

BenchCouncil Transactions

TBench

Volume 5, Issue 1

2025

on Benchmarks, Standards and Evaluations

Original Articles

- 🕒 **LLMs: A game-changer for software engineers?**
Md. Asraful Haque
- 🕒 **Evaluatology's perspective on AI evaluation in critical scenarios: From tail quality to landscape**
Zhengxin Yang
- 🕒 **Tensor databases empower AI for science: A case study on retrosynthetic analysis**
Xueya Zhang, Guoxin Kang, Boyang Xiao, Jianfeng Zhan
- 🕒 **Predicting the number of call center incoming calls using deep learning**
Armaghan Nikfar, Javad Mohammadzadeh

Review Articles

- 🕒 **Regulatory landscape of blockchain assets: Analyzing the drivers of NFT and cryptocurrency regulation**
Junaid Rahman, Hafizur Rahman, Naimul Islam, Tipon Tanchangya, ... Mostafa Ali
- 🕒 **Ethical and regulatory challenges in machine learning-based healthcare systems: A review of implementation barriers and future directions**
Shehu Mohammed, Neha Malhotra

ISSN: 2772-4859

Copyright © 2024 International Open Benchmark Council (BenchCouncil); sponsored by ICT, Chinese Academy of Sciences. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd.

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

LLMs: A game-changer for software engineers?

Md. Asraful Haque 1

Evaluatology's perspective on AI evaluation in critical scenarios: From tail quality to landscape

Zhengxin Yang13

Tensor databases empower AI for science: A case study on retrosynthetic analysis

Xueya Zhang, Guoxin Kang, Boyang Xiao, Jianfeng Zhan 23

Predicting the number of call center incoming calls using deep learning

Armaghan Nikfar, Javad Mohammadzadeh32

Regulatory landscape of blockchain assets: Analyzing the drivers of NFT and cryptocurrency regulation

Junaid Rahman, Hafizur Rahman, Naimul Islam, Tipon Tanchangya, ... Mostafa Ali43

Ethical and regulatory challenges in machine learning-based healthcare systems: A review of implementation barriers and future directions

Shehu Mohammed, Neha Malhotra61


AICB: A benchmark for evaluating the communication subsystem of LLM training clusters

Xinyue Li, Heyang Zhou, Qingxu Li, Sen Zhang, Gang Lu75



Full Length Article

LLMs: A game-changer for software engineers?

Md. Asraful Haque 

Computational Unit, Z.H. College of Engineering & Technology, Aligarh Muslim University, Aligarh-202002, India

ARTICLE INFO

Keywords:

Software engineering
Large language model
AI tools
Coding
Testing
Debugging

ABSTRACT

Large Language Models (LLMs) like GPT-3 and GPT-4 have emerged as groundbreaking innovations with capabilities that extend far beyond traditional AI applications. These sophisticated models, trained on massive datasets, can generate human-like text, respond to complex queries, and even write and interpret code. Their potential to revolutionize software development has captivated the software engineering (SE) community, sparking debates about their transformative impact. Through a critical analysis of technical strengths, limitations, real-world case studies, and future research directions, this paper argues that LLMs are not just reshaping how software is developed but are redefining the role of developers. While challenges persist, LLMs offer unprecedented opportunities for innovation and collaboration. Early adoption of LLMs in software engineering is crucial to stay competitive in this rapidly evolving landscape. This paper serves as a guide, helping developers, organizations, and researchers understand how to harness the power of LLMs to streamline workflows and acquire the necessary skills.

1. Introduction

Software engineering (SE) processes refer to the structured set of activities involved in the development of software systems, including requirements analysis, design, coding, testing, deployment, and maintenance. These processes ensure that software is built systematically and meets user needs while maintaining quality and reliability [1]. Software development follows various models such as the Waterfall, Agile, or DevOps, each outlining different approaches to these phases. Software engineering can be costly and time-consuming for several factors related to the complexity, labor intensity, and long-term maintenance requirements. Fig. 1 illustrates a typical breakdown of the effort or cost allocation throughout the various phases of software development life cycle (SDLC) [2,3]. The primary objective of software engineering is to develop high-quality software at a minimal cost. The software industry faces numerous challenges in developing reliable software, particularly as systems become increasingly complex [4]. The demand for faster development cycles, high-quality code, and the ability to handle large-scale systems has driven the adoption of new tools and technologies. Among these, Large Language Models (LLMs) have emerged as a powerful force, automating and optimizing various aspects of the software engineering process [5]. Large language models are state-of-the-art NLP tools that have been trained on massive amounts of data, allowing them to generate human-like responses and understand complex language patterns. They have gained immense popularity in recent years

because it makes a lot of things easier and quicker. They have the potential to revolutionize various industries and transform the way we interact with technology. They have demonstrated impressive capabilities that are directly applicable to software engineering [6,7]. Some of the key functions include code generation, debugging, testing etc. The integration of LLMs into software engineering (SE) is transforming traditional practices in multiple ways. From altering how developers write, review, and maintain code to revolutionizing collaboration within teams, LLMs are reshaping the landscape of SE [8–10]. The impact of LLMs on software engineering tools and platforms is evident in the growing trend of LLM-powered IDEs. These environments now offer intelligent code suggestions, natural language queries, and automated refactoring, making development more intuitive. While there's much excitement about LLMs in software engineering, significant concerns remain regarding their practical use and ethical implications [11]. LLMs lack true comprehension of the logic behind code, making them prone to generating incorrect or insecure outputs. Additionally, the adoption of these models also brings challenges related to ethics, job roles, and the need for careful human oversight. Thus the question remains: Are these capabilities sufficient to significantly transform the software engineering industry?

In this paper, we aim to explore the transformative potential of Large Language Models in software engineering, assessing whether they represent an overhyped trend or a disruptive innovation capable of reshaping the field. We will delve into the technical strengths and

E-mail address: md_asraf@zhcet.ac.in.

<https://doi.org/10.1016/j.tbench.2025.100204>

Received 4 March 2025; Received in revised form 28 April 2025; Accepted 19 May 2025

Available online 19 May 2025

2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

limitations of LLMs, examine real-world case studies, and discuss the ethical considerations that come with the adoption of AI-driven development tools. Through this comprehensive analysis, we seek to provide a balanced perspective on the role of LLMs in modern software engineering practices.

2. Understanding large language models

Large language models (LLMs) are built on the transformative power of the transformer architecture, a model introduced by Vaswani et al. in 2017 [12] that has since become the foundation of many advanced LLMs. The transformer architecture, unlike its predecessors like recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, excels at handling long-range dependencies in data through its self-attention mechanism. This mechanism enables the model to understand and weigh relationships between all tokens in a sequence simultaneously, rather than processing them in order. This ability to capture both local and global context makes transformers highly effective for tasks that require understanding the structure and flow of text or code. In the context of software engineering, this allows LLMs to not only generate code based on natural language prompts but also to understand the intricate relationships between different parts of a codebase, which is crucial for complex tasks like debugging, code completion, and refactoring. The self-attention mechanism is a key innovation that empowers LLMs to efficiently determine the importance of different parts of input data, whether in a sentence or in a block of code. This helps LLMs better understand the context of programming languages, allowing them to predict the next steps in coding processes or provide useful suggestions during the development cycle. Another vital aspect of transformer models is positional encoding, which helps maintain the order of input data — a necessary feature when processing sequences like code, where the position of elements is critical to functionality. The combination of self-attention and positional encoding allows LLMs to process code sequences with an understanding of both immediate context and overall structure, thus improving their performance in code generation and related tasks.

The development of LLMs involves mainly three stages: pre-training, fine-tuning and reinforcement learning with Human Feedback [13–15]. In the pre-training phase, LLMs are exposed to vast amounts of textual and coded data, learning general language patterns, coding structures, and syntax from diverse sources such as books, websites, and open-source code repositories. The scope of this pre-training enables LLMs to acquire a broad understanding of multiple programming languages and frameworks, making them versatile in handling different software engineering tasks. Once pre-training is complete, the model undergoes fine-tuning on specific datasets tailored to the target application, refining its ability to perform tasks in specialized areas such as

web development, cybersecurity, or enterprise software solutions. This fine-tuning process sharpens the model's ability to generate relevant, high-quality outputs in response to domain-specific inputs. At the end, reinforcement learning is used to further enhance the model's performance by interacting with an environment and receiving human feedback. The feedback, in the form of ratings, rankings, or corrections, is used as a reward signal to guide the model's learning. This is an iterative process and continues until the model meets the desired standards, at which point it can be used in real-world applications. A typical training process of OpenAI's ChatGPT has been shown in Fig. 2 [16]. By leveraging the advantages of pre-training, fine-tuning and RLHF, LLMs become proficient in understanding not only the general syntax and structure of code but also in adapting to specialized coding practices and conventions. This allows LLMs to assist with software engineering tasks such as code generation, debugging, and even testing, making them valuable tools for developers working across a variety of programming languages and problem domains. Despite these strengths, however, LLMs still face challenges, particularly when dealing with complex logic or novel problems outside of their training data. Nevertheless, their advanced architecture and training methodologies have positioned them as powerful, versatile tools in the field of software engineering.

A significant number of LLMs are already in use. Table 1 provides a brief overview of some well-known models [16–22]. The future holds promise for even more powerful and advanced LLMs.

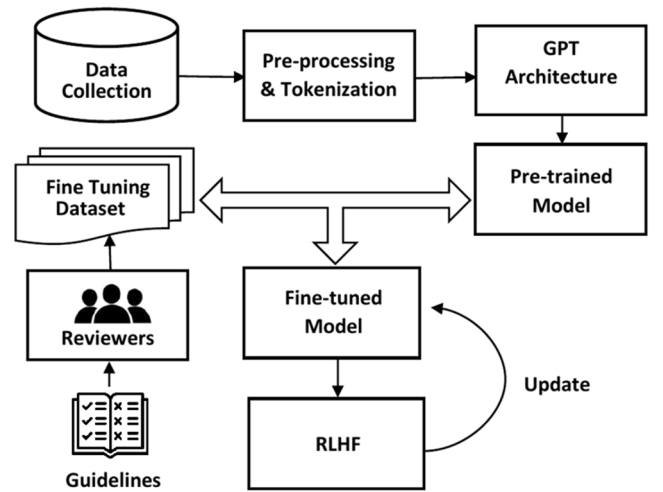


Fig. 2. ChatGPT training process.

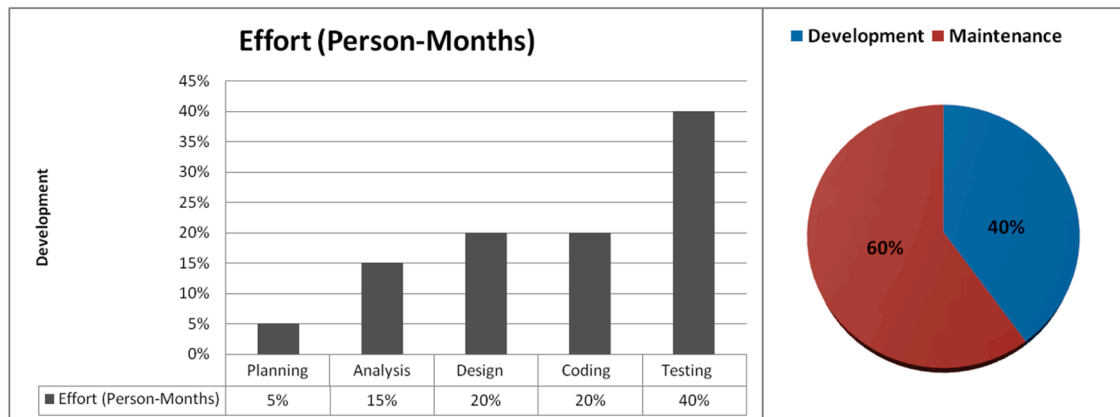


Fig. 1. Typical effort distribution at different phases of SDLC.

Table 1

A brief history of some prominent LLMs.

LLMs	Release Date	Developer	Model Size	
			Number of Parameters	Dimension ($L \times H$)*
BERT	October-2018	Google	110 billion (Base Model)	$L = 12, H = 12$ (Base Model)**
GPT-2	February-2019	OpenAI	1.5 billion	$L = 12, H = 12$ (Small Version)**
XLNet	June-2019	Google & CMU	110 million (Base Model)	$L = 12, H = 12$ (Base Model)**
T5	October-2019	Google	11 billion	$L = 12, H = 12$ (Small Version)**
GPT-3	June-2020	OpenAI	175 billion	$L = 96, H = 96$
Codex	August-2021	OpenAI	12 billion	$L = 24, H = 32$
PaLM	April-2022	Google	540 billion	$L = 118, H = 128$
GALACTICA	November-2022	Meta AI	120-billion	$L = 80, H = 96$
LLaMA	February-2023	Meta AI	65 billion	$L = 80, H = 64$
GPT-4	March-2023	OpenAI	1.76 trillion	Details undisclosed
Gemini 1.5	May-2024	Google DeepMind	Details undisclosed	Details undisclosed

* L =Number of layers, H =Number of attention heads.

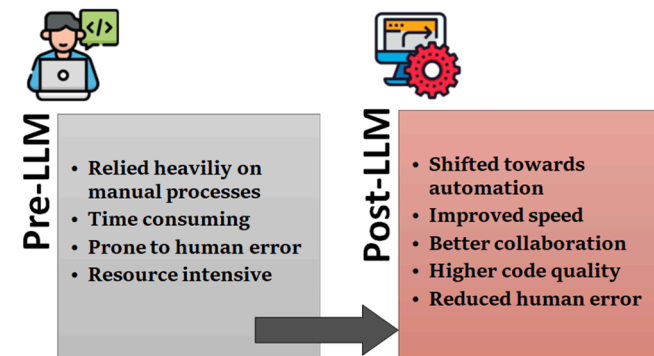
** These models have different variants.

3. Technical strengths and benefits of LLMs in SE

LLMs have brought transformative potential to software engineering, providing a suite of technical strengths and benefits that can drastically enhance productivity, code quality, and innovation (Fig. 3). From improving code generation to automating complex documentation tasks, LLMs are reshaping how developers approach various phases of the software development lifecycle. While the “state-of-the-art (SOTA)” pushes toward autonomous software development, the “state-of-the-practice (SOTP)” is more about augmented intelligence — enhancing human developers rather than replacing them. The SOTA refers to the bleeding edge of what LLMs can do under optimal conditions, typically in research environments or advanced use cases, whereas the SOTP is the reality of how LLMs are being used today by software engineers in real-world environments. Below is a detailed exploration of the technical strengths and benefits of LLMs in software engineering (SE).

3.1. Code generation

One of the most prominent uses is code generation, where models like GitHub Copilot, powered by OpenAI’s Codex, allow developers to describe the functionality they need in natural language, and the LLM generates relevant code snippets [23]. This not only speeds up the

**Fig. 3.** Revolution in SE practices.

coding process but also minimizes repetitive tasks, enabling developers to focus on more complex aspects of software design and architecture [24,25]. The benefit here is a marked improvement in productivity, as LLMs assist in automating routine coding activities like writing boilerplate code, implementing standard algorithms, or creating simple data structures. Moreover, LLM-driven code generation is highly versatile across different programming languages, offering cross-language flexibility that is particularly useful in polyglot development environments where multiple languages are used. The ability to generate code across Python, JavaScript, Java, C++, and other languages adds immense value, reducing the need for developers to switch contexts or master multiple languages to complete tasks efficiently [26,27].

- **State-of-the-Art:** Advanced models like GPT-4, Code Llama, StarCoder, and Claude can synthesize entire functions, classes, and even small applications directly from natural language prompts. These models have achieved high benchmark scores, consistently outperforming earlier generations in evaluations like HumanEval, MBPP, and MultiPL-E, with top models surpassing 50–70 % accuracy on competitive programming tasks [23,28]. In addition, a new frontier is emerging with the development of autonomous code agents, such as Auto-GPT, Smol Developer, and Devin (from Cognition AI), which attempt to autonomously generate multi-file projects with minimal human supervision. These agents are capable of breaking down high-level goals into actionable subtasks, generating project scaffolds, and iterating toward functional software solutions with little direct developer input, pushing the limits of what automated programming can achieve.
- **State-of-the-Practice:** The current state of practice in industry reveals a more cautious and pragmatic adoption of LLM-based code generation tools. Platforms like GitHub Copilot, Codeium, and Amazon CodeWhisperer have seen widespread integration into professional development workflows, primarily assisting with in-line code completion, thus reducing keystrokes and improving overall developer velocity. However, LLMs are still primarily leveraged for writing boilerplate, templated code, and implementing standard patterns rather than handling complex, domain-specific logic or making architecture-level decisions. Limitations become apparent when LLMs are tasked with projects involving intricate dependencies or specialized business rules. As a result, a strong human-in-the-loop paradigm remains necessary; developers must diligently validate, test, and refine AI-generated code to ensure correctness, security, and maintainability, especially in mission-critical or production environments where even minor errors could have significant consequences.

3.2. Code review, debugging and testing

LLMs have the ability to automate code reviews and assist with bug detection [29]. These models can facilitate the knowledge required for high-quality code reviews. Even junior developers, with the assistance of LLM-powered tools, can contribute effectively to the code review process by leveraging the model’s knowledge of industry standards and best practices. Traditionally, debugging requires manual effort from developers, who inspect the code for errors, and once the error is located developers can then implement a fix to correct the error [30]. LLMs can analyze logs, error messages, and code execution paths to suggest potential causes of bugs [31,32]. This helps developers quickly pinpoint the source of an issue, reducing the time spent on manual debugging. LLMs can also suggest potential fixes based on the patterns they have learned from analyzing similar bugs in the past. This capability is particularly useful in large, complex systems where tracking down bugs can be a challenging and time-consuming task. For example, LLMs can identify common mistakes such as unhandled exceptions, resource leaks, or improper variable initializations. They can also flag potential security issues like injection vulnerabilities, insecure data handling, or incorrect

encryption implementations. By catching these issues early, LLMs help developers produce more secure and robust code. Additionally, LLMs have been applied to automated testing, where they generate unit tests, identify edge cases, and suggest test cases based on the functionality described in code [33]. This application can significantly speed up the testing phase of software development, ensuring that code is rigorously tested without developers having to manually write every possible test case. LLMs' capacity to generate exhaustive test suites helps in reducing the likelihood of bugs making it to production, thus increasing the overall robustness and reliability of software systems.

