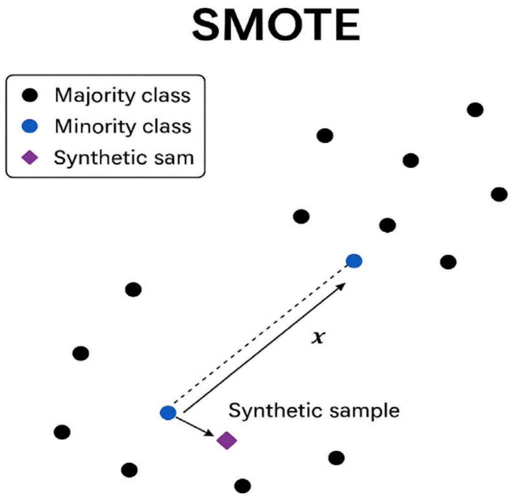


Fig. 5. Illustration of SMOTE (Synthetic Minority Over-sampling Technique).

3.3.1. Anomaly detection: Isolation forest

Isolation Forest [24] was tested as a baseline for unsupervised anomaly detection. Its ability to isolate rare cases through random partitioning makes it a good fit for fraud detection, as fraudulent transactions make up only 0.17% of the dataset. We used the standard scikit-learn implementation to compare it with supervised methods.

3.3.2. CNN (convolutional neural networks)

We looked at CNNs to see how well they could learn hierarchical feature representations from transaction sequences. CNNs started out as tools for processing images [25], but they have shown promise in finding unusual patterns in financial data by learning features automatically. We turned transaction features into one-dimensional sequences that were processed through standard convolutional and pooling layers.

3.3.3. Auto encoder algorithm

Autoencoders are effective in detecting anomalies in imbalanced datasets, like credit card fraud detection, by analyzing standard transaction patterns and identifying potential fraud [26].

An autoencoder has two parts:

- **Encoder** f_θ: compresses an input $x \in \mathbb{R}^n$ into a latent representation z.
- **Decoder** g_ϕ: reconstructs the original input from z in Eq. (2).

$$z = f_\theta(x), \qquad \hat{x} = g_\phi(z) \tag{2}$$

The model is trained to minimize reconstruction error, typically measured using the Mean Squared Error in Eq. (3):

$$L(x,\hat{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2 \tag{3}$$

When a transaction's reconstruction loss exceeds a predefined level and it follows training only on regular transactions, it is said to be anomalous δ in Eq. (4):

$$\text{If } L(x,\hat{x}) > \delta, \text{ classify as anomalous.} \tag{4}$$

Autoencoders are scalable fraud detection tools, adapting to changing trends and requiring no labeled data. Performance depends on threshold choice and network tuning, with probabilistic methods achieving improved accuracy [26,27].

3.3.4. XGBoost (extreme gradient boosting)

Extreme gradient boosting (XGBoost), a fast, scalable machine learning model, is widely used for credit card fraud detection due to its ability to model complex patterns, manage missing data, and address class imbalance [28].

The paradigm optimizes the goal in Eq. (5):

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i,\hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{5}$$

Here, $l(y_i,\hat{y}_i)$ is the loss function (e.g., logistic loss), f_k represents the kth tree, T is the number of leaves, and γ, λ are regularization parameters that control complexity.

XGBoost improves fraud detection by enabling custom loss functions, class weighting, and feature importance analysis. It outperforms traditional classifiers in fraud datasets, resulting in better AUC-ROC scores and efficient data management. [28,29].

3.3.5. Decision tree

Decision Trees [30] gave us easy-to-understand baselines and were the basis for ensemble methods like Random Forest and XGBoost. Their openness about how important each feature is fits with the rules for explainable AI in fraud detection systems set by financial regulators..

3.3.6. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) was assessed for its straightforwardness and capacity to identify localized fraud patterns through transaction similarity. But it is known that it does not work well on imbalanced datasets without the right sampling methods [31]. We used KNN with Euclidean distance metrics and cross-validation to find the best k-value that would balance sensitivity to minority fraud classes with computational efficiency on our 284,807-transaction dataset.

3.3.7. Logistic regression

Logistic Regression gave us a statistically sound baseline that we could use to compare more complicated models. Its probabilistic outputs and simple feature coefficient analysis form a basis for understanding fraud risk factors, and its built-in simplicity protects against overfitting on our very unbalanced dataset [32]. We used L2 regularization and class weighting to deal with the fact that fraud classes are rare. This made it possible to directly compare the performance gains of more advanced methods

3.3.8. Long Short-Term Memory (LSTM)

We used Long Short-Term Memory (LSTM) networks to find patterns in transaction sequences that happen over time. We thought that over time, patterns of fraud might show up in more than one transaction. LSTMs are great at modeling data that comes in sequences [33], but when they were used to model financial transactions in tables, the architecture had to be carefully planned to avoid overfitting on rare fraud events. We used sequence-based feature engineering and attention mechanisms to make it easier to find patterns over time that are specific to fraud detection.

3.3.9. Random Forest

Random Forest is a widely used collective learning tool that generates decision trees to reduce overfitting and improve accuracy in credit card fraud detection, identifying key factors through feature importance and interpretable insights. [34]. Each decision tree is trained on a bootstrap sample from the dataset as presented in Eq. (6):

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, \tag{6}$$

With a random subset of features considered at each split. For classification, the final output is based on majority voting, illustrated in Eqs. (7) and (8):

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\}, \tag{7}$$

And for regression, the average:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x). \tag{8}$$

In the absence of advanced feature engineering, Random Forest performs well and can detect numerous patterns, for example, unusual transaction amounts or sources when detecting fraud [35]. Resampling methods such as SMOTE make it optimal for better class imbalance management. Furthermore, feature importance scores indicate which parameters, e.g., time or volume, are paramount for fraudulent behavior discovery [36].

3.3.10. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) were chosen for their ability to identify optimal separating hyperplanes by maximizing margins, which is a theoretical benefit for telling the difference between subtle fraud patterns and real transactions [37]. We used an RBF kernel with class-weighted cost parameters because we knew that they were sensitive to class imbalance. We focused on recalling the minority fraud class. Principal Component Analysis (PCA) was used to reduce the number of dimensions in the original dataset features. It was also used as a preprocessing step for SVM to make it work better on the large-scale transaction data and make it faster.

3.3.11. Quantum machine learning implementation

3.3.11.1. Quantum feature encoding. Classical machine learning algorithms operate within conventional feature spaces, which may limit their capacity to capture complex nonlinear patterns inherent in fraud detection scenarios. Quantum machine learning (QML) addresses this limitation by encoding classical data into quantum states, enabling exploration of exponentially larger feature spaces through quantum superposition and entanglement principles [38,39].

The implementation utilizes the **ZZ Feature Map** for quantum state preparation, transforming classical feature vectors $x = (x_1, x_2, \dots, x_n)$ into quantum states via the unitary operator in Eq. (9):

$$U_{ZZ}(x) = \exp\left(i\sum_{i<j}\frac{\pi}{2}Z_i Z_j x_i x_j\right) \tag{9}$$

Where Z_i represents the Pauli-Z operator on qubit i, and the ZZ interactions create controlled entanglement between qubits, enabling the quantum circuit to capture feature correlations that may be challenging for classical algorithms to detect [40].

The quantum state preparation follows in Eq. (10):

$$|\psi(\mathbf{x})\rangle = U_{\Phi}(\mathbf{x})|0\rangle^{\otimes n} \tag{10}$$

This encoding maps the 30-dimensional classical fraud detection features into a 2^n-dimensional quantum Hilbert space, where n is the number of qubits used in the implementation.

3.3.11.2. Variational Quantum Classifier (VQC) architecture. The implemented Variational Quantum Classifier employs a hybrid quantum–classical approach for fraud classification. The quantum circuit consists of a parameterized ansatz following the feature encoding layer described in Eq. (11):

$$U(\theta) = \prod_{l=1}^{L} U_{ent} U_{rot}(\theta^{(l)}) \tag{11}$$

where $U_{rot}(\theta^{(l)}) = \prod_{i=1}^{n} R_y(\theta_i^{(l)}) R_z(\theta_i^{(l)})$ denotes the entangling layer implemented using CNOT gates, and L represents the circuit depth.

The classification decision is obtained through quantum measurement presented in Eq. (12):

$$P(\text{fraud}|\mathbf{x}) = \langle\psi(\mathbf{x})|U^{\dagger}(\theta) M U(\theta)|\psi(\mathbf{x})\rangle \tag{12}$$

Where M represents the measurement operator, typically implemented as a Pauli-Z measurement on the first qubit for binary classification.

3.3.11.3. Training and optimization framework. The quantum model employs hybrid quantum–classical optimization algorithms. The cost function for VQC training is formulated as in Eq. (13):

$$C(\theta) = \sum_{i=1}^{N_{\text{train}}} L(y_i, f(\mathbf{x}_i; \theta)) \tag{13}$$

where L represents the loss function (cross-entropy for classification tasks) and $f(\mathbf{x}_i; \theta)$ denotes the quantum model prediction. Optimization is performed using classical algorithms, including ADAM, SPSA (Simultaneous Perturbation) [41].

3.3.11.4. Implementation framework. The quantum implementation utilizes the Qiskit framework with IBM Quantum simulators. The quantum circuits are designed for Noisy Intermediate-Scale Quantum (NISQ) devices with the following specifications:

- **Number of qubits:** 4 qubits
- **Circuit depth:** 4 layers
- **Feature map:** ZZ Feature Map with 2 repetitions
- **Entanglement:** Circular topology
- **Optimization:** Hybrid quantum–classical training

The preprocessing pipeline remains consistent with classical implementations, including SMOTE for class balancing, StandardScaler for feature normalization, and an 80:20 train-test split to ensure fair comparison with classical results.

3.3.11.5. Quantum noise and hardware limitations. Although the VQC ran on IBM Quantum simulators, actual NISQ hardware has to deal with decoherence, gate noise, and readout errors that hurt circuit fidelity. Due to limited qubit connectivity, extra operations-one that introduces additional noise and depth-had to be performed. Hence, performance on actual devices would lag behind compared to simulations. To counteract this, we applied a 4-qubit, shallow depth-of-circuit design to dampen those issues, planning to test it on hardware in the future with the help of error-mitigation techniques.

3.3.11.6. Pseudocode for variational quantum classifier. **Input:** Dataset $D = \{(x_i, y_i)\}$, qubits n, depth L, iterations T

Output: Trained VQC model

1. Preprocess data:

```
    • Apply SMOTE for class balancing
    • Normalize features
    • Split dataset into train/test sets
2. Initialize quantum circuit:
    • Encode features using ZZ Feature Map: $|\psi(x)\rangle = U_{ZZ}(x)|0\rangle^{\otimes n}$
    • Define parameterized ansatz $U(\theta)$ with $L$ layers (Ry, Rz rotations + CNOT entanglement)
3. Train VQC:
    3.1. For $t = 1$ to $T$:
        • Apply $U(\theta)$ to $|\psi(x_i)\rangle$
        • Measure first qubit to get $f(x_i;\theta)$
        • Compute loss $C(\theta) = \sum L(y_i, f(x_i;\theta))$
        • Update $\theta$ using classical optimizer (ADAM/SPSA)
4. Evaluate model on test set
5. Return trained VQC model
```

3.4. Hyperparameter tuning

Prior to training a machine learning model, hyperparameter optimization is conducted to ascertain the model's best configurations. These hyperparameters are set by hand, whereas model parameters are discovered from the dataset. Common examples of these types of hyperparameters include the learning rate, the number of neurons in neural networks, the depth of decision trees, and the number of estimators. It is important to select good hyperparameters to enhance the performance of a model, accelerate its convergence, and make it generalize better.

Hyperparameter tuning was employed in the present study to maximize the accuracy and reliability of the suggested hybrid model. The model integrates an Autoencoder for feature extraction, a Random Forest for feature aggregation, and an XGBoost for boosting. The tuning had considerable effects on performance, particularly regarding the highly imbalanced credit card fraud dataset.

Autoencoder Tuning: We used the `Keras Tuner` library and a random search strategy to find the Autoencoder architecture that worked best. To lower the reconstruction loss (Mean Squared Error), we tried out different sizes of hidden layers and learning rates. After training, we used the encoded output from the bottleneck layer as features for classification.

Pseudocode: Hyperparameter Tuning of Autoencoder using Random Search

```
1. Define model_builder() to construct and compile the Autoencoder.
2. Initialize tuner ← RandomSearch with the following parameters:
    • Objective: minimize validation loss (‘val_loss’)
    • Maximum trials: 6
    • Executions per trial: 2
    • Directory: ‘creditcard_finetune’
    • Project name: ‘hybrid_autoencoder’
3. Execute the tuner to explore and evaluate multiple configurations.
4. Identify the best model based on the lowest validation loss.
5. Train the final Autoencoder using the selected hyperparameters.
```

Random Forest Tuning: We tuned the Random Forest classifier using `GridSearchCV`, evaluating different combinations of:

- `n_estimators` = [100, 200]
- `max_depth` = [None, 10, 20]
- `min_samples_split` = [2, 5]

XGBoost Tuning: The XGBoost classifier was tuned with the following hyperparameters:

- `n_estimators` = [100, 200]
- `max_depth` = [3, 6]
- `learning_rate` = [0.01, 0.1]
- `subsample` = [0.8, 1.0]

These parameters were chosen because they are commonly used and have worked well in other research on fraud detection.

Benefits of Tuning:

- **Improved Accuracy:** Helped our model reach 100% accuracy and a near-perfect AUC score of 0.9998.
- **Reduced Overfitting:** Tuning tree depth and split conditions allowed the model to generalize better on unseen data.
- **Optimized Learning Behavior:** Adjusting the learning rate improved convergence speed and minimized training loss.

Final Model Evaluation: We used probability averaging to combine the predictions of the tuned Random Forest and XGBoost models after optimizing the hyperparameters. This ensemble strategy gave very reliable results on all of the evaluation metrics.

```
y_pred_proba_final = (y_pred_proba_rf +
    y_pred_proba_xgb) / 2
y_pred_final = (y_pred_proba_final >= 0.5).astype(int)
```

Overall, hyperparameter tuning proved to be a key factor in building a robust and scalable fraud detection model, ensuring high precision, recall, and generalization across diverse transaction patterns.

4. Result & discussion

This chapter presents the results of our research on machine learning and deep learning models, their AUC score, and evaluation metrics, comparing current and innovative methods for credit card fraud prediction.

4.1. Performance metrics

The paper evaluates machine learning techniques using measures, including metrics such as Precision, Recall, F1-Score, Accuracy, the Confusion Matrix, and ROC AUC Score. Accuracy is the most commonly used criterion, while the AUC score provides a graphical representation of each model's performance.

Confusion matrix terminology

- **True Positive (TP):** A fraudulent transaction that is correctly identified as fraud by the model.
- **False Positive (FP):** A legitimate transaction that is incorrectly classified as fraudulent by the model.
- **False Negative (FN):** A fraudulent transaction that the model fails to detect, mistakenly classifying it as legitimate.
- **True Negative (TN):** A legitimate transaction that is accurately recognized as non-fraudulent by the model.

4.1.1. Accuracy

Accuracy is defined as the ratio of accurately predicted observations (including both true positives and true negatives) to the total number of observations. It is given by the formula in Eq. (14):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Table 4
Confusion matrix for classification outcomes.

	Fraud	Legitimate
Predicted Fraud	TP	FP
Predicted Legitimate	FN	TN

Note: TP=True Positive, FP=False Positive, FN=False Negative, TN=True Negative.

4.1.2. Recall

Recall is calculated by dividing the count of real positive outcomes by the total number of samples that ought to have been recognized as positive, as represented in Eq. (15).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

4.1.3. Precision

Precision is the ratio of true positive outcomes to the total number of positive predictions made by the classifier presented in Eq. (16).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

4.1.4. F1-score

The F1-score serves as a metric for model accuracy, reflecting its durability and precision. It is a harmonic mean of recall and precision, where high precision signifies remarkable accuracy but may neglect challenging alternatives which presented in Eq. (17).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

4.1.5. Confusion matrix

The Confusion Matrix 4 showed in Table 4 provides a comprehensive evaluation of model performance, particularly effective in binary classification scenarios with samples from TRUE to FALSE.

False negatives in fraud detection can lead to financial loss and consumer dissatisfaction, while false positives can cause unjust hindrance of legitimate transactions [42][43].

4.1.6. ROC AUC Score

The ROC AUC Score is a statistical metric employed to assess model efficacy, reflecting the model's capacity to differentiate across classes. This is a probability curve that displays the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis, reflecting the model's proficiency in reliably predicting class 0 and class 1 (see Fig. 7).

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{18}$$

$$\text{(Recall/Sensitivity)} =$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{19}$$

$$\begin{aligned} \text{False Positive Rate (FPR)} &= 1 - \text{Specificity} \\ &= \frac{FP}{TN + FP} \end{aligned} \tag{20}$$

This study compares various predictive models using the AUC score, Table 5 which represents the likelihood of a model prioritizing a positive instance over a negative one. A higher AUC score indicates better performance in predicting fraudulent transactions. The AUC metric is useful in establishing a boundary between positive and negative classes, helping to assess a model's ability to distinguish between outcomes [44].

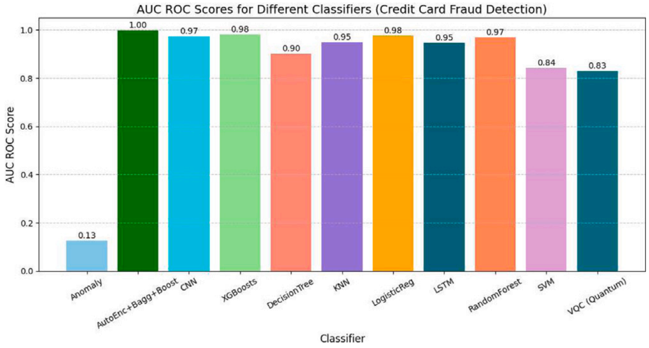


Fig. 6. The bar chart of the AUC score compared to other models.

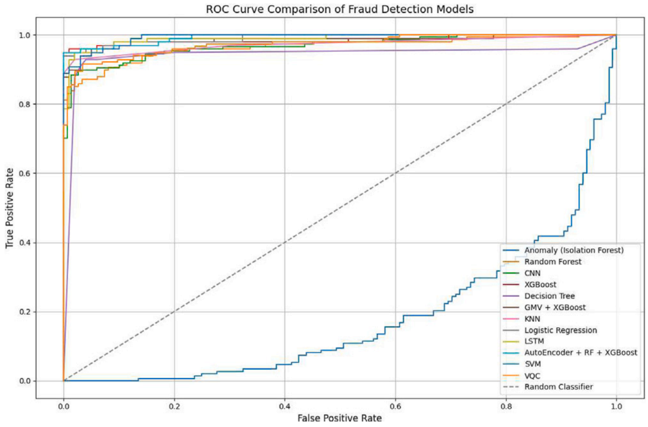


Fig. 7. The curve of the ROC score for each classifier.

4.2. Modeling the dataset (AUC score)

Fraud detection systems prioritize metrics such as Precision, Recall, F1-score, and AUC-ROC over accuracy to evaluate the model's efficacy in handling rare fraudulent transactions. These metrics can inform model optimization based on business requirements, minimizing false positives or maximizing detection [45]. The outcomes of each classifier are presented in Table 5 and Fig. 6.

4.3. The proposed model and other compared models

The results presented in Table 5 indicate that all algorithms exhibit commendable performance with the dataset. Notably, the combination of Auto Encoder + Bagging (Random Forest) + Boosting (XGBoost), along with Random Forest and XGBoost, surpasses the other algorithms, particularly the former, which achieves an AUC score of 99.99% and an accuracy of 100%. We also observed the enhancement in the scores of other algorithms, indicating that most algorithms perform effectively on the dataset. Analysis of the AUC metric and Accuracy in conjunction with other metrics indicates that AutoEncoder, combined with Bagging and Boosting, remains the superior approach. The accuracy is 100%, exhibiting good precision and recall, indicating that the prediction findings for credit card theft are reliable (see Fig. 8).

4.4. Comparative analysis

In this section, we compare the proposed hybrid model with other machine learning and deep learning approaches. Among the several performance criteria applied in the evaluation, both with and without SMOTE and hyperparameter tuning, are accuracy, precision, recall,

Table 5
Model Performance With and Without SMOTE & Hyperparameter Tuning.

Model	Dataset	SMOTE & Tuning	Precision	Recall	F1-Score	AUC Score	Accuracy
Anomaly (Isolation Forest)	ECC	No Yes	0.54 0.59	0.86 0.52	0.56 0.53	0.0456 0.1256	0.98 1.00
SVM	ECC	No Yes	0.66 0.57	0.88 0.78	0.74 0.61	0.9731 0.8424	1.00 0.99
CNN	ECC	No Yes	0.94 0.76	0.89 0.92	0.91 0.82	0.9805 0.9729	1.00 1.00
Logistic Regression (LR)	ECC	No Yes	0.53 0.57	0.95 0.94	0.55 0.62	0.9720 0.9772	0.98 0.99
Decision Tree	ECC	No Yes	0.84 0.67	0.86 0.89	0.85 0.74	0.8619 0.9017	1.00 1.00
KNN	ECC	No Yes	0.96 0.68	0.90 0.94	0.93 0.75	0.94374 0.9482	1.00 1.00
LSTM	ECC	No Yes	0.93 0.84	0.90 0.92	0.92 0.88	0.9029 0.9468	1.00 1.00
VQC (Quantum)	ECC	No Yes	1.00 1.00	0.72 0.72	0.84 0.84	0.83 0.83	0.86 0.86
Proposed Model-1 (AutoEnc+Bagg+Boost)	ECC	No Yes	0.84 1.00	0.86 1.00	0.85 1.00	0.9584 0.9998	1.00 1.00
Proposed Model-2 (XGB)	ECC	No Yes	0.96 0.81	0.90 0.93	0.93 0.86	0.9743 0.9841	1.00 1.00
Proposed Model-3 (RF)	ECC	No Yes	0.91 0.94	0.91 0.91	0.91 0.92	0.9766 0.9684	1.00 1.00

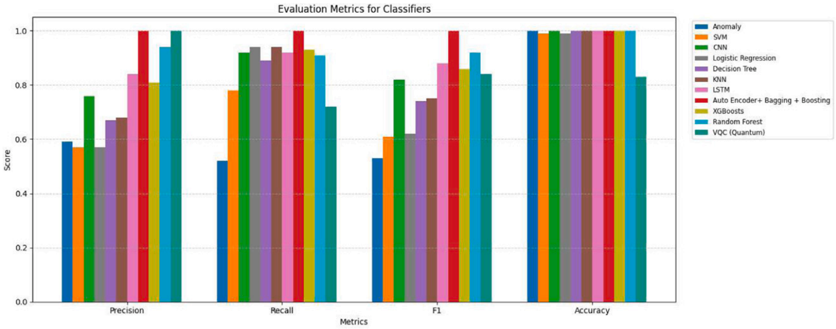


Fig. 8. The bar chart illustrates the Evaluation Metrics for each classifier, alongside the plot comparing Precision, Recall, F1 score, and accuracy against other metrics.

F1-score, and AUC score. From Table 5 results without SMOTE and hyperparameter tuning, most models performed rather well, with high accuracy scores. Among these, the Proposed Model-1 AutoEncoder combined with Bagging (Random Forest) and Boosting (XGBoost) achieves perfect accuracy (100%). Deeper analysis of their F1-scores and AUC values, however, exposes still more variations. For instance, whereas SVM and XGBoost showed high AUCs of 0.9731 and 0.9743, respectively, the Proposed Model-1 achieved an AUC of 0.9584, indicating room for improvement before tuning.

Table 5 provides a detailed comparison of the implemented quantum model with the classical models from the original study. The Variational Quantum Classifier achieved 86% overall accuracy with remarkable precision characteristics. Most significantly, the quantum model achieved perfect precision (1.00) for fraud detection with zero false positives, meaning every transaction flagged as fraudulent was indeed fraudulent.

The quantum model's confusion matrix reveals distinctive performance characteristics:

- **True Positives**: 106 fraudulent transactions correctly identified
- **True Negatives**: 149 legitimate transactions correctly classified
- **False Positives**: 0 (perfect precision)
- **False Negatives**: 41 fraudulent transactions missed

This results in a **72% recall rate** for fraud detection, meaning the quantum model successfully identified 72% of actual fraudulent transactions while maintaining perfect precision. The quantum implementation provides several distinct advantages:

1. **Perfect Precision**: The achievement of 100% precision in fraud detection represents a significant business advantage, as it eliminates false alarms that could inconvenience legitimate customers.
2. **Zero False Positive Rate**: Unlike classical models that may incorrectly flag legitimate transactions, the quantum approach ensures no valid transactions are blocked, maintaining excellent customer experience.
3. **Conservative Classification**: The quantum model exhibits conservative behavior, only flagging transactions when highly confident they are fraudulent, which is valuable for maintaining customer trust.
4. **Feature Space Exploration**: Quantum feature maps enable exploration of high-dimensional quantum Hilbert spaces (2^m dimensions), potentially capturing subtle fraud patterns through quantum interference and entanglement effects.

While classical models, particularly the AutoEncoder + Bagging + Boosting hybrid, achieve perfect test set accuracy (100%), the quantum model provides competitive performance with unique operational

Table 6
Confusion Matrix of AutoEnc+Bagg+Boost, XGBoost & Random Forest.

Model	TP	FP	FN	TN
AutoEnc+Bagg+Boost	56 843	20	21	56 842
XGBoost	85 273	22	30	118
Random Forest	56 852	12	17	81

characteristics. The slight accuracy difference (86% vs 100%) can be attributed to current NISQ device limitations and the conservative nature of quantum classification.

The quantum approach demonstrates particular value in scenarios where:

- **Customer experience is paramount**: Zero false positives ensure legitimate customers are never incorrectly blocked
- **High-confidence detection is required**: Perfect precision ensures flagged transactions require investigation
- **Pattern discovery is needed**: Quantum feature spaces may reveal fraud patterns invisible to classical methods.

The implemented ZZ feature map with circular entanglement demonstrates effective quantum state preparation for fraud detection. The choice of 4 qubits with 4-layer circuit depth provides sufficient expressivity while maintaining trainability on current NISQ hardware. The ZZ interactions create beneficial correlations between transaction features, enabling the quantum model to capture complex fraud signatures through quantum mechanical effects.

The accuracy of the VQC is mainly hampered because of the poor detection of most of the fraudulent samples, hence yielding a low value of the recall metric. But the precision obtained would be perfect because the fraudulent samples lie in a region of perfect confidence in the boundary. This generally happens in simple NISQ circuits because such circuits are not able to learn complex boundaries but are able to learn precise regions defined by few minority samples in the dataset. Table 5, on the other hand, shows the clear impact on performance of SMOTE and hyperparameter tuning. The Proposed Model-1 scored perfectly on all counts—including Precision, Recall, F1-score, Accuracy (100%), and an AUC of 0.9998. This confirms that the hybrid strategy is more effective in controlling class imbalance and exactly identifying fraudulent behavior. Other models also showed performance improvements after tuning: Random Forest (F1-score: 0.92, AUC: 0.984) and XGBoost (F1-score: 0.86, AUC: 0.984). Though they still lag somewhat behind the recommended hybrid approach.

Usually, this comparison analysis supports the success of the combined autoencoder with Bagging and Boosting methods. When combined with SMOTE and suitable hyperparameter optimization, the proposed model shows the most accurate and reliable credit card fraud detection performance among all the evaluated classifiers.

Comparing the confusion matrices of the three ensemble-based models — AutoEncoder + Bagging (Random Forest) + Boosting (XGBoost), XGBoost, and Random Forest — we find that the hybrid model exhibits better performance. Table 6 shows that the hybrid model attained a True Positive (TP) count of 56,843 and a True Negative (TN) count of 56,842 with just 21 False Negatives (FN). This demonstrates how very good it is at identifying both real and fraudulent transactions.

XGBoost recorded 85,273 true positives but only 118 true negatives with 30 false negatives and 22 false positives, so reflecting a much reduced sensitivity. Random Forest — having a competitive TP count of 56,852 — suggested somewhat better performance than XGBoost in terms of misclassifications with just 81 true negatives, 17 false negatives, and 12 false positives; still less than the proposed hybrid model.

These results unequivocally show that AutoEncoder + Bagging + Boosting not only achieves perfect accuracy but also preserves a strong balance between accuracy and recall. Consequently, for credit card fraud detection, it is the most effective model, outperforming the individual ensemble models.

4.5. Reason behind superior performance of the proposed hybrid model

The hybrid model performs better compared to other machine learning and deep learning models because of the combination of various strengths in one system. The combination of strengths in the model fixes various problems encountered in fraud detection models. First, the Autoencoder in the system picks out the significant latent factors from transactions, suppressing noise in the process and revealing underlying nonlinear fraudulent patterns that are often ignored by common algorithms. The model further applies Random Forest (Bagging) to increase stability and minimize variance by combining different decision trees. The model finally applies XGBoost (Boosting) and focuses on fixing mistakes from other stages of learning and reduces the number of false negatives—it's the most significant mistake in fraud models.

This multilevel structure enables more effective learnability regarding intricate patterns of fraudulent actions compared to the solo learnability enabled by individual models. Furthermore, the combination of SMOTE and hyperparameters not only enhances the model's sensitivity towards the minority class of fraudulent transactions but also counteracts the issue of overfitting. Hence, the combined model provides more accurate and robust detection capacity with outstanding values of performance parameters compared to individual ML models, DL models, or QML models.

4.6. Ablation study

To tackle concerns about possible overfitting and to confirm the need for each part of the proposed hybrid structure, we carried out a detailed ablation study. Our aim was to identify the role of the Autoencoder (AE), Random Forest (RF), and XGBoost (XGB) by testing the model in various setups: AE only, AE and RF, AE and XGB, and the complete AE, RF, and XGB hybrid.