- **State-of-the-Art:** Automated code review tools such as CodiumAI, CodeGPT, and SonarLint AI leverage LLMs to provide intelligent suggestions, identifying issues related to logic, security vulnerabilities, and adherence to coding best practices. AI-assisted debugging is also becoming highly sophisticated; for instance, GPT-4 integrations in environments like Visual Studio Code and specialized developer assistants unveiled at OpenAI DevDay can actively parse logs, diagnose errors, and recommend fixes [34]. In testing, LLMs have achieved notable success in generating not just basic unit tests but also more complex integration and property-based tests, sometimes attaining over 90 % code coverage for simple functions, as seen in models like DeepMind's AlphaCode [35,36]. These advancements signal a move towards increasingly autonomous support systems that can substantially augment developers' capabilities throughout the software lifecycle.
- **State-of-the-Practice:** AI tools are primarily used to augment rather than replace human-led code reviews. While they are effective at catching common mistakes and suggesting improvements, they often lack the deep contextual understanding necessary for evaluating architectural choices or domain-specific logic. False positives remain a significant concern, particularly in debugging, where LLMs sometimes propose solutions that seem plausible but fail in real-world execution. This limits the reliability of AI-suggested fixes without human verification. Moreover, while AI-generated test cases can be helpful, integration with Continuous Integration/Continuous Deployment (CI/CD) pipelines remains limited. Most teams continue to rely heavily on manual validation and traditional test frameworks, preferring to use AI-generated tests as a supplementary resource rather than fully trusting them for critical deployments.

3.3. Language and framework agnostic

LLMs have the remarkable ability to work across a wide variety of programming languages and frameworks, making them versatile tools in multi-language environments [37]. Because LLMs are trained on diverse datasets that include code from many different programming languages, they can switch between languages and frameworks with ease. This is especially useful for developers who work in environments that require knowledge of multiple languages, such as Python for backend services, JavaScript for frontend development, and SQL for database management. By supporting a broad range of languages and frameworks, LLMs eliminate the need for developers to switch between different coding assistants or learn new tools for each technology they use [38,39]. This contributes to a more seamless development experience and enhances overall efficiency.

- **State-of-the-Art:** Modern LLMs, trained on vast repositories such as GitHub and Stack Overflow, possess multilingual code understanding, allowing them to generate and translate code across a wide range of languages including Python, JavaScript, Java, C++, Rust, and Go. These models are not only language-flexible but also framework-aware; they can produce code that aligns with the syntax and best practices of popular frameworks such as React, Django, and Spring Boot [40]. Furthermore, cutting-edge AI-assisted refactoring tools are now capable of facilitating cross-language migration — for

instance, helping transition legacy systems from COBOL to Java or modernizing codebases from Python 2 to Python 3, or even migrating JavaScript codebases to TypeScript. These advancements enable significant leaps in productivity, particularly when modernizing or maintaining large, aging code infrastructures.

- **State-of-the-Practice:** In real-world usage, AI-generated code often lacks deep context awareness, meaning it may not consistently adhere to team-specific coding conventions, domain-driven architectural principles, or nuanced project standards. Although LLMs can generate framework-specific code — for example, producing Django, Flask, Express, or FastAPI templates — real-world implementation typically requires manual fine-tuning and adjustment to meet production-grade requirements. Additionally, while cross-language migration tools can offer substantial initial assistance, challenges frequently arise when deep dependencies, intricate business logic, or architecture-specific constraints are involved. In such cases, human expertise remains indispensable to ensure the successful and accurate transition of complex software systems.

3.4. Refactoring and optimization

As software systems grow, they often accumulate technical debt, requiring refactoring and optimization to ensure long-term performance and maintainability. LLMs can assist with these processes by suggesting refactoring opportunities, such as code that can be simplified, duplicated code that can be consolidated, or outdated structures that need updating [41]. This is particularly valuable in large codebases where manually identifying areas for refactoring would be time-consuming and prone to oversight. LLMs can also help optimize code by suggesting more efficient algorithms or design patterns based on established best practices. For instance, if a developer writes a brute-force solution for a problem, the LLM might suggest a more optimal approach using dynamic programming or divide-and-conquer algorithms. Additionally, LLMs can provide performance insights, such as identifying inefficient loops, excessive memory usage, or potential bottlenecks in code execution, which helps ensure that the software remains scalable and performant as it evolves [42].

- **State-of-the-Art:** Automated code refactoring tools powered by LLMs — such as CodiumAI, IntelliCode, and Tabnine — can suggest enhancements aimed at improving readability, modularity, and system performance. Beyond stylistic changes, modern LLMs can engage in AI-driven performance tuning by suggesting optimizations in algorithmic complexity (Big-O improvements), recommending more efficient SQL queries, and proposing memory-optimized solutions based on recognized best practices [43,44]. Some cutting-edge models even demonstrate a level of semantic understanding of code structures, allowing them to detect "code smells" and recommend modularization strategies to improve maintainability and reduce technical debt. These advancements position LLMs as powerful partners for elevating the quality and efficiency of software development.
- **State-of-the-Practice:** While LLMs effectively handle simple refactoring tasks — such as renaming variables, extracting methods, and improving code readability — major architectural refactoring efforts still largely require human expertise and strategic judgment. LLMs often suffer from limited awareness of a project's full architectural context, which can result in inconsistent or suboptimal suggestions, particularly when working across large repositories with complex interdependencies. Trust remains another key issue: developers usually conduct a careful manual check of AI-proposed optimizations before adopting them, wary of potential regressions or performance declines.

3.5. Automated documentation

Generating accurate, up-to-date documentation has long been a challenge for software engineers, as it is often seen as tedious work that lags behind code changes [45]. Developers often deprioritize this due to tight deadlines or a focus on feature development. However, documentation is crucial for ensuring that code is maintainable, understandable, and transferable across teams and developers. LLMs can analyze code and automatically generate documentation for codebases, including explaining the purpose and functionality of specific functions, classes, and modules [46]. This ensures that documentation stays current and can be updated as code evolves, providing developers with easily understandable, well-structured explanations of how various parts of the system work. This capability not only improves team collaboration and knowledge sharing but also supports on-boarding processes by helping new developers quickly understand legacy codebases or complex systems. In agile environments, where requirements and implementations frequently change, the ability of LLMs to update documentation dynamically as the code evolves is an invaluable benefit.

- **State-of-the-Art:** State-of-the-art developments in automated documentation leverage LLMs' deep code understanding to create more contextual and helpful outputs. Modern tools like Codeium and Copilot Chat enable the generation of auto-populated docstrings, providing context-aware inline explanations that enhance code readability. Advanced LLMs can also produce natural language descriptions of APIs, offering suggestions for refining Swagger or OpenAPI specifications, thus bridging gaps between developers and API consumers. Some AI tools, like Sourcegraph Cody, push the boundaries further by summarizing codebases into readable tutorials or blog-style documentation, making complex technical information more accessible to broader audiences. These innovations are redefining how documentation is created and maintained, shifting it from a manual, error-prone task to an automated, continuous process.
- **State-of-the-Practice:** In current development workflows, the output of LLMs often tends to be generic, verbose, or redundant, necessitating human refinement to ensure clarity and relevance. While LLMs excel at describing low-level function behaviors, they typically struggle to capture and explain high-level architectural decisions or intricate system designs. As a result, AI-generated documentation is commonly integrated into internal knowledge bases like Confluence, Notion, or proprietary wikis, where it serves as a helpful supplement rather than a full replacement for human-authored documentation.

4. Challenges

While LLMs offer exciting possibilities in the SE domain, several technical limitations and a range of ethical challenges must be addressed:

4.1. Technical limitations

- **Lack of True Understanding:** LLMs do not "understand" code in the same way humans do [8,44]. While LLMs are powerful at predicting sequences based on statistical patterns learned from vast datasets, they lack a deep understanding of the underlying logic and intent behind a given piece of code. In software engineering, this is particularly problematic because coding often requires not just syntactically correct solutions, but solutions that align with specific business logic, system architecture, and performance requirements [47]. For instance, an LLM may generate syntactically correct code for a sorting algorithm but fail to account for efficiency constraints such as time complexity or memory usage, especially when these concerns are implicit in the task description. The inability of LLMs to grasp these nuances means that while they are useful for generating code snippets or suggesting fixes, developers must rigorously review

and adapt their outputs to ensure they meet the functional and non-functional requirements of the system.

- **Context Sensitivity:** Although LLMs are good at handling short, localized contexts, they often struggle with maintaining long-term context over extended portions of a codebase [48]. Software systems are often composed of multiple files, modules, and libraries that interact with one another in complex ways. Maintaining context across such a large-scale system, where changes in one part of the codebase can have ripple effects across the system, is a challenge for LLMs. For example, an LLM may generate code that works well within a single function but fails to account for broader architectural considerations, such as how this function interacts with others in different modules or libraries. In large enterprise-scale applications, this limitation becomes even more pronounced, as developers need to track dependencies across various subsystems, which LLMs may not handle effectively. The model's understanding tends to deteriorate when it needs to work with codebases that span across multiple files or projects, resulting in incomplete or incorrect suggestions.
- **Inability to Handle Novel or Rare Problems:** LLMs rely heavily on patterns learned from their training data, which means they perform best when tasked with solving common or well-documented problems. However, when faced with novel or rare problems that deviate from established patterns, LLMs often struggle to produce correct or meaningful output. In software engineering, developers frequently encounter unique challenges that require creative problem-solving and a deep understanding of both the problem domain and system architecture. LLMs, constrained by the limitations of their training data, may not have seen enough similar examples to provide an adequate solution. This is especially true for cutting-edge technologies or innovative software designs that have not been widely adopted and thus are not well-represented in public datasets. Furthermore, rare edge cases, which are often the most critical and challenging parts of software development, tend to be poorly handled by LLMs due to the lack of exposure to similar situations during training.
- **Computational Costs:** Large Language Models (LLMs) are computationally intensive, requiring significant hardware resources for training and inference [49]. Training LLMs requires immense computational resources, including large clusters of GPUs or TPUs, leading to high financial costs. Inference, or using the trained model for tasks, can also be computationally expensive, especially for larger models. These high computational costs can be a barrier to entry for organizations, limiting their ability to adopt and utilize LLMs in their software engineering workflows. Another consideration is the energy consumption associated with running LLMs. The large-scale deployment of LLMs in software engineering environments contributes to increased energy usage, which has both economic and environmental implications [50].
- **Transparency and accountability:** LLMs are often seen as black-box models, meaning that their decision-making processes are not easily interpretable by users [51]. When an LLM generates code, it is not always clear how or why it arrived at a particular solution [52]. This lack of transparency becomes problematic in scenarios where LLMs make critical decisions, such as in safety-critical systems or in applications that have legal and regulatory implications. If a software failure occurs due to an LLM's suggestion, it is difficult to assign responsibility — does the fault lie with the developer, the AI, or the organization that provided the AI? This lack of clear accountability creates challenges in governance and compliance, particularly in regulated industries like finance, healthcare, and transportation. Therefore, ensuring that LLMs are explainable and that there are mechanisms in place to track and audit AI-generated outputs is essential for fostering trust and ensuring that ethical guidelines are followed.
- **Security Risks:** If LLMs are trained on large public datasets, including code repositories, they may inadvertently learn insecure or

vulnerable coding practices [53]. For instance, if an LLM is trained on a repository where code contains hard-coded credentials, weak encryption methods, or unpatched vulnerabilities, the model may unknowingly generate code that replicates these flaws. This becomes particularly dangerous in security-critical applications like financial software, healthcare systems, or government infrastructure. The potential for LLMs to suggest insecure code increases the burden on developers to scrutinize the model's output closely, ensuring that it adheres to industry best practices and security standards. Therefore, while LLMs can be helpful for automating routine tasks, they should not be used blindly, especially in areas where security is paramount. Continuous oversight and refinement of the training data, as well as integration with secure coding practices, are essential to mitigate these risks.

4.2. Ethical considerations

- **Copyright and Intellectual Property:** LLMs are trained on publicly available data, but this data may include proprietary or copyrighted code that the model can later reproduce in different contexts [54]. When LLMs generate code that closely resembles or directly replicates code from its training data, it raises serious questions about ownership and accountability. Developers using LLM-generated code may inadvertently violate copyright laws if the generated code mirrors protected material without proper attribution. This could lead to legal disputes and undermine the trust in LLMs as reliable tools in professional software development environments. To address these concerns, companies providing LLM services must implement safeguards that either filter copyrighted material during the training process or ensure that LLM-generated content is appropriately flagged for potential legal issues.
- **Biases in Training Data:** One of the most pressing ethical concerns surrounding LLMs is the issue of bias in training data [55]. LLMs are trained on vast datasets that include both natural language text and code from public repositories, such as GitHub, Stack Overflow, and various forums. However, these sources can contain biased, outdated, or even harmful practices. For example, if the training data includes discriminatory language or biased coding patterns (such as gender or race-based assumptions in user data processing), the LLM may learn and perpetuate these biases in its outputs. In the context of software engineering, biased code generation can lead to inequitable software solutions, unfair user experiences, or even legal and reputational risks for companies. Moreover, models trained on real-world codebases may inherit the biases of past software engineering decisions, such as assumptions about users' technical abilities or geographical location, resulting in software that does not serve all demographics equally. Addressing these biases requires careful dataset curation, as well as developing methods for identifying and mitigating bias in LLM outputs.
- **Impact on the Workforce:** The impact on the workforce is another ethical issue associated with the rise of LLMs in software engineering. By automating tasks like code generation, testing, and debugging, LLMs have the potential to reduce the demand for certain types of coding jobs, particularly entry-level or junior software development roles [17,55]. This could lead to job displacement for new developers or those in low-skilled positions, creating economic inequality within the industry. Additionally, reliance on LLMs may result in a deskilling of the software engineering workforce. If developers become too dependent on AI-generated code and suggestions, they may lose the ability to write complex code or troubleshoot issues independently. This could diminish the overall expertise within the field over time, affecting the quality of software and innovation. To counteract these risks, educational systems and organizations need to evolve, focusing on upskilling developers to work alongside LLMs rather than being replaced by them. Training programs should emphasize

higher-order skills like software architecture, algorithm design, and critical thinking, which cannot be easily replicated by AI.

5. Case studies and recent trends

Initially, AI tools were primarily used for specific tasks like code completion and bug detection. However, the advent of LLMs has ushered in a new era of AI-powered development. This shift promises not only greater efficiency but also new possibilities in adaptive, responsive coding environments. In exploring the impact of LLMs on software engineering, it is essential to examine real-world case studies where LLMs have been applied in different software engineering (SE) environments. By analyzing diverse cases, we can understand how LLMs can be a game-changer in some situations, while potentially overhyped or insufficient in others.

5.1. GitHub Copilot (Powered by OpenAI codex)

GitHub Copilot, powered by LLM technology, has become a widely adopted tool for code generation, suggestions, and auto-completions [56–58]. It is extensively used by developers at organizations such as Microsoft, Shopify, and Datadog, along with many individual developers. A notable case of internal adoption is GitHub's own integration of Copilot into its development workflow [59,60]. This implementation led to a significant increase in developer productivity—up to 55 % in routine coding tasks. Developers experienced faster prototyping and fewer context switches, which streamlined their workflows. Copilot proved especially effective in generating boilerplate code and recurring patterns, enabling engineers to focus more on strategic and creative aspects of development.

Another compelling example comes from Shopify, where engineers utilized Copilot for developing internal tools and web applications [61–63]. The tool contributed to a faster onboarding process for junior developers and significantly reduced cognitive load during development. Moreover, the context-aware code completions offered by Copilot helped developers become more familiar with new libraries and frameworks, enhancing overall efficiency and learning outcomes within the team.

5.2. Amazon CodeWhisperer

Amazon CodeWhisperer is an LLM-based tool designed to provide real-time code suggestions based on natural language prompts, seamlessly integrating with popular IDEs such as Visual Studio Code and JetBrains [64]. A prominent case study of its application involves AWS developer teams, who employed CodeWhisperer internally to automate the generation of API integration code [65]. This implementation led to a reduction in repetitive coding time by approximately 40–50 %, significantly improving developer efficiency. One of the key benefits highlighted by the teams was the tool's high accuracy in generating code specific to AWS SDKs and services, which substantially reduced the need for frequent documentation lookups, thereby streamlining the overall development workflow.

5.3. Tabnine

Tabnine is an LLM-based tool that offers predictive code completions through either local or cloud-hosted models, making it a suitable solution for privacy-focused environments and organizations requiring team-specific model tuning [66]. A case study [67] describes how Tabnine uses Google Cloud to deliver its AI-powered coding tool to one million users. Tabnine's ML models, which help developers autocomplete about 30 % of their code, rely on Google Cloud's GPUs and Google Kubernetes Engine for scalability and performance. Tabnine values its open-source commitment, which aligns with Google Cloud's dedication to the open-source community. They also value the support they have

received from Google Cloud specialists.

5.4. Codium (Now qodo)

Codium uses AI to provide intelligent code completions and automate test generation. It emphasizes code quality and security, offering on-premise deployment options. It aims to boost developer productivity by streamlining coding workflows. It supports over 70 programming languages, and integrates seamlessly with various IDEs and web editors [68]. It is a lightweight alternative to GitHub Copilot. One of its key selling points is that it is available for free to both individuals and teams, making it an accessible solution for a wide range of users. A notable case study involves “Clearwater Analytics”, a fintech SaaS company prioritizing data security, who adopted Qodo (formerly Codium) to enhance developer productivity [69]. Faced with the challenge of maintaining stringent security while leveraging AI-powered coding assistance, they chose Qodo for its unique ability to be deployed within their Enterprise VPC, ensuring code privacy. Developers experienced immediate productivity gains with Qodo’s code completion capabilities, resulting in significant time and cycle savings. The successful implementation was supported by dedicated Qodo team support and training, and the rapid integration of new features like chat integration.

5.5. Replit Ghostwriter

Replit Ghostwriter, an LLM-powered IDE tool, is revolutionizing coding accessibility and efficiency, particularly for students and beginner developers. Seamlessly integrated into Replit’s browser-based platform, it accelerates learning by providing real-time code generation, explanation, and natural language-to-code translation [70]. Educational institutions have witnessed significant improvements in student confidence and assignment completion through its interactive support, reducing reliance on instructors. Beyond education, Ghostwriter boosts productivity for experienced developers by automating tasks, contributing to Replit’s substantial user growth, which surged from 10 million to over 20 million within a year [71]. Ultimately, Ghostwriter democratizes coding, serving as a powerful learning and productivity tool that’s poised to expand its impact as AI technology advances.

5.6. Sourcegraph Cody

Sourcegraph Cody is an AI-powered tool designed to enhance code navigation and documentation understanding, seamlessly integrated

with Sourcegraph’s code intelligence platform. Leidos, a science and technology company facing the challenge of enhancing developer productivity within a complex, security-conscious environment, adopted Sourcegraph Cody [72]. They found Cody’s context-aware assistance and flexible LLM integration to be key differentiators, enabling significant time savings in code understanding, documentation, and debugging. Notably, Cody drastically reduced the time spent answering teammate questions by 75 % and cut code orientation time on legacy systems by 50 %. This resulted in increased efficiency in modernizing and migrating legacy code, with tasks previously taking sprints being completed in minutes. Leidos’s experience demonstrates Cody’s effectiveness in improving developer workflows, particularly in large, complex codebases, and its ability to maintain high security standards.

These case studies reflect that software industries worldwide are leveraging AI tools to streamline processes, increase efficiency, and foster creativity in problem-solving. Table 2 indicates that LLMs can generate code and suggest improvements quickly, but they often lack the precision, ethical insight, and contextual understanding that human developers provide. The AI Index 2024 Annual Report [73] highlights that software developers are among the professionals most likely to incorporate AI in their work. As AI’s role within the economy grows, understanding how developers use and view AI is becoming essential. Stack Overflow, the Q&A platform for programmers, runs an annual survey targeting developers. For the first time in 2023, this survey gathered insights from over 90,000 developers — featured questions on usage of AI tools [73]. It explored how developers employ these tools, which ones they prefer, and their overall perceptions of them. Table 3 shows the developers’ preferences for using AI tools in software engineering tasks. Fig. 4 is the graphical representation of Table 3.

The survey was taken in May 2023, thus it may not reflect the availability of more recent AI technologies such as Gemini and Claude 3. The other findings of that survey were as follows (Fig. 5):

- Most popular AI developer tool among professional developers, 2023 is GitHub Copilot.
- Most popular AI search tool among professional developers, 2023 is ChatGPT.
- Most popular cloud platform among professional developers, 2023 is Amazon Web Services.
- Developers cited higher productivity (32.8 %), quicker learning (25.2 %), and increased efficiency (25.0 %) as the top benefits of AI tools in their work.

Table 2
Comparative analysis.

LLM-powered Tools	Key Impacts	Challenges
GitHub Copilot	<ul style="list-style-type: none"> - Dramatically accelerates coding. - Boosts creativity by suggesting patterns developers may not think of. - Helps junior developers produce higher-quality code. 	<ul style="list-style-type: none"> - Sometimes generates insecure or inefficient code. - Risk of “over-relying” without understanding the logic. - Licensing/legal concerns (e.g., code originality).
Amazon CodeWhisperer	<ul style="list-style-type: none"> - Stronger emphasis on secure coding (e.g., encryption, authentication). - Seamless AWS service integrations save time. - Good for enterprise-grade cloud apps. 	<ul style="list-style-type: none"> - Biased toward AWS ecosystem, less useful for non-AWS projects. - Suggestions can sometimes be more verbose than necessary. - Less flexible across diverse programming stacks.
Tabnine	<ul style="list-style-type: none"> - Highly efficient for boilerplate and repetitive code. - Minimal learning curve — very easy to integrate. - Helps developers “think less” about syntax. 	<ul style="list-style-type: none"> - Limited “deep” understanding of project-specific logic. - Doesn’t recommend security or performance improvements. - Can sometimes offer shallow or redundant completions.
Codium AI	<ul style="list-style-type: none"> - Greatly improves code quality through auto-generated tests. - Encourages a testing culture (important for scaling teams). - Helps identify hidden bugs early. 	<ul style="list-style-type: none"> - Test quality can vary depending on code complexity. - Not a replacement for writing well-thought-out manual tests. - Might generate overly simple test cases if not fine-tuned.
Replit Ghostwriter	<ul style="list-style-type: none"> - Instant environment setup saves huge time (especially for quick experiments). - Ideal for prototyping new ideas without local dependencies. - Very beginner-friendly (low barrier to entry). 	<ul style="list-style-type: none"> - Limited control for large, complex project structures. - Not suitable for full-scale production codebases. - Dependency on Replit ecosystem for best results.
Sourcegraph Cody	<ul style="list-style-type: none"> - Makes navigating and understanding huge codebases faster. - Helps teams maintain consistency across large projects. - Reduces onboarding time for new developers. 	<ul style="list-style-type: none"> - Requires setting up or connecting to indexed repositories. - Effectiveness can drop if code comments/documentations are poor. - Querying the system effectively requires some learning.

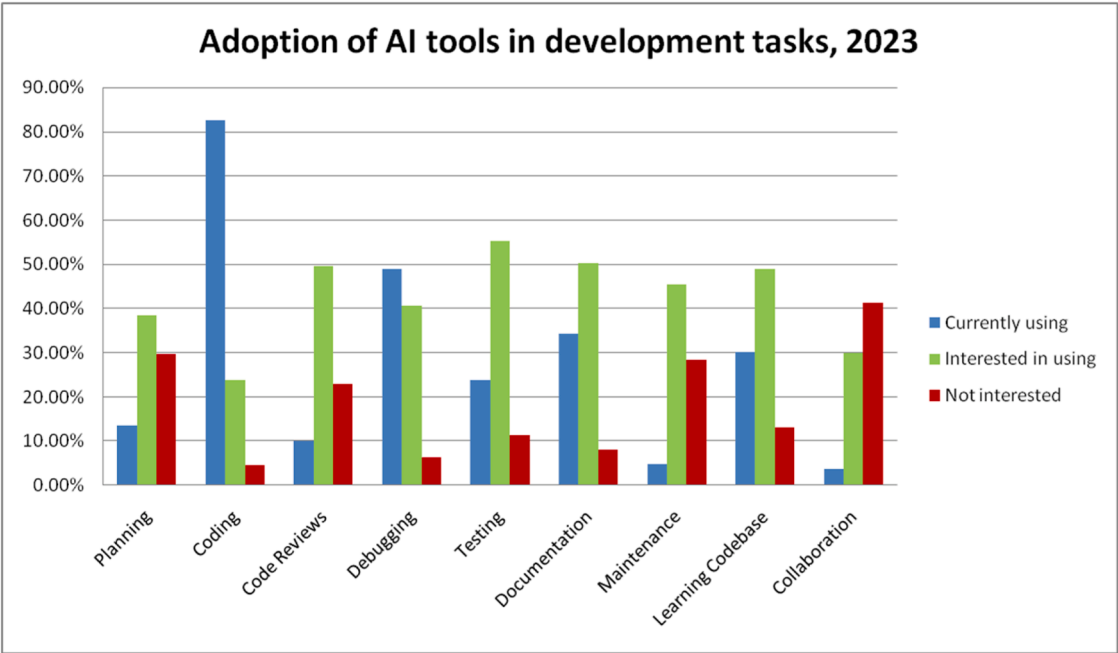


Fig. 4. Developers’ preferences for using AI-tools in development tasks, 2023.

Table 3
Developers’ preferences for using AI tools.

Development Tasks	Currently using	Interested in using	Not interested
Planning	13.52 %	38.54 %	29.77 %
Coding	82.55 %	23.72 %	4.48 %
Code Reviews	10.09 %	49.51 %	22.95 %
Debugging	48.89 %	40.66 %	6.37 %
Testing	23.87 %	55.17 %	11.44 %
Documentation	34.37 %	50.24 %	8.07 %
Maintenance	4.74 %	45.44 %	28.33 %
Learning Codebase	30.10 %	48.97 %	13.09 %
Collaboration	3.65 %	29.98 %	41.38 %

GitHub also conducted a survey [74] from February 26 to March 18, 2024, among 2000 non-student, corporate respondents in the United States, Brazil, India, and Germany who are not managers and work for organizations with 1000 or more employees. According to the survey, developers are increasingly integrating AI tools, with the majority of respondents reporting that AI improves their productivity and coding skills. Fig. 6 represents the respondents view on the benefits of AI tools. It highlights that popularity and use of AI tools varies by region. Fig. 7 displays the current usage of AI coding tools against the corporate endorsement for AI-driven coding. The survey respondents reported that AI tools boost productivity, freeing them up to focus on strategic tasks like system design and client collaboration. To fully leverage AI, organizations should integrate it into every phase of development. AI isn’t a job replacement but an enhancer of human creativity.

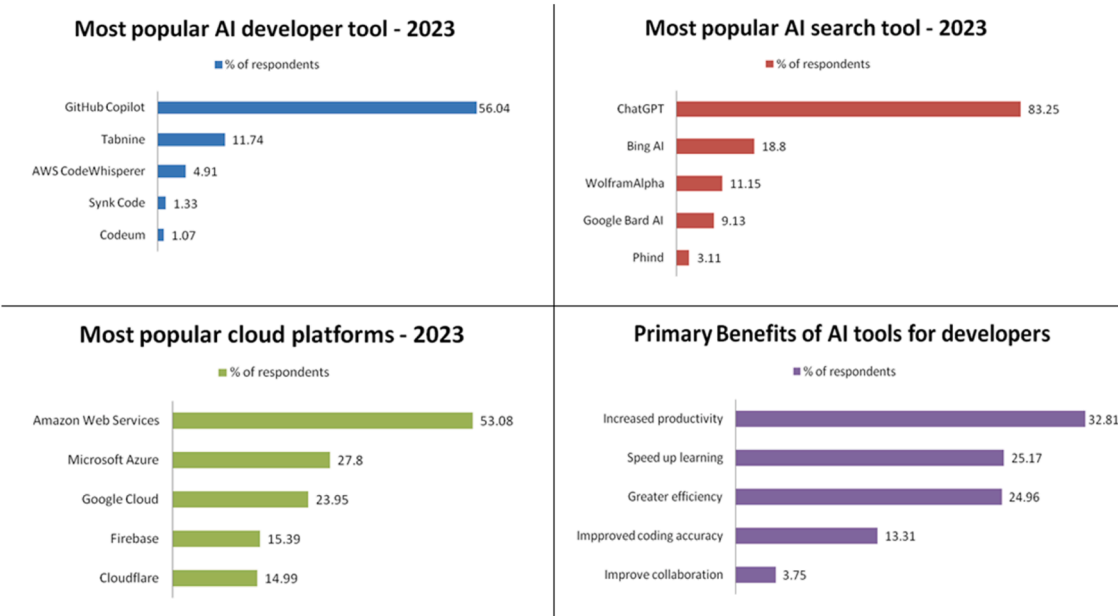


Fig. 5. Popularity of AI-tools among professional developers, 2023.

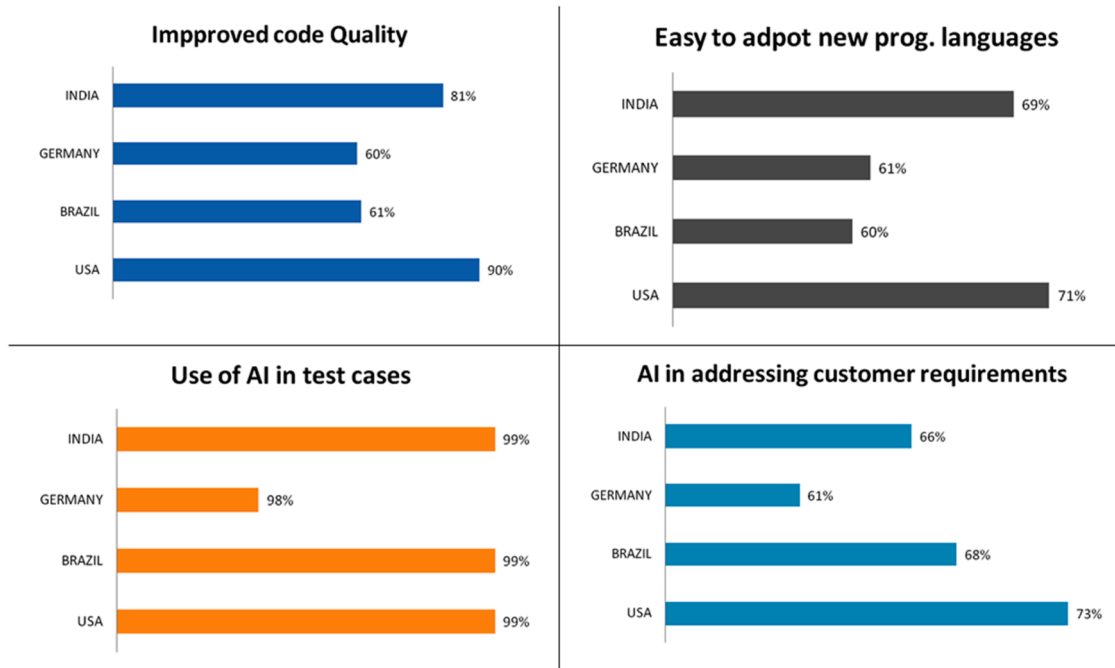


Fig. 6. Respondents view on the benefits of AI tools.

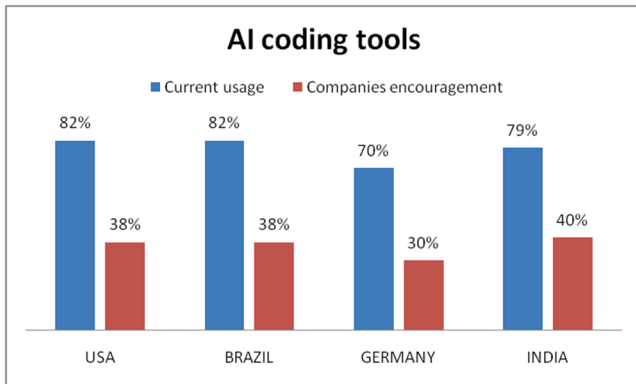


Fig. 7. Usage of AI tools vs. Companies encouragement.

6. Future directions and research opportunities

As Large Language Models (LLMs) continue to evolve and become more deeply integrated into software engineering (SE) processes, the future of this technology holds immense potential. However, there are several areas that still require further exploration, development, and research [75–79]. Understanding the trajectory of LLMs in SE will not only help identify their limitations but also uncover new applications and possibilities for transforming software development practices. In this section, we will explore the key future directions and research opportunities for LLMs in software engineering, ranging from technical advancements to ethical considerations and new ways to collaborate with AI models.

6.1. Specialization and domain-specific LLMs

A major area of research in the future will focus on creating more specialized LLMs tailored for specific domains within software engineering. While general-purpose LLMs like GPT-4 and Codex are highly effective across a wide range of coding tasks, they are often not optimized for niche areas such as embedded systems, real-time applications,

or domain-specific languages like hardware description languages (HDL). Researchers are likely to focus on training LLMs on highly curated, domain-specific datasets, allowing these models to gain deeper expertise in specialized fields. For example, an LLM trained exclusively on medical software code or financial systems might be better equipped to understand the particular regulatory requirements, security needs, and performance constraints of these industries. Such domain-specific models could also include compliance checks that align with industry-specific standards, helping ensure that software adheres to legal and regulatory frameworks. Similarly, LLMs could be fine-tuned for particular programming languages or frameworks, providing deeper insights and optimizations tailored to those specific environments.

6.2. Improved interpretability and explainability

One of the most pressing challenges with the current generation of LLMs is their "black-box" nature, meaning they often provide answers or code suggestions without clear explanations of how or why those suggestions were made. This lack of transparency is problematic, particularly in safety-critical applications like healthcare, finance, or aerospace, where understanding the reasoning behind code is essential for ensuring security and correctness. Research in this area will likely focus on improving the interpretability and explainability of LLMs. Efforts will be made to create models that can not only generate code but also explain the rationale behind their decisions, offering developers more confidence in the accuracy and safety of the suggestions. This could involve developing new methods for LLMs to highlight the key parts of the training data or coding patterns that influenced their output. Explainable AI (XAI) frameworks that allow for deeper interrogation of LLM outputs could become more commonplace in SE environments, helping engineers better understand the suggestions provided by the models.

6.3. Collaborative human-ai programming environments

The future of LLMs in software engineering will likely emphasize collaborative programming environments where humans and AI work together seamlessly. This will involve creating tools and platforms that promote symbiotic relationships between developers and LLMs, allowing both parties to complement each other's strengths. For instance,

while LLMs excel at generating code quickly and efficiently, human developers bring contextual understanding, creativity, and ethical judgment that AI currently lacks. Research opportunities in this area include developing more intuitive, conversational interfaces for LLMs, where developers can interact with models in a fluid and iterative manner. This could involve advancements in multimodal AI, where LLMs can take into account visual inputs, such as system diagrams or wireframes, to better understand the developer's intent and provide more relevant suggestions. Similarly, AI models could be trained to adapt their suggestions based on real-time feedback from developers, improving their effectiveness over time and enabling a more interactive coding process. These collaborative environments could also include AI models acting as "pair programmers," offering continuous feedback, alternative coding approaches, and potential optimizations during the development process.

6.4. Enhanced debugging and automated bug fixing

One of the most promising future directions for LLMs in software engineering is their potential to revolutionize debugging and automated bug fixing. Current LLMs can already identify and suggest solutions for common errors, but future advancements may lead to more sophisticated debugging tools that can understand complex bugs in large, multi-component systems. Future research may focus on training LLMs to detect not just surface-level issues (e.g., syntax errors), but deep-rooted logical bugs, performance bottlenecks, and security vulnerabilities in more extensive codebases. For instance, LLMs of the future could autonomously analyze code dependencies and execution paths to identify the root cause of subtle issues, such as memory leaks or race conditions, which are difficult to detect manually. Moreover, they could propose multiple solutions, weigh the pros and cons of each, and recommend the best course of action, tailored to specific system constraints. Further research could explore the potential for AI to continuously monitor running systems and automatically suggest patches or improvements in real-time, reducing the need for human intervention in maintenance tasks.

6.5. Ethical and security concerns

As LLMs become more prevalent in SE, the ethical and security implications of their use will require ongoing research. For instance, as LLMs generate more and more code, questions about the ownership and licensing of that code will arise, particularly when the models are trained on publicly available, open-source projects. Who owns the code generated by AI models, and how do we ensure that it complies with existing intellectual property laws? Addressing these issues will require interdisciplinary research that involves not just software engineering but also legal scholars, ethicists, and policy makers. Another major area of concern is the security of AI-generated code. Although LLMs can detect certain types of vulnerabilities, they can also inadvertently introduce new ones. Research will need to focus on creating mechanisms that prevent LLMs from generating insecure code, particularly in mission-critical systems. There is also the risk of bias and ethical dilemmas in the datasets used to train LLMs. Models trained on biased or incomplete data may perpetuate harmful stereotypes or make inaccurate decisions, which could have significant consequences in sectors such as healthcare or criminal justice software systems. Future research will need to address ways to mitigate these risks, ensuring fairness and accountability in AI-generated code.