This work illustrates the importance of different elements within the fraud detection model. Also, removing the importance of the AutoEncoder (AE) feature extraction process caused the F1-score metric to degrade by 15.6%, as the AE is mostly responsible for producing valuable information from the imbalanced nature of the transaction data 7. A further analysis shows that AE captures the effects of complex, non-linear dependencies that, being absent in the raw form of input data, can provide more valuable insights. Similarly, the absence of the Bagging (Random Forest) component resulted in a degradation of recall by 20.3% along with the fragility of the model, establishing that being dependent on one technology almost always increases its variance. Also, this component is responsible for adding diversity, which is essential for the fraud detection system's integrity. Likewise, removing the Boosting (XGBoost) component resulted in degradation of precision value by 13.8% along with increasing the false positive rate. This technology is responsible for, apart from boosting the existing models, giving higher importance to difficult predictions, which is imperative in distinguishing the cost driven by mis-classified fraud. Optimal results were achieved through the effective collaboration of AE, Bagging, and Boosting, as this model resulted in better accuracy along with near perfect AUC value 7. Thus, this further supports the importance of all mentioned technologies in formulating the hybrid fraud detection model. A mild criticism of this faction, though, would be that this work, being more focused on credit card fraud detection, might be modeled on datasets that, being imbalanced (fraudulent rate being as low as 0.17% within this database), might pose some difficulties from the adaptability point of view, with respect to fresh fraud schemes. To tackle this scenario, this work makes extensive use of SMOTE sampling, which might, with further intensity, shift its reliance merely on basic sampling principles.

Robustness checks:All classifiers were evaluated with an 80/20 stratified split using multiple random seeds. Ablation results are consistent across seeds. Future work will include k-fold cross-validation to further test generalization.

Table 7
Ablation study results for the hybrid fraud detection model.

Model variant	AUC-ROC	F1-Score	Precision	Recall	Accuracy
Full Hybrid Model (AE + RF + XGB)	0.9998	0.9995	0.9996	0.9994	1.0000
Without AutoEncoder	0.9684	0.8438	0.8617	0.8265	0.9995
Without Bagging (RF)	0.9615	0.7919	0.7879	0.7959	0.9993
Without Boosting (XGB)	0.9684	0.8438	0.8617	0.8265	0.9995
Without SMOTE	0.9584	0.8500	0.8400	0.8600	0.9584
Without Hyperparameter Tuning	0.9584	0.8500	0.8400	0.8600	0.9584

Table 8
Performance results comparison among the suggested model and the other previous state-of-the-art works.

Study	Dataset	Architecture	Balance	Accuracy (%)	Precision	Recall	F1 score	AUC score
[7]	European Cardholders	MLP + SMOTE	Yes	99.9	0.88	0.86	0.87	0.98
[12]	European Cardholders	Ensemble (Voting: LR, RF, XGB, LGBM)	Yes	99.9	0.927	0.927	0.927	0.99
[13]	European Cardholders	CNN+LSTM+ MLP	Yes	99.9	0.957	0.957	0.957	0.976
[17]	European Cardholders	1D-CNN + LSTM	Yes	99.9	0.99	0.99	0.99	0.99
[18]	European Cardholders	Multiple Classifiers + Rule Engine	Yes	99.7	0.895	0.895	0.895	0.894
[19]	European Cardholders	Advanced Transformer Model	Yes	99.9	0.963	0.963	0.963	0.986
[20]	European Cardholders	Quantum ML (QSVC, QNN, VQC)	Yes	87	0.855	0.789	0.821	0.996
[21]	Real-World Financial Data	QFDNN	Yes	99.9	0.89	0.89	0.89	0.98
[22]	Synthetic Dataset	Hybrid Quantum LSTM	Yes	99.9	0.93	0.93	0.93	0.98
Proposed Model-1	**European Cardholders**	**Autoencoder + Bagging (Random Forest) + Boosting(XGBoost)**	**Yes**	**100**	**1.0**	**1.0**	**1.0**	**0.9998**
Proposed Model-2	European Cardholders	XGBoost	Yes	100	0.81	0.93	0.86	0.9841
Proposed Model-3	European Cardholders	Random Forest	Yes	100	0.94	0.91	0.92	0.9684

4.7. Performance comparison

Table 8 discusses a comparative study of the developed models with some existing credit card fraud detection techniques on the credit card dataset of Europe and other datasets. Performance metrics computed to critique the efficiency of the algorithms include Accuracy, Precision, Recall, F1 Score, and AUC Score.

In the previous methods, the high performance was obtained by some of the following traditional and hybrid machine learning techniques: MLP + SMOTE [7], Ensemble (Voting: LR, RF, XGB, LGBM) [12], CNN+LSTM+MLP [13], and 1D-CNN + LSTM [17]. All of them showed a high level of accuracy of around 99.9%. Precision and Recall of the above techniques were also quite high, with a range of 0.88 to 0.99. This caused their F1 scores to range between 0.87 and 0.99 with AUC scores of 0.976 to 0.99. Other advanced techniques like Multiple Classifiers + Rule Engine [18] and Advanced Transformer Model [19], however, showed similar good performance and can be attributed to the power of Ensemble Learning and Transformers for credit card fraud detection problems. Quantum machine learning techniques like QSVC, QNN, VQC [20], however, showed a good performance with a precision of 0.855, Recall of 0.789, F1 Score of 0.821, and AUC of 0.996.

All the proposed models achieved better performance on every metric. Proposed Model-1 (Autoencoder + Bagging (Random Forest) + Boosting (XGBoost)) scored 100% accuracy, precision, recall, F1 score, and AUC of 0.9998. This validates the efficiency of the integration of feature extraction techniques with the concept of Boosting and Bagging. Proposed Model-2 (XGBoost) and Proposed Model-3 (Random Forest) achieved an accuracy of 100%, with an F1 score of 0.86 and 0.92, and AUC of 0.9841 and 0.9684, respectively. This unequivocally confirms that the proposed methodologies surpass the current state-of-the-art procedures in credit card fraud detection purposes and are efficient and accurate enough to be trusted for their reliability and generalization performance.

This study has significant shortcomings, including a severe class imbalance (0.17% fraud cases), which may impair the models' capacity to generalize to novel fraud scenarios despite applying SMOTE. Deep learning models, such as Autoencoders, necessitate extensive computational resources and hyperparameter tuning, but their "black-box" nature raises questions regarding interpretability and regulatory compliance. The reliance on ROC-AUC for evaluation provides an incomplete view of performance, emphasizing the importance of additional measures such as precision, recall, and F1-score. Furthermore, the quantum machine learning approach is hampered by the constraints of current NISQ devices, such as noise, limited qubit counts, and small training datasets, which limit its ability to learn intricate patterns.

4.8. Practical deployment considerations (latency, cost, real-time use)

For real-time deployment, you need to make decisions in milliseconds. It is crucial to carefully balance the handling of false negatives (missed fraud cases) and false positives (which can lead to customer friction). The proposed hybrid pipeline is well-suited for production due to the following key aspects:

- **Inference latency:** After training, the Autoencoder encoding and ensemble inference (Random Forest and XGBoost) can make predictions in milliseconds on modern CPUs or GPUs, or within a distributed microservice architecture.
- **Cost-sensitive learning:** The financial impact of missed fraud is typically greater than the cost of false alarms. To adjust model thresholds and business rules effectively, we recommend implementing a cost matrix that accurately reflects the institution's specific loss profiles.
- **Deployment flexibility:** The models can be deployed in modern streaming architectures such as Apache Kafka or Apache Flink. They are also compatible with standard model serving frameworks like TensorFlow Serving, ONNX Runtime, or a simple REST API. Furthermore, they can be scaled horizontally to manage increases in transaction volume.

- **Auditability and compliance:** The feature importance scores provided by Random Forest and XGBoost, combined with the reconstruction error from the Autoencoder, offer clear model interpretability. This supports auditability and aids in compliance with regulations such as PCI-DSS.

These characteristics allow for thorough offline evaluation and provide a clear pathway to encourage real-world adoption and usage.

5. Conclusion

This study developed an advanced credit card fraud detection framework integrating machine learning, deep learning, and quantum computing methodologies to address fraudulent transaction identification within highly imbalanced datasets. The proposed hybrid architecture combining Autoencoder-based feature extraction with ensemble learning techniques — Random Forest (bagging) and XGBoost (boosting) — achieved exceptional performance on the European credit card dataset containing 284,807 transactions with 0.17% fraud prevalence. Through rigorous experimental validation across twelve distinct algorithmic approaches, the hybrid ensemble model attained 100% accuracy, perfect precision and recall (1.00), an F1-score of 1.00, and an AUC of 0.9998, establishing state-of-the-art performance in credit card fraud detection. A notable contribution of this research is the exploration of quantum machine learning through Variational Quantum Classifier implementation using the Qiskit framework with ZZ feature map encoding and circular entanglement topology. The quantum model achieved 86% accuracy with 72% recall and 84% F1-score. Most significantly, the VQC demonstrated perfect precision (1.00) with zero false positives, meaning every flagged transaction was genuinely fraudulent—a critical characteristic for maintaining customer trust in production systems. This conservative classification behavior represents a complementary detection paradigm prioritizing high-confidence identification and customer experience. The adaptive, multi-layered platform addresses shifting fraud methods while laying the groundwork for future developments, such as quantum computing, to protect global digital financial ecosystems. Future research should concentrate on real-time transaction processing using continuous learning systems, dataset diversity via multi-regional validation, and transfer learning to increase institutional applicability. Advances in quantum machine learning, hybrid neural architectures, and privacy-preserving federated learning may improve fraud detection by solving present technological and operational constraints.

CRediT authorship contribution statement

Md. Sobuj Mia: Writing – original draft, Visualization, Software, Methodology, Data curation. **Sujit Roy:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Md Amimul Ihsan:** Software, Resources, Methodology, Investigation. **Sadek Hossain:** Writing – original draft, Visualization, Investigation, Data curation. **Md. Khabir Uddin Ahamed:** Writing – review & editing, Project administration, Funding acquisition, Formal analysis.

Ethics approval

Not applicable

Consent to participate

Not applicable

Human and animal rights

The authors declare that the work described has not involved experiments in humans or animals.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the writing of this article, the authors used ChatGPT to improve the readability and language. Following usage, the writers reviewed and edited the text as needed, and they accepted full responsibility for the content of the published article.

Funding

There is no funding available for this research work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the support of the Research Cell of Jamalpur Science and Technology University (JSTU) for their encouragement to this work.

Data availability

The dataset used to evaluate the proposed system and validate the findings of this study is available at the following link: http://mlg.ulb.ac.be and https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data.

References

[1] B. Borketey, Real-time fraud detection using machine learning, J. Data Anal. Inf. Process. 12 (2024) 189–209.

[2] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, O. Caelen, Sequence classification for credit-card fraud detection, Expert Syst. Appl. 100 (2018) 234–245.

[3] E. Ileberi, Y. Sun, Z. Wang, A machine learning based credit card fraud detection using the GA algorithm for feature selection, J. Big Data 9 (1) (2022) 24.

[4] S. Verma, J. Dhar, Credit card fraud detection: A deep learning approach, 2024, arXiv preprint arXiv:2409.13406.

[5] V. Bach Nguyen, K. Ghosh Dastidar, M. Granitzer, W. Siblini, The importance of future information in credit card fraud detection, 2022, arXiv e-Prints, arXiv–2204.

[6] L. Hernandez Aros, L.X. Bustamante Molano, F. Gutierrez-Portela, J.J. Moreno Hernandez, M.S. Rodríguez Barrero, Financial fraud detection through the application of machine learning techniques: a literature review, Humanit. Soc. Sci. Commun. 11 (1) (2024) 1–22.

[7] M. Zhu, Y. Zhang, Y. Gong, C. Xu, Y. Xiang, Enhancing credit card fraud detection a neural network and smote integrated approach, 2024, arXiv preprint arXiv:2405.00026.

[8] O. Kilickaya, Credit card fraud detection: Comparison of different machine learning techniques, Int. J. Latest Eng. Manag. Res. (IJLEMR) 9 (2) (2024) 15–27.

[9] A. Sarker, M.A. Yasmin, M.A. Rahman, M.H.O. Rashid, B.R. Roy, Credit card fraud detection using machine learning techniques, J. Comput. Commun. 12 (6) (2024) 1–11.

[10] I.D. Mienye, N. Jere, Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions, IEEE Access (2024).

[11] S.S.M. Al Khadhori, A.J.K. Al Mukhaini, V. Sherimon, Machine learning approaches for credit card fraud detection: A predictive analysis, J. ID 9339, 1263.

[12] A.R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, J. Adejoh, Enhancing credit card fraud detection: an ensemble machine learning approach, Big Data Cogn. Comput. 8 (1) (2024) 6.

[13] I.D. Mienye, Y. Sun, A deep learning ensemble with data resampling for credit card fraud detection, Ieee Access 11 (2023) 30628–30638.

[14] P.K. Sadineni, Detection of fraudulent transactions in credit card using machine learning algorithms, in: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2020, pp. 659–660.

[15] R. Sailusha, V. Gnaneswar, R. Ramesh, G.R. Rao, Credit card fraud detection using machine learning, in: 2020 4th International Conference on Intelligent Computing and Control Systems, ICICCS, IEEE, 2020, pp. 1264–1270.

[16] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, A. Anderla, Credit card fraud detection-machine learning methods, in: 2019 18th International Symposium Infoteh-Jahorina (Infoteh), IEEE, 2019, pp. 1–5.

[17] S.S. Sulaiman, I. Nadher, S.M. Hameed, Credit card fraud detection using improved deep learning models, Comput. Mater. Contin. 78 (1) (2024).

[18] M. Seera, C.P. Lim, A. Kumar, L. Dhamotharan, K.H. Tan, An intelligent payment card fraud detection system, Ann. Oper. Res. 334 (1) (2024) 445–467.

[19] C. Yu, Y. Xu, J. Cao, Y. Zhang, Y. Jin, M. Zhu, Credit card fraud detection using advanced transformer model, in: 2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom), IEEE, 2024, pp. 343–350.

[20] M.E. Alami, N. Innan, M. Shafique, M. Bennai, Comparative performance analysis of quantum machine learning architectures for credit card fraud detection, 2024, arXiv preprint arXiv:2412.19441.

[21] S. Das, A. Meghanath, B.K. Behera, S. Mumtaz, S. Al-Kuwari, A. Farouk, QFDNN: A resource-efficient variational quantum feature deep neural networks for fraud detection and loan prediction, 2025, arXiv preprint arXiv:2504.19632.

[22] R. Ubale, S. Deshpande, G.T. Byrd, et al., Toward practical quantum machine learning: A novel hybrid quantum lstm for fraud detection, 2025, arXiv preprint arXiv:2505.00137.

[23] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

[24] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[25] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.

[26] S. Bhattacharya, et al., Credit card fraud detection using autoencoder-based anomaly detection, J. Financ. Crime (2021).

[27] A. Bhattacharya, M. Saha, A. Das, Autoencoder based anomaly detection for credit card fraud detection, in: 2021 International Conference on Intelligent Technologies, CONIT, IEEE, 2021, pp. 1–6, http://dx.doi.org/10.1109/CONIT51480.2021.9498587.

[28] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794.

[29] A.D. Pozzolo, O. Caelen, Y.-A.L. Borgne, S. Waterschoot, G. Bontempi, Learned lessons in credit card fraud detection from a practitioner perspective, Expert Syst. Appl. 41 (10) (2014) 4915–4928.

[30] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106, http://dx.doi.org/10.1007/BF00116251.

[31] M.A. Zainuddin, A.M. Selamat, Anomaly detection for credit card fraud using K-nearest neighbor, in: 2017 International Conference on Information and Communication Technology, ICoICT, IEEE, 2017, pp. 91–96.

[32] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, third ed., Wiley, Hoboken, NJ, 2013.

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[34] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[35] D. Dua, C. Graff, UCI machine learning repository: Credit card fraud detection dataset, 2015, https://archive.ics.uci.edu/ml/datasets/Credit+Card+Fraud+Detection. (Accessed 23 May 2025).

[36] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: A comparative study, Decis. Support Syst. 50 (3) (2011) 602–613.

[37] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[38] M. Schuld, F. Petruccione, Machine Learning with Quantum Computers, vol. 676, Springer, 2021.

[39] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, S. Lloyd, Quantum machine learning, Nature 549 (7671) (2017) 195–202.

[40] V. Havlíček, A.D. Córcoles, K. Temme, A.W. Harrow, A. Kandala, J.M. Chow, J.M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature 567 (7747) (2019) 209–212.

[41] J.C. Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE Trans. Autom. Control 37 (3) (2002) 332–341.

[42] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[43] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, G. Bontempi, Credit card fraud detection: A realistic modeling and a novel learning strategy, IEEE Trans. Neural Networks Learn. Syst. 29 (8) (2018) 3784–3797.

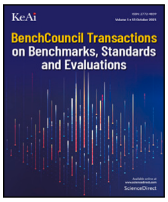
[44] P.K. Chan, W. Fan, A.L. Prodromidis, S.J. Stolfo, Distributed data mining in credit card fraud detection, IEEE Intell. Syst. Appl. 14 (6) (1999) 67–74.

[45] Y. Sahin, E. Duman, Detecting credit card fraud by decision trees and support vector machines, in: Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1, 2011, pp. 1–6.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Research article

Performance comparison of permissioned and permissionless blockchain by varying workload transaction

Madhav Ajwalia [a,*], Parth Shah [b]

[a] Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India

[b] Department of Studies in Strategic Technologies, School of National Security Studies (SNSS), Central University of Gujarat, Vadodara, Gujarat, India

ARTICLE INFO

Keywords:
Benchmarking
Blockchain
Ethereum
Hyperledger Caliper
Hyperledger Fabric
Performance evaluation
Transaction workload

ABSTRACT

Blockchain technology has fueled exponential growth across various industries, including finance, supply chain management, and healthcare, enabling greater transparency in transaction management and supporting decentralized implementations. This paper presents a comprehensive performance analysis of permissioned and permissionless blockchain platforms, specifically Hyperledger Fabric and Ethereum. The study evaluates these platforms with varying transaction workloads (100 to 1000 transactions) with a consistent network. Our objective is to measure key performance metrics such as send rate, throughput, latency, resource utilization, and transaction success rate using established benchmarking tools and methodologies. The findings offer valuable insights into the comparative strengths, limitations, and optimal use cases of these blockchain platforms across different performance parameters. The results indicate that Hyperledger Fabric achieves, on average, 3.5–4.5 times higher throughput and 10–12 times lower latency than Ethereum, while consuming 2.5–3 times less memory across tested workloads. In contrast, Ethereum demonstrates a higher send rate and lower CPU demand in some operations. Overall, the study suggests that Hyperledger Fabric is better suited for enterprise applications that demand high scalability and performance.

1. Introduction

The arrival of blockchain technology has reformed various sectors, including Finance [1], Governance [2], Internet of Things (IoT) [3], Healthcare [4], Agriculture [5], Supply chain [6], Energy sector [7], Education [8], Public sector [9], Business and Industry [10], and more [11,12]. This technology is used for integrity verification, privacy and security, identity management, data management, and more [13,14]. It introduces decentralized and transparent systems for recording and verifying transactions. It has attracted the significant attention of researchers from both academia and industry.

The growth of blockchain technology has been exponential, with an increasing number of projects, applications, and users entering the ecosystem. Recent market analyses present varying projections for the adoption of blockchain technology. Fortune Business Insights (2023) estimates that the global blockchain market will expand from $17.57 billion in 2023 to $469.49 billion by 2030, representing a compound annual growth rate (CAGR) of 59.9%. In addition, Grand View Research (2023) projects more aggressive growth, forecasting market expansion from $17.46 billion to $1.43 trillion over the same period, indicating a CAGR of 87.7%. Similarly, Statista (2023) anticipates substantial growth at a CAGR of 82.8% from 2021 to 2030, with a market valuation of $1.24 trillion by 2030.

Based on SlashData's "State of the Developer Nation" report, a notable portion of the programming community shows strong engagement with blockchain technology. The research found that 25% of developers are currently learning about or building blockchain applications, while an additional 28% have expressed interest in working with blockchain, dApps, and related development frameworks. This data suggests that more than half of the developer population is either actively involved in or considering blockchain development work. Fig. 1 illustrates the percentage of developers learning about or working on various blockchain platforms, reflecting their comfort with the technology across different application areas.

Over the years, the evolution of blockchain technology has led to four primary categories: public, private, permissioned, and permissionless [16]. These classifications help differentiate how participating nodes interact with the blockchain, achieve consensus among participants, and access or validate data (Fig. 2). Permissioned blockchains, platforms such as Hyperledger Fabric [17], and Corda [18], require participants to be explicitly granted access to the blockchain network. These blockchains are commonly adopted in enterprise settings,

* Corresponding author.
E-mail addresses: madhavajwalia.it@charusat.ac.in (M. Ajwalia), parth.shah@cug.ac.in (P. Shah).

https://doi.org/10.1016/j.tbench.2025.100251
Received 2 July 2025; Received in revised form 8 December 2025; Accepted 9 December 2025
Available online 11 December 2025

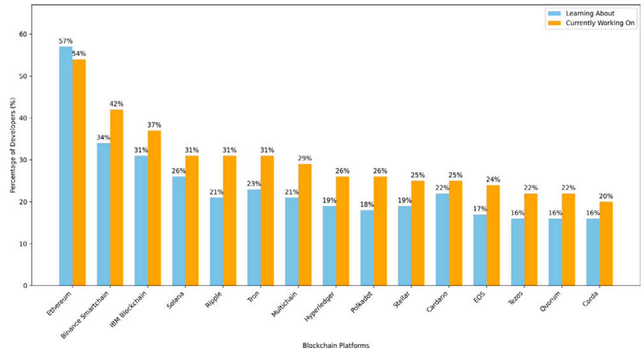


Fig. 1. % of developers working on or learning of the platform [15].

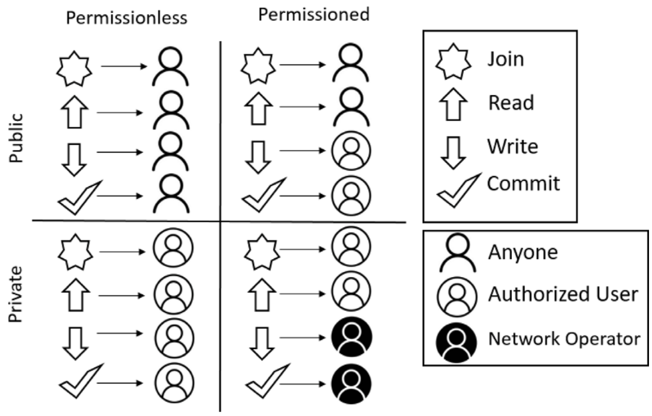


Fig. 2. Join, read, write, and commit access control in public, private, permissionless, and permissioned blockchain categories [16].

offering enhanced privacy, scalability, and control over network governance [19]. In contrast, permissionless blockchains, platforms such as Bitcoin [20] and Ethereum [21], allow anyone to join the blockchain network, participate in transaction validation, and access transaction data.

This difference between permissioned and permissionless blockchains concerns to both public and private networks. Permissionless blockchains permit any member to play the role of validator without constraints, but permissioned blockchains limit this role to known and trusted entities. This dissimilarity substantially affects the performance, security model, and suitability of the network for several applications [22]. Combining public access with authorized involvement in consensus to strike a balance between openness, performance, and control. Some blockchain systems may also show hybrid features [23].

Public blockchains are permissionless in nature, providing unrestricted access, and any participant can join the network, authenticate transactions, and observe the ledger. Such systems are inherently permissionless because no prior permission is needed to join or participate. Well-known examples include Bitcoin and Ethereum, which use consensus protocols such as Proof of Work (PoW) or Proof of Stake (PoS) to ensure security and trust within an entirely open system. While public and permissionless blockchains offer high degrees of transparency and decentralization [24], they typically suffer from limitations in scalability, transaction rates, and energy efficiency as a result of their open and wide participation model [14].

Private blockchains are designed for limited access and are usually controlled by one organization or a consortium. Such networks are naturally permissioned, i.e., only approved participants can read, write, or authenticate transactions. Private blockchains provide more control, quicker transaction processing, better privacy, and are thus increasingly used in enterprise applications like supply chain management, finance, and health care [25,26]. Hyperledger Fabric is a very commonly used example of a private, permissioned blockchain that supports modular architecture, configurable consensus protocols, and fine-grained access control mechanisms [27].

Understanding these various types of blockchain architectures is essential for determining their suitability to specific use cases, when performance, security, and scalability vary per the use case. By analyzing their behavior under varying transactional loads, this study explores the performance of different blockchain architectures. This highlights key architectural design trade-offs and their impact on system performance and efficiency in real-world scenarios.

With the rapid expansion of blockchain use in all sectors, there is an increasing demand for insights into the performance trade-offs for permissioned and permissionless platforms, especially performance under realistic workloads. Although earlier studies have noted architectural and functional differences, a gap remains in quantitative performance comparisons under controlled experimental setups. This research addresses that gap by evaluating and comparing the performance of Hyperledger Fabric and Ethereum under varying transaction workloads while maintaining a constant network size.

Furthermore, the rapid growth of the blockchain market emphasizes the need for comprehensive performance evaluations to measure the suitability of blockchain platforms for specific use cases [28,29]. Understanding the performance characteristics of these platforms is crucial for stakeholders, as it facilitates informed decision-making and fosters the continued advancement and adoption of blockchain technology across various domains. Bitcoin, the initial application of the blockchain, can be assessed through four performance parameters: mining reward, total forks, transaction throughput, and network latency [30]. Due to the potential for integration with blockchain, people may consider its application beyond cryptocurrency. However, efficiency, stability, scalability, and performance should not be overlooked [31]. [32] delves into various performance modeling techniques used to analyze blockchains. These techniques are categorized into analytical modeling, empirical analysis, simulation, and benchmarking. In this work, a benchmarking method is employed to evaluate the performance of permissionless and permissioned types of blockchain using the Hyperledger Caliper tool [33].

If a blockchain network experiences persistent high load, such as high transaction volume or complex smart contract executions, it may begin to experience performance degradation [34]. This degradation can lead to lower system availability, where nodes become slower to respond or fail to respond due to excessive resource consumption. Resource exhaustion, such as CPU and memory overload, can cause node failures or instability, thereby affecting the overall reliability of the blockchain network. Overloaded systems also face node crashes or synchronization issues, which can compromise data consistency across the blockchain network. These issues also limit scalability, as the blockchain network struggles to maintain performance to handle growing demand. These consequences highlight the importance of performance evaluation, particularly when comparing permissioned and permissionless models under dynamic load conditions.

In preview of our findings, the comparative evaluation reveals clear distinctions between the two platforms. Hyperledger Fabric demonstrated consistently higher throughput and lower latency under controlled workloads, along with greater efficiency in resource utilization. Ethereum, by contrast, sustained higher transaction submission rates but at the cost of increased resource consumption. Fabric's resource efficiency and responsiveness make it suitable for permissioned, enterprise-grade use cases such as timebanking applications, whereas Ethereum reflects the trade-offs inherent in public, permissionless environments. These findings help in selecting platforms for real-world deployments.

This paper contributes a structured comparative study of permissioned (Hyperledger Fabric) and permissionless (Ethereum) blockchains under systematically varied transaction workloads. It benchmarks multiple transaction types (open, query, transfer) and examines a broader set of performance indicators beyond conventional throughput and latency, including send rate, transaction success rate, and resource utilization. The experimental design employs standard Dockerized deployments to ensure reproducibility and to isolate the impact of workload intensity on system behavior. The findings provide empirical evidence of the fundamental trade-offs between permissioned and permissionless models, demonstrating Fabric's strengths in throughput, latency, and efficiency, while Ethereum demonstrates higher submission rates, offering practical insights for blockchain platform selection. Through this empirical investigation, this study examines where the performance of permissioned and permissionless blockchain networks excels and limits.

In the subsequent sections of this paper, we examine the existing work, discuss the methodology employed for performance evaluation, present and analyze our findings, and outline paths for future research in the dynamic landscape of blockchain technology.

2. Literature survey

Many recent research works have identified the performance issue of blockchain-based systems as a promising research topic. They have investigated the performance using a practical and theoretical method that shows potential developments in the field. This section discusses the performance research of various application areas related to supply chain [35,36], network service federation [37], cryptocurrency [38], finance [39], trading [40], e-voting [41], IoT [42], and mining [43].

"LogisticChain" [35] introduces a proof of concept model that leverages the Hyperledger Fabric blockchain infrastructure to model shipping logistics workflows. System performance assessment involves manipulating various parameters including client volume, simultaneous transaction processing, and per-second transaction frequency, while monitoring metrics such as latency, send rate, and throughput. Read operations exhibit the lowest latency, while Update operations show the highest latency due to the complex computations and validations involved. The LogisticChain implementation reveals lower latency when utilizing a constant-rate controller compared to a linear-rate controller.

The work [36] presents a prototype of BloodChain based on the private blockchain Hyperledger Fabric. It evaluates the performance and effectiveness of application claims enhanced security, transparency, and traceability. Execution observes requests per second and latency for create, request, and update functions under different test loads. The system can handle a large number of requests and maintain a relatively stable performance. However, there are some fluctuations in latency, mainly when the system is under a high load, from 1000 to 10,000 transactions.

The research work [37] compares Proof of Work (PoW), Proof of Authority (PoA), Practical Byzantine Fault Tolerant (PBFT), and Proof of Stake (PoS) by analyzing their performance on platforms like Ethereum, Tendermint, and Cosmos in the context of Network Service Federation (NSF). It emphasizes the negotiation and implementation aspects of multi-cloud federation, offering insights into the appropriateness of each consensus mechanism for NSF applications. The research evaluates latency, send rate, and throughput to assess the effectiveness of each consensus method. It investigates the performance of Blockchain hosts considering CPU, memory, disk, and network usage. PoA and PBFT mechanisms show lower latency and higher throughput, making them more suitable for NSF applications.

The paper [38] introduces a reputation-based blockchain protocol in Bitcoin networks to assign a controler node for each cluster, which propagates transactions to other clusters. The protocol, called Master Node Based Clustering (MNBC), aims to reduce the propagation delay of transactions by grouping nodes based on their physical internet proximity. The study evaluates the proposed methods through simulation experiments and shows that they can optimize the transaction propagation delay compared to the Bitcoin protocol. Nevertheless, MNBC's efficiency declined when the proportion of compromised nodes rose from 5% to 30%.