6.6. Continual learning and model adaptation

As software development environments evolve, so too must the LLMs that support them. One area of research is continual learning, where LLMs can update their knowledge in real-time as they are exposed to new coding patterns, languages, or technologies. This would eliminate

the need for retraining models from scratch and allow LLMs to stay relevant in dynamic environments. Future LLMs could potentially learn from real-world codebases as they evolve, adapting to new trends in development practices and adjusting their suggestions accordingly. Moreover, research into adaptive LLMs may explore models that can fine-tune themselves based on specific user needs or project contexts. For instance, a developer working on a web application might receive different types of suggestions from an LLM compared to someone working on an embedded system. Models could be fine-tuned not just for specific industries but also for individual developers, offering personalized feedback based on past interactions, coding styles, and preferred development frameworks.

6.7. Cross-Language and multimodal development

With the rise of LLMs in software engineering, there is growing interest in models that can understand and generate code across multiple programming languages. This capability would be especially useful for projects that involve integrating systems built in different languages or for teams with diverse language preferences. Research opportunities in this area include developing LLMs that are fluent in cross-language development, offering seamless transitions between languages and ensuring that code components written in different languages can work together efficiently. Additionally, multimodal LLMs that can integrate text, code, and even visual information (such as UI wireframes or architectural diagrams) offer exciting possibilities for the future of SE. These models could enable more comprehensive understanding of complex software systems, allowing developers to describe features in natural language while the LLM generates code, suggests optimizations, and aligns it with the visual or architectural elements of the project.

6.8. Education and training in the AI era

Lastly, the rise of LLMs in software engineering will have a profound impact on how future developers are trained and educated. As LLMs take over more of the rote coding tasks, the focus of SE education will likely shift toward higher-level problem-solving, system design, and ethical decision-making. Researchers will explore new pedagogical models that emphasize the collaboration between humans and AI, teaching developers not only how to code but also how to work effectively with AI tools. Future research in education will likely investigate how to integrate LLMs into software engineering curricula, ensuring that developers are well-prepared to work with AI-enhanced development tools. There will also be a need to develop new metrics for assessing coding skills, as the traditional focus on syntax and manual coding proficiency may become less relevant in a world where LLMs handle much of the low-level programming work.

7. Conclusion

The integration of LLMs into software engineering represents a significant turning point in how software is developed, maintained, and optimized. This article has explored the potential of LLMs to both enhance and challenge the current practices within the field of software engineering. Throughout the discussion, several important findings have emerged regarding the use of LLMs. They have proven to be game-changers across various phases of the software development lifecycle, including requirement analysis, code generation, testing, and debugging. By automating routine tasks and improving code quality, LLMs allow developers to focus on more complex and creative aspects of their work. Furthermore, LLMs can ensure consistency across large codebases and assist in maintaining legacy systems, thereby addressing technical debt effectively. However, while the potential of LLMs is vast, ethical concerns surrounding data bias, intellectual property, and job displacement must be carefully managed. The computational costs associated with training and deploying these large-scale models can also

be prohibitive, particularly for smaller organizations. In light of these findings, it is clear that LLMs are not merely a product of overhyped marketing; they represent a profound shift in how software is engineered. They should be seen as powerful tools that augment human capabilities rather than replace them. Human oversight remains crucial for ensuring that AI-generated code aligns with project goals, is secure, and is free from biases. Therefore, the verdict is that LLMs indeed are game-changers in software engineering, but their true potential can only be unlocked when combined with human expertise and ethical safeguards. For developers and organizations, embracing the rise of LLMs is not just a choice but a strategic imperative. Developers must become familiar with how LLMs can assist in coding, testing, debugging, and maintenance while continuing to refine their higher-level skills such as system design and ethical decision-making. Organizations should invest in integrating LLMs into their development environments, starting with pilot projects to gauge effectiveness, as this can reduce development costs, accelerate time-to-market, and enhance software quality. Educational institutions, too, should revise their software engineering curricula to prepare the next generation of developers for the future of AI-driven development.

The impact of this article extends beyond simply presenting the advantages and challenges of LLMs in software engineering; it provides a balanced and nuanced perspective that allows stakeholders to make informed decisions about adopting these technologies. By highlighting real-world case studies, technical strengths, ethical considerations, and future research opportunities, the article contributes to the growing discourse on AI-driven development tools and their place in the future of software engineering. Ultimately, it serves as a guide for developers, organizations, and researchers, helping them understand how LLMs can enhance workflows and the skills needed to remain competitive in an AI-driven landscape. As LLMs continue to evolve, their integration into software engineering practices will redefine what is possible in software development, pushing the boundaries of automation, creativity, and collaboration. Thus, this article offers a foundational understanding of how LLMs are poised to change the software engineering landscape, encouraging stakeholders to embrace these tools thoughtfully and strategically. In conclusion, LLMs hold the potential to significantly disrupt and enhance the software engineering process, and as developers and organizations adapt to these changes, they will find themselves at the forefront of a new era in software development—one that is faster, more efficient, and more collaborative than ever before.

Funding

The authors declare that no funds, grants, or other supports were received during the preparation of this manuscript.

Data availability

No datasets were generated or analyzed during the current study.

CRediT authorship contribution statement

Md. Asraful Haque: Writing – original draft, Methodology, Investigation, 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.

References

- [1] Guide to the Software Engineering Body of Knowledge (SWEBOK Guide), in: H. Washizaki (Ed.), *Guide to the Software Engineering Body of Knowledge (SWEBOK Guide)*, IEEE Computer Society, 2024.
- [2] R.S. Pressman, *Software Engineering: A Practitioner's Approach*, 5th Edition, McGraw-Hill Higher Education, 2001, ISBN- 0073655783.
- [3] P. Jalote, *Software Engineering: A Precise Approach*, Wiley, 2010, ISBN- 9788126523115.
- [4] M.A. Haque, N. Ahmad, Key issues in software reliability growth models, *Recent Adv. Comput. Sci. Commun.* 15 (5) (2022) 741–747.
- [5] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, Li Li, X. Luo, D. Lo, J. Grundy, H. Wang, Large language Models for Software engineering: a systematic literature review, *ACM Trans. Soft. Eng. Methodol.* (2024), <https://doi.org/10.1145/3695988>.
- [6] I. Ozkaya, Application of large language models to software engineering tasks: opportunities, risks, and implications, *IEEE Softw.* 40 (3) (2023) 4–8.
- [7] S. Rasnayaka, G. Wang, R. Shariffdeen, G.N. Iyer, An empirical study on usage and perceptions of LLMs in a software engineering project, in: *Proceedings of the 1st International Workshop on Large Language Models for Code*, 2024, pp. 111–118, <https://doi.org/10.1145/3643795.3648379>. Pages.
- [8] Y. Li, T. Zhang, X. Luo, H. Cai, S. Fang, D. Yuan, Do pretrained language models indeed understand software engineering tasks? *IEEE Trans. Software Eng.* 49 (10) (2023) 4639–4655.
- [9] Z. Liu, Y. Tang, X. Luo, Y. Zhou, L.F. Zhang, No need to lift a finger anymore? Assessing the quality of code generation by ChatGPT, *IEEE Trans. Software Eng.* 50 (6) (2024) 1548–1584.
- [10] A. Tarassow, 2023. The potential of llms for coding with low-resource and domain-specific programming languages. *arXiv preprint arXiv:2307.13018*.
- [11] J. Sallou, T. Durieux, A. Panichella, Breaking the silence: the threats of using LLMs in software engineering, in: *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, 2024, pp. 102–106. Pages.
- [12] A. Vaswani, et al., Attention Is All You Need, 30, *Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.
- [13] T.B. Brown, et al., Language models are few-shot learners, in: *Proc. of the 34th Int. Conf. on Neural Information Processing Systems (NIPS '20)*, 2020, pp. 1877–1901. Article 159.
- [14] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: *56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018, pp. 328–339, pagesMelbourne, Australia.
- [15] L. Ouyang, et al., Training language models to follow instructions with human feedback, in: *36th Conf. on Neural Information Processing Systems 35*, 2022, pp. 27730–27744.
- [16] M.A. Haque, S. Li, 2024. Exploring ChatGPT and its impact on society. AI and ethics. doi:10.1007/s43681-024-00435-4.
- [17] M. Lubbad, 2023. GPT-4 parameters: unlimited guide NLP's game-changer. Medium (March 19, 2023). Available online: <https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a>.
- [18] W.X. Zhao et al. 2023. A survey of large language models. *arXiv preprint arXiv: 2303.18223v11*.
- [19] R. Taylor, M. Kardaş, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic. 2022. Galactica: a large language model for science. *arXiv preprint arXiv:2211.09085*.
- [20] H. Touvron et al. 2023. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- [21] M.A. Haque, A brief analysis of 'ChatGPT' – A revolutionary tool designed by OpenAI, *EAI Endorsed Trans. AI and Robotics* 1 (1) (2023) e15.
- [22] Y. Chang, et al., A survey on evaluation of large language models, *ACM Trans. Intell. Syst. Technol.* 15 (3) (2024) 1–45. VollssueArticle No. 39Pages.
- [23] Jiang, J., et al. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.
- [24] S. Zhang, J. Wang, G. Dong, J. Sun, Y. Zhang, G. Pu. 2024. Experimenting a new programming practice with llms. *arXiv preprint arXiv:2401.01062*.
- [25] R.A. Husein, H. Aburajouh, C. Catal, Large language models for code completion: a systematic literature review, *Comput. Stand. Interf.* 92 (2025) 103917.
- [26] M. Welsh, The end of programming, *Commun. ACM* 66 (1) (2023) 34–35, <https://doi.org/10.1145/3570220>.
- [27] Z. Wang, J. Li, Ge Li, Z. Jin. 2023. Chatcoder: chat-based refine requirement improves llms' code generation. *arXiv preprint arXiv:2311.00272*.
- [28] Z. Yu et al. 2024. HumanEval Pro and MBPP Pro: evaluating large language models on self-invoking code generation. *arXiv preprint arXiv:2412.21199v2*.
- [29] K. Huang, et al., An empirical study on fine-tuning large language models of code for automated program repair, in: *Proceedings 38th IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE)*, 2023, pp. 1162–1174.
- [30] M.A. Haque, S. Li, The potential use of ChatGPT for debugging and bug fixing, *EAI Endorsed Trans. AI and Robot.* 2 (1) (2023) e4.
- [31] S. Kang, J. Yoon, N. Askarbekkyzy, S. Yoo, Evaluating diverse large language models for automatic and general bug reproduction, *IEEE Trans. Software Eng.* 50 (10) (2024) 2677–2694.
- [32] Z. Fan, X. Gao, M. Mirchev, A. Roychoudhury, S. Hwei Tan, Automated repair of programs from large language models, in: *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, Melbourne, Australia, 2023, pp. 1469–1481.

- [33] M. Schafer, S. Nadi, A. Eghbali, F. Tip, An empirical evaluation of using large language models for automated unit test generation, *IEEE Trans. Software Eng.* 50 (1) (2024) 85–105.
- [34] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, Q. Wang, Software testing with large language models: survey, landscape, and vision, *IEEE Trans. Software Eng.* 50 (4) (2024) 911–936.
- [35] Y. Wang et al. 2025. ProjectTest: a project-level LLM unit test generation benchmark and impact of error fixing mechanisms. arXiv preprint arXiv: 2502.06556v4.
- [36] J.A. Pizzorno and E.D. Berger. 2025. CoverUp: coverage-guided LLM-based test generation. arXiv preprint arXiv:2403.16218v3.
- [37] T. Xue, X. Li, T. Azim, R. Smirnov, J. Yu, A. Sadrieh, B. Pahlavan. 2024. Multi-programming language ensemble for code generation in large language model. arXiv preprint arXiv:2409.04114.
- [38] J. Zhang, P. Nie, J.J. Li, M. Gligoric, Multilingual code Co-evolution using large language models, in: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023)*, 2023, pp. 695–707.
- [39] Q. Peng, Y. Chai, X. Li, Humaneval-xl: a multilingual code generation benchmark for cross-lingual natural language generalization, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024, pp. 8383–8394, pagesTorino, Italia.
- [40] J. Xing, M. Bhatia, S. Phulwani, D. Suresh, R. Matta. 2025. HackerRank-ASTRA: evaluating correctness & consistency of large language models on cross-domain multi-file project problems. arXiv preprint arXiv:2502.00226v1.
- [41] A. Shirafuji, Y. Oda, J. Suzuki, M. Morishita, Y. Watanobe, in: *Refactoring Programs Using Large Language Models with Few-Shot Examples*. 30th Asia-Pacific Software Engineering Conference (APSEC-23), Seoul, South Korea, 2023, pp. 151–160.
- [42] S. Ishida et al. 2024. LangProp: a code optimization framework using large Language Models applied to driving. arXiv preprint arXiv:2401.10314.
- [43] P. Akiyamen. 2024. The unreasonable effectiveness of LLMs for query optimization. arXiv preprint arXiv:2411.02862v1.
- [44] Z. Yao et al. 2025. A query optimization method utilizing large language models. arXiv preprint arXiv:2503.06902v1.
- [45] A.D. Porta, et al., Using large language models to support software engineering documentation in waterfall life cycles: are we there yet?, in: *Proceedings of the 4th National Conference on Artificial Intelligence*, organized by CINI, May 29-30, Naples, Italy, 2024.
- [46] L. Belzner, T. Gabor, M. Wirsing, Large language model assisted software engineering: prospects, challenges, and a case study, in: *Bridging the Gap Between AI and Reality: First International Conference, AISoLA 2023, Crete, Greece, 2023*, pp. 355–374.
- [47] H. Jin, L. Huang, H. Cai, J. Yan, Bo Li, H. Chen. 2024. From LLMs to LLM-based Agents for Software Engineering: a survey of current, challenges and future. arXiv preprint arXiv:2408.02479.
- [48] F. Errica, G. Siracusano, D. Sanvito, R. Bifulco. 2024. What did I do wrong? Quantifying LLMs' Sensitivity and consistency to prompt engineering. arXiv preprint arXiv:2406.12334v1.
- [49] Y. Xia, J. Kim, Y. Chen, H. Ye, S. Kundu, C. Hao, N. Talati. 2024. Understanding the performance and estimating the cost of LLM fine-tuning. arXiv preprint arXiv: 2408.04693.
- [50] M.C. Rillig, M. Agerstrand, M. Bi, K.A. Gould, U. Sauerland, Risks and benefits of large language models for the environment, *Environ. Sci. Technol.* 57 (9) (2023) 3464–3466.
- [51] C. Tantithamthavorn, J. Cito, H. Hemati, S. Chandra, Explainable AI for SE: experts' interviews, challenges, and future directions, *IEEE Softw.* 40 (4) (2023).
- [52] B. Kou, S. Chen, Z. Wang, L. Ma, and T. Zhang. 2023. Do large language models pay similar attention like Human programmers when generating code?. arXiv preprint arXiv:2306.01220.
- [53] N. Perry, M. Srivastava, D. Kumar, D. Boneh, Do users write more insecure code with AI assistants?, in: *ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, 2023, pp. 2785–2799.
- [54] C. Kirchhubel, G. Brown, Intellectual property rights at the training, development and generation stages of Large language Models, in: *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024*, pp. 13–18, pages.
- [55] J. Jiao, S. Afroogh, Y. Xu, C. Phillips. 2024. Navigating LLM Ethics: advancements, challenges, and future directions. arXiv preprint arXiv:2406.18841.
- [56] S. Imai, Is GitHub copilot a substitute for human pair-programming? An empirical study, in: *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering (ICSE '22)*, Pittsburgh, Pennsylvania, 2022, pp. 319–321.
- [57] M. Jaworski, D. Piotrkowski. 2023. Study of software developers' experience using the GitHub Copilot Tool in the software development process. arXiv preprint arXiv: 2301.04991.
- [58] S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer. 2023. The impact of AI on developer productivity: evidence from GitHub copilot. arXiv preprint arXiv: 2302.06590.
- [59] D. Smit, H. Smuts, P. Louw, J., Pielmeier, C. Eidelloth, The impact of GitHub Copilot on developer productivity from a software engineering body of knowledge perspective, in: *AMCIS 2024 Proceedings*, 2024, p. 10.
- [60] A. Ziegler, E. Kalliamvakou, X.A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam, E. Aftandilian, Measuring GitHub copilot's impact on productivity, *Commun. ACM* 67 (2024) 54–63, 3 (March 2024).
- [61] R. Salva. 2025. Essentials of GitHub copilot. Available online: <https://resources.github.com/learn/pathways/copilot/essentials/essentials-of-github-copilot/>.
- [62] Revolutionizing Shopify's Engineering: The Power of GitHub Copilot, Available online, <https://www.toolify.ai/ai-news/revolutionizing-shopifys-engineering-the-power-of-github-copilot-821646>, 2024.
- [63] A. Ambo. Empowering devs with AI: how Shopify made GitHub Copilot core to its culture. 2023. Available online: <https://medium.com/@AmboAtsushi/empowering-devs-with-ai-how-shopify-made-github-copilot-core-to-its-culture-d237ba09d61b>.
- [64] Impactful work: helping developers around the world improve productivity with AI. 2023. Available online: <https://aws.amazon.com/careers/life-at-aws-impactful-work-helping-developers-around-the-world-improve-productivity/>.
- [65] V.K. Sikha, AI fueled transformation in application development & coding, *Int. J. Commun. Networks Inf. Security* 15 (04) (2023).
- [66] V. Joshi, I. Band. 2024. Disrupting test development with AI assistants. arXiv preprint arXiv:2411.02328.
- [67] Tabnine: using Google Cloud powered AI to help developers code faster. 2023. Available online: <https://cloud.google.com/customers/tabnine>.
- [68] K.-A. Marvel. 2025. Codeium: the best github copilot alternative. Available online: <https://semaphore.io/blog/codeium>.
- [69] Clearwater Analytics on Codeium, Available online, <https://codeium.com/blog/clearwater-analytics-case-study>, 2024.
- [70] A. Masood, S. Dahal, A. Cai, G. Burtini. 2022. Ghostwriter AI & Complete Code Beta. Replit Blog: <https://blog.replit.com/ai>.
- [71] A. Masad. 2023. Replit's path to product-market fit — The \$1 billion side project. Available online: <https://review.firstround.com/replits-path-to-product-market-fit/>.
- [72] Cody + Leidos: maximizing efficiency with heightened security in the AI race. 2025. Available online: <https://sourcegraph.com/case-studies/cody-leidos-maximizing-efficiency-heightened-security-ai-race>.
- [73] N. Maslej, L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J.C. Niebles, Y. Shoham, R. Wald, J. Clark, The AI Index 2024 Annual Report, in: *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, CA, 2024*.
- [74] Kyle Daigle & GitHub Staff. 2024. Survey: the AI wave continues to grow on software development teams. GitHub blog. Available online: <https://github.blog/news-insights/research/survey-ai-wave-grows/#key-survey-findings>.
- [75] D. Russo, Navigating the complexity of generative AI adoption in software engineering, *ACM Trans. Softw. Eng. Methodol.* 33 (2024) 50, <https://doi.org/10.1145/3652154>, 5, Article 135 (June 2024)pages.
- [76] A. Drake. 2024. The future of software engineering: LLMs and beyond. Comet (February 28, 2024). Available online: <https://www.comet.com/site/blog/the-future-of-software-engineering-llms-and-beyond/>.
- [77] J. Monti. 2024. The future of AI-driven software development. Medium (Mar 8, 2024). Available online: <https://joemonti.org/the-future-of-ai-driven-software-development-0dec24759a71>.
- [78] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekk, D. Doermann, Future of software development with generative AI, *Automat. Software Eng.* 31 (2024) 26, <https://doi.org/10.1007/s10515-024-00426-z>.
- [79] V. Terragni, P. Roop, K. Blincoe. 2024. The future of software engineering in an AI-driven world. arXiv preprint arXiv:2406.07737.



Full length article

Evaluatology's perspective on AI evaluation in critical scenarios: From tail quality to landscape

Zhengxin Yang *

University of Chinese Academy of Sciences, Beijing, China

Research Center Of Distributed Systems, State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:

Tail Quality
Evaluatology
AI inference

ABSTRACT

Tail Quality, as a metric for evaluating AI inference performance in critical scenarios, reveals the extreme behaviors of AI inference systems in real-world applications, offering significant practical value. However, its adoption has been limited due to the lack of systematic theoretical support. To address this issue, this paper analyzes AI inference system evaluation activities from the perspective of Evaluatology, bridging the gap between theory and practice. Specifically, we begin by constructing a rigorous, consistent, and comprehensive evaluation system for AI inference systems, with a focus on defining the evaluation subject and evaluation conditions. We then refine the Quality@Time-Threshold (Q@T) statistical evaluation framework by formalizing these components, thereby enhancing its theoretical rigor and applicability. By integrating the principles of Evaluatology, we extend Q@T to incorporate stakeholder considerations, ensuring its adaptability to varying time tolerance. Through refining the Q@T evaluation framework and embedding it within Evaluatology, we provide a robust theoretical foundation that enhances the accuracy and reliability of AI system evaluations, making the approach both scientifically rigorous and practically reliable. Experimental results further validate the effectiveness of this refined framework, confirming its scientific rigor and practical applicability. The theoretical analysis presented in this paper provides valuable guidance for researchers aiming to apply Evaluatology in practice.

1. Introduction

With the rapid advancement of artificial intelligence (AI) technologies, evaluating AI inference systems has become increasingly critical. These systems operate in dynamic and unpredictable environments, ranging from the online deployment of large-scale language models such as ChatGPT [1,2] to real-time applications in autonomous driving [3–5] and smart healthcare [6,7]. The increasing reliance on AI in these critical domains introduces significant challenges. For instance, online real-time recommendation systems are crucial in e-commerce and content streaming platforms, directly affecting user engagement and satisfaction. Delays in inference can lead to user impatience and churn, impacting the overall effectiveness of these systems. Similarly, autonomous driving, a safety-critical domain, requires real-time decision-making for tasks such as object detection and lane detection [8], with even minor errors posing catastrophic risks to safety [9,10]. These challenges necessitate the development of reliable and objective methods to evaluate the inference performance of AI systems.

However, traditional evaluation methods often face two major issues. First, evaluations frequently rely on isolated metrics, such as accuracy or inference throughput, without accounting for the complex interactions between various factors [11]. These single-dimensional evaluations fail to capture the true performance of AI inference systems, especially when real-world, dynamic conditions are taken into account. Second, AI inference system evaluation remains a complex and uncertain process due to the inherent intricacies of computer systems [12] and the absence of well-established, interpretable theories—particularly for systems equipped with neural networks [13,14]. Without clear, theoretically grounded evaluation criteria, existing methods often fall short, relying on industry standards without comprehensive theoretical analysis [15–18]. This theoretical gap poses challenges for researchers trying to evaluate AI inference systems, often leading to questions about the reliability of existing evaluation methods.

To address the issues, **Quality@Time-Threshold (Q@T)** was introduced by Yang et al. [11] as a metric designed to measure how an AI inference system's quality fluctuates under strict time constraints. By

* Correspondence to: Research Center Of Distributed Systems, State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

E-mail address: yangzhengxin17z@ict.ac.cn.

<https://doi.org/10.1016/j.tbench.2025.100203>

Received 24 January 2025; Received in revised form 6 April 2025; Accepted 28 April 2025

Available online 27 May 2025

2772-4859/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

considering both inference time and quality, Q@T provides a statistical evaluation framework that captures the “Tail Quality” phenomenon—extreme fluctuations in inference quality that are often overlooked by traditional methods [15,18]. This is especially crucial in critical applications such as autonomous driving and medical diagnostics, where poor performance could lead to severe consequences [11]. While Q@T offers an important step forward in AI evaluation, it still faces challenges due to the absence of a solid theoretical foundation, particularly the lack of guidance from Evaluatology. This theoretical gap raises concerns about the reliability and rigor of Q@T, especially when applied in diverse scenarios. Furthermore, Q@T’s applicability is limited in contexts where flexible time constraints are necessary, as it is more suited for strict time thresholds. Thus, the need for a more robust theoretical framework becomes evident, one that can guide the evaluation process in a scientifically grounded way.

In this paper, we aim to address these challenges by analyzing Q@T from the perspective of **Evaluatology**, a formal science of evaluation. Evaluatology offers a systematic methodology for modeling and understanding complex systems, providing a theoretical foundation for improving evaluation methods. By applying Evaluatology’s five core axioms and standard evaluation methodology, we redefine and clarify the core components of the evaluation system in Q@T—namely, the evaluation subject and evaluation conditions. This approach helps overcome the limitations of Q@T, ensuring that it is both scientifically rigorous and adaptable to different application scenarios, such as autonomous driving, healthcare, and e-commerce.

This paper’s main contributions are as follows:

- **Theoretical Validation of Q@T:** We integrate Evaluatology’s principles with Q@T to provide a stronger theoretical foundation for evaluating AI inference systems.
- **Bridging Theory and Practice:** We bridge the gap between theoretical foundations (Evaluatology) and the practical evaluation of AI systems (Q@T).
- **Refining Q@T:** We refine the Q@T framework by introducing a more flexible approach that incorporates stakeholder needs and time constraints, enhancing its versatility across real-world applications.

2. Background

This section introduces two key aspects: Evaluatology [19–21] and Q@T [11]. Section 2.1 introduces the five core axioms of Evaluatology and the standard evaluation methodology. Section 2.2 briefly introduces the definition of the Q@T evaluation metric and the statistical evaluation framework for assessing Q@T in AI inference systems.

2.1. Evaluatology

This section introduces the five core axioms of Evaluatology and the standard evaluation methodology. These are the two most fundamental components of Evaluatology, providing a rigorous theoretical foundation for evaluating complex systems such as AI inference systems. Applying these principles, we can reanalyze Q@T and ensure its evaluation framework is scientifically sound and practically applicable.

2.1.1. The five core axioms of Evaluatology

The five core axioms of Evaluatology form the foundation of its evaluation methodology, detailed as follows:

- **The Axiom of the Essence of Composite Evaluation Metrics:** A composite evaluation metric either has inherent physical meaning or is defined by a value function that combines base quantities.
- **The Axiom of True Evaluation Outcomes:** For a well-defined evaluation system, when well-defined evaluation conditions are applied to a well-defined evaluation subject, its evaluation outcomes must possess true values.

- **The Axiom of Evaluation Traceability:** For the same evaluation subject, differences in evaluation outcomes can be attributed to variations in the evaluation conditions (ECs), ensuring traceability and interpretability of evaluation outcomes.
- **The Axiom of Comparable Evaluation Outcomes:** Evaluation outcomes are comparable only if evaluation subjects are evaluated under equivalent evaluation conditions (EECs).
- **The Axiom of Consistent Evaluation Outcomes:** Evaluation outcomes from different samples within a population of evaluation conditions consistently converge toward the true evaluation outcomes of the entire population.

By adhering to these axioms, the evaluation methodology provides a rigorous, reliable, and comparable framework for evaluating complex systems, including AI inference systems, ensuring consistency, accuracy, and reliability across various application scenarios.

2.1.2. The standard evaluation methodology

In Evaluatology, the standardized evaluation methodology consists of four key steps. These are summarized as follows:

Defining and characterizing the subject: The first step is clearly defining and describing the subject to be evaluated. A well-defined subject is essential for valid comparisons between different instances of the same subject definition. This stage also involves modeling the subject, which includes outlining its detailed structure. A rigorous definition and consensus on the subject’s model among stakeholders are crucial for ensuring the validity of the evaluation process.

Defining and clarifying the evaluation system (ES): The second step is constructing a minimal yet complete Evaluation System (ES) that operates autonomously. This is a crucial component of the evaluation process and must meet two main criteria: it must function independently and encompass all the essential factors needed for the evaluation task. It is important to note that any changes to the independent factors in the ES will impact the final evaluation results. Defining the ES comprehensively is challenging because too many factors can lead to excessive evaluation costs, while too few may fail to capture all the critical influences on the results.

Acquiring the evaluation conditions (ECs): Once the ES is defined, the next step is establishing Evaluation Conditions (ECs). These conditions are derived by isolating the subject from the ES. Defining these conditions ensures the evaluation process is based on realistic and controlled parameters.

Determining the evaluation methodologies: After defining the ES and ECs, the final step is to analyze the nature of the ES and determine the appropriate evaluation methodologies. The key goal in this phase is to ensure that the evaluation methods meet the standards of Evaluatology’s five core axioms, ensuring the comprehensiveness and reliability of the evaluation process.

In conclusion, the standardized evaluation methodology in Evaluatology involves clearly defining the evaluation subject, establishing the evaluation system, defining the evaluation conditions, and selecting the appropriate evaluation methods. These steps ensure that complex systems, such as AI inference systems, are evaluated rigorously and systematically. The evaluation process can produce consistent, accurate, and comparable results by following these steps.

2.2. Q@T and tail quality

The Quality@Time-Threshold (Q@T) metric was proposed to measure the ability of an AI inference system to maintain stable high inference quality under strict time constraints. This is of practical significance, as an AI inference system with high-quality predictions should achieve high and stable inference quality even under lower time thresholds.

2.2.1. Definition of Q@T

Q@T evaluates AI inference systems by balancing inference quality and time constraints. Given a dataset $D = \{x_i, y_i\}_{i=1}^n$, where x_i represents the input and y_i represents the ground truth, the model M generates prediction $y'_i = M(x_i)$. The overall inference quality q is determined by comparing $Y' = \{y'_i\}_{i=1}^n$ with the ground truth $Y = \{y_i\}_{i=1}^n$ using a quality evaluation function (e.g., accuracy or F-score). To account for the impact of inference time, the validity of each inference result is determined by whether the inference time exceeds a specified time threshold θ . This leads to the following equation for Q@T:

$$q_\theta = \text{evaluate}(\{M(x_i) \cdot \mathbf{1}_\theta + \text{error} \cdot (1 - \mathbf{1}_\theta)\}_{i=1}^n, \{y_i\}_{i=1}^n), \quad (1)$$

where $\mathbf{1}_\theta$ is an indicator function that returns 1 if the inference of x_i completes within the threshold θ , and 0 otherwise. The placeholder **error** denotes a default output substituted when the inference time exceeds the threshold, thereby effectively invalidating the corresponding model output. The function $\text{evaluate}(\cdot, \cdot)$ computes a quality metric — such as accuracy — between the (potentially **error**-substituted) model outputs and the corresponding ground-truth labels.

Q@T offers a comprehensive evaluation that is especially useful for real-time applications where quality and time are critical.

2.2.2. Statistical evaluation framework for Q@T

In AI inference systems, inference time can vary significantly due to factors such as hardware configurations, deep learning frameworks, and data processing pipelines. This variability impacts the estimation of Q@T, as fluctuations in inference time influence quality under specific time constraints.

To evaluate Q@T accurately, the statistical framework models inference time as a random variable T , which follows an unknown distribution D and is influenced by various system components. The Q@T metric becomes a random variable Q dependent on T and the system components C' , expressed as a conditional probability distribution:

$$Q_\theta = f(T \mid \theta, C^1, C^2, \dots), \quad T \sim D. \quad (2)$$

The framework uses the Monte Carlo simulation to collect inference time samples and Kernel Density Estimation (KDE) to estimate the distribution of these times. This non-parametric approach avoids assumptions about the form of the distribution. Convergence is monitored using Jensen–Shannon Divergence (JSD), stopping the simulation when the distribution stabilizes.

The steps include:

- **Sampling Inference Time:** Collect inference time samples using Monte Carlo simulations across multiple rounds.
- **Kernel Density Estimation (KDE):** Apply KDE to estimate the probability density function $f(t)$ of inference time.
- **Convergence Check using Jensen–Shannon Divergence (JSD):** Calculate the JSD between distributions from different sample sizes, stopping when the JSD is sufficiently small, indicating convergence.

After convergence is achieved, the framework performs N independent evaluation trials to quantify the quality metric under the stabilized distribution. Each trial produces a sample q_i from the random variable Q , resulting in a set of observations $\{q_i\}_{i=1}^N$.

The final Q@T metric is computed as a statistical characterization of Q_θ when the time threshold is set to $\theta = T$, based on these samples:

$$Q@T = S_{\theta=T}(\{q_i\}_{i=1}^N), \quad q_i \sim Q_{\theta=T}, \quad (3)$$

where $S_{\theta=T}(\cdot)$ denotes a set of statistical characteristics such as the sample mean, variance, and quantiles. This formulation enables a comprehensive representation of inference quality under time constraints, going beyond reliance on a single observation or the expected value alone.

Once a reliable distribution is obtained, the framework computes the final Q@T values, reflecting both the variability in inference time and the extreme fluctuations in inference quality.

2.2.3. The extreme tail quality phenomenon

The Tail Quality phenomenon arises from the statistical evaluation framework in Q@T, which generates a distribution of inference quality values instead of a single-point estimate. Tail Quality specifically refers to the extremely low-quality values observed at the tail of this distribution. This highlights how Q@T can reveal performance variations that traditional evaluation methods might overlook.

This phenomenon underscores the importance of Q@T in evaluating AI systems, especially in critical or real-time applications, where such extreme deviations could have serious consequences. This makes Q@T a valuable tool for assessing AI systems, it provides a more comprehensive understanding of the system's inference ability under strict time constraints.

However, a limitation of Q@T is that it is focused on evaluating systems within a predefined time threshold. In cases where time constraints are less stringent, Q@T may not be as applicable. This limitation will be addressed in Section 4, where potential extensions and improvements to the Q@T framework are discussed, making it more versatile and suitable for a wider range of scenarios.

3. Reanalyzing AI inference evaluation from the perspective of evaluatology

In the context of Evaluatology, an evaluation system (ES) is defined as the smallest autonomous, self-contained system capable of operating automatically. The evaluation system can be broken down into two parts: evaluation subject and evaluation conditions. The evaluation subject refers to the “thing” being evaluated [20], which can be either an individual or a system. This is the core of the entire evaluation framework [19]. By precisely defining the subject, we can distinguish which part of the evaluation system the final evaluation outcomes belong to. The evaluation conditions, on the other hand, encompass all factors of the evaluation system other than the subject, serving as the primary determinants that influence the evaluation outcomes. When evaluating AI inference systems, defining both the evaluation subject and conditions presents challenges due to the inherent complexity of AI and computer systems. Therefore, in this section, we will follow a process where we first clearly define the evaluation system (Section 3.1), then separate the evaluation subject from the system (Section 3.2), and finally, establish the evaluation conditions (Section 3.3).

3.1. Clarifying primary components constitute evaluation system

This section provides a detailed discussion and analysis of the components that make up the evaluation system (ES) in the context of evaluating AI inference systems. The primary purpose of AI inference systems is to make predictions across various AI inference activities. Therefore, the key to constructing the evaluation system (ES) is to understand the structure of these inference activities. AI inference activities are complex and multifaceted, involving various components that interact with one another. These activities are hierarchical and require several components to work in concert to produce accurate predictions. This inherent complexity is a central challenge in evaluating AI inference systems. Below, we provide a concrete definition of the evaluation system, the primary framework we use to define the evaluation systems for AI inference tasks. Fig. 1 also illustrates the general structure of the entire framework.

Application Scenarios & Tasks: First, we must define the highest level of the evaluation system, which involves clarifying the problem definition for the AI inference activities. This includes identifying the application scenario and the tasks the AI inference activities are expected to perform. The definition of the application scenario and tasks is crucial in determining the evaluation criteria and metrics for the entire evaluation. For example, in e-commerce applications, the most important task for the AI inference system is often recommendation, where recommendation accuracy and latency are the core evaluation

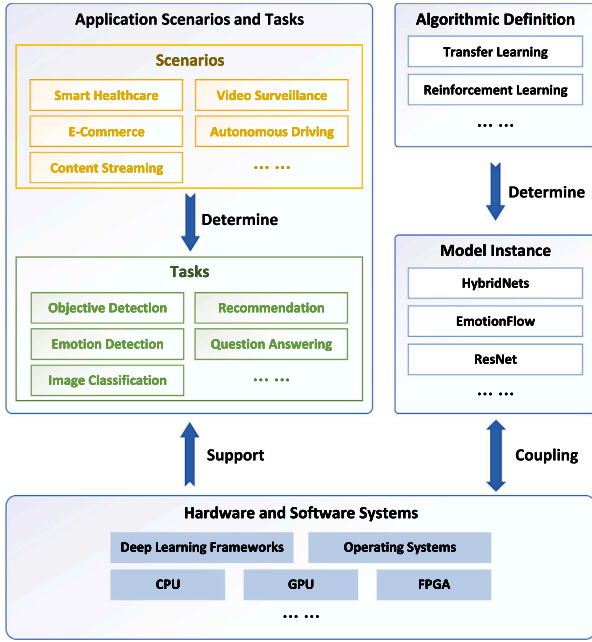


Fig. 1. An overview of the main components that make up the evaluation system in the context of evaluation AI inference systems. The diagram illustrates a hierarchical relationship, progressing from application scenarios and tasks (top left) to algorithmic definitions (top right) and then to specific model instances (middle right) and their execution environments (bottom). Each layer influences the instantiation of the next: application requirements determine which tasks are relevant, which in turn guide the selection of algorithmic approaches and models. The arrow labeled “Determine” indicates downward decisions or constraints, while the arrow labeled “Support” denotes foundational infrastructure (e.g., hardware/software platforms supporting model execution). “Coupling” highlights tight interdependencies between components, such as how models and algorithms must adapt to the capabilities and constraints of the hardware/software stack. This structure underscores the complexity of defining the evaluation subject, where inference performance emerges from interactions across all layers.

metrics [22]. However, in autonomous driving, the evaluation system must account for real-time tasks such as object detection, drivable area segmentation, and lane line detection [8]. Given the critical nature of real-time decision-making in this scenario, time constraints become particularly important, and the evaluation system must emphasize the inference time.