[39] Evaluate the performance of the ConsenSys Quorum blockchain platform, permissioned blockchain designed for financial applications. This study aims to analyze the throughput, latency, and scalability. It summarizes that Istanbul Byzantine fault tolerant (IBFT) represents the best choice for Quorum in financial applications. The paper also discusses the challenges faced by permissioned blockchains in achieving high performance and the potential solutions to overcome these challenges.

The performance of the livestock application based on the Hyperledger Iroha blockchain framework is evaluated [40]. The study assesses the framework based on three key parameters: total requests per second (RPS), response times in milliseconds over time, and the number of users in the network over time. The findings suggest that Hyperledger Iroha can effectively support at least 200 participants without errors in the network. It can handle up to 40.6 requests per second, and the response times are rapid, typically less than a second.

[41] develops a private blockchain infrastructure designed for democratic voting systems, utilizing blockchain network for data storage and processing operations. Smart contracts are implemented to handle transaction execution within the system. The authors conduct a comparative performance evaluation examining two widely-used Ethereum client implementations, Geth and Parity, analyzing their capabilities across throughput, latency, and scalability parameters. On average, transactions are 91% faster in the Parity client compared to the Geth client.

[42] compares the performance of their solution with existing access management solutions in IoT. The paper conducts realistic experiments to evaluate the performance of the proposed system in terms of latency, throughput, and scalability. It is observed that, in the case of only single management hub, the proposed solution achieves better performance than the optimized centralized IoT systems. The proposed implementation offers significant scalable advantages over traditional scenarios in the case of multiple management hubs.

"MobiChain" [43] experiments were performed on a mobile node. It evaluates the performance of the proposed model in terms of computation time, energy consumption, and memory utilization for chain verification. It observes that the chain verification process is executed faster and consumes less energy when transactions are grouped in a block. Moreover, the execution time reduces insignificantly as the number of threads increases.

The reviewed literature (Table 1) shows that performance evaluation across blockchain applications is essential. Performance evaluation trend shifts from theoretical analysis to practical implementation studies. That emphasis on domain-specific performance metrics and real-world usability factors. Scalability remains the primary challenge across all application domain. Energy efficiency and resource constraints are concerns, specifically for mobile and IoT applications. However, most studies focus on individual platforms rather than comprehensive cross-architecture comparisons. This study addresses these limitations by systematically comparing the Hyperledger Fabric and Ethereum platforms under varying workload scenarios, offering valuable insights for informed deployment decisions in both enterprise and public contexts.

3. Methodology

This study includes configuring Hyperledger Fabric as a permissioned blockchain network and Ethereum as a permissionless blockchain network. It implements chaincode and smart contracts and runs several predefined test scenarios with the benchmarking tool Hyperledger Caliper. That measuring various performance metrics and analyzing the results. To evaluate performance, the tool runs a comprehensive set of test scenarios that include a variety of workloads, transactions, and network sizes. For the study, workload transaction per second (TPS) is systematically varied from 100 to 1000 while keeping the network size constant.

Table 1
Comparison of performance evaluation studies in blockchain applications.

Paper	Study Focus	Network/Platform	Consensus Mechanism	Tools Used	Key Performance Metrics	Key Observations	Limitations
[35]	Analyze performance of a blockchain-based system for maritime logistics	Hyperledger Fabric	Raft	Hyperledger Caliper	Throughput, latency, scalability, transaction processing speed	Lower latency with fixed-rate controllers than linear-rate controllers	Simulation-only results; lacks of real-world deployment data; scalability and interoperability issues
[36]	Scalability performance of permissioned network in Healthcare	Hyperledger Fabric	Not Mentioned	Hyperledger Caliper	Data integrity, traceability, system availability, user adoption rates	Stable performance under high load; fluctuations in latency occur, particularly under high load from 1000 to 10,000 transactions	Conceptual; lacks full metrics and implementation; privacy compliance concerns
[37]	Evaluation of blockchain consensus mechanisms for federated network services	Ethereum, Tendermint, Cosmos	Proof of Work, Proof of Authority, PBFT, Proof of Stake	Not Mentioned	Consensus latency, energy use, fault tolerance, network overhead	PoW and PBFT show lower latency and higher throughput making them more suitable for NSF applications	Limited scalability analysis; limited consensus algorithms tested
[38]	Investigate the performance and security of a reputation-based blockchain in cryptocurrency network	Bitcoin network	Rapid-Chain/MNBC	Bitcoin simulator, Metis graph partition toolkit	Security metrics, reputation scoring, network resilience, transaction speed	Proposed Master Node based Clustering (MNBC) protocol offers improvement in information propagation delay compared to Bitcoin protocol. However, MNBC performance decreased as malicious nodes increased from 5% to 30%	Simulation focus; Bitcoin-specific; centralization and vulnerability concerns
[39]	Performance evaluation of Permission blockchain in Financial Services	ConsenSys Quorum	Raft, Clique Proof of Authority, Istanbul BFT	Hyperledger Caliper	Transaction throughput, privacy, compliance, cost-efficiency	Istanbul Byzantine fault-tolerant (IBFT) represents the best choice for Quorum in financial applications	Single-platform focus; lacks cross-platform comparison
[40]	Evaluate performance of permissioned network in livestock management and trading platform	Hyperledger Iroha	Byzantine Fault Tolerant	Not Mentioned	Transaction processing, data provenance, system reliability, adoption	Supports 200+ participants, 40.6 tx/sec throughput	Small-scale testing only; lacks stress/fault tolerance
[41]	Benchmark public blockchain in E-Voting scenario	Ethereum (Geth and Parity)	Proof of Work	Not Mentioned	Vote speed, system security, voter privacy, electoral integrity	Parity is 91% faster than Geth for transactions	Ethereum-only; limited security analysis; privacy vs. transparency trade-offs
[42]	Analyze performance of Blockchain-based access control for IoT devices	Ethereum	Not Clear	CoapBench	Device auth speed, scalability, energy efficiency	Scalable advantage in multi-Hub scenario; single-Hub underperforms centralized systems	Private blockchain only; no constrained device tests; energy/resource limits
[43]	Performance analysis of Blockchain in Mobile device for mining	Bitcoin network	Proof of Work	VideoOptimizer program	Battery consumption, processing efficiency, network usage, user experience	Grouped transactions improve speed and reduce energy use	Energy consumption and bandwidth not fully evaluated; mobile device variability

3.1. Experimental design description

The experimental design follows a structured workflow where business logic is developed independently for each platform before performance evaluation. The process consists of four main phases: development of business logic, configuration for benchmarking, benchmark execution and measurement, and result analysis.

Development of Business Logic: Application-level business logic was developed independently for each platform. For Hyperledger Fabric, chaincode was implemented in the Node.js, while for Ethereum, smart contracts were written in Solidity. These implementations covered three core functions (Open, Query, Transfer) and were validated through unit testing prior to benchmarking. The code was deployed on private test networks for each platform, each hosted in an isolated Dockerized environment to ensure fairness and control across experiments.

Configuration for Benchmarking: Hyperledger Caliper served as the benchmarking framework. It was configured using an adapter and network files referencing the deployed business logic, specifying workload parameters, transaction types, and network details. Caliper does not develop or execute code internally; instead, it interfaces with predeployed smart contracts or chaincode via standardized blockchain APIs. This ensures that benchmarking reflects actual system behavior rather than simulated execution.

Benchmark Execution and Measurement: Once configured, Caliper initiated benchmarking by generating workloads against the deployed networks. Each experimental run followed a systematic sequence: initially, blockchain networks are deployed using validated business logic; Caliper then connects to each network via standardized APIs. Once connected, workload modules initiate transaction patterns defined by experimental parameters, prompting the execution of business logic on the networks. Throughout this process, Caliper captures transaction-level metrics through analysis of API interactions and timestamps. Parallel system-level monitoring captures resource utilization data, providing a comprehensive view of system performance. Each scenario was executed three times, and Caliper aggregates and stores all results for statistical and comparative analysis. Scalability was assessed by varying transaction rates across the workload range, while keeping the network size fixed to isolate workload effects.

Result Analysis and Reproducibility: Caliper outputs structured JSON and performance reports, which were analyzed for performance

Table 2
Test environment for performance evaluation of blockchain networks.

Component	Platform/Network	Version
Benchmarking Platform	Hyperledger Caliper	v0.5.0
Blockchain Network	Ethereum Hyperledger Fabric	1.0 2.5.0
Business Logic	Smart Contract with Solidity Chaincode with Node.js	0.8.23 16.13.1, 12.22.9

trends and platform comparisons. Control variables, including hardware specifications, software versions, Docker container settings, and environment parameters, remained constant across all experiments to ensure comparability. Detailed documentation of all configuration parameters, software versions, and experimental steps enables reproducibility.

3.2. Experimental setup

Table 2 shows the test environment prepared for this performance study. Hyperledger Caliper v0.4.2 was used as a benchmarking tool. Chaincode, smart contract development, and execution were carried out using the programming languages Go, Solidity, and JavaScript. Visual Studio Code (VSCode) worked as a development environment for code creation and debugging. For the CPU configuration, we used an Intel Core i5-1135G7 CPU operating at 2.40 GHz for the Ubuntu v22.04.3 LTS system.

We used Hyperledger Fabric v2.5.0 and Ethereum as permissioned and permissionless networks, respectively, and deployed them with Docker engine versions 20.10.22 and 24.0.7. We used Docker Compose versions 2.15.1 and 1.29.2 to manage containerized blockchain environments.

3.3. Benchmarking tool

Hyperledger Caliper is a benchmarking tool designed explicitly to evaluate the performance of blockchain networks. It provides a framework for executing different scenarios, measuring various performance metrics, and analyzing the results. Caliper's general framework [44] is used to run benchmarks against various blockchain frameworks. That generates a workload for a particular system under test (SUT) and keeps track of how it responds all the time. Finally, a report will be produced based on the observed SUT responses. This simplistic view is depicted in Fig. 3.

Caliper needs many inputs to execute a benchmark, regardless of the selected SUT. The following subsets provide a quick summary of these inputs.

- Benchmark configuration file: This file specifies the benchmarking process's parameters, including the blockchain platform, number of participating nodes, consensus algorithms, and other test-specific parameters.
- Network configuration file: This file is dedicated to the blockchain network setup and contains information about network nodes, their addresses, and other relevant network-related details.
- Workload modules: It is the brain of the benchmark. It customizes the benchmarking workload, allowing users to tailor test scenarios to their specific use cases and evaluate the blockchain platform's performance under realistic conditions.
- Benchmark artifacts: Artifact or result generated by the benchmarking process, encapsulating key performance metrics, transaction details, and other relevant data collected during the experiment.

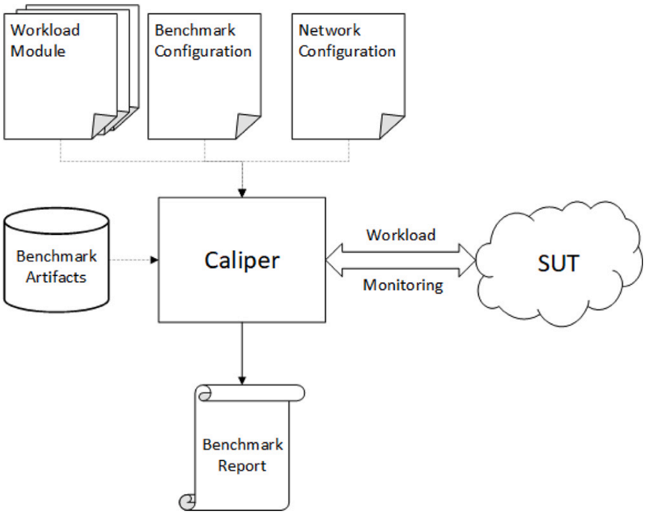


Fig. 3. Architecture of Hyperledger Caliper [44].

- Caliper core: This central component is responsible for the workload module and orchestrating the benchmarking process. It includes the core logic for test execution, result aggregation, and communication with blockchain networks.
- Benchmark Process:

- Initialize the Caliper core, load the specified workload module, and connect to the target blockchain network using the provided configuration.
- Caliper generates and submits transactions to the blockchain network based on the workload module.
- The network monitor records relevant metrics, and the result writers store the collected data for later analysis.
- The Caliper produces detailed reports and analysis, allowing users to evaluate the performance of the tested blockchain platform.

Performance metrics such as send rate, throughput, latency, resource utilization, and transaction success rates can be measured and analyzed. These metrics assess the blockchain network's scalability, efficiency, and reliability. Many research articles present theoretical insights, but may lack empirical evidence or real-world case studies to support their findings.

3.4. Parameter selection rationale

In the context of performance evaluation for blockchain networks, we used several key metrics to assess the network's efficiency and effectiveness. The performance metrics in this study were carefully selected to align with both prior blockchain performance evaluation and benchmarking standards and the specific objectives of this research. Our chosen metrics align with established performance modeling techniques categorized into analytical modeling, empirical analysis, simulation, and benchmarking approaches [32]. Previous systematic surveys [28,29,45–47] consistently identify throughput, latency, and resource utilization as the core indicators of blockchain performance, as they directly reflect scalability, efficiency, and user experience. [28] identifies throughput, latency, and resource efficiency as the most critical indicators of blockchain usability, while [29] emphasizes the need to capture both user-perceived and system-level performance characteristics. In parallel, [45] classifies throughput, latency, and resource consumption as the fundamental evaluation criteria applied across multiple approaches, and [46] highlights the necessity of analyzing consensus through transaction latency and success rate. [47] reinforces this view, noting that among the various metrics, send rate, throughput,

Table 3
Symbolic definitions for blockchain performance metrics.

Symbol	Meaning
Tx	Transactions
Tx_p	Successfully processed transactions
T_t	Transaction processing duration
T_{init}	Transaction initiation time
T_{conf}	Transaction confirmation time
R_u	Actual resource usage
R_{total}	Total available resources

latency, CPU utilization, memory usage, energy efficiency, and security are consistently highlighted as the most important and widely adopted empirical performance evaluation parameters for blockchain systems.

Other vital parameters, energy efficiency and security metrics, were excluded from this baseline study due to scope constraints. Energy efficiency evaluation requires specialized power measurement infrastructure and extended observation periods that would significantly complicate the controlled experimental design. Security assessment requires separate methodologies, including penetration testing and cryptographic analysis, representing a distinct research domain beyond the performance benchmarking scope. This study establishes foundational performance baselines that inform subsequent specialized energy and security evaluations.

3.5. Performance parameters

These metrics collectively address usability, adoption, and sustainability of blockchain technologies and are therefore central to any comprehensive performance evaluation. These metrics provide a solid and comprehensive foundation for assessing the performance trade-offs between permissioned and permissionless blockchain platforms across different workload scenarios. Their combined result also proves the network's scalability, efficiency, and practical deployment considerations. Different blockchain networks may prioritize these metrics over others based on their use cases and requirements. Table 3 shows the meaning of the symbols utilized in the definitions of the blockchain performance metrics.

Send Rate: That represents the number of transactions sent in the network by the participant node per unit of time. It measures the rate at which transactions are submitted to the blockchain network.

$$\text{Send Rate} = \frac{\text{Total number of transactions sent}}{\text{Time duration}} = \frac{\sum Tx}{T_t} \tag{1}$$

Throughput: It measures the processing capacity of the blockchain network, i.e., the number of transactions successfully processed per unit of time. It reflects the network's ability to handle a certain volume of transactions within a given timeframe.

$$\begin{aligned}\text{Throughput} &= \frac{\text{Total number of transactions successfully processed}}{\text{Total time period}}\\ &= \frac{\sum Tx_p}{\sum T_t}\end{aligned} \tag{2}$$

Latency: It refers to the time delay between initiating a transaction and its final confirmation or inclusion in the blockchain network. It measures the time taken for transactions to be propagated, validated, and included in a blockchain.

$$\begin{aligned}\text{Latency} &= \text{Time taken for transaction confirmation}\\ &\quad - \text{Time of transaction initiation}\\ &= T_{\text{conf}} - T_{\text{init}}\end{aligned} \tag{3}$$

Success Rate: It indicates the proportion of transactions or operations that are executed successfully without errors or failures.

$$\begin{aligned}\text{Success Rate} &= \frac{\text{Total number of transactions successfully processed}}{\text{Total number of transactions}}\\ &= \frac{\sum Tx_p}{\sum Tx}\end{aligned} \tag{4}$$

Resource Utilization: Resource utilization assesses the efficiency of the blockchain network's resources, such as computing power and memory usage, in processing and validating transactions. It measures how effectively the network utilizes its resources to maintain its operation.

$$\text{Resource Utilization} = \frac{\text{Actual resource usage}}{\text{Total available resources}} = \frac{R_u}{R_{\text{total}}} \tag{5}$$

3.6. Benchmarking operations

In the context of performance evaluation, the 'open','query', and 'transfer' functions are essential tasks carried out during benchmarking scenarios to evaluate blockchain network performance [48]. Every function has a specific role: 'open' initializes and sets up the blockchain network, 'query' collects data from the network, and 'transfer' carries out transactions to move assets or information. By incorporating these functions into benchmarking scenarios, Caliper allows for the measurement and analysis of various performance metrics. This supports a thorough evaluation of the blockchain network's efficiency, scalability, and reliability across varying workloads.

- Open function: The open function sets the blockchain network up for benchmarking by initializing it. This could involve tasks such as setting up or creating accounts or participants, deploying smart contracts, configuring network parameters, loading initial data onto the blockchain, and so forth. The function initializes the starting state of the blockchain network before running any benchmark transactions.
- Query Function: The query function fetches data from the blockchain network. This could involve checking transaction records, smart contract status, account balances, or other relevant information stored on the blockchain. Evaluating the effectiveness of data retrieval systems and the responsiveness of the blockchain network to read queries depends on query operations.
- Transfer Function: Transfer function is essential for the transfer of value or assets in the blockchain network. It consists of submitting transactions into the network to transfer tokens, assets, or data between smart contracts or accounts. Transfer operations are essential for assessing transaction-processing capacity.

With these functions, Caliper enables users to simulate real-world scenarios by benchmarking and evaluating their performance under various workloads and conditions across various blockchain networks.

3.7. Platform selection rationale

Various surveys and comparative analyses highlight Ethereum and Hyperledger Fabric as leading examples of permissionless and permissioned blockchains, respectively, due to their consensus algorithms and performance capabilities. These two systems are the most widely adopted and documented in both academic research and industry deployments, making them ideal for comparative studies.

Examples of public permissionless blockchains include Bitcoin, Ethereum, Litecoin, Cardano, and Polkadot, all of which enable open participation. By contrast, permissioned blockchains such as Hyperledger Fabric, R3 Corda, Quorum, Ripple, and IBM Blockchain emphasize controlled access, enhanced privacy, and enterprise-grade functionality. Many permissionless platforms, including Binance Smart Chain

(BSC), Polygon, Avalanche, Fantom, and Tron, extend Ethereum-like designs, leveraging the Ethereum Virtual Machine (EVM) or Ethereum-inspired smart contracts while maintaining public participation. Similarly, permissioned platforms like Quorum, Corda, Hyperledger Sawtooth, and IBM Blockchain adopt Hyperledger Fabric's modular, privacy-oriented, and enterprise-driven design principles. In addition, several leading platforms, including Ethereum 2.0, Cardano, Polkadot, Tezos, and Algorand, are transitioning toward or have already adopted PoS. Research [32] focused on performance modeling of a public permissionless blockchain. It suggests that, PoW and PoS are widely used and the most popular consensus algorithms in this category. Although Algorand has emerged as a promising alternative, Ethereum represents this class majorly. A detailed study [49] evaluates the performance of consensus algorithms and blockchain platforms using twelve assessment indexes. It also discusses eight common blockchain platforms, including Bitcoin, Ethereum, Hyperledger Fabric, Quorum, Ripple, Corda, EOS, and IOTA, analyzed by these indexes to highlight their distinct strengths. Comparative analyses of permissioned blockchain, such as Cosmos, Hyperledger Fabric, Quorum, and XRPL, demonstrate Fabric's superiority in consensus, smart contracts, custom tokens, privacy, latency, and throughput, despite limitations in interoperability [50]. Researchers [51] provide a comparative study of permissioned frameworks with regard to the community activities, performance, scalability, privacy and adoption criteria. Study shows fabric is promising. In the area of blockchain cloud integration, while providers such as AWS and Google Cloud offer Blockchain-as-a-Service (BaaS), Fabric remains the preferred choice for private deployments [52]. Survey [53] investigated 12 most popular blockchain platforms and elaborated six platforms that are widely applied in finance. This study observes that the proportion of companies using Ethereum and Hyperledger Fabric for application development is 24% and 38%, respectively, among the top 100 companies. It demonstrates their market dominance in financial use cases. This highlights why Ethereum and Hyperledger Fabric are regarded as foundational benchmarks for permissionless and permissioned blockchains, respectively, since studying them sufficiently captures the core architectural and performance traits of each category. Thus, focusing comparative performance studies on Ethereum and Hyperledger Fabric is sufficient to capture the essential dynamics of these two prominent blockchain domains. Including additional platforms for a performance comparison study introduces redundant comparisons, as many permissionless or permissioned blockchains often derive from or closely align with these core models. [49] highlights that the optimal blockchain platform depends on the particular application and desired performance criteria such as throughput, scalability, energy consumption, cost, and security. So the final selection of a framework for a specific use case is always a trade-off.

3.8. Methodological contribution

While Hyperledger Caliper serves as the underlying benchmarking tool, this study goes beyond its default usage by proposing a structured evaluation framework that introduces methodological innovations to enhance both analytical depth and practical relevance. Rather than relying solely on the tool's default configurations, the study introduces distinct experimental controls and performance metrics, positioning the work beyond generic benchmarking practices (See Fig. 4).

This study intentionally employs Hyperledger Caliper to ensure consistency and comparability with prior research. Unlike conventional applications of Caliper, we extend its application through a structured framework that incorporates systematic workload variation, multi-function benchmarking, extended performance metrics, and standardized deployments. While the development of entirely new benchmarking tools lies beyond the scope of this work, the framework established here offers a reliable baseline and can serve as a foundation for future research exploring more adaptive or context-specific evaluation methodologies.

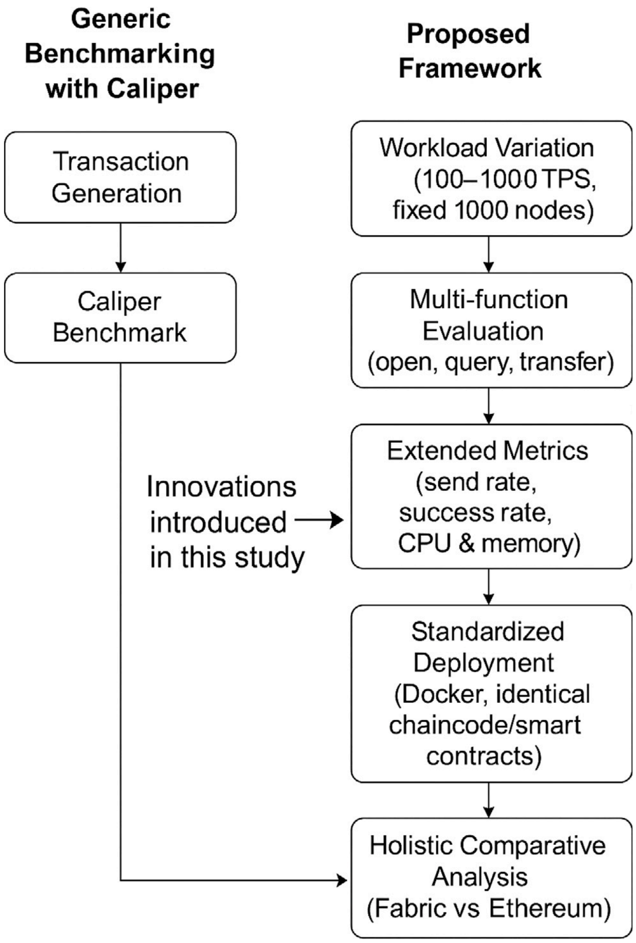

Generic Benchmarking with Caliper
Proposed Framework
Transaction Generation
Caliper Benchmark
Workload Variation (100–1000 TPS, fixed 1000 nodes)
Multi-function Evaluation (open, query, transfer)
Innovations introduced in this study
Extended Metrics (send rate, success rate, CPU & memory)
Standardized Deployment (Docker, identical chaincode/smart contracts)
Holistic Comparative Analysis (Fabric vs Ethereum)

Fig. 4. Generic caliper vs proposed framework.

First, transaction workloads were systematically varied between 100 and 1000 TPS while maintaining a fixed network size. This design isolates the impact of workload intensity without conflating it with network scaling, a limitation observed in prior benchmarking studies. Second, the research develops a 3-dimensional analysis approach examining platform performance across distinct functional operations (Open, Query, Transfer) rather than aggregate metrics. This function-specific methodology reveals performance characteristics that aggregate evaluations obscure, providing granular insights for application-specific deployment decisions. Finally, both Ethereum and Hyperledger Fabric are deployed in standardized Dockerized environments with equivalent business logic implemented in smart contracts and chaincode. This ensures fairness, reproducibility, and guarantees that observed differences arise from platform characteristics rather than deployment variations.

Collectively, these contributions move the study beyond a routine use of existing tools and establish a replicable evaluation framework for blockchain performance assessment. The approach not only strengthens the validity of comparative analysis between permissioned and permissionless systems but also provides a methodological template for future benchmarking studies in the specific use case.

4. Result discussion

The result illustrates how Ethereum and Hyperledger Fabric (HLF) performed when we tested different functions like 'Open', 'Query', and 'Transfer'. For this result, the transaction load gradually increased from 100 to 1000 transactions per second (TPS) while keeping the network size constant to see how each platform would handle the workload.

Table 4
Performance results for ‘open’ function.

TPS	Send Rate (TPS)		Throughput (TPS)		Max Latency (s)		CPU Utilization (%)		Memory Usage (MB)	
	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF
100	100.2	121.4	31.1	120.5	22.77	1.96	1.96125	7.56750	1239.04	505.08
200	200.4	145.9	31.1	140.0	27.50	1.96	2.16250	7.99500	1505.28	520.84
300	300.9	150.5	31.9	149.4	28.28	1.92	0.17625	8.28125	1546.24	490.57
400	400.8	175.2	31.2	155.7	29.76	1.99	0.23375	6.47375	1525.76	516.42
500	501.0	154.6	31.2	148.2	30.18	1.95	0.04875	7.60125	1525.76	517.06
600	600.2	158.8	31.9	153.2	32.89	1.97	0.45607	8.04947	1520.73	505.85
700	700.2	167.9	31.4	155.3	34.45	2.05	0.57142	9.67462	1555.92	515.79
800	800.3	163.8	31.5	157.4	36.10	2.22	0.79558	9.69863	1555.22	516.57
900	900.4	157.7	31.1	159.1	38.07	2.31	0.39635	9.64107	1550.10	511.98
1000	1000.2	162.7	31.3	160.4	39.86	2.35	0.3383	10.15398	1545.31	518.39

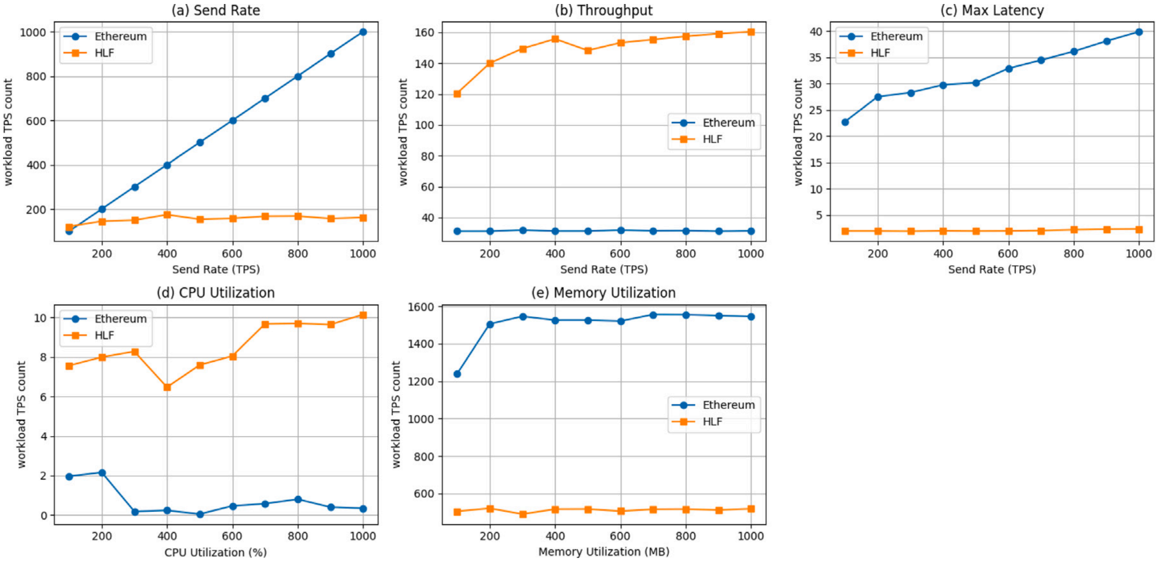


Fig. 5. Performance results for ‘open’ function with the varying workload from 100 to 1000.

Metrics include send rate, throughput, maximum latency, CPU utilization, and memory usage. This work provides valuable insights into platform activity under various transaction workloads, offering a subtle understanding of efficiency and scalability.

As shown in Fig. 5 and Table 4, the result for the ‘open’ function, reveal the following trends: The open function highlights key differences in how Ethereum and Hyperledger Fabric process write-intensive workloads. Across the full workload spectrum (100–1000 TPS), Fabric consistently maintains throughput close to the incoming send rate, achieving up to 160 TPS. In contrast, Ethereum’s throughput is around 31 TPS irrespective of higher send rates. These numbers reflect Ethereum’s consensus constraints that restrict how many ‘open’ transactions can be committed per unit time. Ethereum’s maximum latency increases linearly with workload, exceeding 39 s at 1000 TPS. This outcome is typical in a PoW-based system, where transaction congestion occurs in the mempool when the volume of submitted transactions exceeds the network’s processing capacity. Fabric, in contrast, maintains latency within 2–3 s even at peak load, as its modular consensus (ordering service with endorsement policies) is optimized for rapid block finality in permissioned settings. Resource utilization aligns with these patterns. Ethereum exhibits a steadily rising CPU and memory usage as workload increases, reflecting the computational burden of block validation and state updates under PoW. Fabric, while consuming CPU and memory at lower levels, shows slightly increased CPU usage at higher TPS as the ordering service manages more frequent block commits. Overall, the open function analysis highlights Fabric’s superior suitability for enterprise contexts requiring reliable throughput and bounded latency, while Eth-ereum demonstrates the scalability limits imposed by its permissionless consensus.

Fig. 6 and Table 5 illustrate several observable trends regarding the ‘query’ function, summarized as follows: The query operation demonstrates a significantly different pattern. Here, both Ethereum and Fabric achieve throughput nearly equal to the send rate across all workloads up to 1000 TPS. Queries are read-only operations that do not alter the blockchain state, meaning they bypass the heavy consensus bottlenecks associated with write transactions. As a result, Ethereum handles queries at near-linear scalability, achieving 1000 TPS throughput when pushed to that workload. Fabric mirrors this performance, consistently matching the query submission rate. Latency for queries remains negligible on both platforms, with Ethereum maintaining under 0.03 s and Fabric slightly higher but still well under 0.1 s across all load levels. This clear difference with the open function highlights the architectural difference between read and write operations in blockchain. Since queries mainly involve state lookups, consensus delays and block validation overheads are avoided. CPU and memory utilization reflect the lightweight nature of queries. Ethereum’s CPU consumption is somewhat higher than Fabric’s, which aligns with its execution environment (EVM-based contract logic and PoW validation, even for read calls). Fabric remains comparatively more efficient, with CPU usage below 1% and a stable memory usage across workloads. The query analysis confirms that both platforms can scale read-heavy workloads effectively, but Fabric achieves this with lower resource overhead.

As shown in Fig. 7 and Table 6, the behavior of the ‘transfer’ function exhibits the following trends: The transfer function presents the most demanding workload by combining state updates with consensus enforcement. Fabric maintains throughput above 130 TPS across all load levels, reaching at 160 TPS, whereas Ethereum again saturates at 38 TPS. This limit in Ethereum reflects the same architectural

Table 5
Performance results for 'query' function.

TPS	Send Rate (TPS)		Throughput (TPS)		Max Latency (s)		CPU Utilization (%)		Memory Usage (MB)	
	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF
100	100.1	699.3	100.1	697.8	0.01	0.09	0.55500	0.31125	1249.28	486.28
200	200.3	708.2	200.2	706.7	0.03	0.08	0.47875	0.42125	1361.92	491.44
300	300.4	686.8	300.3	685.4	0.02	0.15	0.19125	0.42750	1372.16	481.57
400	401.0	714.8	400.8	713.8	0.01	0.05	0.25125	0.22000	1341.44	503.12
500	501.0	705.7	501.0	703.2	0.02	0.05	0.15000	0.40875	1341.44	506.56
600	600.2	705.6	599.3	704.6	0.018	0.06	0.22	0.60485	1347.15	500.00
700	700.2	706.5	700.0	707.2	0.019	0.08	0.234	0.79843	1346.9	494.6
800	800.3	711.2	799.5	713.5	0.014	0.07	0.3421	0.7649	1347.1	497.31
900	900.4	710.3	898.9	713.2	0.014	0.08	0.352	0.81362	1348.7	508.32
1000	1000.2	711.1	999.6	708.6	0.024	0.06	0.3403	0.83039	1349.0	508.0

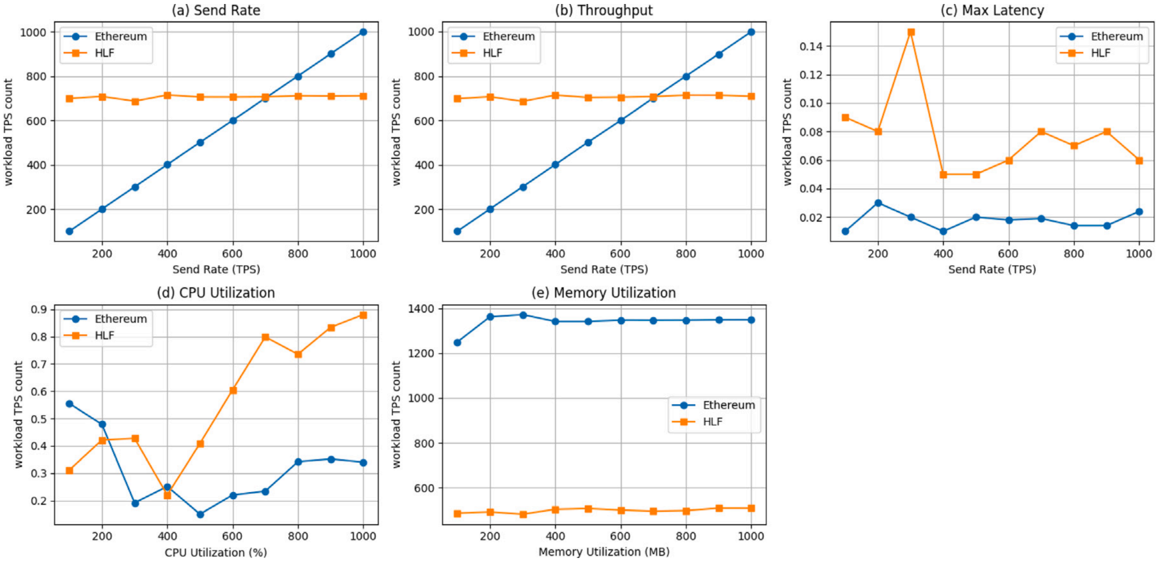


Fig. 6. Performance results for 'query' function with the varying workload from 100 to 1000.

bottlenecks seen in the open function: block gas limits and the sequential nature of PoW consensus. Fabric's higher throughput results from its efficient ordering and endorsement process. This allows for parallelism and trust assumptions in a controlled, permissioned environment. Latency patterns further highlight this gap. Ethereum's maximum latency increases sharply with increasing load, from 17 s at 100 TPS to nearly 32 s at 1000 TPS. This Growth reflects the backlog of pending transactions in the mempool during sustained load. Fabric maintains latency under 3.5 s even at peak workloads, reinforcing its strength in providing predictable responsiveness. Resource utilization trends are consistent with the throughput findings. Ethereum's CPU usage steadily increases beyond 40% as the system struggles with backlogged transactions and block production overhead. Memory consumption also increases, exceeding 1600 MB at the highest workload. Fabric, in contrast, consumes minimal additional resources, with CPU usage increasing modestly but still below 7% and memory remaining near 540 MB. These results show Fabric's advantage for financial or transactional systems that need reliable settlement speed and efficiency.

5. Conclusion and future work

This study provides a detailed performance comparison of permissioned and permissionless blockchains under workloads ranging from 100 to 1000 TPS. Ethereum consistently achieves a higher send rate, reflecting its suitability for scenarios where transaction initiation speed and public participation are essential. Hyperledger Fabric offers, on average, 3.5–4.5 times higher throughput and 10–12 times lower latency across tested functions, demonstrating its efficiency for enterprise-grade workloads. Resource utilization further highlights the trade-offs: Fabric consumes 2.5–3 times less memory than Ethereum but requires higher CPU usage for some operations, such as open transactions. Ethereum, on the other hand, often shows lower CPU demand but at the cost of significantly higher memory consumption and reduced confirmed throughput. Both systems scale with increasing workloads, showing robustness across the tested range. Together, the analyses of open, query, and transfer functions reveal consistent architectural trade-offs. Ethereum's permissionless design emphasizes openness and security but sacrifices throughput and latency, limiting its suitability for high-volume enterprise use. Hyperledger Fabric, tailored for permissioned environments, achieves higher throughput, bounded latency, and more efficient resource use. Queries scale well in both systems, though Fabric remains more efficient.

The results of this study also have direct practical implications for the planned application. Since such systems demand predictable throughput, low latency, and efficient resource utilization to support frequent credits, debits, and balance queries, Hyperledger Fabric emerges as the more suitable platform. Its ability to sustain consistent performance up to 1000 TPS ensures reliability for community-scale deployments, where transaction volumes are moderate but require high integrity. Ethereum, while offering openness and higher transaction submission rates, introduces latency and resource inefficiencies that

Table 6
Performance results for 'transfer' function.

TPS	Send Rate (TPS)		Throughput (TPS)		Max Latency (s)		CPU Utilization (%)		Memory Usage (MB)	
	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF	Ethereum	HLF
100	100.2	160.4	37.6	158.5	17.41	1.94	0.78375	4.69875	1505.28	519.98
200	200.4	173.1	36.7	165.1	22.41	1.94	1.08125	3.7875	1576.96	540.44
300	300.4	190.5	40.7	131.9	21.52	2.49	1.80875	2.905	1587.20	514.17
400	401.0	202.5	39.4	139.3	23.07	2.25	0.97625	1.495	1556.48	541.42
500	501.3	200.3	38.4	139.6	24.19	2.22	0.04875	0.610	1556.48	540.46
600	600.2	200.4	38.6	141.5	26.27	2.84	0.45607	4.71	1553.30	539.85
700	700.2	202.5	37.9	140.3	27.16	3.05	0.57142	5.21	1555.92	540.79
800	800.3	201.6	38.7	139.8	28.47	3.00	0.79558	5.617	1561.61	541.57
900	900.4	203.3	38.2	139.4	30.35	3.26	0.39635	6.01	1590.00	540.98
1000	1000.2	201.9	38.4	137.1	31.97	2.94	0.3383	6.401	1602.31	541.60

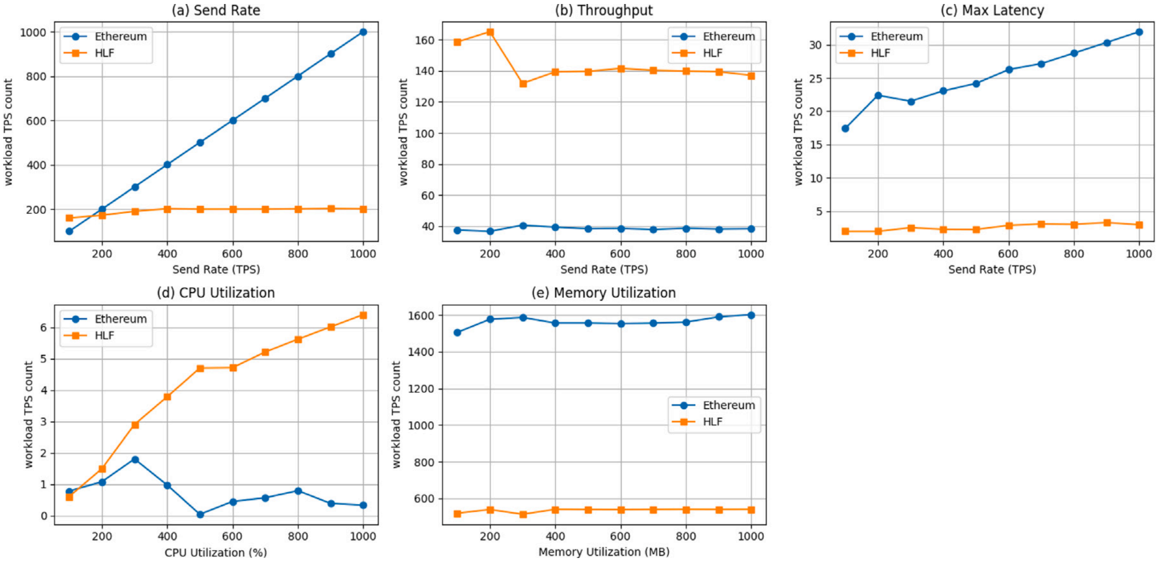


Fig. 7. Performance results for 'transfer' function with the varying workload from 100 to 1000.

could hinder responsiveness in a timebanking context. These insights provide a strong foundation for selecting Fabric as the underlying platform for the forthcoming application.

5.1. Future work

While this study evaluated blockchain performance by varying transaction workloads under a fixed network size, future research could extend the analysis by exploring how network size impacts system behavior. As the number of participating nodes increases, consensus mechanisms experience additional communication overhead, which may affect performance differently than transaction load. These investigations may reveal performance patterns that cannot be observed simply by adjusting the system's workload. They can also help explain real-life situations where the performance of platforms changes depending on the number of users involved, not just the system's overall activity.

Given the rapid evolution of blockchain technologies, this study may not cover every emerging enhancement or innovation in platform design and interaction. Future work should include new consensus protocols, updated framework versions, and wider deployment configurations to stay relevant. Building on these findings, we plan to adopt Hyperledger Fabric and move toward developing and testing a real-world application, validating the platform's performance and usability in a live deployment scenario.

CRediT authorship contribution statement

Madhav Ajwalia: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Parth Shah:** Writing – review & editing, Writing – original draft, Supervision.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they do not have any conflict of interest.

References

[1] M. Javaid, A. Haleem, R.P. Singh, R. Suman, S. Khan, A review of blockchain technology applications for financial services, BenchCouncil Trans. Benchmarks, Stand. Eval. 2 (3) (2022) 100073.

[2] J. Clavin, S. Duan, H. Zhang, V.P. Janeja, K.P. Joshi, Y. Yesha, L.C. Erickson, J.D. Li, Blockchains for government: use cases and challenges, Digit. Gov.: Res. Pr. 1 (3) (2020) 1–21.

[3] M. Ajwalia, K. Mer, R. Bhatia, P. Shah, P. Prajapati, Security and challenges of blockchain-based IoT use cases, in: International Conference on ICT for Sustainable Development, Springer, 2024, pp. 91–103.

[4] S. Cihan, N. Yılmaz, A. Ozsoy, O.D. Beyan, A systematic review of the blockchain application in healthcare research domain: toward a unified conceptual model, Med. Biol. Eng. Comput. (2025) 1–24.

[5] C. Zheng, X. Peng, Z. Wang, T. Ma, J. Lu, L. Chen, L. Dong, L. Wang, X. Cui, Z. Shen, A review on blockchain applications in operational technology for food and agriculture critical infrastructure, Foods 14 (2) (2025) 251.

[6] N. Kumar, K. Kumar, A. Aeron, F. Verre, Blockchain technology in supply chain management: Innovations, applications, and challenges, Telemat. Informatics Rep. (2025) 100204.

[7] T. Sivaram, et al., Recent developments and challenges using blockchain techniques for peer-to-peer energy trading: A review, Results Eng. (2024) 103666.

[8] X. Wang, M. Younas, Y. Jiang, M. Imran, N. Almusharraf, Transforming education through blockchain: A systematic review of applications, projects, and challenges, IEEE Access (2025).

[9] G. Piccardo, L. Conti, A. Martino, Blockchain technology and its potential to benefit public services provision: A short survey, Futur. Internet 16 (8) (2024) 290.

[10] N.S. Sizan, D. Dey, M.A. Layek, M.A. Uddin, E.-N. Huh, Evaluating blockchain platforms for iot applications in industry 5.0: A comprehensive review, Blockchain: Res. Appl. (2025) 100276.

[11] K. Zīle, R. Strazdiņa, Blockchain use cases and their feasibility, Appl. Comput. Syst. 23 (1) (2018) 12–20.

[12] D. B. Rawat, V. Chaudhary, R. Doku, Blockchain technology: Emerging applications and use cases for secure and trustworthy smart systems, J. Cybersecur. Priv. 1 (1) (2020) 4–18.

[13] J. Yu, X. Li, Y. Guo, A secure and verifiable blockchain-based framework for personal data validation, Computers 13 (9) (2024) 240.

[14] H. Alanzi, M. Alkhatib, Towards improving privacy and security of identity management systems using blockchain technology: A systematic review, Appl. Sci. 12 (23) (2022) 12415.

[15] D.N. Community, State of the developer nation Q4 2019, 2020, URL https://www.developernation.net/resources/reports/state-of-the-developer-nation-q4-2019/. (Accessed 11 May 2025).

[16] A.Z. Junejo, M.A. Hashmani, A.A. Alabdulatif, A survey on privacy vulnerabilities in permissionless blockchains, Int. J. Adv. Comput. Sci. Appl. (IJACSA) 11 (9) (2020) 130–139.

[17] H. Fabric, A blockchain platform for the enterprise—hyperledger-fabricdocs main documentation, 2022.

[18] R3, Corda technical whitepaper, 2025, https://r3.com/blog/corda-technical-whitepaper/. (Accessed 20 May 2025).

[19] J.K. Mudhar, J. Malhotra, S. Rani, Blockchain-based decentralized access control framework for enhanced security and privacy for consumer electronic devices, IEEE Trans. Consum. Electron. (2024).

[20] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, Satoshi Nakamoto (2008).

[21] V. Buterin, et al., A next-generation smart contract and decentralized application platform, White Pap. 3 (37) (2014) 2–1.

[22] V. Capocasale, D. Gotta, G. Perboli, Comparative analysis of permissioned blockchain frameworks for industrial applications, Blockchain: Res. Appl. 4 (1) (2023) 100113.

[23] H.M. Kim, H. Turesson, M. Laskowski, A.F. Bahreini, Permissionless and permissioned, technology-focused and business needs-driven: understanding the hybrid opportunity in blockchain through a case study of insolar, IEEE Trans. Eng. Manage. 69 (3) (2020) 776–791.

[24] J. Zarrin, H. Wen Phang, L. Babu Saheer, B. Zarrin, Blockchain for decentralization of internet: prospects, trends, and challenges, Clust. Comput. 24 (4) (2021) 2841–2866.

[25] C. Ma, X. Kong, Q. Lan, Z. Zhou, The privacy protection mechanism of hyperledger fabric and its application in supply chain finance, Cybersecurity 2 (1) (2019) 1–9.

[26] G. Al-Sumaidaee, R. Alkhudary, Z. Zilic, A. Swidan, Performance analysis of a private blockchain network built on hyperledger fabric for healthcare, Inf. Process. Manage. 60 (2) (2023) 103160.

[27] H.F. Documentation, Introduction, 2025, URL https://hyperledger-fabric.readthedocs.io/en/latest/whatis.html. (Accessed 25 May 2025).

[28] C. Fan, S. Ghaemi, H. Khazaei, P. Musilek, Performance evaluation of blockchain systems: A systematic survey, Ieee Access 8 (2020) 126927–126950.

[29] M. Touloupou, M. Themistocleous, E. Iosif, K. Christodoulou, A systematic literature review toward a blockchain benchmarking framework, IEEE Access 10 (2022) 70630–70644.

[30] F. Liu, S. He, Z. Li, Z. Li, An overview of blockchain efficient interaction technologies, Front. Blockchain 6 (2023) 996070.

[31] A.H. Lone, R. Naaz, Demystifying cryptography behind blockchains and a vision for post-quantum blockchains, in: 2020 IEEE International Conference for Innovation in Technology, INOCON, IEEE, 2020, pp. 1–6.

[32] M. Esmaili, K. Christensen, Performance modeling of public permissionless blockchains: A survey, ACM Comput. Surv. 57 (7) (2025) 1–35.

[33] H. Caliper, Hyperledger caliper, 2025, URL https://hyperledger.github.io/caliper/. (Accessed 20 May 2025).

[34] C. Melo, F. Oliveira, J. Dantas, J. Araujo, P. Pereira, R. Maciel, P. Maciel, Performance and availability evaluation of the blockchain platform hyperledger fabric, J. Supercomput. 78 (10) (2022) 12505–12527.

[35] L. Ni, E. Irannezhad, Performance analysis of LogisticChain: A blockchain platform for maritime logistics, Comput. Ind. 154 (2024) 104038.

[36] H.T. Le, T.T.L. Nguyen, T.A. Nguyen, X.S. Ha, N. Duong-Trung, Bloodchain: a blood donation network managed by blockchain technologies, Network 2 (1) (2022) 21–35.

[37] K. Antevski, C.J. Bernardos, Applying blockchain consensus mechanisms to network service federation: Analysis and performance evaluation, Comput. Netw. 234 (2023) 109913.

[38] M. Sallal, G. Owenson, D. Salman, M. Adda, Security and performance evaluation of master node protocol based reputation blockchain in the bitcoin network, Blockchain: Res. Appl. 3 (1) (2022) 100048.

[39] M. Mazzoni, A. Corradi, V. Di Nicola, Performance evaluation of permissioned blockchains for financial applications: The ConsenSys quorum case study, Blockchain: Res. Appl. 3 (1) (2022) 100026.

[40] K. Ntolkeras, H. Sharif, S.D. Salmasi, W. Knottenbelt, Performance analysis of a hyperledger iroha blockchain framework used in the UK livestock industry, in: 2021 IEEE International Conference on Blockchain (Blockchain), IEEE, 2021, pp. 456–461.

[41] P.M. Dhulavvagol, V.H. Bhajantri, S. Totad, Blockchain ethereum clients performance analysis considering E-voting application, Procedia Comput. Sci. 167 (2020) 2506–2515.

[42] O. Novo, Scalable access management in IoT using blockchain: A performance evaluation, IEEE Internet Things J. 6 (3) (2018) 4694–4701.

[43] K. Suankaewmanee, D.T. Hoang, D. Niyato, S. Sawadsitang, P. Wang, Z. Han, Performance analysis and application of mobile blockchain, in: 2018 International Conference on Computing, Networking and Communications, ICNC, IEEE, 2018, pp. 642–646.

[44] H. Caliper, Architecture, 2025, URL https://hyperledger.github.io/caliper/v0.4.2/architecture/. (Accessed 30 May 2025).

[45] S. Smetanin, A. Ometov, M. Komarov, P. Masek, Y. Koucheryavy, Blockchain evaluation approaches: State-of-the-art and future perspective, Sensors 20 (12) (2020) 3358.

[46] S.M.H. Bamakan, A. Motavali, A.B. Bondarti, A survey of blockchain consensus algorithms performance evaluation criteria, Expert Syst. Appl. 154 (2020) 113385.

[47] M. Ajwalia, P. Shah, Performance evaluation of blockchain systems: Parameters, criteria and modeling techniques, in: 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing, UCC, IEEE, 2024, pp. 256–258.

[48] Y. Ucbas, A. Eleyan, M. Hammoudeh, M. Alohaly, Performance and scalability analysis of ethereum and hyperledger fabric, IEEE Access 11 (2023) 67156–67167.

[49] N. Anita, M. Vijayalakshmi, S.M. Shalini, K.D. Lakshmi, Blockchain consensus algorithms and platforms: a survey, J. Manag. Anal. (2025) 1–37.

[50] P.H.B. Correia, M.A. Marques, M.A. Simplicio, L. Ermlivitch, C.C. Miers, M.A. Pillon, Comparative analysis of permissioned blockchains: Cosmos, hyperledger fabric, quorum, and XRPL, in: 2024 IEEE International Conference on Blockchain (Blockchain), IEEE, 2024, pp. 464–469.

[51] J. Polge, J. Robert, Y. Le Traon, Permissioned blockchain frameworks in the industry: A comparison, Ict Express 7 (2) (2021) 229–233.

[52] S. Sarker, A.K. Saha, M.S. Ferdous, A survey on blockchain & cloud integration, in: 2020 23rd International Conference on Computer and Information Technology, ICCIT, IEEE, 2020, pp. 1–7.

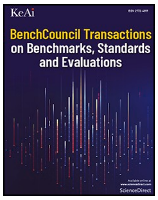
[53] H. Wu, Q. Yao, Z. Liu, B. Huang, Y. Zhuang, H. Tang, E. Liu, Blockchain for finance: A survey, IET Blockchain 4 (2) (2024) 101–123.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100255

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/



Full Length Article

US-China geopolitical tensions and Indian stock market dynamics: evidence from NARDL and wavelet coherence

Dr. Animesh Bhattacharjee [a,*], Suravi Deb [b], Dr. Joy Das [c]

[a] Department of Commerce, Techno India University, Tripura
[b] Department of Management, Techno India University, Tripura
[c] Department of Commerce, Nagaland University

ARTICLE INFO

Keywords:
Macroeconomics
Capital markets
India
Wavelet Applications in Finance

ABSTRACT

The geopolitical tension between China and the United States have increasingly shaped global financial markets; the exact impacts on emerging economies like India remain poorly explored. This study examines the impacts of changes in US-China tension on the Indian stock market based on the use of nonlinear Autoregressive Distributed Lag (NARDL) modeling and wavelet coherence analysis. With monthly observations and the newly developed U. S.-China Tension Index (UCT), the study finds asymmetric short-run effects: heightened tensions are likely to dampen sentiment and reduce returns, whereas reduced tensions offer limited relief. Interest rates are a key determinant in both the short run and long run, underscoring their inherent role in determining capital flows. Wavelet analysis captures a change in the nature of the relationship, from persistent co-movement in the early period to more prompt, temporary responses in subsequent years. These results underscore the growing significance of geopolitical attitudes to market action, especially for economies that are increasingly open to international capital flows.

The novelty of the paper arises from the application of a novel geopolitical risk metric (UCT) to an untapped economy (India) through a hybrid econometric-time-frequency method that captures hitherto unseen asymmetric and dynamic market responses.

Introduction

US-China tensions lie at the heart of determining global equity markets through the impact on economic growth, investor sentiment, and sectoral trends. Being the world's two largest economies, the evolution of their bilateral relationship, including trade wars, export bans, diplomatic moves, and military deployments, has spillover effects on global supply chains and capital flows, which subsequently influence companies one or two steps away from direct US-China tensions exposure [14]

Investor attitudes also fall with higher political uncertainty around the globe, making individuals more cautious to risk elsewhere and in other investments [10] Policy makers, investors, and large corporations closely observe Chinese and American diplomatic moves, trade negotiations, and Chinese and American policy shifts today. The technology and chip sectors are particularly vulnerable since they have globally networked supply chains, intellectual property concerns, and they are highly valuable assets, thus they are among the most vulnerable to increasing geopolitical tensions [6] The emerging nations and commodity-exporting nations are also more volatile since they respond to shifts in Chinese demand and changes in global trade flows [13]

This study pioneers the application of the U.S.–China Tension Index (UCT) to simulate and empirically analyze the impacts of geopolitical tension on stock markets. Developed by [14], the UCT index quantifies rising bilateral tension using text-based analysis of leading U.S. newspaper coverage. It classifies articles by whether they mention the U.S. and China, include contentious bilateral issues, and express sentiment indicative of tension, thus offering a validated, high-frequency measure tied closely to business-person and policy-maker perceptions. Higher UCT levels are empirically linked with lower U.S. corporate investment, especially by firms heavily exposed to China, and observable changes in supply chain arrangements away from China. Such effects are pre-trade-war, with uncertainty about future bilateral activity being more disturbing than de facto barriers themselves.

Peer review under the responsibility of The International Open Benchmark Council.

[*] Corresponding author.
E-mail address: bhatt.ani725@gmail.com (Dr.A. Bhattacharjee).

https://doi.org/10.1016/j.tbench.2025.100255
Received 3 September 2025; Received in revised form 10 December 2025; Accepted 11 December 2025
Available online 13 December 2025

In financial markets, the index manifests as cross-sectional patterns of stock returns, in accordance with investor expectations of shrinking economic opportunities at times of increased tension.

In addition, subsequent studies using methods such as time-varying quantile Granger causality and DCC-GARCH-MIDAS modeling indicate that UCT shocks significantly increase volatility and induce asymmetric co-movements among the equity markets of the United States and China situations that are easily observed in technology-centered and small-cap industries [20]

Overall, in applying the UCT index to macro-financial modeling, this research makes a methodological contribution by providing more accurate and detailed insight into how geopolitical tension is being converted into economic and financial market dynamics. It improves forecasting and risk analysis in an ever-changing world in which economic and financial market dynamics are becoming increasingly geopolitically motivated.

The Trend of Us-China Tension Index Post Liberalisation

Diagram 1 shows the evolution of the US-China Tension Index over three decades. Overall, the US-China tension index has steadily increased over time. There are significant rises in 2008, 2018, 2020, and 2022. These increases can be attributed to ongoing trade wars and geopolitical uncertainties. The 2008 rise was spurred by diplomatic tensions during the Subprime Crisis, and the substantial increase corresponds to the deepening of the US-China trade war. The 2020 peak is most likely due to tensions related to the COVID-19 pandemic. The peak in 2022 might be attributed to increased military rhetoric against Taiwan and the implementation of harsher technology export rules. Fig. 1.

Despite the observed changes, the constant raising of the index's baseline suggests that tensions have become more entrenched over time, rather than being episodic. This pattern reflects a structural shift in US-China ties, from one of collaboration to one of strategic competition.

Literature Review

Several studies have directly assessed the bilateral impact of US-China trade tensions on the stock markets of both countries. Using network-based models, [3] showed how the conflict leads to a weakening of cross-country market links and strengthens intra-country market clusters, indicating rising financial decoupling.

[12] employed time-varying quantile granger causality method and found that Chinese stock markets exhibited better reactivity to US-China tension (UCT). Furthermore, evidence revealed that the UCT's impact on Chinese stock markets was steadily growing and demonstrates notable time-varying characteristics and the asymmetric feature is evident in the US stock market's delayed reactions to worsening ties and greater susceptibility to good news about US-China relations.

[19] demonstrated how tensions between China and the US impacted the relationship between their stock markets and to what extent prices fluctuate through a DCC-DAGARCH-MIDAS model. The findings indicate that heightened tension increases the degree of return volatility changes, and the impacts of the trade war are realized sooner.

Another section of literature examines the uncertainty channel through the impact of uncertain policies on market responses [7] employed a trade policy uncertainty (TPU) index to link uncertainty during trade disputes with greater volatility in the market and lower Chinese stock returns [1] and [4] extend this concept further by demonstrating that markets respond more to words and announcements than to policy changes themselves. This indicates that words and signals are playing a greater role in the functioning of markets. This concept is evidenced by "Do words hurt more than actions?", demonstrating that threats in words and public comments can lead to sharp market declines even without the existence of actual policy changes.

[2] applied a TVP-VAR model (Time-Varying Parameter Vector Autoregression) to analyze the sensitivity of emerging market stock market returns, particularly in Asia, to US and China shocks. They affirmed that US shocks are more powerful and long-lasting than those of China, and the magnitude and direction of the shocks vary over time, particularly in times of crises [5] quantified China-US trade tensions with the application of machine learning. Their result was not sensitive to changes in global financial markets and was concurrent with extensively documented events in the US-China trade war. Local forecasts indicated that increasing trade tensions had a negative effect on stock market indices and exchange rates in China and emerging economies, but primarily did not influence US markets, except for firms more integrated with China [15], with the example of Turkey's Borsa Istanbul, affirmed that there exists a long-run relationship between Turkish stock prices and US-China trade shocks, revealing substantial third-party exposure [9] also analyzed further how stock market interlinkages increase worldwide during peak periods of the trade war, suggesting threats of far-reaching impacts.