Algorithmic Definition: After identifying the AI inference system’s application scenario and tasks, the next step is to clarify the algorithmic definition required to accomplish these tasks. This is a critical layer in the evaluation system as the algorithms form the foundation for the system’s inference capabilities. Since this evaluation is focused on assessing the inference level of AI systems, we concentrate on neural network-based algorithms. Additionally, these algorithms must be clearly defined, including the input and output data requirements, which are typically influenced by both the application scenario and the specific tasks.

For example, in the driver assistance scenario, the system needs to identify abnormalities in the driver’s facial expressions and voice to prevent potential accidents [23,24]. In the smart healthcare scenario, emotion recognition might focus on monitoring the patient’s bioelectrical signals to avoid unforeseen incidents during medical procedures [23,25]. In both cases, while the emotion recognition remains the same, the specific data sources and algorithmic needs vary due to the differing application contexts.

Model Instances: Next, the core implementation of the algorithm must be defined. This involves specifying which neural network model will implement the task objectives. The choice of neural network model directly impacts how the algorithm processes input and output

data. The chosen model must also account for environmental factors and data variances in real-world applications. For instance, if the system encounters rain or snow in autonomous driving, the model trained on a general dataset might require additional complex data processing for the images captured by the vehicle’s cameras. However, suppose the model is trained using a specialized dataset that includes weather-related variations [5,26,27]. It may not require these additional processing steps and could still achieve accurate object detection in adverse weather conditions.

Hardware & Software Systems: Finally, to support the algorithm’s functioning, complete hardware and software infrastructure are necessary to run the neural network model efficiently. The hardware configuration typically includes CPUs, GPUs, and potentially specialized hardware such as FPGA or TPU. In real-time performance evaluations, hardware is critical because it directly affects the inference speed. For instance, an AI model might perform well on a powerful GPU but may face delays on a CPU, especially in time-sensitive applications [28]. Furthermore, hardware selection must also consider other factors such as energy consumption, size, and form factor, particularly for embedded systems. On the software side, this includes libraries, frameworks, and operating systems that facilitate the execution of AI models, such as TensorFlow [29], PyTorch [30], and others. These frameworks optimize the computation processes of the neural network model but interact directly with the hardware, and variations in the specific configurations and versions used can result in differences in accuracy and performance [31].

In summary, we have analyzed several primary factors the evaluation system contains: the application scenario, evaluation metrics, data processing, algorithm selection, and computer hardware/software infrastructure. These factors are interdependent and form a complete evaluation system for AI inference activities [15,32]. However, a key challenge remains in clearly delineating the boundaries between the evaluation subject and the evaluation conditions. This is a critical issue for understanding how different components of the evaluation system interact and how their influence on the final evaluation outcome can be isolated and measured.

3.2. Defining and identifying evaluation subject

Clarifying the evaluation subject is essential for a rigorous evaluation, as it directly determines the boundaries within which the evaluation outcomes will be interpreted. However, this task is not trivial in AI inference systems evaluation. This is particularly evident in the work of Yang et al. [11], where the definition of the AI inference system itself is somewhat ambiguous. This lack of a clear and universally agreed-upon definition poses challenges in accurately assessing the performance of AI inference systems in complex environments.

The reason for this ambiguity stems from the inherent complexity of the evaluation system [12,32]. As shown in Fig. 1, the evaluation system comprises multiple interrelated components, each of which affects the system’s overall performance. A neural network model can be seen as an instantiation of an algorithm designed to solve a particular task and as a component that must be integrated seamlessly with the computational hardware and software environment. The interaction between these components creates a complex, tightly coupled system. Therefore, a central question in evaluatology for AI inference systems is whether or not the neural network model should be considered part of the evaluation subject.

3.2.1. Definition of evaluation subject: Excluding model instance

One potential approach to defining the evaluation subject is to exclude the neural network model instance and treat only the computational hardware and software system as the subject being evaluated. As illustrated in Fig. 2, the model instance is considered an input to the evaluation subject, which processes it within a defined task.

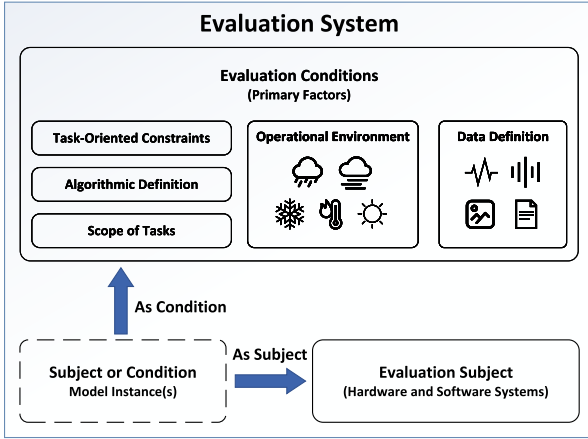


Fig. 2. The composition of the Evaluation System, including the Primary Factors as Evaluation Conditions, the Hardware and Software system as the Evaluation Subject, and the Model Instance, which may be considered part of the Subject or the Condition, depending on the specific evaluation objectives.

From this perspective, the evaluation is directed at understanding the general capability of the AI inference system to perform inference tasks rather than measuring the specific performance of the neural network model and a computer system. This means that different implementations of the same task — such as two distinct models solving the same problem — would not affect the evaluation, as the focus is on the hardware–software system and its ability to process data efficiently. For example, in autonomous driving, the evaluation would focus on the system’s ability to process input data from sensors and generate timely decisions independent of the specific model implementation used for object detection.

3.2.2. Definition of evaluation subject: Including model instance

The second definition of the evaluation subject involves treating both the neural network model and the hardware/software system as part of the evaluation subject, as shown in Fig. 2. Here, the specific model implementation (e.g., a particular deep learning network) becomes a key factor in determining the system’s ability to execute inference tasks. In this case, the evaluation would measure the system’s overall inference capability, including inference quality and inference time, while considering the effects of different model implementations. This does not mean that the first definition cannot measure the system’s overall performance, but rather that eliminating the impact of model instances would require greater costs to measure the overall performance that is universally applicable under specific tasks. For example, a model trained with a specialized dataset for handling weather-related changes in autonomous driving would perform differently than a model trained on a generic dataset.

3.3. Establishing evaluation conditions

According to Evaluatology, evaluation conditions (ECs) are determined by isolating the evaluation subject within the evaluation system. ECs encompass all external factors that influence the performance of the AI inference system. Therefore, in this study, the evaluation conditions (ECs) naturally include practical application scenarios, tasks, algorithmic definitions, and the model itself, depending on whether the model instance is considered part of the evaluation subject. Specifically, based on the analysis in Section 3.1, these components define the primary factors that must be considered within the EC, as detailed in the following sections.

Operational Environment: In autonomous driving applications, factors such as weather (rain, snow, fog), traffic density, and road

conditions can significantly affect the evaluation outcomes. The system must process complex sensor data in real time while also meeting the computational demands of the model’s inference, partly due to the intrinsic complexity of the subject itself. To address this, by isolating the subject and removing confounding factors through statistical methods within the ECs, we can focus on understanding how specific environmental factors impact the evaluation results. As a result, ECs must account for these environmental variations and simulate different conditions to assess the system’s robustness.

Data Definition: In real-world applications, AI systems must handle data that can be multimodal, noisy, or incomplete. For example, in autonomous driving, the system may receive data from multiple sensors, such as cameras and LiDAR systems. Thus, the ECs must clearly define the form of input data in the evaluation process, ensuring that data quality is consistently considered. How data from each sensor is processed, combined, and interpreted is crucial for accurately assessing the system’s overall outcomes.

Task-Oriented Constraints: Tasks like object detection in autonomous driving have strict real-time performance requirements, whereas tasks like clinical diagnostics may not have stringent time constraints but must still meet high-precision standards. In the context of Q@T, the time threshold θ is a central evaluation condition used to determine whether the system’s inference time is acceptable. If the inference time exceeds this threshold, the system’s performance may still be considered inadequate, even if the inference quality is high. Therefore, ECs must include these real-time constraints, which may vary depending on the application’s requirements.

Scope of Tasks: Additionally, we propose that the scope of tasks for evaluation should be appropriately limited. While some researchers may wish to assess the general inference capability of an AI system across all possible tasks, doing so would result in an explosion of the EC space, significantly increasing evaluation costs. Moreover, we argue that evaluating an AI inference system’s general processing ability across a broad range of tasks does not align with real-world application needs. Instead, focusing on specific, well-defined tasks relevant to the system’s intended deployment is more practical and efficient.

Algorithmic Definition: ECs should specify the AI inference algorithm used, including whether it is based on transfer learning, reinforcement learning, or other approaches. The choice of algorithm significantly influences how input data is processed and how inference results are generated. Therefore, it is essential to include this information in the ECs to ensure consistency and comparability when evaluating the system’s performance under different algorithmic conditions.

Model Instances: Lastly, if the model instance is not included in the evaluation subject, the ECs must be designed to eliminate the effects of variations in model instances on evaluation outcomes. This requires carefully controlling model configurations and ensuring that the evaluation reflects the underlying AI inference system’s performance rather than the idiosyncrasies of a specific model instance.

In conclusion, by clearly defining the evaluation conditions (ECs), we ensure that the AI inference system is evaluated within comparable contexts, guaranteeing the consistency and reliability of the evaluation outcomes. This structured approach allows for a more accurate, reproducible, and objective evaluation of AI systems in complex, real-world applications, ensuring that the evaluation is scientifically rigorous and practically applicable.

4. Refining Q@T evaluation framework

In this section, we refine the Q@T evaluation framework by formalizing the evaluation subject and conditions based on the earlier analysis using Evaluatology 4.1. This formalization ensures clear definitions of the evaluation subject and conditions, providing a more rigorous and consistent framework for evaluating AI inference systems. Through experiments, we further emphasize the importance of tail quality as an indicator of extreme performance in AI inference system evaluation 4.2.

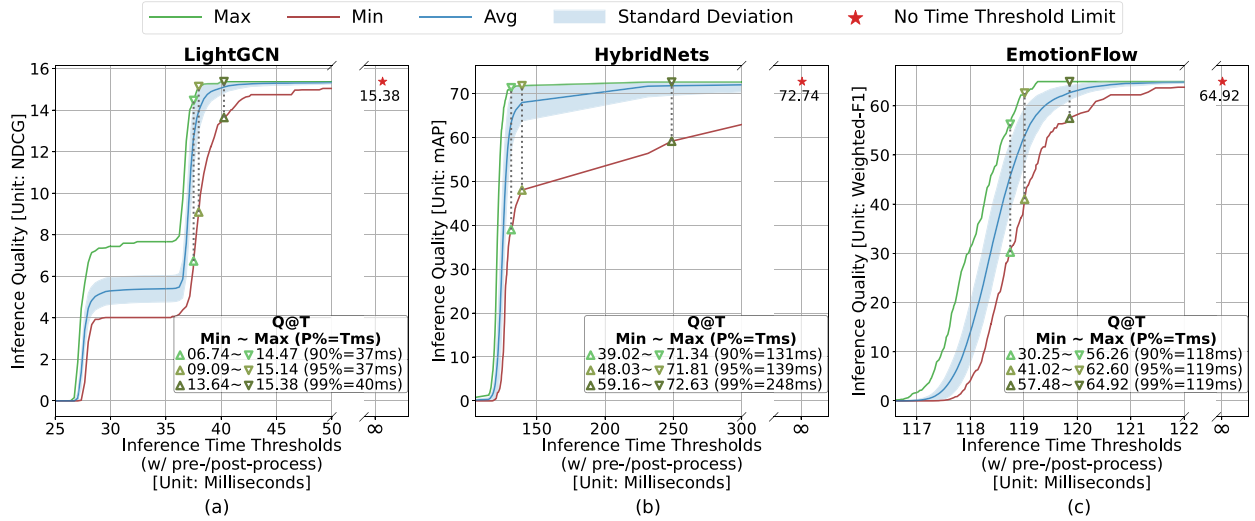


Fig. 3. Comparative Q@T evaluation under multiple inference time thresholds for three representative AI models: LightGCN (e-commerce recommendation), HybridNets (autonomous driving), and EmotionFlow (emotion recognition). The horizontal axis indicates inference latency (including data pre- and post-processing), while the vertical axis reports task-specific quality metrics (NDCG, mAP, and weighted-F1, respectively). For each model, Q@T is computed at different percentiles of inference latency, with key tail-latency points (e.g., 90%, 95%, 99%) annotated. Max/min/average Q@T scores and standard deviation are also indicated, revealing the relative performance and stability of each system under varying time constraints.

Finally, we discuss the landscape of stakeholder considerations, highlighting how incorporating the time tolerance of different stakeholders into the Q@T framework enhances the adaptability of the evaluation process 4.3.

4.1. Formalization of evaluation subject and conditions

From the perspective of the Q@T framework, Yang et al. [11] adopts the second definition of the evaluation subject, as described in Section 3.2, where both the model instance and the computational infrastructure (hardware/software system) are considered a unified entity. This approach aligns with Q@T's focus on evaluating the overall AI inference system, especially where the model and hardware are tightly integrated. However, the Q@T framework is flexible enough to accommodate the first definition of the evaluation subject, where the model instance is excluded. By sufficiently sampling the model instance space, Q@T can eliminate the impact of specific model instances on the evaluation outcomes, inspired by Evaluatology. This adaptability allows Q@T to evaluate AI systems from different perspectives, depending on the goals of the evaluation.

Building on this flexibility, the Q@T evaluation framework establishes a causal relationship between the system configuration C^i , inference time T , and quality Q , as shown in Eq. (2). This relationship allows for the assessment of the AI inference system's performance across varying configurations, even though Yang et al. [11] does not explicitly define the evaluation subject.

Formalization Given the inherent complexity of the evaluation subject, the evaluation system must operate as a whole to objectively assess the inference capability of AI systems. This requires running the system multiple times to isolate the effects of this complexity. Statistical indicators, such as averages and confidence intervals, are essential to represent the final evaluation outcomes, mitigating the influence of confounding factors and ensuring the consistency and reliability of the evaluation process. The Q@T evaluation framework addresses this challenge by incorporating statistical methods that improve the rigor of the evaluation.

While the Q@T metric was originally designed to explore the relationship between quality Q and time T , quantifying the trade-off between inference quality and time, the evaluation framework in Yang et al. [11] lacks clear definitions of the evaluation subject and conditions. Therefore, we propose distinguishing the elements within C^i into

primary factors of evaluation conditions and **inherent factors** of the subject. Specifically, primary factors should be considered as factors P^i that influence the evaluation outcomes and can be actively identified and controlled. In contrast, elements belonging to the subject should be regarded as inherent factors I^j , which may affect the evaluation results due to system complexity but are difficult to observe or isolate explicitly. Therefore, the original evaluation framework, as modeled by Eq. (2) can be reformulated as the following equation:

$$Q_\theta = f(T | \theta, \{P^i\}_{i=1}^n; \{I^j\}_{j=1}^m), \quad T \sim \mathcal{D}. \quad (4)$$

This distinction reflects a fundamental challenge in evaluation design: Although the identification and control of primary factors enable the construction of a self-contained evaluation system, the presence of inherent factors introduces variability that cannot be eliminated through experimental control alone. Instead, their influence must be mitigated through statistical methods. The statistical evaluation framework for Q@T thus plays a critical role in ensuring that the evaluation results remain robust and generalizable, despite the uncontrollable complexity of the inference system.

Based on the revised formulation in Eq. (4), we can derive practical principles for carrying out the evaluation process. When analyzing these factors, it is essential to vary one condition at a time — either from P^i or I^j — to isolate the effects of each. In practice, however, because inherent factors I^j are often unobservable or difficult to control, statistical methods must be applied first to reduce their influence. This requires keeping the ECs P^i fixed during repeated measurements, enabling the identification and mitigation of inherent variability. Only after this stabilization can we vary the ECs to meaningfully compare evaluation outcomes across different settings. Following this principle ensures that the evaluation process remains consistent and comparable, adhering to the five axioms of Evaluatology and guaranteeing the reliability and objectivity of the evaluation results.

Experiments We conducted three groups of experiments across diverse AI tasks to evaluate the effectiveness and generalizability of our proposed evaluation framework. As shown in Fig. 3, these experiments illustrate how Q@T varies under different inference time thresholds:

- **E-commerce Recommendation (LightGCN [22]):** As shown in Fig. 3(a), Q@T effectively tracks the changes in NDCG (Normalized Discounted Cumulative Gain), which measures the ranking quality of

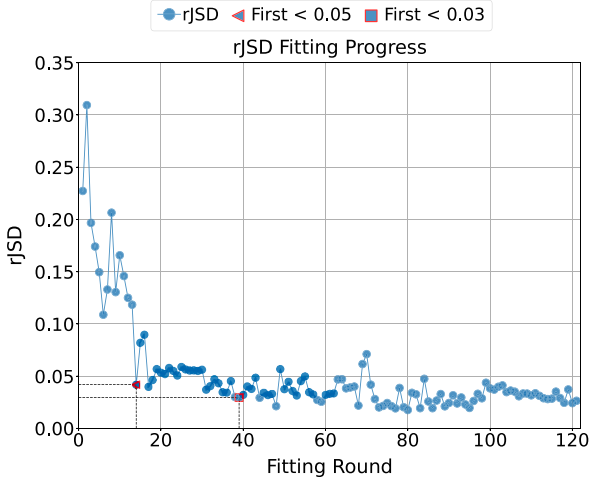


Fig. 4. Convergence validation of the revised Q@T evaluation framework, conducted on the HybridNets model for autonomous driving tasks, running on an A100 GPU. The curve shows the square root of Jensen-Shannon divergence (rJSD) over successive sampling rounds, which measures the stability of Q@T estimation as sampling progresses. The first occurrences of rJSD falling below 0.05 and 0.03 are annotated, indicating fast convergence, with rJSD dropping below the two thresholds at around the 20th and 40th rounds, respectively. This confirms the statistical robustness of the evaluation under the revised formulation.

recommendation results—the higher the value, the better the performance. These results demonstrate the applicability of our framework in time-sensitive recommendation scenarios.

- **Autonomous Driving (HybridNets [8]):** As shown in Fig. 3(b), the model performs lane line detection, traffic object detection, and drivable area segmentation. For clarity, we focus on the mAP (mean Average Precision), a widely used metric for object detection that reflects precision across categories—higher values indicate better detection accuracy. Q@T reveals how performance degrades as inference time constraints become stricter, highlighting potential quality loss in safety-critical scenarios.