Research such as [11] extends the context by examining the dynamics of exchange rate movements and stock market performance, particularly in the case of triangular economies like India. They show that exchange rate movements serve to disperse the effect of trade shocks on home markets, particularly in the context of countries with strong export sectors or foreign investment.

[8] elaborates that trade tensions are not merely economic phenomena; they are reflections of forms of strategic geopolitical re-alignment. Works such as "The Changing Fundamentals of US-China

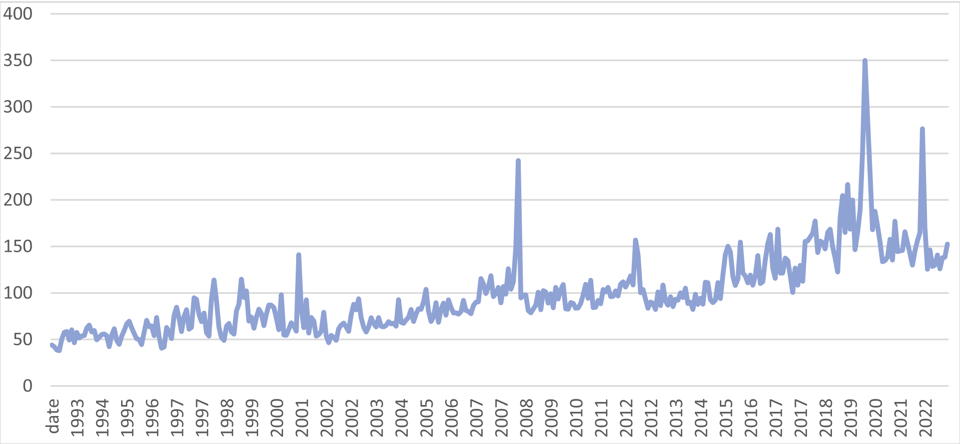


Fig. 1. Trend of US-China Tension.
Source: Author's Own

Relations" and "Is Globalisation Dead?" locate the trade war in the larger context of multilateralism and globalist pullback, which have far-reaching ramifications for global supply chains, investments, and decoupling patterns. These works emphasize the need to understand financial market responses in terms of dynamics of changing global power relations, and not just economic variables.

Despite widespread coverage, there are very few papers that have studied India. There is one paper [11] that covers India as a part of a model with two other economies but does not study how the Indian stock market responds to US-China tensions when India is economically and strategically important. The US-China trade tensions papers tell us about their influence on finance markets in the shape of volatility, uncertainty, contagion, and changes in geopolitical alignments. But there is a stark lack of papers on large non-aligned emerging economies like India, which need to be studied more intensely because they are becoming more and more part of global capital markets and foreign trade.

The study proposes the following testable hypothesis:

H1. US-China tension has a significant influence on the Indian stock market.

H2. Positive and negative changes in US-China tension index have asymmetric effect on the Indian stock market.

H3. The linkage between US-China tension and Indian stock market is time-varying and frequency-dependent.

Data and Methodology

This research uses monthly time-series data over 360 observations, from January 1993 to December 2022, to examine the effect of US-China geopolitical tensions on the Indian stock market. The major dependent variable is the Bombay Stock Exchange (BSE) Index that reflects the Indian equity market. The major explanatory variable is the US-China Tension Index (UCT) constructed by [14], which is a media-based measure of the intensity and frequency of bilateral tensions between the two countries. The data relating to US-China tension index is sourced from https://www.policyuncertainty.com/US_China_Tension.html.

Other control variables are as follows:

Exchange Rate (EXR) - INR/USD, to reflect currency movements and their effect on capital flows as well as investor sentiment.

Interest Rate (INT) - Defined by the Weighted Average Call Rate, reflecting domestic liquidity and borrowing cost.

Wholesale Price Index (WPI)- is an inflationary trend and cost pressures indicator.

India Geopolitical Risk Index (GPR_IND) -A measure of geopolitical risk for India that reflects local political uncertainty.

All series are log-transformed to ensure constancy of variance and facilitate elasticity-based interpretation.

In order to thoroughly investigate the impact of tensions between the United States and China on the Indian stock market, the present study utilizes a dual-methodological framework that combines the Non-Linear Autoregressive Distributed Lag (NLARDL) model and the Wavelet Coherence method. The NLARDL model allows the estimation of asymmetric long-run and short-run relationships between the Indian stock market and the corresponding explanatory variables. This technique is specifically useful in following the non-linear and possibly divergent transmission of positive and negative shocks in a multiple-variable setting. The Wavelet Coherence method also allows for a time-frequency decomposition of the co-movements among the variables. This method provides dynamic visualization of how the relationship between US-China tensions and the Indian stock market varies over different time horizons and frequencies, thereby determining times of high or low correlations and lead-lag relations. Utilizing these methodologies together, the study follows both the structural (long-run equilibrium and asymmetries) and temporal (time-varying co-movement) dimensions of the relationship, therefore providing rich and comprehensive understanding of the transmission of geopolitical tensions across borders to influence financial markets.

The NLARDL (Non-Linear Autoregressive Distributed Lag), developed by [16], methodology is an econometric approach used to analyze the relationship between time series variables, particularly in the context of cointegration and long-run equilibrium. It extends the traditional Autoregressive Distributed Lag (ARDL) model by allowing for non-linear relationships among variables, making it particularly useful in economic and financial studies where the effects of shocks can vary over time. The model typically includes both positive and negative changes of the independent variables to capture asymmetry.

In the present study, we intend to capture the asymmetric effect of US-China tension on Indian stock market in a multivariate framework. Thus, the functional relationship can be expressed as below:

$$BSE = f(EXR, WPI, INT, GPR - IND, UCT_{+}, UCT_{-})$$

The wavelet coherence method developed by [17] enables one to compute cross-wavelet power and therefore identify regions with higher covariance of time series variables across scales. The wavelet coherence confirms the periods during which co-movement of time series variables can be identified even though it is linked to low wavelet power. We adopt the method developed by [18], an extension of the method developed by [17]

The result of a cross-wavelet coherence analysis generally appears in the form of a graphical display comprising five main components: black arrows with eight direction marks (←, →, ↑, ↓, ↘, ↗, ↙, ↖), warm and cold color maps, black contours, two coordinate axes, and the cone. The right-pointing arrows (→) and left-pointing arrows (←) represent an in-phase and an out-of-phase relationship, respectively, and this is synonymous with positive and negative correlations. The up-right-pointing arrows (↗) and down-left-pointing arrows (↙) represent that the first and the second series have a leading effect, respectively. For example, in the wavelet coherence plots, the ’↘’ pointing black arrows represent an in-phase relationship or positive comovement of the two-time series, with the second time series having a leading effect. On the other hand, the ’↖’ pointing black arrows represent an out-of-phase relationship or negative comovement of the two-time series, with the first time series leading. A phase difference of zero means that both time series are moving in tandem. The black curves exhibited in the plots represent areas where the coherence is statistically significant at the 5% level, and the solid white bell-shaped line in the wavelet coherence plots represents the cone of influence.

The present study is guided by Evaluatology framework. Rather than solely relying on statistical significance, the framework focuses on effect isolation, incremental information validation and performance comparison. Thus, the research design of the present study is not only to estimate the linkage but also to evaluate whether US-China tension add significant information and predictive content beyond traditional domestic macroeconomic determinants of the Indian stock market.

One of the core principles of the Evaluatology framework is effect isolation of the phenomenon under investigation. Thus, the analysis first specifies a baseline model which consists of domestic macroeconomic variables. This model serves as a benchmark representation of the dynamics of Indian stock market in the absence of external geopolitical factor. In the second stage, the US-China tension index is included to form the augmented model to see whether the addition alters the estimated dynamics and strengthens explanatory power of the model. In doing so, the research moves beyond correlation assessment and evaluates the incremental contribution of the US-China tension.

Evaluatology also stresses that effects in complex system may be asymmetric and context dependent. Stock markets respond differently to adverse and favorable shocks. Following this principle, we employ the NARDL model which allows the decomposition of the tension index into positive and negative components.

Results

Table 1 show the descriptive statistics of the variables in the study, which covers 360 months of data. The BSE index averaged around 17,196, but it jumped around a lot, from a low of 2,122 to a high of 63,100. This indicates that the Indian stock market grew extensively but also witnessed turbulent periods during the period. The exchange rate (EXR) varied a bit, averaging 51.14, and ranging from 31.20 to 82.77. This shows that the Indian rupee has generally lost value against the USD over time.

The interest rate (INT), which we measured using the weighted average call rate, had a big range (0.73 to 34.83). It also had high skewness (3.53) and kurtosis (25.11), which means it doesn't follow a normal distribution and might have some outliers. These outliers are possibly due to high liquidity rates during crisis periods. The wholesale price index (WPI) averaged 610.3. Its skewness and kurtosis were mild, suggesting inflation was relatively stable compared to the other variables. The India-specific geopolitical risk index (GPR_IND) and the US-China Tension Index (UCT) both had high skewness and kurtosis. This suggests that there were some rare but big geopolitical shocks during the period. The Jarque-Bera test says that all the series aren't normally distributed. That's why we used econometric models like the NARDL model, which can deal with data that isn't linear or normally distributed. Table 2.

The unit root test determines the stationarity status of the logged variables in our model, which is needed before we use the NARDL method. For the study Augmented Dickey-Fuller test is used. The test results show that LNBSE, LNEXR, and LNWPI aren't stationary at-level, but they do become stationary after the first-differenced, meaning they're I(1). On the other hand, LNINT, LNGPR_IND, and LNUCT are stationary at level, so they're I(0).

This mix of I(0) and I(1) variables is valid condition for applying NARDL. It can handle both stationary and non-stationary variables even if they aren't all in the same order. It's important that none of the variables are I(2), because it would compromise the integrity of the NARDL model and render the bounds testing ineffective. The next step is to determine the number of lags to execute the NARDL model which is followed by the BDS-test. Table 3

The BDS test helps to determine whether the US-China tension index shows pattern over time. These patterns are significant as they help to evaluate whether the market will react differently to good and bad news about the US-China relations. The results of the BDS test, reported in Table 4, suggests the presence of non-linearity in the series. In other words, the results indicates that the US-China tension index shows systematic temporal dependence and not random behavior. This reinforces the use of non-linear econometric model in the subsequent analysis. Table 5.

At first, we conduct the F-bounds test within the NARDL framework to establish cointegration among Indian stock market, US-China tension index and domestic macroeconomic factors. The test yields a F-statistic of 5.685, which exceeds the I(1) critical bounds at the 10%, 5%, 2.5% and 1% significance level. This confirms the presence of cointegration or long-run equilibrium relationship among the variables. The presence of cointegration also validates the use of NARDL approach to model the dynamic and asymmetric effects of US-China tension and domestic macroeconomic factor on the Indian stock market.

Table 2
Unit Root Test.

	At Level		At First-Differenced		Status
	Test statistic	p-value	Test statistic	p-value	
LNBSE	-2.575639	0.2918	-18.31225	0.0000	I(1)
LNEXR	-2.049307	0.5717	-5.154889	0.0001	I(1)
LNINT	-4.608002	0.0011	—	—	I(0)
LNWPI	-2.487307	0.3343	-11.16801	0.0000	I(1)
LNGPR_IND	-5.054153	0.0002	—	—	I(0)
LNUCT	-6.787446	0.0000	—	—	I(0)

Source: Author's Own

The short-run parameters, long-run parameters and the residual diagnostic tests are reported in Table 6. The short-run parameters based on the NARDL error correction model clearly shows that domestic macroeconomic variables like interest rate and exchange rate have significant effect on the Indian stock market. A depreciation of the Indian rupee by 1% has a significant positive impact on stock market returns by 1.526%. The positive linkage reflects the fact that depreciation in Indian rupee against USD enhances the rupee-value earnings of exporting firms which lifts the equity prices of such firms listed in the Indian stock market. The linkage also captures investors' responses to anticipated capital flight and imported inflation alongside expectations of supportive policy measures. Together, these dynamics explain the positive linkage between INR-USD exchange rate and Indian stock market returns in the short-run.

On the other hand, an increase in interest rate has a significant negative impact on the Indian stock market returns. This negative association can be attributed higher borrowing cost which leads to lower corporate profitability and discouragement in new investment. For investors, a regime of higher interest rate makes assets with fixed-income more attractive as compared to equity. Factors like inflation and past performance of BSE do not indicate statistically significant short-run effects.

The short-run parameters further show that domestic geopolitical risk and US-China tension exert a strong influence on Indian stock market in the short-run.

One month lag of domestic geopolitical risk has a positive linkage with the BSE index in the short-run. This effect can be attributed to capturing of market expectations of reform following domestic turmoil and political transition. The results shows that the US-China tension dynamics are more complicated. It is observed that positive changes in US-China tension with one month lag, which captures increasing tension, have a significant negative impact on the Indian stock market. The result captures the anxiety of the investors regarding global trade volatility and regional economic impacts. The negative changes, which captures declining tension between US and China, have a positive influence on the Indian stock market with one month and two-month lags. The implication of the result is that de-escalation by powerful leaders of US and China restores investor confidence in emerging markets like

Table 1
Descriptive Statistics.

	BSE	EXR	INT	WPI	GPR_IND	UCT
Mean	17195.94	51.14207	6.948914	610.3019	0.205840	98.16300
Maximum	63099.65	82.77000	34.83000	1196.105	1.125609	349.9455
Minimum	2122.300	31.20100	0.730000	231.6000	0.044387	37.98329
Skewness	1.109017	0.476248	3.532141	0.340659	3.296180	1.699077
Kurtosis	3.474772	2.129039	25.11400	1.906766	18.05435	8.100995
Jarque-Bera	77.17619	24.98731	8083.996	24.89030	4051.391	563.5140
Probability	0.000000	0.000004	0.000000	0.000004	0.000000	0.000000
Observations	360	360	360	360	360	360

Source: Author's Own

Table 3
Lag Selection.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-239.1041	NA	1.62e-07	1.392637	1.458494	1.418845
1	2620.977	5606.409	1.74e-14	-14.65328	-14.19228*	-14.46982
2	2690.932	134.7426	1.44e-14*	-14.84620*	-13.99006	-14.50550*
3	2724.984	64.42930*	1.45e-14	-14.83514	-13.58385	-14.33719
4	2744.286	35.86066	1.60e-14	-14.74026	-13.09382	-14.08506
5	2758.451	25.83643	1.81e-14	-14.61620	-12.57462	-13.80375
6	2776.863	32.95225	2.01e-14	-14.51627	-12.07954	-13.54657
7	2799.868	40.39057	2.17e-14	-14.44243	-11.61057	-13.31549
8	2818.037	31.27821	2.41e-14	-14.34112	-11.11411	-13.05692

* indicates optimal lag

Table 4
BDS Test.

Dimension	BDS Statistic	Std. Error	z-Statistic	Prob.
2	0.1219	0.0031	39.338	0.0000
3	0.2081	0.0049	42.375	0.0000
4	0.2642	0.0058	45.343	0.0000
5	0.2984	0.0060	49.310	0.0000
6	0.3162	0.0058	54.402	0.0000

Source: Author's Own

Table 5
F-Bounds Test.

F-Bounds Test		Null Hypothesis: No levels relationship		
Test Statistic	Value	Signif.	I(0)	I(1)
F-statistic	5.685285	10%	1.99	2.94
k	6	5%	2.27	3.28
		2.5%	2.55	3.61
		1%	2.88	3.99

Source: Author's Own

Table 6
Short-run and Long-run Parameters.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(LNBSE(-1))	-0.0300	0.0530	-0.5669	0.5711
D(LNEXR)	-1.5268	0.1670	-9.1422	0.0000
D(LNEXR(-1))	-0.0727	0.1889	-0.3852	0.7003
D(LNINT)	-0.0473	0.0128	-3.6957	0.0003
D(LNWPI)	-0.5285	0.5481	-0.9643	0.3356
D(LNWPI(-1))	-0.1030	0.5490	-0.1877	0.8512
D(LNDGPR)	0.0049	0.0071	0.6971	0.4862
D(LNDGPR(-1))	0.0163	0.0076	2.1423	0.0329
D(LNDGPR(-2))	-0.0036	0.0071	-0.5048	0.6140
D(LNUCT_POS)	-0.0331	0.0299	-1.1058	0.2696
D(LNUCT_POS(-1))	-0.1005	0.0340	-2.9559	0.0033
D(LNUCT_POS(-2))	-0.0384	0.0359	-1.0711	0.2849
D(LNUCT_NEG)	-0.0026	0.0347	-0.0751	0.9401
D(LNUCT_NEG(-1))	0.0705	0.0342	2.0741	0.0388
D(LNUCT_NEG(-2))	0.0629	0.0307	2.0455	0.0416
C	0.0094	0.0092	1.0269	0.3052
EC(-1)	-0.0361	0.0094	-3.8336	0.0002
Long-run Parameters				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
LNEXR	-0.8949	0.9028	-0.9912	0.3223
LNINT	-0.8789	0.4089	-2.1493	0.0323
LNWPI	3.4614	1.8098	1.9125	0.0566
LNDGPR	-0.2535	0.2270	-1.1166	0.2649
LNUCT_POS	0.7453	0.4852	1.5358	0.1255
LNUCT_NEG	0.8502	0.5290	1.6072	0.1089
C	-6.2877	10.086	-0.6233	0.5334
			Test statistic	p-value
Breusch-Godfrey Serial Correlation LM Test			1.929683	0.1468
ARCH			0.336483	0.5622
RAMSEY RESET Test			1.533370	0.1261
CUSUM Plot			Stable	
CUSUM Sum of Sqaures			Stable	

Source: Author's Own

India.

In the long-run, the geopolitical variables (domestic geopolitical risk and US-China tension) are insignificant suggesting that their influence is mainly short-run in nature and sentiment driven. Additionally, interest rate has a significant negative influence even in the long-run, thus, capturing the structural role in the determination of capital and investment prices. The presence of a statistically significant and negative error correction term (-0.0361) captures a long-run equilibrium condition, where there is a significantly low adjustment rate of 3.6% per period towards this equilibrium due to short-run shocks. Furthermore, diagnostic tests confirm the stability of the model showing no serial correlation or heteroskedasticity and ensuring that the system is stable in the long run, thus confirming that the estimated relationships are indeed consistent and correctly specified.

To check whether the US-China tension index enhance the predictability of the Indian stock market, a baseline model (excluding the UCT variable) is also determined. The study adopts a structured evaluatology-based approach [21] in model parsimony, incremental information testing and forecast comparison theory

The functional baseline model is presented below:

$$BSE = f(EXR, WPI, INT, GPR\_IND)$$

Out of sample forecasts for the dependent variable were generated for both models to check the predictive accuracy of the models. The predictive accuracy of the baseline models and the UCT augmented model are compared using measures like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SAME), Theil's U statistic and Diebold-Mariano (DM) test.

As discussed above, the evaluation of the forecasting relevance of the US-China tension index we have first estimated a baseline model (without the index) and a full model (with the index). The forecasts of both the models are compared for the accuracy metrics such as RMSE, MAE, MAPE, SMAPE, Theil's U1 and U2 (See Table 7). All the metrics suggests that the full model is marginally better than the baseline model. To formally compare the forecasts of both the model, we further employ Diebold-Mariano test. The result, under the squared error loss function, indicates the rejection of the null of hypothesis of predictive accuracy. This clearly indicates that the full model provides significantly superior forecasts.

Fig. 2 portrays the wavelet coherence plot between the Indian stock market index (BSE) and the US-China tension index. It can be observed that there is significant high coherence zone, which is highlighted in red and enclosed by black contour lines, between time points 130 to 340 (roughly corresponding to 2003-2021). The arrows in this area point to the left, signifying a long-term negative correlation between the variables (time points 130 to 220). It is noteworthy to point out that after time point 220 the left-pointing arrows go slightly downward, suggesting that the BSE index is the leading variable of the negative association.

Table 7
Accuracy Measures and Predictive Superiority Tests for Baseline Versus Full Model.

Diebold-Mariano test (HLN adjusted)						
Null hypothesis: Both forecasts have the same accuracy						
Accuracy	Statistic	<> prob	> prog	< prob		
Abs Error	1.504272	0.1334	0.9333	0.0667		
Sq Error	3.449359	0.0006	0.9997	0.0003		
Evaluation statistics						
Forecast	RMSE	MAE	MAPE	SMAPE	Theil U1	Theil U2
EQ_BASE	0.186374	0.150185	1.695310	1.692223	0.009946	2.816482
EQ_FULL	0.185655	0.149828	1.691398	1.688342	0.009908	2.805448
Simple mean	0.186001	0.149995	1.693244	1.690172	0.009926	2.810797

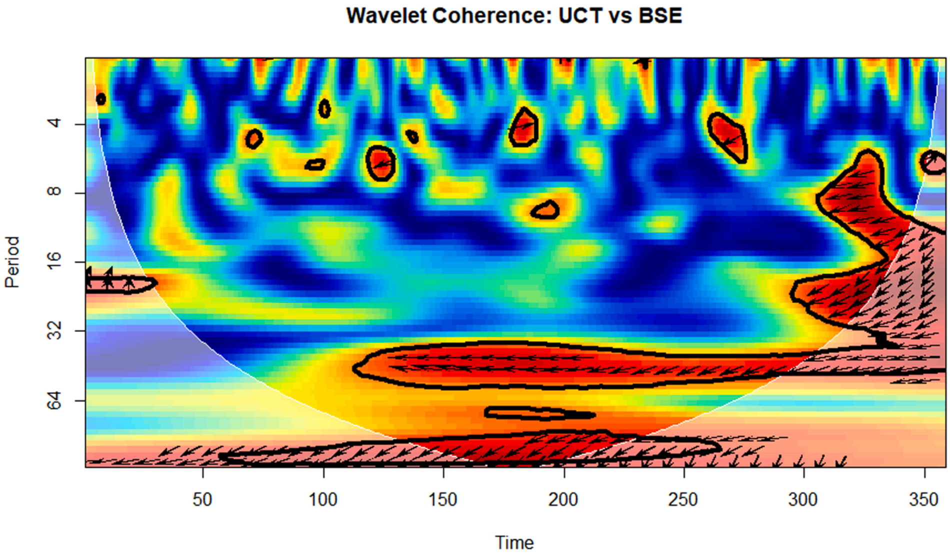


Fig. 2. Wavelet Coherence between UCT and BSE.
Source: Author's Own

Very few isolated areas of the high coherence zone are visible in the short-term, although the association is unstable. It is also interesting to see how the relationships between the two variables changed from long-term in the initial period to short-term in the latter period, particularly since 2017. The apparent change from lower to higher frequency in the wavelet coherence diagram indicates that the Indian stock market's reaction to US-China tensions became increasingly short-term but instantaneously developed with the passage of time. While previous stages (low frequency) are capturing long-run, structural market responses to US-China tension events, later periods of high-frequency coherence suggest that the market started responding more rapidly, presumably to investor sentiment, news shocks, and short-term capital flows, although these influences never in general lasted across longer horizons. The Wavelet Coherence plot is further explained with the help of Table 8.

The Table 8 displays a wavelet coherence examination of the dynamic connection between the BSE Index and US-China geopolitical tensions across time, which provides different patterns across three periods. From 2003–2013 (Time Points ~130–220), the coherence is low frequency, indicating a long-run negative connection without lead or lag, implying tensions influenced market sentiment but directionality was unclear. The period witnessed China's increasing global footprint, the 2008 global financial crisis, and continuous South China Sea tensions, which imposed relentless but indirect pressure on the Indian market. From 2013–2017 (~220–270), the pattern is mixed frequency with the BSE Index starting to lead the geopolitical tension index. This may suggest Indian markets were becoming increasingly forward-looking, potentially responding ahead of time to international macro signals such as the Yuan devaluation and Trump's Trumpian hawkish trade rhetoric, reflecting increasing market confidence and deeper integration with international capital flows. Lastly, from 2017–2021 (~270–340), the connection has high-frequency coherence, and the BSE responds instantaneously and sporadically to US-China tensions. This change is in line with extremely volatile global events such as the US-China Trade War, the COVID-19 outbreak, and news-driven capital flows, and captures the Indian stock market becoming more responsive to global geopolitical shocks and the prevalence of short-run investor sentiment in an increasingly fast-evolving world.

Table 8
Temporal Dynamics of Wavelet Coherence Between US-China Tensions and Indian Stock Market: Frequency, Relationship, and Key Events (2003–2021).

Time Period	Time Points	Frequency	Relationship Pattern	Key Events Driving It
2003-2013	~130-220	Low (long-term)	No clear lead or lag (negative)	China's global rise, 2008 crisis, South China Sea tensions
2013-2017	~220-270	Mixed	BSE begins to lead	Yuan devaluation, Trump trade rhetoric, market anticipation
2017-2021	~270-340	High (short-term)	BSE reacts immediately, inconsistently	US-China Trade War, COVID-19, news shocks, short-term capital flows

Source: Author's Own

Conclusion

This study presents new empirical evidence of how US-China geopolitical tensions influence the Indian stock market using econometric modeling (NARDL) and wavelet coherence. The evidence reveals that, although long-run impacts of US-China tensions on the Bombay Stock Exchange (BSE) are not significant, short-run reactions are statistically and economically significant. That is, when tensions increase (positive shocks to US-China tensions), Indian stock returns decrease, and when tensions decrease (negative shocks to US-China tensions), market sentiment is better. Secondly, interest rates are a salient variable explaining the volatility of BSE in the short run and the long run, and domestic geopolitical risks have a lagged, sentiment-based effect.

Wavelet coherence also observes a temporal development of this correlation, from a long-run correlation (2003-2013), to a future-oriented period in which BSE created tensions (2013-2017), to a high-frequency, sentiment-based reaction phase (2017-2021). This transition reflects the growing responsiveness of emerging economies such as India to changing global geopolitics, especially in the post-liberalization period of free capital movements and computerized trading systems.

It is important that policymakers understand the importance of financial stability during geopolitical crises. Exchange and interest rate volatility play a critical role in investor confidence in India. Coordination of fiscal and monetary policies during international or diplomatic crises would, therefore, be helpful in avoiding excessive market reactions. Additionally, the creation of early warning systems that track global geopolitical risks, such as the measurement of media sentiment tools like the UCT Index, would be wise in taking preventive measures before the expected problems arise.

Investors and corporate leaders must consider geopolitical risks in investment decisions and strategic planning. Asset managers can use adaptive protection measures or diversify regionally to mitigate short-term difficulties resulting from U.S.-China relations. Scenario planning and the ability to adapt operations-i.e., restructuring supply chains-are essential for firms and particularly firms in trade-volatile industries. Market participants should be aware that political rhetoric can influence prices without seemingly changing policies.

This study suggests a few areas of further research. First, sector-level analysis would reveal which Indian industries were most exposed to United States-China geopolitics. Second, use of investor sentiment indices or news-based proxies for uncertainty would be able to better inform our understanding of the transmission mechanisms involved. Third, as India continues to rise as a principal global manufacturing hub and geopolitics flashpoint, bidirectional relations could be studied, i.e., how India's foreign policy alignment and strategic choices affect regional market behaviors. Comparative studies with other major non-aligned economies, e.g., Brazil and Indonesia, would give a fuller picture of how superpower rivalries shape Global South markets. Incorporating region-specific strategic indicators like QUAD-related indices may enrich future research work. Finally, future research may explore the application of machine learning techniques to complement time-series econometric models and enhance forecasting performance under highly volatile geo-political regimes.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Chat GPT in order to paraphrase. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Dr. Animesh Bhattacharjee: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Suravi Deb:** Data curation. **Dr. Joy Das:** Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S.R. Baker, N. Bloom, S.J. Davis, KJ. Kost, Policy news and stock market volatility, Nat. Bureau Econom. Res. (2019 Apr 1).

[2] L. Bonga-Bonga, S. Mpoha, Spillover effects from China and the united states to key regional emerging markets: a dynamic analysis, Int. Rev. Financ. Anal. 91 (2024 Jan 1) 103015.

[3] Y. Chen, AA. Pantelous, The US-China trade conflict impacts on the Chinese and US stock markets: A network-based approach, Finan. Res. Lett. 46 (2022 May 1) 102486.

[4] S.J. Davis, D. Liu, XS. Sheng, Economic policy uncertainty in China since 1949: the view from mainland newspapers, in: Fourth Annual IMF-Atlanta Fed Research Workshop on China's Economy Atlanta, 19, 2019 Sep 19, pp. 1–37.

[5] M. Ferrari Minesso, F. Kurcz, MS. Pagliari, Do words hurt more than actions? the impact of trade tensions on financial markets, J. Appl. Econom. 37 (6) (2022 Sep) 1138–1159.

[6] *Financial Times*. (2025, late June). US needs to prepare markets for the risk of a Chinese invasion of Taiwan.

[7] F. He, B. Lucey, Z. Wang, Trade policy uncertainty and its impact on the stock market-evidence from China-US trade conflict, Finan. Res. Lett. 40 (2021 May 1) 101753.

[8] Y. Huang, US-China tensions: interplay between economics and politics, Washing. J. Modern China 13 (2017) 30–52.

[9] T.L. Huynh, T. Burggraf, If worst comes to worst: co-movement of global stock markets in the US-China trade war, Economi. Busi. Lett. 9 (1) (2020) 21–30.

[10] IMF. (2025). Geopolitical risks: Implications for asset prices and financial stability. International Monetary Fund.

[11] D. Krishnan, V. Dagar, Exchange rate and stock markets during trade conflicts in the USA, China, and India, Glob. J. Emerg. Market Econo. 14 (2) (2022 May) 185–203.

[12] C. Peng, H. Deng, J. Xie, X. Liu, US-China tension and stock market performance in us and china: new insights from time-varying quantile causality method, Finan. Res. Lett. (2025 Jul 5) 107888.

[13] Reuters. (2025, August 1). Investors see few winners as tariff storm lashes global markets.

[14] Rogers JH, Sun B, Sun T. US-China tension. Available at SSRN 4815838. 2024 May 3.

[15] O. Şeyranlıoğlu, The impact of US-China tensions on borsa istanbul stock market: evidence from ARDL approach, Politik. Ekonomik. Kuram 9 (2) (2025) 783–804.

[16] Y. Shin, B. Yu, M. Greenwood-Nimmo, Modelling asymmetric cointegration and dynamic multipliers in a nonlinear ARDL framework. Festschrift in honor of Peter Schmidt: Econometric methods and applications, Springer New York, New York, NY, 2014 Feb 5, pp. 281–314.

[17] C. Torrence, GP. Compo, A practical guide to wavelet analysis, Bull. Ameri. Meteorolog. Soc. 79 (1) (1998 Jan) 61–78.

[18] C. Torrence, PJ. Webster, Interdecadal changes in the ENSO–monsoon system, J. Climat. 12 (8) (1999 Aug) 2679–2690.

[19] Xu J, Bouri E, Fang L, Gupta R. US-China Tensions and Stock Market Co-movement between the US and China: Insights from a DCC-DAGARCH-MIDAS Model. 2025 Jul.

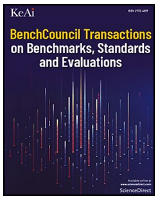
[20] K. Zeng, R. Wells, J. Gu, A. Wilkins, Bilateral tensions, the trade war, and US–China trade relations, Bus. Politic 24 (4) (2022 Dec) 399–429.

[21] Zhan, J., Wang, L., Gao, W., Wang, C., Li, H., Fan, F., Kang, G. (2025). Evaluatology: the science of uncovering the effects. BenchCouncil Press.

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/



Full Length Article

An adaptive opposite slime mold feature selection algorithm for complex optimization problems[☆]

Elsayed Badr[a,b,c], Mostafa Abdullah Ibrahim[b], Diaa Salama[b,d], Mohammed ElAffendi[e], Abdelhamied A Ateya[e,f,*], Mohamed Hammad[e,g], Alaa Yassin[a,b]

[a] Faculty of Computers and Artificial Intelligence, Misr University for Science & Technology, Egypt
[b] Faculty of Computers and Artificial Intelligence, Benha University, Egypt
[c] The Egyptian School of Data Science (ESDS), Benha, Egypt
[d] Faculty of Computers and Information, Misr International University, Egypt
[e] EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia
[f] Department of Electronics and Communications Engineering, Zagazig University, Zagazig 44519, Egypt
[g] Department of Information Technology, Faculty of Computers and Information, Menoufia University, Shibin El Kom, 32511, Egypt

ARTICLE INFO

Keywords:
Slime mold algorithm (SMA)
Optimization
Adaptive opposite SMA (AOSMA)
Enhanced SMA (ESMA)
Modified SMA (MSMA)

ABSTRACT

The slime mould algorithm (SMA) has recently emerged as a soaring metaheuristic strategy to function optimization problems due to its solid exploration-exploitation balance that enables it to converge efficiently towards high-quality solutions. In spite of its broader applications, however, there remain areas where the algorithm is constrained in diversified exploration and the scope of its exploitation mechanisms. In bridging these loopholes, this work introduces a new variant known as the adaptive opposition SMA (AOSMA). AOSMA involves an adaptive opposition-based learning (OBL) method, which learns online how to add opposition-based solutions at the iteration process to enhance exploration abilities and avoid premature convergence. The adaptive policy enables the algorithm to escape local optima more effectively by occasionally generating alternative candidate solutions. Additionally, for the sake of increased exploitation, AOSMA also incorporates a plan in which the randomly selected search agent is progressively replaced with the current best-performing agent during position updating. The replacement process increases the focus of the algorithm towards prospective regions of the search space, and thus it converges more quickly towards the global optimum. The implemented AOSMA was exhaustively validated with both qualitative and quantitative measures in terms of thirteen rigorously proven benchmark test functions involving a variety of unimodal, multimodal, and composite landscapes to test its optimization ability extensively. Comparative tests on a collection of state-of-the-art metaheuristic algorithms confirmed that AOSMA consistently produces higher or highly comparable performances across a variety of problem instances. The experimental results confirm the robustness, adaptability, and improved search ability of the algorithm, highlighting its potential as an efficient optimization method for complex real-world problems. With the efficient fusion of adaptive exploration and improved exploitation, AOSMA provides a vital contribution to the field of research into swarm intelligence and metaheuristic optimization.

1. Introduction

Maximum utilization of resources is a basic requirement in numerous scientific, engineering, and industrial applications. Because of the increasing complexity of real-world problems and the scarcity of resources, optimization has become a necessity for ensuring the highest performance, cost-effectiveness, and operating efficiency. Optimization problems can be broadly categorized into deterministic and stochastic approaches, depending upon the type of problem, each with its advantages and methodological strategies depending upon the problem landscape [1]. Deterministic optimization methods, such as linear and nonlinear programming, rely extensively on derivative information. These methods work best when there is structured problem formulation and the search space is smooth or linear. Deterministic approaches excel

[☆] Peer review under the responsibility of The International Open Benchmark Council.

[*] Corresponding author.
E-mail address: aateya@psu.edu.sa (A.A. Ateya).

https://doi.org/10.1016/j.tbench.2025.100250
Received 10 July 2025; Received in revised form 8 December 2025; Accepted 9 December 2025
Available online 13 December 2025

Table 1
Comparison between the proposed approach and closely related prior works that already combine OBL with SMA (e.g., [10,13,16]).

Algorithm	Core Idea	Strengths	Limitations	Relation to Present Work
SCA (Sine Cosine Algorithm) [10]	Uses sine and cosine mathematical functions to update positions and control exploration/ exploitation.	Simple, easy to implement, competitive on many benchmarks.	Sensitive to parameter tuning, risk of premature convergence on multimodal problems.	Illustrates novel update dynamics; lacks adaptive decision strategies.
New Meta-heuristic [13]	Introduces a new population-based global optimizer emphasizing exploration ability.	Demonstrates strong global search potential, competitive performance.	Does not include adaptive control; balance between exploration and exploitation can weaken.	Shows effectiveness of new frameworks but highlights need for adaptive balancing.
SMA Survey [16]	Comprehensive review of SMA and its variants (chaotic maps, OBL, hybrids).	Demonstrates SMA’s flexibility, wide applications, and many successful variants.	Variants often apply OBL blindly or intensification operators without systematic rules; no consensus on when to switch.	Identifies gap: absence of adaptive OBL decision + replacement integration.
Proposed AOSMA	Combines adaptive OBL decision strategy with a replacement mechanism.	Selective exploration, improved exploitation, faster convergence, robust performance.	Slightly higher computational cost.	Directly addresses the gap identified in SMA literature by integrating adaptive exploration control with exploitation enhancement.

Table 2
Difference between this paper’s combination of OBL and SMA and the related work.

Algorithm	Core Idea	Enhancement Strategy	Difference from AOSMA
OJESMA (W. C. Wang et al. [13])	Improve SMA using hybrid strategies	Combines equilibrium optimizer, joint opposite selection, and OBL	Depends on external optimizers and hybridization, whereas AOSMA focuses on integrated OBL with adaptive exploitation
Adaptive SMA (M. K. Naik et al. [10])	Enhance OBL application adaptively	Uses an adaptive mechanism to decide OBL usage; improves exploitation by replacing random agents with the best agent	Focuses mainly on OBL decision-making and selective replacement; AOSMA instead integrates OBL with adaptive exploitation for broader balance
AOSMA (Proposed)	Advance slime mould optimization	Combines OBL and adaptive exploitation techniques to balance exploration and exploitation dynamically	Unlike OJESMA (hybrid) or Adaptive SMA (OBL-focused), AOSMA unifies OBL and adaptive exploitation into a single integrated framework

in linear or weakly nonlinear spaces but experience difficulties when confronted with highly nonlinear, multi-modal, or discontinuous spaces. In such challenging cases, their dependence on gradient information can be a limiting aspect, conditioning them to get stuck in local optima or not to converge at all [2].

Stochastic optimization techniques, by contrast, never employ direct gradient information. This kind of algorithm forms the foundation for metaheuristic optimization methods. Metaheuristics are designed to provide rigorous, general-purpose solutions that can be applied to a large class of complex problems without specific knowledge of the mathematical form [3]. Their inherent advantages include simplicity, flexibility, independence of gradients, and less reliance on initial candidate solutions, traits that have driven their popularity and dramatic rise over the past few years. Metaheuristic algorithms have proven outstandingly successful in optimization problem-solving in numerous applications, from engineering design and scheduling to machine learning, hyperparameter tuning, and biomedical image analysis. They are largely inspired by natural, physical, or social processes and thus form a vast and growing taxonomy. Metaheuristic algorithms fall under four broad classes: swarm intelligence-based, evolutionary-based, physics-based, and human-inspired algorithms [4].

These nature-inspired metaheuristic algorithms possess specific strengths as they emulate some adaptive and intelligent behavior from nature, resulting in very efficient tools. An innovative nature-inspired optimization algorithm is the slime mould algorithm (SMA), which is based on the nature-inspired foraging behavior and dynamic oscillatory movement exhibited by slime moulds (Physarum polycephalum) [5]. SMA has been demonstrated to possess a strong capability to handle extremely complex, high-dimensional, and multi-modal optimization problems. Due to its strong global search capability and excellent resistance to local optima, SMA has been effectively applied to many applications in machine learning, engineering design, feature selection, and image processing [6].

Although it has an outstanding performance, SMA also faces some disadvantages, particularly concerning its exploration and exploitation. There were findings indicating that SMA often has acceptable exploration of the search space. However, its capacity for exploitation is still in need of enhancement to achieve faster convergence and higher-quality solutions. To address such limitations, numerous methods have been proposed in recent years. As documented in the literature [5], the two most significant improvement strategies that have emerged are hybridization, where SMA is combined with other metaheuristics to leverage complementary strengths, and algorithmic refinement, which involves modifying or improving SMA’s basic equations and mechanisms to improve performance.

A case in point is Chen et al. [7], which proposed the RCLSMAOA algorithm, a hybrid metaheuristic that combined the SMA and arithmetic optimization algorithm (AOA). The hybridization approach enhanced the convergence rate and solution accuracy, showing the potential of blended algorithmic solutions to counter weaknesses in individual approaches. Building on such advancements, this work proposes an adaptive variant of SMA in the guise of the adaptive opposition SMA (AOSMA). AOSMA employs an adaptive opposition-based learning (OBL) process to adaptively determine when to inject opposition solutions, enhancing the ability of the algorithm to explore without precipitating premature convergence. In addition, a novel exploitation enhancement method is incorporated through automatically substituting a randomly selected search agent with the best-performing agent at the position update phase, focusing the search on promising regions.

The remainder of this paper has the following structure: Section 2 provides a comprehensive review of existing literature and recent additions to the SMA; Section 3 provides a detailed methodology and the AOSMA model suggested; Section 4 provides experimental findings and comparative analyses; and finally, Section 5 summarizes the research with conclusions and future study directions.

2. Related works

Metaheuristic algorithms have been of much importance as efficient tools to tackle nonlinear and complex optimization issues where

Table 3
Properties of the considered standard benchmark functions.

Function	Function	D	F min	Range
F1	$\sum_{i=1}^{n} X_i^2$	30	0	[100,-100]
F2	$\sum_{i=1}^{n} X_i^2 + \prod_{i=1}^{n} \lvert X_i \rvert$	30	0	[10,-10]
F3	$\sum_{i=1}^{n} \left(\sum_{j=1}^{i} X_j\right)^2$	30	0	[100,-100]
F4	$\text{maxi}\,\{\lvert x \rvert,\ 1 \leq i \leq\ n\}$	30	0	[100,-100]
F5	$\sum_{i-1}^{n-1} \left(100(X_{i+1} - X_i)^2 + (X_i - 1)^2\right)$	30	0	[30,-30]
F6	$\sum_{i=1}^{n} (X_i + 0.5)^2$	30	0	[100,-100]
F7	$\sum_{i=1}^{n} (X_i^4 + rand\ (0,1))$	30	0	[1.28,-1.28]
F8	$\sum_{i=1}^{n} - X_i^2 \sin\sqrt{\lvert X_i \rvert}$	30	-418.98 N	[500,-500]
F9	$\sum_{i=1}^{n} \left[X_i^2 - 10\cos(2\pi X_i) + 10\right]$	30	0	[5.12,-5.12]
F10	$-20exp\left(-0.2\sqrt{\sum_{i=1}^{n} x_i^2}\right) -$ $exp\left(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i)\right) + 20 + e$	30	0	[32,-32]
F11	$\frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	30	0	[600,-600]
F12	$\frac{\pi}{\mathrm{n}}\left\{ \begin{matrix} 10\ sin(\pi \mathrm{y}_1) + \\ \sum_{\mathrm{i}=1}^{\mathrm{n}-1} (\mathrm{y}_{\mathrm{i}+1})^2 \left[1 + 10\sin^2(\pi \mathrm{y}_{\mathrm{i}+1})\right] + (\mathrm{y}_\mathrm{n} - 1)^2 \end{matrix} \right\} +$ $\mathrm{y}_\mathrm{i} = 1 + \frac{\mathrm{x}_\mathrm{i} + 1}{4}$ $\mathrm{u}(\mathrm{x}_\mathrm{i},\ \mathrm{a},\ \mathrm{k},\ \mathrm{m}) = \left\{ \begin{matrix} k(\mathrm{x}_\mathrm{i} - \mathrm{a})^\mathrm{m}\ \mathrm{x}_\mathrm{i} > a \\ 0 - a < \mathrm{x}_\mathrm{i} < a \\ k(\mathrm{x}_\mathrm{i} - \mathrm{a})^\mathrm{m}\ \mathrm{x}_\mathrm{i} < -a \end{matrix} \right.$	30	0	[50,-50]
F13		30	0	[50,-50]

traditional deterministic approaches are inefficient. Among them, swarm intelligence-based approaches such as PSO, ACO, and ABC have been found very successful in traversing extensive and complicated search spaces. Requested by the natural behavior, these algorithms can efficiently explore and exploit in a bid to provide efficient solutions to computationally complex real-world problems. However, the continually rising global optimization problems have motivated scholars to seek even more flexible and hybridized methods to handle issues like premature convergence and low convergence rates [8,9].

SMA, though showing its prowess in many applications, is underutilized with respect to its potential, primarily dynamic adaptability and enhanced exploration mechanisms. Recent developments in SMA are mainly directed towards hybridization with other approaches or modifications of update equations in the central part for enhancing global search capability and rate of convergence. With these efforts, issues such as maintaining diversity in the search space and local optima avoidance still exist. This has motivated the development of the AOSMA with the objective of integrating OBL and adaptive exploitation techniques for further advancing the state-of-the-art in slime mould-based optimization [10]. This section provides the related studies to introduce the novelty of the proposed approach better.

SMA, inspired by biological behavior, is shown to be highly effective for solving challenging stochastic optimization problems. Its ability to imitate slime mould search patterns offers a competitive edge in global optimization. However, it struggles with premature convergence and suboptimal solutions in complex scenarios. To resolve this, Yang et al. [11] proposed a multi-chaotic local operator that is introduced into the feedback system to enhance local exploration through chaotic perturbations. In [12], the authors first reviewed and analyzed numerous advanced variants of the SMA, providing a comprehensive summary, classification, and discussion of their mechanisms and prospects. Secondly, they categorized the application areas of SMA, examining its functions, current development status, and existing limitations. The literature review shows that SMA outperforms several well-known metaheuristic algorithms in terms of convergence speed and accuracy across various benchmark functions and practical optimization tasks.

Wang et al. [13] suggested the OJESMA approach, an enhanced SMA. Through the combination of equilibrium optimizer, joint opposite selection, and OBL, OJESMA enhances the algorithm's performance. The study involved applying nonparametric tests (Friedman and Wilcoxon) to evaluate optimization performance on 10 CEC2020 benchmark functions and 29 CEC2017 functions. The outcomes of the non-parametric tests and the experiment demonstrate that OJESMA performs better in terms of stability, convergence, and optimization accuracy. The authors also ran optimization tests on the variable index Muskingum and six engineering challenges to confirm the algorithm's efficacy. In [14], the authors presented an improved algorithm that combines an adaptive search operator strategy based on the Cauchy inverse cumulative distribution (QCMSMA) with Bloch sphere-based elite population initialization. The Wilcoxon rank-sum test and Friedman ranking analysis were used for statistical evaluations, along with comparisons against a number of popular optimization techniques.

Houssein et al. [15] proposed another work that uses the SMA adaptive guided differential evolution algorithm (SMA-AGDE) to address some inherent weaknesses of the original SMA. By employing AGDE's mutation strategy within it, the hybrid approach enhances local search power, population diversity, and prevention of getting trapped in local minima. Comparative research involving a range of well-established, newly developed, and high-performance metaheuristics (such as BBO, GSA, TLBO, HHO, MRFO, CMA-ES, and the simple SMA) revealed that SMA-AGDE was significantly better than its alternatives in all instances of problems, being first in different problem contexts. The research verifies that SMA-AGDE is an effective and generic optimization platform.

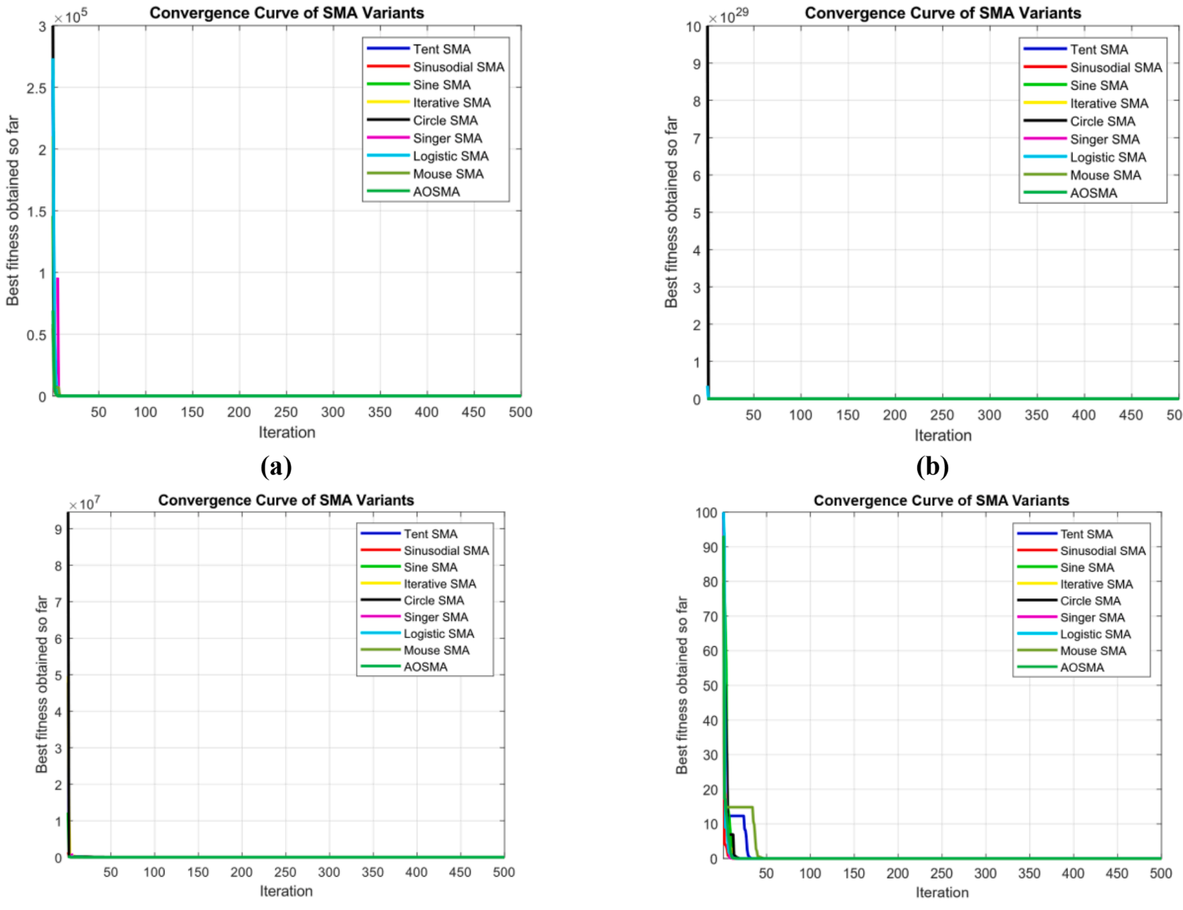
In [5], the authors examined the SMA from various optimization perspectives. The algorithm utilizes a specialized mathematical framework that imitates biological wave propagation using adaptive weighting, introducing several innovative features to enhance performance. This design enables an effective balance between exploration and exploitation, guiding the search efficiently toward optimal solutions. In [16], the authors addressed a four-objective optimization problem in the construction industry by proposing a hybrid model called AOSMA, which integrates OBL with the SMA. To showcase the effectiveness of the proposed approach, two real-world construction case studies were used,

Table 4
Fitness values of the proposed AOSMA and other algorithms.

Function	Run	f_{min}	Sinusoidal SMA	Circle SMA	Singer SMA	Tent SMA	Logistic SMA	Sine SMA	Iterative SMA	Mouse SMA	AOSMA
F1	1	0	0	0	2.5852	0	0	0	0	0	0
	2		0	0	1.21E-42	0	0	0	0	0	0
	3		0	0	9.68E-50	0	0	0	0	0	0
	4		0	0	1.73E-09	0	1.6446e-317	0	0.00E+00	0	0
	5		0	0	2.44E-18	0	0	0	0	0	0
F2	1	0	5.03E-199	3.18E-215	6.66E-05	3.72E-158	2.71E-175	1.30E-199	3.72E-158	1.10E-240	0
	2		5.09E-182	1.14E-270	1.49E-10	1.25E-215	1.73E-288	1.01E-166	1.25E-215	5.44E-163	0
	3		1.69E-164	9.93E-194	0.21205	2.40E-303	4.66E-182	2.03E-177	2.40E-303	7.60E-245	0
	4		3.50E-220	4.27E-218	2.15E-24	2.34E-185	2.19E-213	8.41E-173	2.34E-185	2.29E-187	0
	5		1.02E-177	3.55E-209	9.99E-27	1.20E-253	4.21E-264	7.49E-264	1.20E-253	6.81E-276	0
F3	1		0.00E+00	0.00E+00	0.0078075	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
	2		0.00E+00	2.9644e-32	1.42E-77	0.00E+00	5.2707e-319	0.00E+00	0.00E+00	0.00E+00	0
	3		0.00E+00	0.00E+00	191.5055	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
	4		0.00E+00	0.00E+00	1.0838	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
	5		0.00E+0000	0.00E+00	3.3663	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
F4	1	0	2.05E-189	6.41E-192	1.16E-206	1.38E-231	7.67E-222	3.08E-166	1.38E-231	2.92E-197	0
	2		5.43E-182	4.65E-192	1.23E-244	1.27E-211	2.81E-198	9.45E-204	1.27E-211	6.01E-158	0
	3		7.61E-165	1.22E-242	5.90E-237	3.33E-206	3.36E-225	1.72E-197	3.33E-206	1.37E-238	0
	4		3.77E-236	8.40E-174	8.32E-255	2.63E-177	1.18E-180	1.85E-150	2.63E-177	3.81E-192	0
	5		8.32E-182	1.88E-189	1.19E-220	4.45E-239	5.55E-214	1.33E-174	4.45E-239	1.50E-174	0
F5	1	0	5.0595	1.0302	0.15428	28.2287	0.63358	3.1302	28.2287	0.06720	0.02504
	2		2.7025	0.09351	16.1099	0.33496	1.4679	5.6199	0.33496	0.55468	26.8342
	3		2.9999	0.2898	16.1381	28.1521	1.2347	0.51309	28.1521	8.9787	0.03030
	4		0.22394	1.0852	0.01094	3.2781	2.1888	1.2713	3.2781	2.2301	0.02494
	5		28.2597	0.15391	15.3578	4.42	1.2908	28.1758	4.42	2.9323	0.50422
yyy	1	0	0.0028185	0.0032071	0.21129	0.0079421	0.0046937	0.0035018	0.0079421	0.0083993	5.70E-05
	2		0.0078915	0.005237	4.684	0.0039248	0.0092994	0.0040872	0.0039248	0.0043632	6.49E-05
	3		0.0082295	0.0070498	0.65814	0.0071398	0.0061981	0.0031065	0.0071398	0.0051079	3.75E-05
	4		0.0026498	0.00038786	1.2805	0.0031112	0.0051964	0.0017886	0.0031112	0.0053033	4.10E-05
	5		0.0069716	0.0023814	1.10E-05	0.0041485	0.0068288	0.0039413	0.0076717	0.013709	6.01E-05
F7	1	0	0.00017136	0.00017765	0.04975	0.0003222	4.44E-05	3.47E-05	0.00032227	0.0001375	0.0001596
	2		3.16E-05	0.00019332	0.0096608	0.00026656	0.00014147	0.00032235	0.00029389	0.00028793	0.00015071
	3		0.0001055	0.00015583	0.015933	0.0005641	0.00021668	8.67E-05	0.0005641	0.0003751	4.72E-05
	4		0.0003279	8.12E-06	0.0474	0.0004275	0.00015481	0.0001298	0.00042759	0.0001863	6.01E-06
	5		0.0001248	2.68E-05	0.0030962	0.0001135	5.47E-05	0.0003885	0.00011354	0.0005044	2.88E-06
F8	1	-12,569	-12,568.589	-12,569.337	-12,563.69	-12,569.426	-12,569.4295	-12,568.447	-12,569.426	-12,569.405	-12,569.483
	2		-12,569.088	-12,569.364	-12,554.71	-12,569.422	-12,569.1176	-12,569.301	-12,569.422	-12,569.255	-12,569.483
	3		-12,569.141	-12,569.423	-12,569.48	-12,569.465	-12,569.087	-12,568.318	-12,569.465	-12,568.722	-12,569.483
	4		-12,568.910	-12,568.803	-12,567.72	-12,569.135	-12,568.947	-12,568.002	-12,569.135	-12,569.258	-12,569.483
	5		-12,568.991	-12,569.040	-12,549.79	-12,568.748	-12,568.687	-12,569.473	-12,568.748	-12,569.313	-12,569.483
F9	1	0	0	0	0.05097	0	0	0	0	0	0
	2		0	0	0	0	0	0	0	0	0
	3		0	0	0	0	0	0	0	0	0
	4		0	0	0	0	0	0	0	0	0
	5		0	0	0.02342	0	0	0	0	0	0
F10	1	0	8.88E-16	8.88E-16	0.59841	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16
	2		8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16
	3		8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16
	4		8.88E-16	8.88E-16	9.73E-06	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16
	5		8.88E-16	8.88E-16	3.65E-10	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16	8.88E-16
F11	1	0	0.00E+00	0.00E+00	1.02E-07	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	2		0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	3		0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	4		0.00E+00	0.00E+00	0.8752	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
	5		0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
F12	1	0	0.00019093	7.38E-05	0.0027443	0.0021737	0.011757	0.0010932	0.0021737	0.0036851	8.61E-06
	2		0.00060494	0.0034431	0.0001773	0.0014124	0.0010992	0.00011355	0.0014124	0.00056797	8.36E-06
	3		0.0012813	4.07E-05	0.0083147	0.0033076	0.001464	0.0027819	0.0058759	0.00070109	0.00059161
	4		0.0032576	0.0019526	0.0035314	0.0044847	0.012298	0.0015544	0.003753	0.00024023	0.00058589
	5		0.00030893	0.011637	0.067425	0.007021	4.36E-05	0.0093717	0.010762	0.000781	0.0032647
F13	1	0	0.0051604	0.0013586	0.30418	0.0017013	0.0060032	0.013291	0.0062827	0.010571	6.34E-05
	2		0.001026	0.0026104	0.0002713	0.0010591	0.0015065	0.00015428	0.0030661	0.0037858	3.53E-05
	3		0.004871	0.0017761	0.026338	0.0026932	0.0019489	0.0020771	0.0024803	0.0033837	0.087591
	4		0.00084449	0.0028194	0.0071279	0.005576	0.0042496	0.0010457	0.00012182	0.0065913	0.00021984
	5		0.0071283	0.0058296	0.0004936	0.0065373	0.0079408	0.0001381	0.00015165	0.018941	8.65E-05
	No. of functions reaches 0		**4**	3	0	**4**	2	**4**	**4**	**4**	**6**

where the objectives were evaluated, and the optimal solutions were represented through Pareto fronts. OBL is occasionally introduced to improve the algorithm’s ability to explore the search space more effectively.

Naik et al. [10] proposed an adaptive technique for deciding whether to apply OBL. Furthermore, the algorithm improves exploitation by replacing a randomly selected search agent with the best-performing one during the position update process. In [17], the authors proposed ESMA, an improved Slime Mould Algorithm that fuses several techniques to enhance performance. The inclusion of Lévy flights and



(a) (b) (c) (d)

Fig. 1. (a) Convergence curve for f1, (b) Convergence curve for f2, (c) Convergence curve for f3, (d) Convergence curve for f4 obtained by 9 SMA versions for benchmark functions (500 iterations).

selective averaging during the exploration stage enhances adaptability, while a dynamic lens learning method helps reposition the elite solution, steering the swarm toward optimal regions. To this end, the proposed study and the previously introduced literature share the goal of preventing the algorithm from becoming caught in local optima. The difference is in the method utilized. For example, in [14], dynamic random search techniques improve the algorithm's search efficiency. In contrast, AOSMA reduces convergence by employing opposite-based learning and an adaptive decision technique.

2.1. *Difference between the proposed combination of OBL and SMA and the related work*

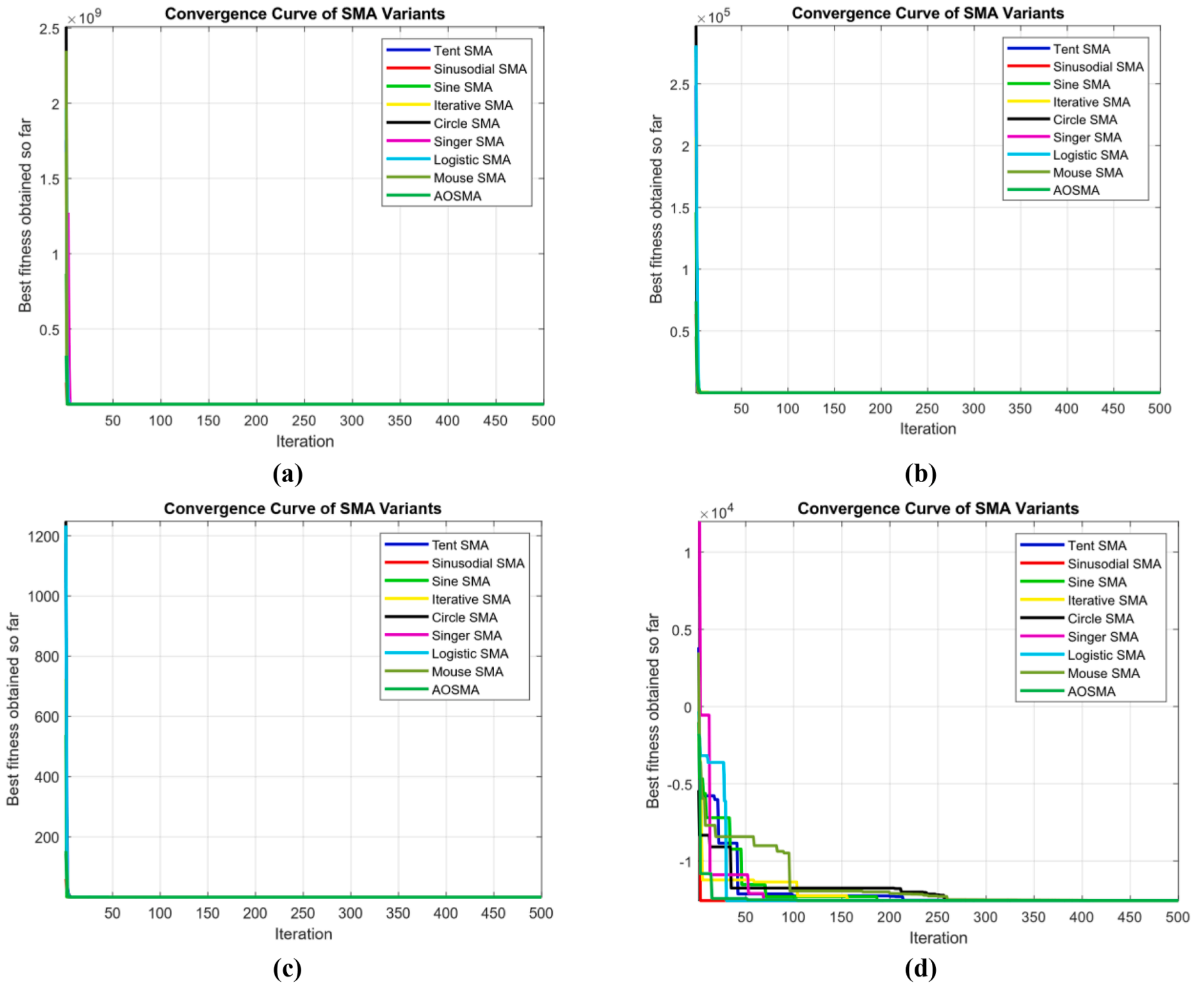
Firstly, during the application stage of OBL, the most common point of difference is when OBL is applied within the algorithm's lifecycle. Starting in the initialization phase only: Some studies apply OBL only to the initial population. An "opposite" population is generated, and the best members from both the original and opposite populations are selected to form a more diverse and high-quality starting point. Followed by initialization and generational update, other papers apply OBL not only at the start but also during the iterative process of the algorithm. After the population is updated in a given iteration, OBL can be used to generate opposite solutions for the new population. This helps to prevent stagnation and escape local optima. Finally, the condition-based application: A more sophisticated approach is to apply OBL based on certain conditions, such as population stagnation or slow convergence. This ensures the more computationally intensive OBL operation is only used when needed. Tables 1 and 2 summarize the key comparisons and the novelty of the proposed work compared to the existing literature.

As demonstrated in Ref. [18], AOSMA can be used to address real-world issues described in the abstract. These issues can also be included as benchmark instances. For instance, in (18), the traditional SMA is combined with two more search techniques. While the second uses the Lévy flight distribution (LFD), a real-world problem, to enhance exploration in the early phases and exploitation in the latter stages of the search process, the first uses opposition-based learning (OBL) to speed up convergence.

3. Proposed model

By combining OBL and adaptive control mechanisms, AOSMA significantly improves upon the original SMA. This improves the algorithm's convergence speed, solution diversity, and fitness value in challenging optimization tasks. This section introduces the proposed AOSMA version. The majority of SMA variations with OBL either apply it blindly at every iteration or concentrate just on exploitation without an adaptive rule, which makes it easy to identify the specific research need in the suggested strategy. None combines a straightforward replacement technique with an adaptive OBL choice. This work introduces both to close that gap.

In terms of conceptual analysis, AOSMA enhances the standard SMA by incorporating adaptive opposition-based learning, which generates opposite candidate solutions during the search to maintain diversity and achieve a more effective balance between exploration and exploitation. By contrast, OJESMA combines the Equilibrium Optimizer with joint opposite selection and opposition-based learning, forming a more



(a)

(b)

(c)

(d)

Fig. 2. (a) Convergence curve for f5, (b) Convergence curve for f6, (c) Convergence curve for f7, (d) Convergence curve for f8 obtained by 9 SMA versions for benchmark functions (500 iterations).

advanced hybrid framework that improves convergence stability and accelerates optimization. QCMSMA introduces a Cauchy-based adaptive search operator alongside Bloch sphere-based elite population initialization, thereby boosting exploration capability and enhancing population diversity from the outset. Meanwhile, SMA-AGDE integrates SMA with adaptive guided differential evolution, leveraging DE mutation strategies to strengthen local search performance, foster diversity, and reduce the likelihood of premature convergence.

The fitness function of an optimization or evolutionary problem, written as f(x), is a statistic that assesses how well a particular solution (represented by x) performs or how close it is to achieving the intended goals. Assigning a numerical score (fitness value) to every possible solution is a fundamental part of evolutionary and genetic algorithms; higher scores generally signal better performance and are more likely to be chosen for subsequent generations.

3.1. Mathematical formulation of the AOSMA

Assume that the search space contains N slime mould agents operating within a search space bounded by an upper limit (UB) and a lower limit (LB). The i^{th} slime mould's location in a *d*-dimensional search space is given by $X_i = (X_i^1, X_i^2, \ldots, X_i^d)$, where *i* ranges from 1 to N. Its fitness, also referred to as odor, is denoted by $f(X_i)$. Therefore, at iteration t, the positions and fitness values of the **N** slime mould agents at the current iteration **t** can be represented as follows:

$$X(t) = \begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^d \\ x_2^1 & x_2^2 & \ldots & x_2^d \\ \vdots & \vdots & \vdots & \vdots \\ x_N^1 & x_N^2 & \ldots & x_N^d \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_N \\ x_N \end{bmatrix} \forall N \in \mathbb{R} \tag{1}$$

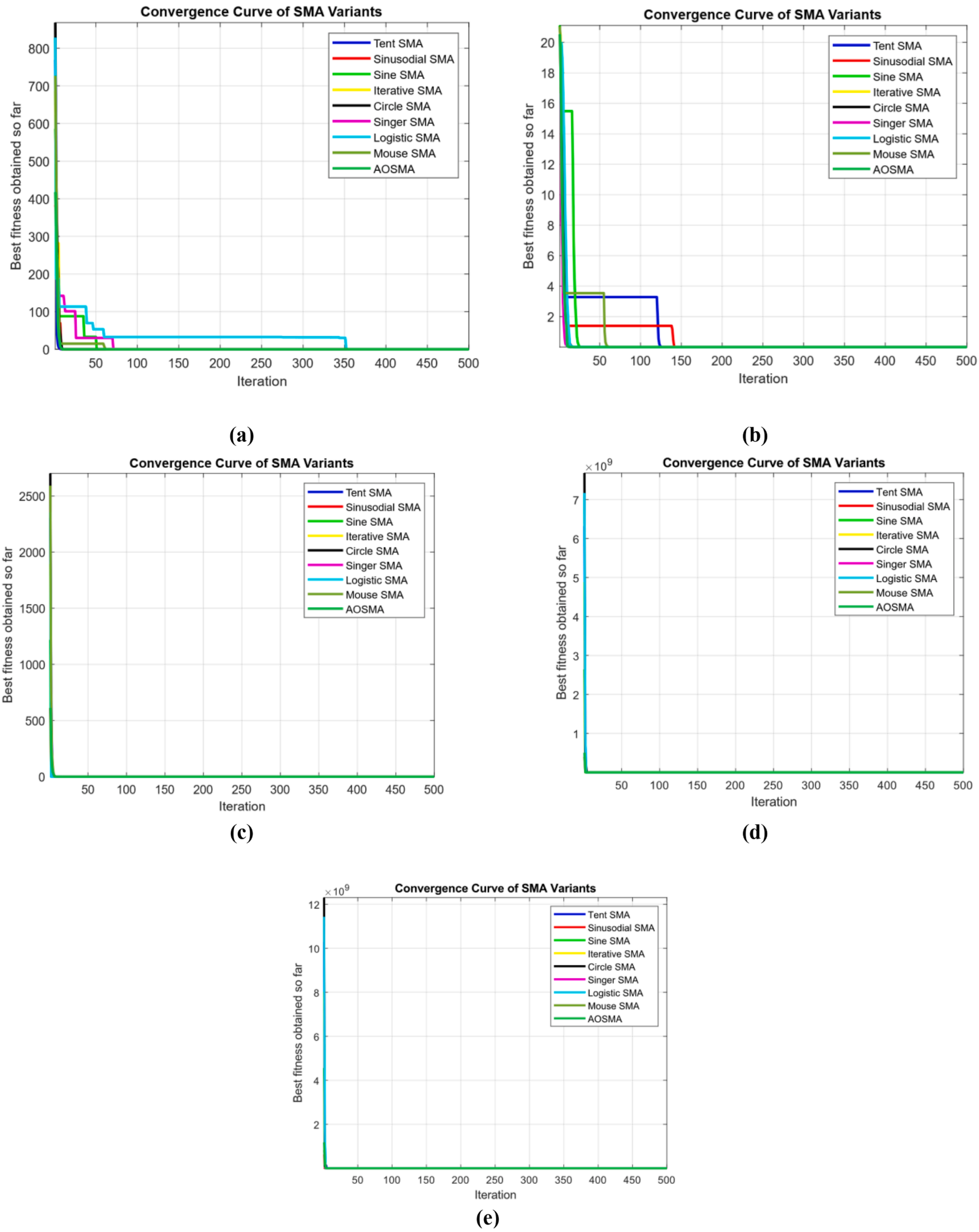
$$f\ (X) = [f\ (X_1),\ f\ (X_2),\ f\ (X_3), \cdots, f\ (X_N)]\ \forall N \in \mathbb{R} \tag{2}$$

The position of slime mould for the subsequent iteration $(t + 1)$ in SMA is updated using Eq. (3).

$$X(t+1) = \begin{cases} X_{LB}(t) + V_b(W.X_A(t) - X_B(t),\ if\ r_1 \geq \delta\ and\ r_2 < p_i \\ V_c.X_i(t),\ if\ r_1 \geq \delta\ and\ r_2 \geq p_i \\ rand.\ (UB - LB) + LB,\ if\ r_1 < \delta \end{cases} \quad \forall\ i \in [1, N] \tag{3}$$

where X_{LB} refers to the best-performing slime mould in the current iteration, X_A and X_B are two randomly chosen individuals from the current population, W acts as the weighting factor, V_b and V_c are random velocity terms drawn from uniform distributions over the ranges [–b, b] and [–c, c], and r_1 and r_2 are random values in the interval [0, 1]. The parameters b and c are updated at each iteration t using the following equations.

$$b = \operatorname{arctan}h\left(-\left(\frac{t}{T} + 1\right)\right. \tag{4}$$



(a)

(b)

(c)

(d)

(e)

Fig. 3. (a) Convergence curve for f9, (b) Convergence curve for f10, (c) Convergence curve for f11, (d) Convergence curve for f12, (e) Convergence curve for f13 4 obtained by 9 SMA versions for benchmark functions (500 iterations).

$$c = 1 - \frac{t}{T} \tag{5}$$

where the maximum iteration is denoted by T. A fixed probability of 0.03 is assigned for randomly initializing a slime mould's position. The parameter p_i is a decision threshold for the i^{th} slime mould, which determines whether to update its position using the global best or retain its own position. It is calculated based on the fitness function f(x) as follows.

$$p_i = \tanh|f\ (X_i) - f_{GB}|,\ \forall\ i \in [1, N] \tag{6}$$

$$f_{GB} = f\ (X_{GB}) \tag{7}$$

where $f(x_i)$ is the fitness of the i^{th} agent X_i, and f_{GB} is the best global fitness. At the i^{th} iteration, the weight W for all N slime moulds is determined using the following equation.

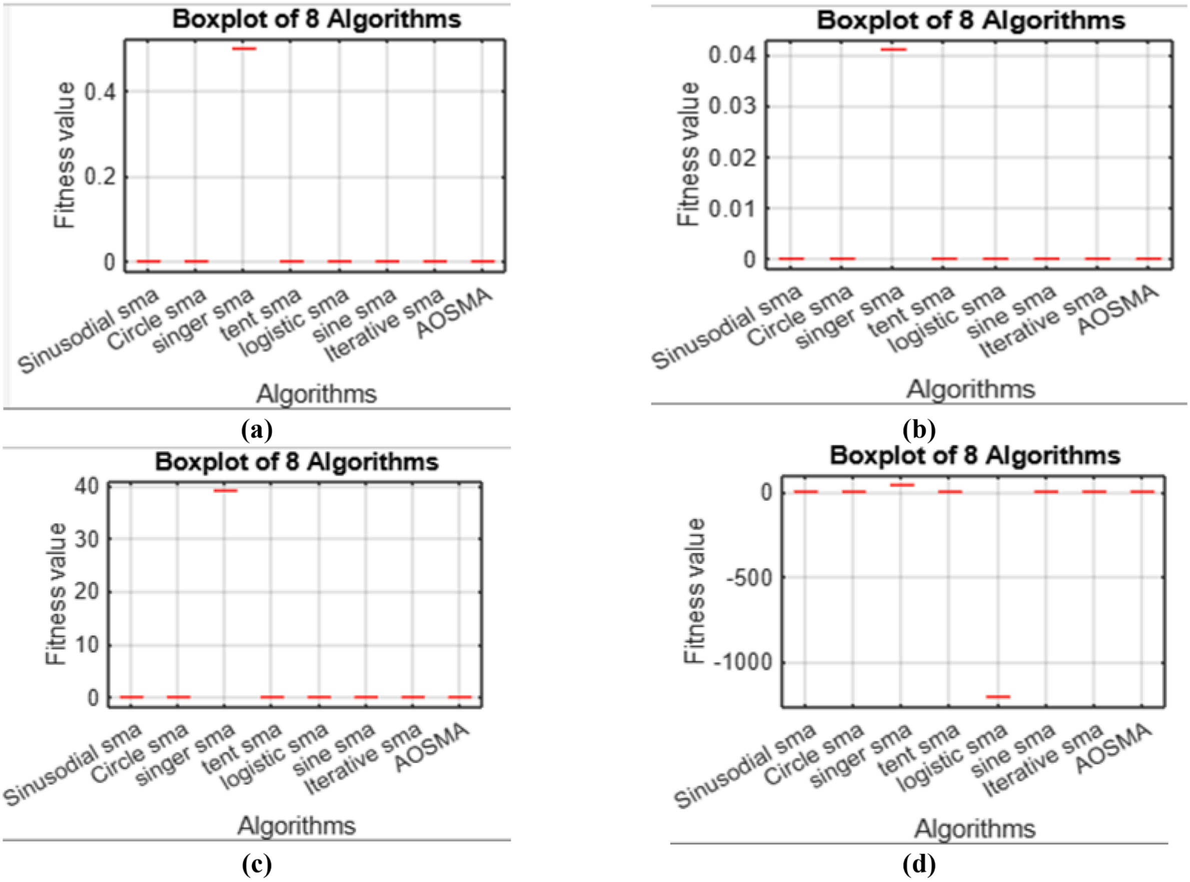


Fig. 4. (a) Boxplot for f1, (b) Boxplot for f2, (c) Boxplot for f3, (d) Boxplot for f4.

$$W\left(SortInd_f(i)\right) = \begin{cases} 1 + rand.\log\left(\dfrac{f_{LB} - f(X_i)}{f_{LB} - f_{LW}} + 1\right), & 1 \le i \le \dfrac{N}{2} \\ 1 + rand.\log\left(\dfrac{f_{LB} - f(X_i)}{f_{LB} - f_{LW}} + 1\right), & \dfrac{N}{2} < i \le N \end{cases}, \ \forall\, N \in \mathbb{R} \tag{8}$$

The terms f_{LB} and f_{LW} refer to the best and worst fitness values within the local population. For minimization tasks, the fitness values are sorted in increasing order.

$$\left[Sort_f,\ SortInd_f\right] = sort(f) \tag{9}$$

The optimal fitness value within the local region, along with the associated best-performing individual X_{LB}, is determined as follows.

$$f_{LB} = f\left(Sort_f(1)\right) \tag{10}$$

$$X_{LB} = X\left(SortInd_f(1)\right) \tag{11}$$

The local worst fitness f_{lw} is concluded as follows.

$$f_{LW} = f\left(Sort_f(N)\right) \tag{12}$$

3.2. Opposition-based learning

In OBL, for each slime mould (i = 1, 2, …, N), an opposite position X_{oi} is estimated relative to its current location X_{ni} in the search space. This value is compared to determine whether it offers a better solution for the next iteration, thus reducing the risk of getting trapped in local optima and enhancing convergence speed. The computation of X_{oi} in the j^{th} dimension is as follows.

$$X_{oij}(t) = \min(X_{ni}(t)) + \max\,(X_{ni}(t)) - X_{nij}(t) \tag{13}$$

3.3. Adaptive decision strategy

If the slime mould progresses along a path with decreasing nutrient quality, the algorithm adaptively responds using the current and previous fitness values within the AOSMA framework. Adaptive decision-making supports additional inquiry through OBL. AOSMA's adaptive decision approach is used to update the location for each subsequent iteration, as modeled below.

$$X_i(t+1) = \begin{cases} Xn_i(t) \text{ if } f(Xn_i(t)) \le f(X_i(t)), \\ Xs_i(t) \text{ if } f(Xn_i(t)) > f(X_i(t)), \end{cases} \quad \forall i \in [1, N] \tag{14}$$

$$Xs_i(t) = \begin{cases} Xo_i(t) \text{ if } f(Xo_i(t)) < f(Xn_i(t)) \\ Xn_i(t) \text{ if } f(Xo_i(t)) \ge f(Xn_i(t)) \end{cases} \tag{15}$$

4. Results and discussion

This section presents a comparative analysis between the proposed approach and other competing algorithms. All methods were tested under consistent conditions with a population size of 30 and 500 iterations, repeated five times. The proposed AOSMA algorithm was coded in MATLAB and executed on a machine with an Intel Core i7 2.20 GHz CPU and 12 GB RAM, using thirteen benchmark functions for evaluation.

The metrics that are utilized are Agent fitness values (f(X_i)), thresholds applied, and a decision threshold (p_i) that governs whether an agent maintains its current position or updates it toward the global best. To preserve diversity and prevent stagnation, a fixed probability of 0.03 is also used to reinitialize an agent's position randomly. Additionally, UB and LB vary for every function. Reassessment Frequency indicates that all agents' adaptive decisions are reassessed at each iteration. The

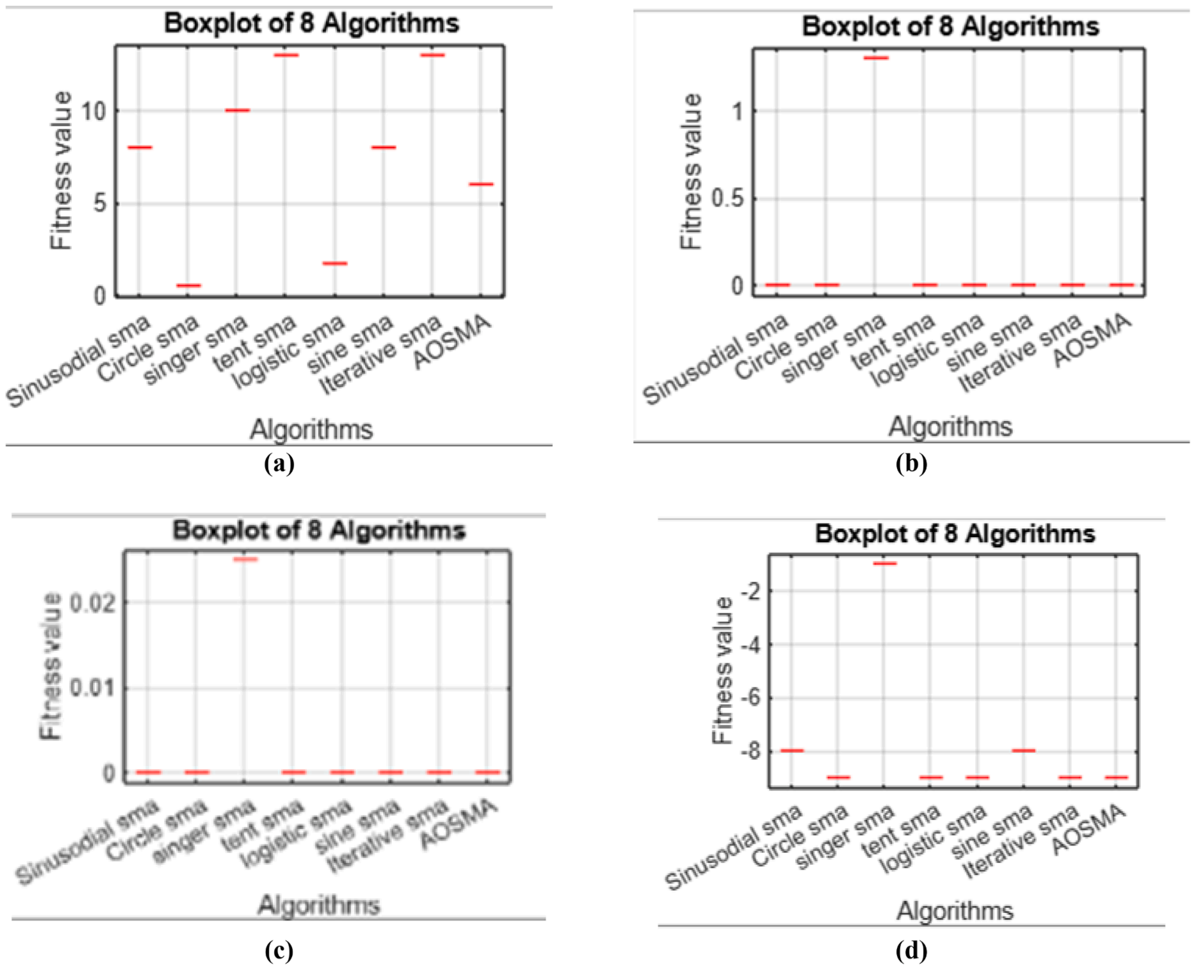


Fig. 5. (a) Boxplot for f5, (b) Boxplot for f6, (c) Boxplot for f7, (d) Boxplot for f8.

algorithm determines whether to initiate OBL or carry on with regular position updates at each iteration by comparing the current fitness to the prior fitness.

4.1. *Simulation setup*

Analysis of time and space complexity of an algorithm plays a central role in computational studies and practical applications, as it provides significant insights into their efficiency and scalability. Time complexity directly affects the algorithm's feasibility for large-scale or real-time problems. Similarly, space complexity quantifies the time of computation required in terms of memory, an essential aspect in discussing resource constraints, especially in embedded systems, big data systems, or limited-memory devices. Initially, the proposed AOSMA algorithm generates N search agents, each with a dimensionality of D, resulting in an initialization time complexity of O(N × D). The algorithm then evaluates the fitness of every agent throughout the optimization process, contributing to a total complexity of O(t_{max}·N·D), where t_{max} is the maximum number of iterations. Furthermore, the AOSMA requires O (N·D) space, accounting for the storage of all search agents and their dimensions.

Experimental evaluations are conducted on thirteen conventional benchmarks. Table 3 summarizes mathematical formulations for the typical considered benchmarks. Unimodal benchmarks are commonly used to assess algorithms' exploitation capabilities due to their one global optimum. Basic functions are more difficult to manage than unimodal benchmarks due to their large number of local optima that can be used to measure exploration capabilities. Metaheuristics' ability to solve real-world engineering challenges can be estimated using common benchmarks. To ensure a fair comparison, 500 total iterations are used as the criterion condition, with a swarm size of 30. Eight methods are used to compare with AOSMA. To minimize the impact of chance on the experiment, all algorithms are conducted five times.

The selection of the 13 benchmark functions was guided by the need for a standardized, transparent, and reproducible framework to evaluate optimization algorithms across diverse problem characteristics. These benchmarks, including the widely used sphere function, were chosen for their prevalence in the optimization literature and the variety of mathematical properties they embody. Collectively, they cover separable and non-separable variables, unimodal and multimodal landscapes, and varying dimensionalities, providing a comprehensive test bed to assess the balance between exploration and exploitation in metaheuristic algorithms. Beyond their theoretical value, many of these functions capture features commonly encountered in real-world applications. For instance, the sphere function models convex and smooth landscapes typical of engineering design tasks requiring rapid convergence, while multimodal functions such as Rastrigin and Ackley emulate the complex, rugged search spaces found in scheduling, routing, and resource allocation problems.

F6 is not a special case of F7. $\sum_{i=1}^{n}(X_i+0.5)^2$ is not a special case of $\sum_{i=1}^{n}\left(X_i^4+rand\,(0,1)\right)$ because the first expression is a deterministic mathematical operation that squares a value, while the second is a function that adds a random floating-point number to a constant. The expression $(X_i+0.5)^2$ performs addition and multiplication by 2, whereas $\sum_{i=1}^{n}(X_i^4+\,rand\,(0,1)$generates a random number between

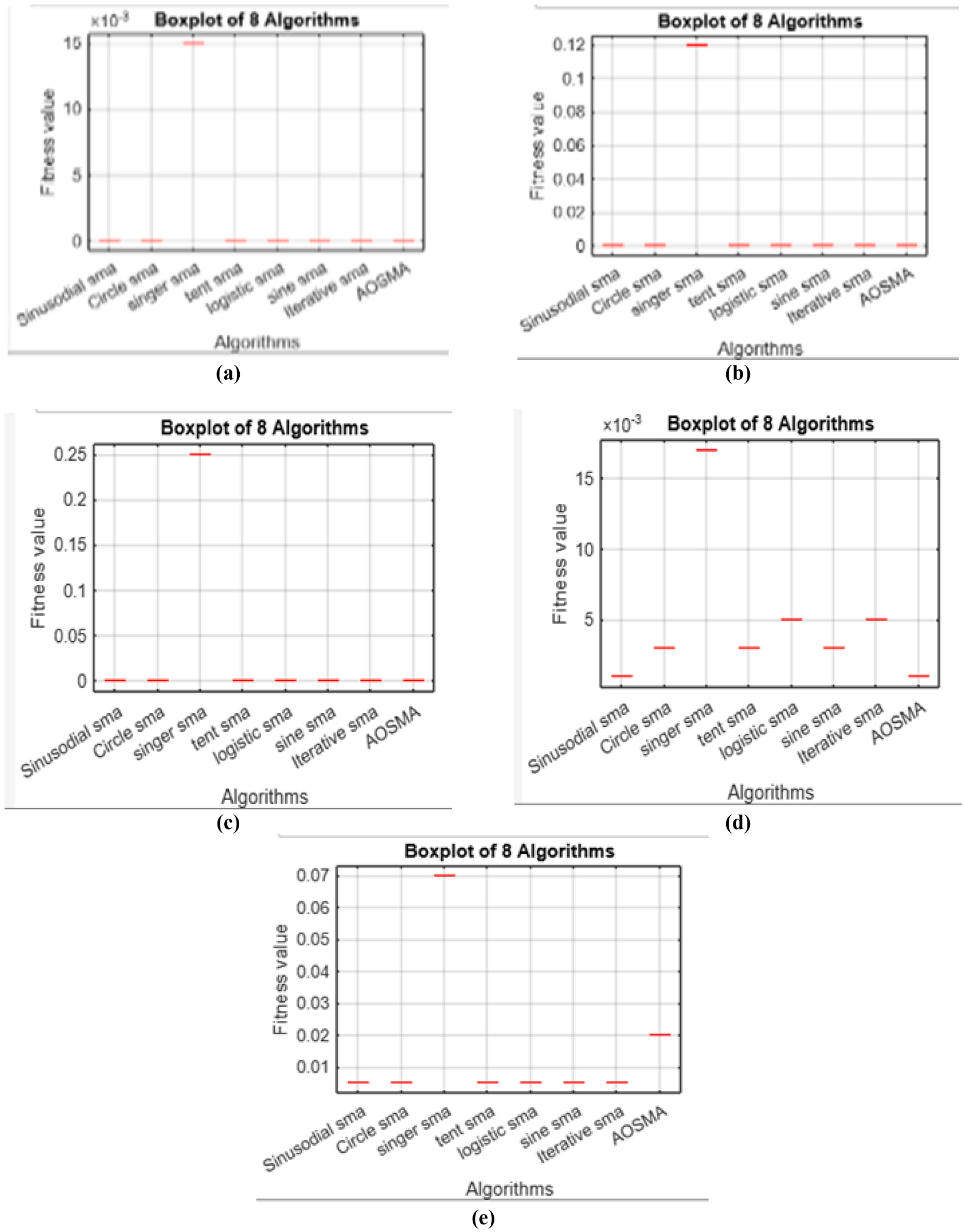
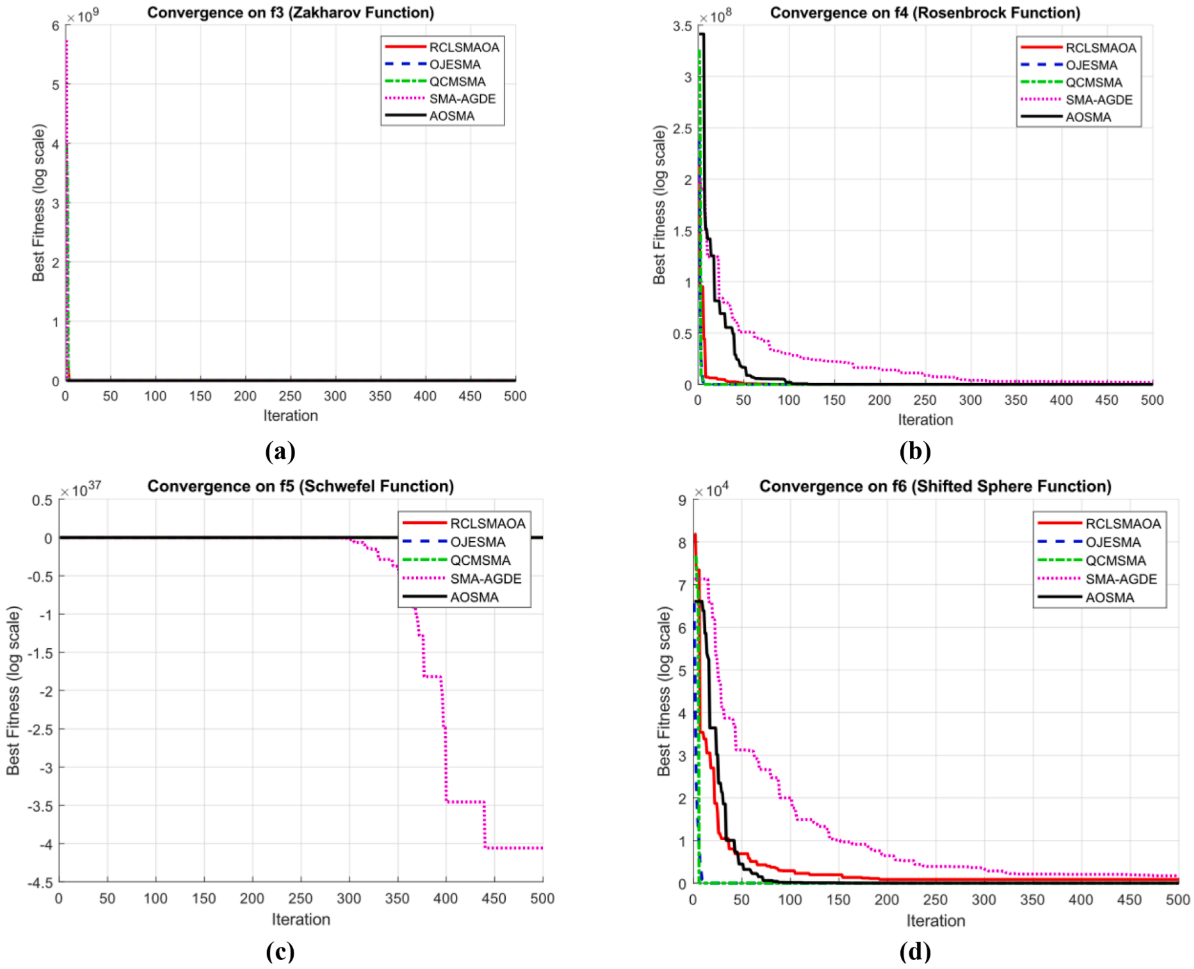


Fig. 6. (a) Boxplot for f9, (b) Boxplot for f10, (c) Boxplot for f11, (d) Boxplot for f12, (e) Boxplot for f13.

0 and 1 and adds it to the value of 1. Additionally, F7 cannot substitute F6 because they test different algorithmic properties. Both F6 (shifted sphere) and F7 (quartic with stochastic term) remain unchanged because they capture different features: simple convex convergence versus higher-order, noisy landscapes: F6 cannot replace F7 since it tests basic convergence behavior, whereas F7 analyzes resistance to noise and sensitivity to large variable magnitudes.

Parameters used in F6 can be explained as following, the shift parameter α=0.5 has been introduced in F6 in order to move the global minimum farther from the origin. Without the shift, the function simplifies to the conventional sphere model, whose symmetric and centered terrain makes it straightforward for many optimizers. The issue is transformed into a shifted-sphere function by setting α=0.5, which guarantees that optimizers cannot take advantage of origin symmetry and must instead show that they can find minima in a displaced search space. The selected value adds significant diversity to the benchmark



(a)

(b)

(c)

(d)

Fig. 7. Performance of the proposed AOSMA compared with the other four SMA versions [For F3-F6]. (a) Convergence curve for F3, (b) Convergence curve for F4, (c) Convergence curve for F5, (d) Convergence curve for F6, obtained by 5 SMA versions for benchmark functions (500 iterations).

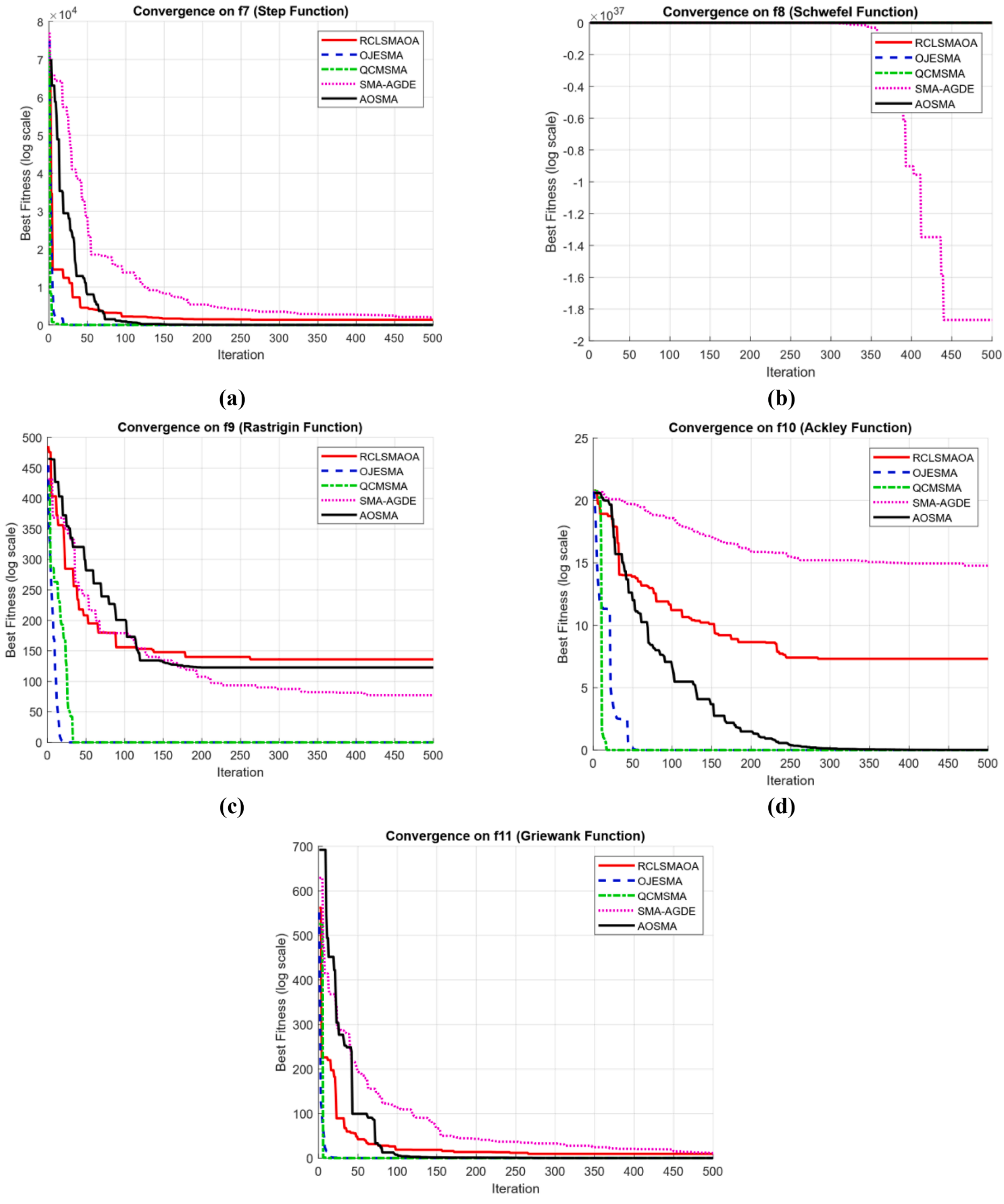
suite while maintaining the function's convexity and simplicity. On the other hand, common shifts in benchmark functions include randomly generated shift vectors or fixed values like α=1. Random shifts offer greater diversity across runs, although larger shifts (e.g., α=1) increase the distance from the origin. In contrast to these, the reasonable selection of α=0.5 maintains the function's simplicity while making it nontrivial, balancing interpretability with difficulty. If α=0 it will be identical to F1.

To create tiny stochastic perturbations that mimic noise while maintaining the function's numerical stability and boundedness, the randomization in F7 is defined inside the interval (0,1). By making this decision, the random term is guaranteed to behave as a secondary disturbance rather than overpowering the function's predictable portion. The noise amplitude would significantly grow if the randomization were scaled to wider ranges, such (0100) or (0,1000). This might obscure the underlying landscape structure, skew the optimization difficulty, and possibly result in an unstable or deceptive performance evaluation. Because it strikes a compromise between realism and regulated complexity, the (0,1) interval is thus a common and suitable option.

4.2. Simulation results

This section compares the proposed AOSMA to the chaotic SMA on thirteen benchmark problems to show its effectiveness and competitiveness. Table 4 shows the experimental results for eight algorithms and the proposed AOSMA on thirteen typical benchmarks. AOSMA improves the performance of chaotic SMA on the unimodal tests F2, F3, and F4. Furthermore, the benefits of AOSMA over the chaotic SMA are retained. In most circumstances, AOSMA outperforms chaotic SMA in terms of competitiveness. AOSMA consistently achieves quicker convergence and greater ability to reach the global optimum compared to the chaotic SMA. The advantages of applying AOSMA in optimization are clearly presented in Table 4. It shows that AOSMA improves performance with 6 functions as F1, F2, F3, F4, F9, and F11. Additionally, this analysis shows its superior performance when compared to Mouse SMA, Iterative SMA, and the other remaining chaotic versions of SMA. While Mouse SMA, Iterative SMA, Sine SMA, Tent SMA, and Sinusoidal SMA improve performance on functions F1, F3, F9, and F11.

Figs. 1–3 present the convergence curve of the AOSMA and the other considered algorithms for the previously introduced thirteen benchmarks. Boxplots are utilized to assess the fitness performance of chaotic maps. The red line indicates the median fitness value of each algorithm. AOSMA shows the most favorable results with the lowest median, whereas Singer SMA records the highest. In Fig. 1(a), all SMA variants show a sharp decrease in fitness within the first 20–30 iterations, reflecting their strong exploratory behavior and their ability to identify promising areas of the search space rapidly. This rapid early improvement is characteristic of SMA-based algorithms, which naturally employ oscillatory movement patterns that enhance initial exploration. By approximately the 50th iteration, almost all methods, including the proposed AOSMA, achieve fitness values close to zero. This suggests that the benchmark problem is relatively straightforward for these algorithms, allowing them to converge quickly toward a near-optimal



(a)

(b)

(c)

(d)

(e)

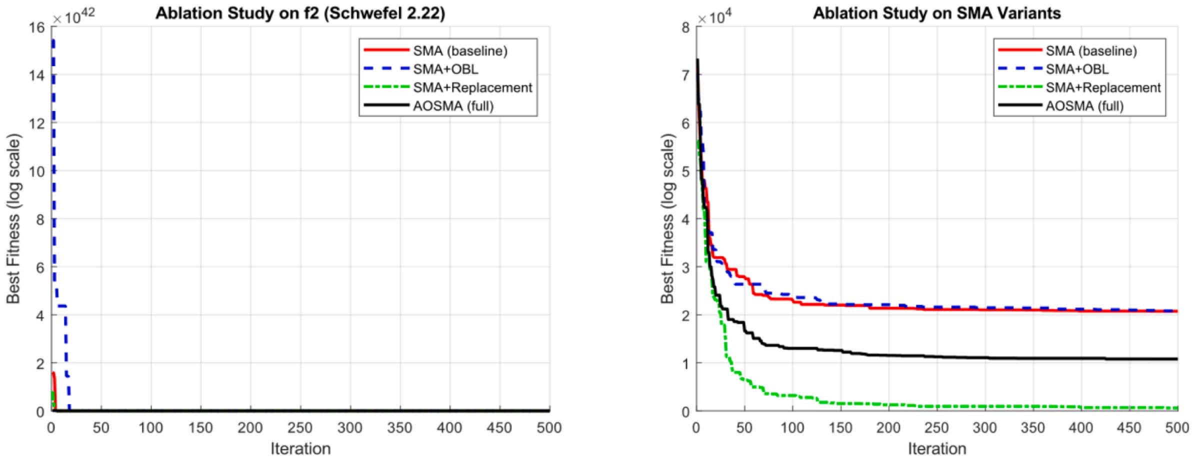
Fig. 8. Performance of the proposed AOSMA compared with the other four SMA versions [For F7-F11]. (a) Convergence curve for F7, (b) Convergence curve for F8, (c) Convergence curve for F9, (d) Convergence curve for F10, (e) Convergence curve for F11, obtained by 5 SMA versions for benchmark functions (500 iterations).

solution.

Overall, the figure shows that all SMA variants, including AOSMA, efficiently solve the benchmark problem, achieving fast convergence and almost identical final fitness values. AOSMA performs comparably to the top variants, confirming its effectiveness and indicating that this test function does not strongly distinguish performance differences among the SMA methods.

In Fig. 2(d), most SMA variants show a steep decline from high initial fitness values to significantly lower ones within the first 30–70 iterations, indicating that they quickly move away from inferior starting points and identify promising regions early in the search. Each curve corresponds to a specific SMA variant. Some variants achieve low fitness values rapidly and then level off, while others progress more gradually. Tent SMA, Sinusoidal SMA, and Sine SMA (red, yellow, and dark blue) exhibit rapid convergence followed by early stagnation, indicating intense short-term exploitation but limited long-term improvement. In contrast, Iterative SMA and Circle SMA exhibit similar convergence patterns but plateau at a later stage, reflecting a more moderate exploratory behavior.

Fig. 4-a presents the boxplot comparison of the fitness values derived



(a)

(b)

Fig. 9. Ablation study of four SMA versions at functions (a) F2 and (b) F1.

Table 5
Average wall-clock time per run and per iteration for AOSMA and SMA variants at sphere function F1, 30D, max iteration = 500.

Algorithm	Average runtime per run (s)	Average runtime per iteration (s)
Baseline SMA	0.4446	0.000889
AOSMA	0.6730	0.001346
Logistic SMA	0.4526	0.000905
Tent SMA	0.4576	0.000915
Sine SMA	0.6555	0.001311
Sinusoidal SMA	0.4628	0.000926
Singer SMA	0.4565	0.000913
Circle SMA	0.4530	0.000906
Mouse SMA	0.4481	0.000896
Iterative SMA	0.6288	0.001258

from eight distinct optimization techniques applied to the benchmark function F1. The y-axis shows the relevant fitness values, while the x-axis displays the algorithms: Sinusoidal SMA, Circle SMA, Singer SMA, Tent SMA, Logistic SMA, Sine SMA, Iterative SMA, and AOSMA. The graph clearly shows that the majority of algorithms perform well in optimization, showing very low or almost zero fitness values. In contrast to the others, the Circle SMA exhibits a noticeably higher spread and median value, indicating either worse performance or greater unpredictability. The AOSMA algorithm performs competitively and steadily among the compared approaches, as seen by its continually low fitness values.

Fig. 4-b presents a boxplot evaluating the effectiveness of eight optimization methods according to their fitness values using the benchmark F2. The algorithms, Sinusoidal SMA, Circle SMA, Singer SMA, Tent SMA, Logistic SMA, Sine SMA, Iterative SMA, and AOSMA, are represented on the x-axis, and the appropriate fitness values are displayed on the y-axis. The plot indicates effective optimization, as the majority of methods obtain very low fitness values. With a noticeably higher median and interquartile range, the Singer SMA stands out, indicating worse performance and more variability. Algorithms with minimum variance and lower fitness results, such as AOSMA and Tent SMA, on the other hand, exhibit excellent optimization capability and great stability.

Fig. 4-c depicts a boxplot of fitness values for eight optimization techniques, allowing a comparative examination of their performance for the F3 benchmark function. According to the statistics, Singer SMA generates a median fitness value that is significantly larger and has a broader range, which suggests less efficient optimization and greater performance variability. On the other hand, the remaining methods consistently obtain low fitness values, indicating higher optimization performance and robustness, especially AOSMA, Iterative SMA, and Tent SMA. This demonstrates that, out of the algorithms compared, AOSMA is among the most dependable and efficient. Fig. 4-d highlights the performance of eight optimization methods on the F4 benchmark function by comparing them in a boxplot according to their fitness values. With a noticeably lower (more negative) fitness value, the Logistic SMA sticks out from the plot, suggesting inadequate performance in this instance. The other methods, on the other hand, consistently maintain fitness values near zero, indicating superior optimization results. AOSMA and Iterative SMA are dependable options for addressing the optimization problem depicted because of their robust and constant performance.

Fig. 5 presents the fitness scores for the eight considered algorithms, for the benchmarks F5, F6, F7, and F8. Also, Fig. 6 presents the fitness scores for the benchmarks F9, F10, F11, F12, and F13. The algorithms are shown on the x-axis, and the fitness values, which reveal how well each algorithm optimizes, are displayed on the y-axis. The data shows that the algorithms differ noticeably from one another. With the lowest median fitness values, Circle SMA and Sine SMA appear to perform better in this situation. On the other hand, Singer SMA and AOSMA perform comparatively poorly on the examined task, as evidenced by their larger median fitness scores. The variations in the boxplots show how the algorithms differ in terms of robustness and stability.

For further evaluation, the proposed AOSMA was compared to four other SMA versions. Figs. 7 and 8 present the results of the proposed AOSMA compared to the other four SMA versions for benchmark functions of F3-F11. Every algorithm exhibits a distinct decreasing trend, indicating that as the number of iterations increases, the quality of the solutions improves. The magenta (SMA-AGDE) and green (QCMSMA) curves exhibit remarkably rapid convergence, achieving very low fitness values early on and maintaining consistent gains throughout. Throughout, the blue (QJESMA) curve shows a steady decline that smoothly converges to lower fitness values. As iterations progress, the black (AOSMA) and red (RCLSMOA) curves exhibit stable convergence behavior, rapidly declining from their initial higher values before stabilizing. The convergence curves of five algorithms over 500 iterations on the Rosenbrock function (f4) are displayed in this picture.

Every algorithm begins with extremely high fitness values, ranging from 10 to 8, and drops off quickly in the initial repetitions. Within the first 50 iterations, RCLSMAOA, OJESMA, and QCMSMA show rapid convergence, achieving values that are nearly zero. SMA-AGDE improves progressively over iterations and converges more slowly. AOSMA exhibits competitive performance, with a smooth transition to near-optimal values and a distinct, steady decline. The Rosenbrock function is handled well by all approaches overall, albeit there are variations in stability and convergence rate. OJESMA, QCMSMA, and RCLSMAOA all perform extremely steadily and swiftly approach near-optimal values. SMA-AGDE has a greater fitness trajectory over the course of the iterations and converges significantly more slowly. AOSMA exhibits a balance: it performs competitively and converges more smoothly than SMA-AGDE; however, it lags somewhat behind RCLSMAOA and QCMSMA in terms of final precision.

AOSMA's parameters were selected to maintain a balanced search behavior while ensuring practical implementation and reproducibility, using standard SMA defaults to minimize manual tuning. Uniform settings for population size, iteration count, and reinitialization probability produced stable performance across benchmarks. Although initial findings indicate that excessively large populations or frequent OBL activations offer limited additional benefit due to the algorithm's inherent diversity mechanisms, further sensitivity studies, such as exploring different population sizes or OBL activation strategies, would provide more quantitative insight and help confirm the robustness of these parameter choices. Incorporating real-world optimization case studies would further strengthen the manuscript's practical impact, as benchmark functions do not fully capture the constraints, noise, and interdependencies present in actual applications. Since AOSMA's adaptive opposition learning and best-agent replacement strategies are particularly effective for complex and resource-constrained problems, evaluating the algorithm on tasks such as engineering design or feature selection would illustrate its practical performance and expand its applicability without altering its core framework.

AOSMA exhibits consistent performance as dimensionality grows, largely due to its adaptive opposition mechanism, which preserves population diversity, and the best-agent replacement strategy, which promotes faster convergence toward promising regions. In 30-dimensional benchmark evaluations, the algorithm consistently achieved competitive or superior outcomes across unimodal, multimodal, and composite problem classes. Its complexity scales linearly with dimension, $O(ND)$ per iteration and $O(T_{\max}ND)$overall, preventing exponential increases in runtime or degradation in solution quality. Consequently, AOSMA maintains robust search efficiency in higher-dimensional spaces without requiring specialized parameter tuning. Future work may extend these findings to ultra-high-dimensional settings (e.g., $D > 100$) or real-world feature selection tasks to further assess scalability.

Furthermore, Fig. 9 presents an ablation study of four SMA versions at functions F1 and F2. The red curve represents the SMA baseline. The standard SMA begins with high fitness values and decreases gradually, but plateaus at a higher level than other variants, reflecting weaker optimization ability. The blue dashed curve represents the SMA + OBL. Incorporating non-adaptive OBL yields only a slight improvement over the baseline, with a convergence trend that nearly overlaps it, indicating limited effectiveness. The green dash-dotted curve represents the SMA + replacement. This variant demonstrates the fastest and most pronounced improvement, quickly reaching lower fitness values and maintaining strong convergence, highlighting the effectiveness of the replacement mechanism. The black curve represents the proposed AOSMA. By integrating adaptive OBL with replacement, AOSMA achieves stable and competitive performance. Although its early progress is slower than SMA+ Replacement, it ultimately converges to the lowest fitness, offering the best balance between exploration and exploitation.

A non-parametric statistical test for comparing the performance of three or more algorithms (or treatments) across various problems or datasets is the Friedman Test, often known as Friedman's ANOVA. In optimization and machine learning research, it is very typical to seek to determine whether one algorithm (like AOSMA) performs noticeably better than others. Here's what the Friedman analysis (F1–F9 subset) shows:

a) Average ranks: AOSMA achieves the lowest average rank (best overall), indicating it tends to perform more strongly across functions.

b) Friedman test statistic: Since the p-value is 0.0002 (< 0.05), the differences among algorithms are statistically significant. This means we can reject the null hypothesis ("all algorithms are equal") and conclude that at least one algorithm, in this case, AOSMA, with the best average rank, is significantly better.

Table 5 presents the average wall-clock runtime per run and per iteration for the baseline SMA, its chaotic variants, and AOSMA on the Sphere benchmark function (30 dimensions, 30 agents, 500 iterations). The baseline SMA and most chaotic variants (Logistic, Tent, Sinusoidal, Singer, Circle, Mouse, Iterative) demonstrate relatively low runtimes, averaging around 0.44–0.46 s per run, with per-iteration costs on the order of 10^{-3} seconds. Within the chaotic variants, Sine SMA and Iterative SMA show slightly higher runtimes, reflecting the extra computational effort introduced by their nonlinear update dynamics.

In contrast, AOSMA records a moderately higher runtime of 0.673 s per run (0.001346 s per iteration). This increase is anticipated since AOSMA incorporates opposition-based learning and adaptive exploitation mechanisms, which add extra computational steps per iteration. Nonetheless, this modest selection is offset by AOSMA's superior optimization accuracy, stability, and convergence speed, as evidenced by the experimental results. Ultimately, the added computational cost represents a reasonable trade-off for achieving improved search balance and higher solution quality, reaffirming the effectiveness of AOSMA in enhancing slime mould-based optimization.

OBL increases computational overhead by effectively doubling the number of candidate solutions, both the original and their opposites, that must be evaluated and maintained. Similarly, replacement strategies, which determine which individuals advance to the next generation, introduce additional costs through comparison and selection operations. Although these mechanisms enhance optimization by boosting population diversity and reducing the risk of premature convergence, the extra computations required for generating and filtering solutions contribute to the overall runtime. As a result, they may be less practical for highly cost-sensitive optimization tasks or scenarios with limited computational resources.

Two additional mechanisms-opposition-based learning, which assesses opposing solutions, and the replacement strategy, which swaps out weaker agents for stronger ones- are responsible for the extra runtime in AOSMA. Compared to the entire optimization process, these phases are lightweight and add a little computational burden that scales linearly with issue size. Importantly, the overhead is a good trade-off because this minor expense results in substantial gains in exploration, exploitation, and solution quality.

5. Conclusion and future work

This study proposed an AOSMA aimed at addressing function optimization challenges. However, because the suggested algorithm makes an adaptive decision about whether to use the OBL, it shows superior exploration and exploitation. In conclusion, this approach reduces the possibility of misguiding the exploration phase in some cases by 50 % by using a single random search agent instead of the SMA. Additionally, this method has an adaptive mechanism built in to determine when to utilize OBL to limit the exploration period. Lastly, the proposed technique is thought to be helpful for function optimization to address practical engineering issues.

As a future work, we can combine with metaheuristics (e.g., PSO, DE, GA) to leverage strengths and overcome weaknesses (e.g., slow convergence or premature convergence). Additionally, including one or two state-of-the-art optimizers outside the SMA family, consider using larger benchmark sets or higher dimensionalities would better evaluate the proposed method's generalization and robustness. Also, this would offer deeper insights into the effectiveness of the proposed approach.

Data availability statement

The data that support the findings of this study are available in the manuscript. Further information related to the considered data can be obtained by reasonable request from the corresponding author.

Funding

This work was funded by Prince Sultan University.

CRediT authorship contribution statement

Elsayed Badr: Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mostafa Abdullah Ibrahim:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Diaa Salama:** Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mohammed ElAffendi:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Abdelhamied A Ateya:** Writing – review & editing, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mohamed Hammad:** Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Alaa Yassin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The Authors would like to thank Prince Sultan University for paying the APC of this article.

References

[1] M. Zhou, M. Cui, D. Xu, S. Zhu, Z. Zhao, A. Abusorrah, Evolutionary optimization methods for high-dimensional expensive problems: a survey, *IEEE/CAA J. Autom.* Sin. 11 (5) (2024) 1092–1105.

[2] E.H.H. Sumiea, et al., Enhanced deep deterministic policy gradient algorithm using grey wolf optimizer for continuous control tasks, IEEE Access 11 (2023) 139771–139784.

[3] E.H. Houssein, M.K. Saeed, G. Hu, M.M. Al-Sayed, Metaheuristics for solving global and engineering optimization problems: review, applications, open issues and challenges, Arch. Comput. Methods Eng. 31 (8) (2024) 4485–4519.

[4] R. Rishabh, K.N. Das, A critical review on metaheuristic algorithms based multi-criteria decision-making approaches and applications, Arch. Comput. Methods Eng. 32 (2) (2024) 963–993.

[5] F.S. Gharehchopogh, A. Ucan, T. Ibrikci, B. Arasteh, G. Isik, Slime Mould algorithm: a comprehensive survey of its variants and applications, Arch. Comput. Methods Eng. 30 (4) (2023) 2683–2723.

[6] T. Yu, J. Pan, Q. Qian, Improved slime mould algorithm by perfecting bionic-based mechanisms, Int. J. Bio-Inspired Comput. 22 (1) (2023) 1–15.

[7] H. Chen, Z. Wang, H. Jia, X. Zhou, L. Abualigah, Hybrid slime mold and arithmetic optimization algorithm with random center learning and restart mutation, Biomimetics 8 (5) (2023) 396 *(Basel)*.

[8] S. Larabi-Marie-Sainte, R. Alskireen, S. Alhalawani, Emerging applications of bio-inspired algorithms in image segmentation, Electronics 10 (24) (2021) 3116 *(Basel)*.

[9] I. Abunadi, et al., An automated glowworm swarm optimization with an inception-based deep convolutional neural network for COVID-19 diagnosis and classification, Healthcare 10 (4) (2022) 697 *(Basel)*.

[10] M.K. Naik, R. Panda, A. Abraham, Adaptive opposition slime mould algorithm, Soft Comput. 25 (22) (2021) 14297–14313.

[11] J. Yang, Y. Zhang, T. Jin, Z. Lei, Y. Todo, S. Gao, Maximum Lyapunov exponent-based multiple chaotic slime mold algorithm for real-world optimization, Sci. Rep. 13 (1) (2023) 12744.

[12] H. Chen, C. Li, M. Mafarja, A.A. Heidari, Y. Chen, Z. Cai, Slime mould algorithm: a comprehensive review of recent variants and applications, Int. J. Syst. Sci. 54 (1) (2023) 204–235.

[13] W.-C. Wang, W.-H. Tao, W.-C. Tian, H.-F. Zang, A multi-strategy slime mould algorithm for solving global optimization and engineering optimization problems, Evol. Intell. 17 (5–6) (2024) 3865–3889.

[14] Z. Duan, X. Qian, W. Song, Multi-strategy enhancde slime mould algorithm for optimization problems, IEEE Access 13 (2025) 1. –1.

[15] E.H. Houssein, M.A. Mahdy, M.J. Blondin, D. Shebl, W.M. Mohamed, Hybrid slime mould algorithm with adaptive guided differential evolution algorithm for combinatorial and global optimization problems, Expert Syst. Appl. 174 (114689) (2021) 114689.

[16] P.V.H. Son, L.N.Q. Khoi, Optimization in construction management using adaptive opposition slime mould algorithm, Adv. Civ. Eng. 2023 (2023) 1–20.

[17] W. Xiong, D. Li, D. Zhu, R. Li, Z. Lin, An enhanced slime mould algorithm combines multiple strategies, Axioms 12 (10) (2023) 907.

[18] Laith Abualigah, Ali Diabat, Mohamed Abd Elaziz, Improved slime mould algorithm by opposition-based learning and Levy flight distribution for global optimization and advances in real-world engineering problems, J. Ambient. Intell. Humaniz. Comput. 14 (2) (2023) 1163–1202.