- **Emotion Recognition (EmotionFlow [33]):** As shown in Fig. 3(c), Q@T captures changes in weighted-F1 score, a harmonic mean of precision and recall, under different latency thresholds—higher values represent better balance between accuracy and completeness. The experiment shows that emotion detection models can suffer from quality drops in low-latency settings.

These cross-domain experiments demonstrate the generalizability and practical utility of the revised Q@T framework under the guidance of Evaluatology, particularly in capturing time-sensitive performance variations across different AI applications.

In addition, we further verified the statistical stability of the proposed evaluation method using HybridNets in the autonomous driving domain. As shown in Fig. 4, the square root of Jensen-Shannon divergence (rJSD) gradually decreases and stabilizes as the number of sampling rounds increases. The convergence threshold of 0.05 was reached in fewer than 20 rounds, indicating that reliable statistical evaluation can still be efficiently achieved under the revised definition.

Finally, we conducted a focused analysis of EmotionFlow to explore Q@T's ability to capture tail quality. As shown in Fig. 5, we set the inference time threshold to 90% tail latency (118.75 ms) and observed the corresponding quality fluctuation. The results show that Q@T effectively reflects system performance at critical latency thresholds, offering a more comprehensive perspective for optimizing the quality-latency trade-off in AI system deployment.

Furthermore, we observed that all three models exhibited significant quality fluctuations under strict inference time constraints. This indicates that current AI software and hardware systems may still lack sufficient stability in critical scenarios, emphasizing the need for further optimization to ensure robust inference performance in real-world applications.

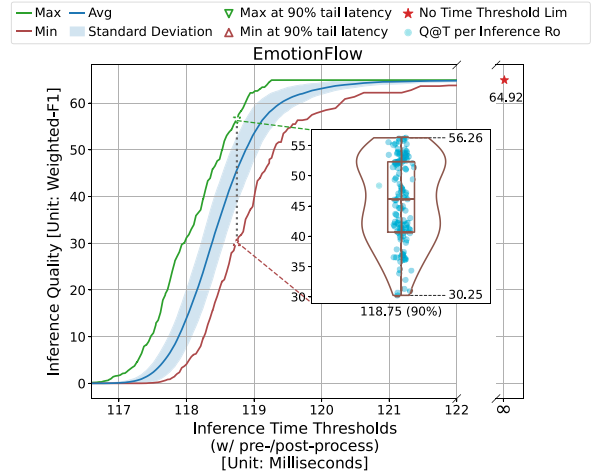


Fig. 5. Q@T evaluation results across multiple inference time thresholds using the revised Q@T framework, with the subject composed of the A100 GPU and the EmotionFlow model for emotion recognition. The violin plot represents the distribution of Q@T (measured by weighted-F1) at a specific time threshold. The box highlights the distribution when the time threshold is set to 90% tail latency (118.75 ms). Key statistics — including maximum, minimum, average, and standard deviation — are annotated, with the baseline “No Time Threshold Limit” included for reference.

4.2. The importance of tail quality

The mean value alone is insufficient to accurately capture the overall performance of AI inference systems, particularly in scenarios where extreme behavior is more indicative of the system's true performance. In real-world applications, the performance of AI systems often deviates significantly from the average in critical applications. This is the tail quality phenomenon, which Q@T specifically aims to address.

For example, in the MLPerf inference benchmark [15], the evaluation primarily focused on average- or best-quality metrics. However, this approach overlooks the performance in extreme cases where the system might fail or perform poorly, often the most critical aspect for real-world applications. Incorporating statistical methods, such as the Monte Carlo Simulation, into evaluating Q@T ensures that these extreme cases are not overlooked, providing a more comprehensive understanding of the system's true capability.

Through our experiments, we further emphasize the importance of Tail Quality in AI inference system evaluation. As shown in Fig. 5, the refined Q@T effectively captures the full spectrum of evaluation quality fluctuations, including extreme performance variations at the tail end of the distribution. For instance, when the inference time threshold is set to 118.75 ms, the lowest observed inference quality is 30.25 weighted-F1, significantly lower than the value of 64.92 obtained using traditional isolated quality metrics. The difference between these two values is approximately 2.15 times, highlighting how Q@T can reveal extreme quality fluctuations that traditional evaluation methods cannot capture. This capability ensures that the evaluation reflects not only the system's overall quality but also how it behaves under the most demanding conditions, which is critical for ensuring reliability in high-risk or real-time applications.

4.3. The landscape of stakeholder consideration

In the original Q@T evaluation framework, the focus has largely been on strict time thresholds, where inference time is a critical factor in determining system performance. However, real-world applications often involve varying levels of tolerance for inference delay, depending on the specific needs of different stakeholders. These stakeholder needs can significantly influence how the system is evaluated and what performance metrics are prioritized.

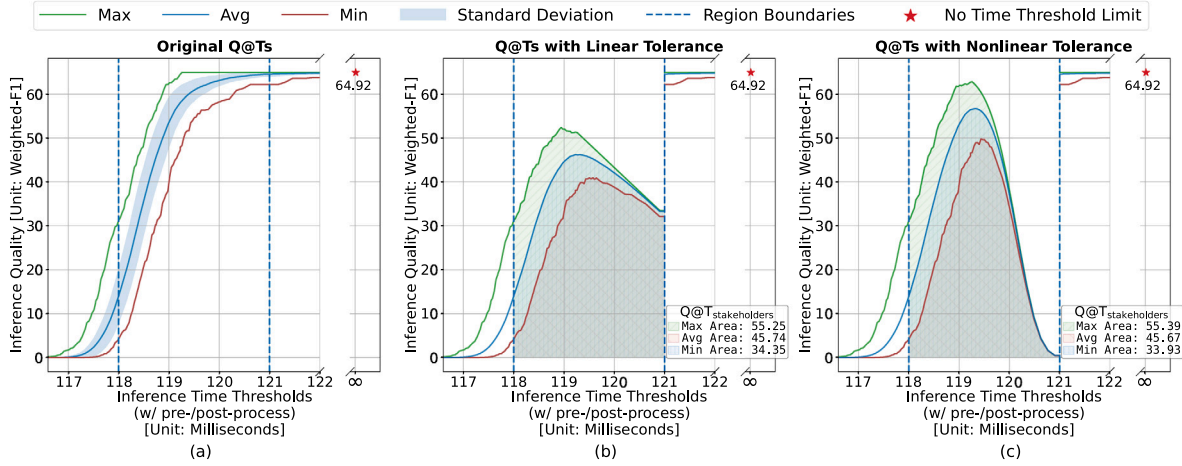


Fig. 6. Stakeholder-aware Q@T evaluation on the EmotionFlow model (weighted-F1). (a) shows the original Q@T scores under different inference time thresholds; as the threshold increases, the model is allowed more time, leading to higher Q@T values. (b) illustrates Q@T variation after applying a linear tolerance function over the tolerance interval from 118 ms to 121 ms. (c) shows the corresponding result with a nonlinear tolerance function, where high-latency points are penalized more severely.

To better address these diverse needs, we propose expanding the Q@T framework by incorporating a tolerance function that adjusts the weight of the Q@T metric based on the acceptable time thresholds for different stakeholders. This allows the evaluation framework to reflect time tolerance based on the level of urgency or flexibility that stakeholders require. The new approach ensures that the evaluation process is more adaptable to real-world constraints, where not every scenario demands the same level of urgency.

For instance, a real-time autonomous driving application may impose strict time constraints on inference, whereas in e-commerce, a slight delay beyond the threshold might be tolerable, though it could lead to user churn. However, user churn is not necessarily directly correlated with a strict time threshold, as exceeding the threshold does not automatically result in user loss. Different users have varying levels of tolerance for delay, and the relationship between user patience and inference time is not as binary or stepwise as the rigid time thresholds might suggest. In this context, the tolerance function for time delay in e-commerce applications must reflect the more gradual and non-linear impact of delay, where mild delays may still be acceptable but could lead to different levels of consequences depending on the user's tolerance.

Building upon the original formulation of Q@T in Eq. (4), we propose a generalized version that incorporates stakeholder-specific tolerance to inference delay, referred to as Stakeholder-aware Q@T:

$$Q@T_{\text{stakeholders}} = \int_{\hat{T} \in R} Q@T \times \text{Tolerance}(\hat{T}) d\hat{T}, \quad (5)$$

where $Q@T$ denotes the quality metric evaluated at time threshold \hat{T} , $R = [R_{\min}, R_{\max}]$ is the stakeholder-defined tolerance interval for inference time, and $\text{Tolerance}(\hat{T})$ is a user-defined weighting function that expresses the degree to which inference delays at \hat{T} are acceptable. The integration domain R is determined by the specific requirements of stakeholders, reflecting the time thresholds they consider relevant or acceptable for their application scenarios. For strict time constraints, the tolerance value will be low, reducing the weight of $Q@T$ at those time intervals. On the other hand, for more flexible time requirements, the tolerance value will be higher, increasing the weight of $Q@T$ at those intervals.

Note that Eq. (5) integrates the weighted Q@T across a stakeholder-defined tolerance interval R . It does not represent a normalized average, but rather an aggregate evaluation score that emphasizes performance in regions preferred by the stakeholder. A normalized version can be obtained by dividing by $\int_{\hat{T} \in R} \text{Tolerance}(\hat{T}) d\hat{T}$, if desired. We intentionally leave the expression unnormalized to maintain flexibility,

allowing users to interpret the result either as a total weighted score or to apply normalization as needed for their specific use cases.

This adaptation of Q@T allows us to tailor the evaluation process to the specific time tolerance of the task at hand. In practice, this enables the Q@T evaluation to be much more flexible, reflecting the landscape of stakeholder needs. For example, in autonomous driving, any delay could be catastrophic. Here, Q@T would prioritize faster inference times and impose stricter thresholds. In contrast, in healthcare applications, where accurate diagnosis is critical but minor delays may be acceptable, Q@T would adjust the tolerance to allow for longer inference times while still ensuring high-quality outputs.

By incorporating stakeholder-driven tolerance for inference time into the evaluation process, the Q@T framework can provide a more realistic, adaptable, and comprehensive evaluation metric that reflects the diversity of real-world applications. This adjustment enhances the flexibility of the evaluation, making it more suitable for a variety of use cases where performance criteria differ significantly based on the context and needs of the stakeholders.

Experiments To empirically validate the stakeholder-aware extension of Q@T, we designed two experiments using synthetic tolerance functions. Due to the lack of large-scale data on stakeholder demands, we manually constructed two representative forms of the $\text{Tolerance}(\hat{T})$ function to simulate varying tolerance for inference time. These functions define how much weight each evaluation result $Q@T$ receives at a given inference time threshold \hat{T} , thereby reflecting hypothetical stakeholder preferences across different application contexts.

The first form is a linear decay function, controlled by a slope parameter $\alpha \in (0, 1]$:

$$\text{Tolerance}_l(\hat{T}) = 1 - \alpha \cdot \frac{\hat{T} - R_{\min}}{R_{\max} - R_{\min}}. \quad (6)$$

The second form is a nonlinear exponential decay function, controlled by shape parameters $\lambda > 0$ and $p > 1$:

$$\text{Tolerance}_n(\hat{T}) = \exp\left(-\lambda \cdot \left(\frac{\hat{T} - R_{\min}}{R_{\max} - R_{\min}}\right)^p\right). \quad (7)$$

In both cases, R_{\min} and R_{\max} define the range of inference time thresholds under consideration. This range represents the domain in which stakeholders are assumed to express meaningful tolerance variations. For instance, in safety-critical scenarios, R may be narrow and focused on low-latency thresholds; in contrast, in offline reasoning tasks, a broader range may apply.

We applied both tolerance functions to two tasks: (1) Emotion recognition using EmotionFlow, evaluated with weighted-F1; and (2)

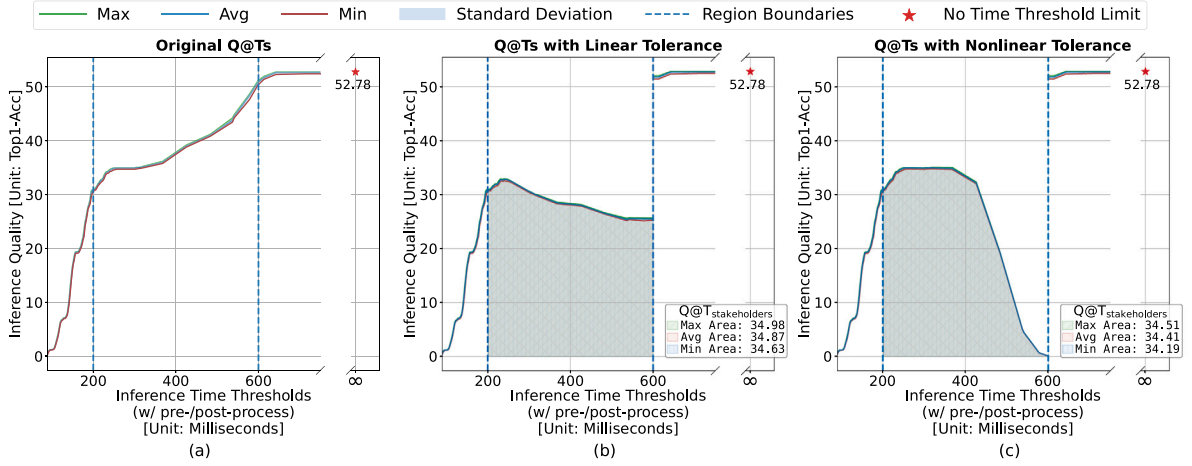


Fig. 7. Stakeholder-aware Q@T evaluation on the Vicuna model (top-1 accuracy). (a) presents the original Q@T curve under varying inference time constraints. (b) depicts the evaluation using a linear tolerance function across the defined interval from 200 ms to 600 ms. (c) applies a nonlinear tolerance function, emphasizing low-latency performance and reducing the impact of slower responses.

Question answering using Vicuna, evaluated with top-1 accuracy. For each case, we compared the original Q@T formulation with the stakeholder-aware variant $Q@T_{\text{stakeholders}}$, using both linear and nonlinear tolerance.

Figs. 6 and 7 demonstrate two key implications of incorporating stakeholder-aware tolerance functions. First, rather than focusing on Q@T under a single time threshold, this framework encourages evaluators to consider a broader range of thresholds, each weighted according to stakeholder preferences. This allows Q@T scores at higher-latency points — often overestimated in unconstrained evaluations — to be reasonably discounted, reflecting more realistic expectations.

Second, the stakeholder-aware score $Q@T_{\text{stakeholders}}$ can be interpreted as a weighted average of Q@T over a tolerance-defined interval R , capturing the system's overall quality across multiple thresholds. This integrates not only peak performance but also performance stability over time, which is crucial for reliable deployment.

Comparing Figs. 6 and 7, we observe that although the nonlinear tolerance function penalizes high-latency Q@T more severely, EmotionFlow exhibits a smoother degradation curve than Vicuna. As a result, its $Q@T_{\text{stakeholders}}$ shows a relative gain under nonlinear weighting. Additionally, Vicuna demonstrates more consistent behavior across the tolerance range, with minimal difference between the maximum area (i.e., best-case Q@T) and the minimum area (i.e., worst-case Q@T) regions—an important indicator of stability in deployment. In contrast, EmotionFlow shows significant performance variance, with a 21.46-point gap in average weighted-F1 between the best and worst Q@T segments.

These results illustrate that the stakeholder-aware Q@T can adapt evaluation outcomes based on context-specific tolerance levels, offering a more flexible and realistic assessment approach. Even with synthetic tolerance profiles, the influence on evaluation is evident, providing preliminary support for the practical applicability of this framework in critical tasks.

5. Conclusion

This paper reanalyzes and extends the Quality@Time-Threshold (Q@T) framework from the perspective of Evaluatology, offering a robust theoretical foundation for evaluating AI inference systems. By applying Evaluatology's five core axioms and its universal evaluation methodology, we redefine the components of the evaluation system, ensuring precise and consistent assessments of AI systems across various tasks and environments. Additionally, we enhance Q@T by incorporating a stakeholder-driven tolerance function, making the framework more adaptable to diverse real-world requirements. This work

also emphasizes the importance of tail quality, demonstrating how Q@T captures extreme performance variations overlooked by traditional metrics. Overall, we bridge the gap between theoretical evaluation frameworks and practical AI evaluation, providing a comprehensive, adaptable, and scientifically grounded approach for assessing AI systems in complex applications.

CRedit authorship contribution statement

Zhengxin Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, 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

Supported by the Innovation Funding of ICT, CAS under Grant No. E461070.

References

- [1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E.P. Xing, H. Zhang, J.E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023, <http://dx.doi.org/10.48550/arXiv.2306.05685>, arXiv:2306.05685.
- [2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H.W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S.P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S.S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L. Kaiser, A. Kamali, I. Kanitscheider, N.S. Keskar, T. Khan, L. Kilpatrick, J.W. Kim, C. Kim, Y. Kim, H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, L.u. Kondradiuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C.M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R.

- Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S.M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F.d.B. Peres, M. Petrov, H.P.d. Pinto, Michael, Pokorny, M. Pokrass, V. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F.P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J.F.C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J.J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C.J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report, 2023, <http://dx.doi.org/10.48550/arXiv.2303.08774>, URL <http://arxiv.org/abs/2303.08774>.
- [3] C. Chen, A. Seff, A. Kornhauser, J. Xiao, DeepDriving: Learning affordance for direct perception in autonomous driving, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2722–2730, <http://dx.doi.org/10.1109/ICCV.2015.312>.
- [4] H. Xu, Y. Gao, F. Yu, T. Darrell, End-to-End learning of driving models from large-scale video datasets, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3530–3538, <http://dx.doi.org/10.1109/CVPR.2017.376>.
- [5] G. Li, Y. Yang, X. Qu, Deep learning approaches on pedestrian detection in hazy weather, IEEE Trans. Ind. Electron. 67 (10) (2020) 8889–8899, <http://dx.doi.org/10.1109/TIE.2019.2945295>.
- [6] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F.J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C.J. Kelly, D. King, J.R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J.J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K.C. Young, J. De Fauw, S. Shetty, Addendum: International evaluation of an AI system for breast cancer screening, Nature 586 (7829) (2020) <https://www.nature.com/articles/s41586-020-2679-9>, E19–E19. URL <https://www.nature.com/articles/s41586-020-2679-9>.
- [7] P. Rajpurkar, J. Irvin, R.L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C.P. Langlotz, B.N. Patel, K.W. Yeom, K. Shpanskaya, F.G. Blankenberg, J. Seekins, T.J. Amrhein, D.A. Mong, S.S. Halabi, E.J. Zucker, A.Y. Ng, M.P. Lungren, Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLOS Med. 15 (11) (2018) e1002686, <http://dx.doi.org/10.1371/journal.pmed.1002686>, URL <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002686>.
- [8] M.F. Ijaz, G. Alfian, M. Syafrudin, J. Rhee, Hybrid prediction model for Type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest, Appl. Sci. 8 (8) (2018) 1325, <http://dx.doi.org/10.3390/app8081325>, URL <https://www.mdpi.com/2076-3417/8/8/1325>.
- [9] Y. Zhang, A. Carballo, H. Yang, K. Takeda, Autonomous driving in adverse weather conditions: A survey, 2021, CoRR, [arXiv:2112.08936](https://arxiv.org/abs/2112.08936).
- [10] D. Wang, W. Fu, Q. Song, J. Zhou, Potential risk assessment for safe driving of autonomous vehicles under occluded vision, Sci. Rep. 12 (2022) <http://dx.doi.org/10.1038/s41598-022-08810-z>, URL <https://api.semanticscholar.org/CorpusID:247628492>.
- [11] Z. Yang, W. Gao, C. Luo, L. Wang, F. Tang, X. Wen, J. Zhan, Quality at the tail of machine learning inference, 2024, [arXiv:2212.13925](https://arxiv.org/abs/2212.13925).
- [12] W. Gao, L. Wang, M. Chen, J. Xiong, C. Luo, W. Zhang, Y. Huang, W. Li, G. Kang, C. Zheng, B. Xie, S. Dai, Q. He, H. Ye, Y. Bao, J. Zhan, High fusion computers: The IoTs, edges, data centers, and humans-in-the-loop as a computer, BenchCouncil Trans. Benchmarks, Stand. Eval. 2 (3) (2022) 100075, <http://dx.doi.org/10.1016/j.tbench.2022.100075>, URL <https://www.sciencedirect.com/science/article/pii/S277248592200062X>.
- [13] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Trans. Radiat. Plasma Med. Sci. 5 (6) (2021) 741–760, <http://dx.doi.org/10.1109/TRPMS.2021.3066428>.
- [14] Y. Zhang, P. Tiño, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Trans. Emerg. Top. Comput. Intell. 5 (5) (2021) 726–742, <http://dx.doi.org/10.1109/TETCI.2021.3100641>.
- [15] V.J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Diamos, J. Duke, D. Fick, J.S. Gardner, I. Hubara, S. Idrunji, T.B. Jablin, J. Jiao, T.S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A.T.R. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, Y. Zhou, MLPerf inference benchmark, in: 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), 2020, pp. 446–459, <http://dx.doi.org/10.1109/ISCA45697.2020.00045>.
- [16] W. Gao, F. Tang, L. Wang, J. Zhan, C. Lan, C. Luo, Y. Huang, C. Zheng, J. Dai, Z. Cao, D. Zheng, H. Tang, K. Zhan, B. Wang, D. Kong, T. Wu, M. Yu, C. Tan, H. Li, X. Tian, Y. Li, J. Shao, Z. Wang, X. Wang, H. Ye, AIBench: An Industry Standard Internet Service AI Benchmark Suite, 2019, <http://dx.doi.org/10.48550/arXiv.1908.08998>, [arXiv:1908.08998](https://arxiv.org/abs/1908.08998).
- [17] W. Gao, C. Luo, L. Wang, X. Xiong, J. Chen, T. Hao, Z. Jiang, F. Fan, M. Du, Y. Huang, F. Zhang, X. Wen, C. Zheng, X. He, J. Dai, H. Ye, Z. Cao, Z. Jia, K. Zhan, H. Tang, D. Zheng, B. Xie, W. Li, X. Wang, J. Zhan, AIBench: Towards scalable and comprehensive datacenter AI benchmarking, in: C. Zheng, J. Zhan (Eds.), Benchmarking, Measuring, and Optimizing, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 3–9, http://dx.doi.org/10.1007/978-3-030-32813-9_1.
- [18] C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Ré, M. Zaharia, DAWNbench: An end-to-end deep learning benchmark and competition, in: Workshop on ML Systems At Advances in Neural Information Processing Systems, 2017.
- [19] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatolgy: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162, <http://dx.doi.org/10.1016/j.tbench.2024.100162>.
- [20] J. Zhan, Five axioms of things, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (3) (2024) 100184, <http://dx.doi.org/10.1016/j.tbench.2024.100184>.
- [21] J. Zhan, A short summary of evaluatolgy: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (2) (2024) 100175, <http://dx.doi.org/10.1016/j.tbench.2024.100175>.
- [22] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 639–648, <http://dx.doi.org/10.1145/3397271.3401063>.
- [23] D. Ayata, Y. Yaslan, M.E. Kamasak, Emotion recognition from multimodal physiological signals for emotion aware healthcare systems, J. Med. Biological Eng. 40 (2) (2020-04-01) 149–157, <http://dx.doi.org/10.1007/s40846-019-00505-7>.
- [24] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R.W. Picard, Driver emotion recognition for intelligent vehicles: A survey, ACM Comput. Surv. 53 (3) (2020-07-04) 64:1–64:30, <http://dx.doi.org/10.1145/3388790>, URL <https://dl.acm.org/doi/10.1145/3388790>.
- [25] M. Dhuheir, A. Albaser, E. Baccour, A. Erbad, M. Abdallah, M. Hamdi, Emotion recognition for healthcare surveillance systems using neural networks: A survey, in: 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021-06, pp. 681–687, <http://dx.doi.org/10.1109/IWCMC51323.2021.9498861>.
- [26] T. Turay, T. Vladimirova, Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey, IEEE Access 10 (2022) 14076–14119, <http://dx.doi.org/10.1109/ACCESS.2022.3147495>.
- [27] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE Access 8 (2020) 58443–58469, <http://dx.doi.org/10.1109/ACCESS.2020.2983149>.
- [28] Q. Guo, S. Chen, X. Xie, L. Ma, Q. Hu, H. Liu, Y. Liu, J. Zhao, X. Li, An empirical study towards characterizing deep learning development and deployment across different frameworks and platforms, in: 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2019, pp. 810–822, <http://dx.doi.org/10.1109/ASE.2019.00080>.
- [29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, URL <https://www.tensorflow.org/> Software available from tensorflow.org.
- [30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, 2019, <http://dx.doi.org/10.48550/arXiv.1912.01703>, [arXiv:1912.01703](https://arxiv.org/abs/1912.01703).
- [31] Y. Wu, L. Liu, C. Pu, W. Cao, S. Sahin, W. Wei, Q. Zhang, A comparative measurement study of deep learning as a service framework, IEEE Trans. Serv. Comput. 15 (1) (2022-01) 551–566, <http://dx.doi.org/10.1109/TSC.2019.2928551>.
- [32] J. Zhan, A BenchCouncil view on benchmarking emerging and future computing, BenchCouncil Trans. Benchmarks, Stand. Eval. 2 (2) (2022) 100064, <http://dx.doi.org/10.1016/j.tbench.2022.100064>, URL <https://www.sciencedirect.com/science/article/pii/S2772485922000515>.
- [33] X. Song, L. Zang, R. Zhang, S. Hu, L. Huang, Emotionflow: Capture the dialogue level emotion transitions, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 8542–8546, <http://dx.doi.org/10.1109/ICASSP43922.2022.9746464>.



Full Length Article

Tensor databases empower AI for science: A case study on retrosynthetic analysis

Xueya Zhang^a, Guoxin Kang^b, Boyang Xiao^c, Jianfeng Zhan^b^a University of Chinese Academy of Sciences, Beijing, China^b Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China^c The University of Manchester, Manchester, United Kingdom

ARTICLE INFO

Keywords:

Tensor database
Approximate retrieval
Retrosynthetic

ABSTRACT

Retrosynthetic analysis is highly significant in chemistry, biology, and materials science, providing essential support for the rational design, synthesis, and optimization of compounds across diverse Artificial Intelligence for Science (AI4S) applications. Retrosynthetic analysis focuses on exploring pathways from products to reactants, and this is typically conducted using deep learning-based generative models. However, existing retrosynthetic analysis often overlooks how reaction conditions significantly impact chemical reactions. This causes existing work to lack unified models that can provide full-cycle services for retrosynthetic analysis, and also greatly limits the overall prediction accuracy of retrosynthetic analysis. These two issues cause users to depend on various independent models and tools, leading to high labor time and cost overhead.

To solve these issues, we define the boundary conditions of chemical reactions based on the Evaluatology theory and propose BigTensorDB, the first tensor database which integrates storage, prediction generation, search, and analysis functions. BigTensorDB designs the tensor schema for efficiently storing all the key information related to chemical reactions, including reaction conditions. BigTensorDB supports a full-cycle retrosynthetic analysis pipeline. It begins with predicting generation reaction paths, searching for approximate real reactions based on the tensor schema, and concludes with feasibility analysis, which enhances the interpretability of prediction results. BigTensorDB can effectively reduce usage costs and improve efficiency for users during the full-cycle retrosynthetic analysis process. Meanwhile, it provides a potential solution to the low accuracy issue, encouraging researchers to focus on improving full-cycle accuracy.

1. Introduction

Retrosynthetic analysis is an important method for exploring efficient synthetic pathways for target molecules. It holds significant importance in fields such as chemistry, materials science, and pharmaceuticals [1]. The main goal of retrosynthetic analysis is to identify the appropriate reactants and reaction conditions for the efficient synthesis of the target molecule. For example, when a target product is inputted, the work aims to obtain the correct reactants, reaction conditions, and multiple pathways for its synthesis.

People are always committed to developing efficient and user-friendly tools to help scientists conduct retrosynthetic analysis more quickly and conveniently. Since the 1960s, computer-aided synthesis planning (CASP) has been a key tool in this area [2]. Particularly, in today's era of rapid development of artificial intelligence (AI), AI for Science (AI4S) brings about significant changes in many fields. It also injects new vitality into retrosynthetic analysis technology.

More and more machine learning-based methods, especially deep learning models, are being used in the field of retrosynthetic analysis. These AI technologies significantly improve the efficiency and accuracy of it [3–10]. In 2022, Liu et al. [11] first highlighted three major contradictions facing machine learning in materials science: data characteristics, model interpretability, and result authenticity [12–15]. These also apply to retrosynthetic analysis. On further analysis, existing prediction models are found to treat reactants and reaction conditions as separate factors, which gives rise to two key issues as follows:

- a. **Lacking a unified model that can provide full-cycle service for retrosynthetic analysis.** The full-cycle service for retrosynthetic analysis involves a step-by-step prediction of reactants and reaction conditions, ultimately yielding complete candidate chemical reaction equations that are ready for direct experimental validation. Current research on AI-based retrosynthetic

* Corresponding author.

E-mail addresses: zhangxueya21@mails.ucas.ac.cn (X. Zhang), kangguoxin@ict.ac.cn (G. Kang), boyang.xiao@postgrad.manchester.ac.uk (B. Xiao), zhanjianfeng@ict.ac.cn (J. Zhan).<https://doi.org/10.1016/j.tbench.2025.100216>

Received 28 January 2025; Received in revised form 20 April 2025; Accepted 28 May 2025

Available online 16 June 2025

2772-4859/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

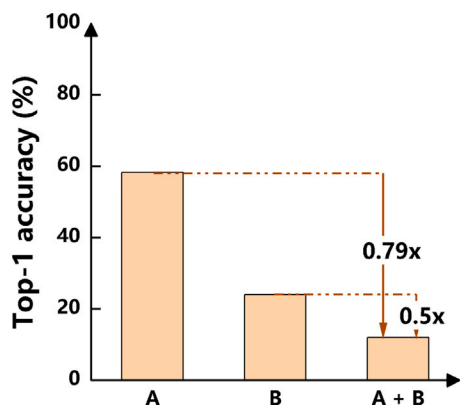


Fig. 1. The entire process of retrosynthetic analysis involves a step-by-step prediction of reactants and reaction conditions. The prediction of reactants requires a target molecule as input, while the prediction of reaction conditions requires a reaction equations without reaction conditions. We choose model RetroTRAE [16] as model A for reactants prediction and model Parrot [17] as model B for reaction conditions prediction. For detailed experimental descriptions and workload information, please refer to Sections 4.3.2 and 3.2. The figure indicates that the individual accuracies of Model A and Model B are quite low, at 58.3% and 24%, respectively. When Model A's output is used as input for Model B, the combined accuracy of the prediction results (Model A + B) drops even further to a mere 12%. This represents a significant decrease of 79% for Model A and 50% for Model B compared to their original accuracies. Given these substantial declines in performance, there is an urgent need for a unified model to improve the overall accuracy.

analysis models mainly focuses on one direction of single-step or multi-step prediction of synthetic routes. These directions can be categorized into two types based on their prediction targets, which usually include reactants and reaction conditions. However, when scientists use retrosynthetic analysis tools, they hope to directly obtain a set of complete reaction equation candidates. Therefore, it is necessary to design a full-cycle service that can provide a set of complete reaction equation candidates, including both reactants and reaction conditions.

- b. Significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.** We have noticed that the accuracy of individual prediction models is often low, and when used together, their combined accuracy tends to decrease even more. For each prediction models, the prediction accuracy is defined as the proportion of correct results among the Top-k generated predictions. However, as shown in Fig. 1, when the Top-1 accuracy of the reactant prediction model RetroTRAE [16] (A) is 58.3% and that of the reaction condition prediction model Parrot [17] (B) is 24%, the overall Top-1 accuracy of the final predicted result (A+B) is only 12%. Therefore, optimizing the accuracy of a specific type of model is insufficient for full-cycle retrosynthetic analysis. And there is an urgent need for new work focusing on overall prediction accuracy.

To address above issues, this paper proposes BigTensorDB. BigTensorDB is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. It integrates storage, prediction generation, search and analysis functions. Its main contributions are summarized as follows:

- (1) We design a tensor format for storing chemical reactions. This format efficiently stores all key information related to chemical reactions, including reactants, products, and reaction conditions (such as solvents and reagents). We provide users with services for storing and retrieving chemical reactions.
- (2) We integrate multiple retrosynthetic analysis prediction models including reactants prediction and reaction conditions prediction. We

also integrate SMILES embedding models. These integrations offer full-cycle retrosynthetic analysis services to users. Our work reduces usage costs and improves the pipeline's efficiency.

(3) We provide search and analysis services. Through these services, we re-rank and analyze the final prediction results. At the same time, we provide real reaction equations similar to the predicted results for user reference. These works can enhance the accuracy and interpretability of the final outcomes.

2. Background and related work

2.1. Background

Retrosynthetic analysis was established by E.J. Corey [2]. It involves identifying target molecules, deconstructing them in a reverse manner, and devising synthetic routes to achieve the desired synthesis. The purpose of retrosynthetic analysis is to assist scientists in finding more efficient synthetic routes to synthesis more useful molecules. The target molecules often originate from diverse application scenarios. They typically cannot be synthesized via known reactions or have very inefficient existing synthetic pathways. Thus, there is a need to explore new and more efficient synthetic routes by retrosynthetic analysis.

To solve this problem, scientists usually need to search through a vast space of possible transformations of the target molecules. This can be done either by hypothetically disconnecting bonds or by converting one functional group into another which goal is to match existing reaction templates. However, this process demands that scientists have a rich knowledge base and extensive synthetic experience. Additionally, it requires significant time and material costs for experiments to verify the correctness of hypotheses.

The earliest computer-aided tools work by first enumerating possible reaction types for the target molecule. Then, they use search algorithms to recursively enumerate and search for potential reaction pathways. This process continues until viable starting materials are identified. However, these methods essentially do not create new reactions but rather rearranged existing knowledge. Today, with the rapid development of AI, an increasing number of machine learning-based models are being applied to retrosynthetic analysis. This helps scientists become more creative in their retrosynthetic analysis works.

These machine learning models, particularly deep learning models, mainly fall into two categories: prediction reactants and prediction reaction conditions. In real-world scenarios, scientists must select from a wide range of models first. Then they use the chosen reactant prediction model to generate a set of reaction equations without reaction conditions. It is essential to emphasize that the reaction conditions play an important role in chemical reactions. The same reactants can undergo different reactions under different conditions, leading to different products. Therefore, scientists must also select a reaction condition prediction model for another round of predictions. Afterward, they need to conduct theoretical analysis and experimental verification manually. Each experimental verification of a reaction requires substantial time and material costs. Thus, the accuracy of prediction models is vital for real experiments. It also determines the efficiency of retrosynthetic analysis.

2.2. CASP's related work

2.2.1. Prediction models

Researchers have developed the Simplified Molecular Input Line Entry System [18] (SMILES) notation, a text-based method that encodes molecular graphs into simple, human-readable character sequences. Some prediction models extract functional groups from SMILES expressions to analyze a target molecule's reaction-related structural, spatial, and functional group features, achieving prediction. Others ignore reaction structures, spatial configurations, and functional groups, instead directly using SMILES for sequence-to-sequence [19–25] (Seq2Seq)

prediction. Generative AI, including the Seq2Seq method, is crucial for discovering new substances or materials. It enables predictive models to transcend existing knowledge and create new knowledge. However, these models show lower prediction accuracy on United States Patent and Trademark Office (USPTO) datasets. Notably, some models achieve high accuracy and better user-friendliness by calling large language model Application Programming Interfaces (APIs) or fine-tuning these models [26] to generate products and recommend reaction conditions. As discussed earlier, deep learning-based machine learning models in this field can be categorized into reactants prediction models and reaction condition prediction models. We will detail these models based on this classification.

From the perspective of the methods used, reactant prediction models can be classified as template-based, template-free, and semi-template-based models [27,28].

Template-based methods play an important role in retrosynthesis prediction. These methods employ reaction templates extracted from chemical databases to guide the retrosynthesis process through template-target molecule matching. The templates, which can be manually curated or automatically generated, enable models to identify optimal chemical transformations [29]. Multiple approaches [30–33] have been developed for template prioritization [29]: RetroSim [30] ranks candidate templates using molecular fingerprint comparisons, Neural-Sym [31] employs a deep neural network classifier, and GLN [32] evaluates template-reactant compatibility with a conditional graph logic network. While template-based models provide interpretability and ensure molecule validity, their practical applications [34] are constrained by limited generalization capability and scalability [29].

Template-free methods aim to eliminate dependency on predefined templates. It achieves retrosynthesis prediction through data-driven or innovative architectural design, opening up a new direction of exploration in this field. Most existing methods turn the task into a Seq2Seq problem [19–25], using the SMILES [18] format to represent molecules. This is first to use by Liu et al. [19] who proposed a long short-term memory [35] (LSTM)-based Seq2Seq model to change the SMILES of a product into the SMILES of reactants. Meanwhile, there are some studies treat this task as a graph-to-sequence problem, using molecular graphs as input [36]. For example, Graph2SMILES [36] combined a graph encoder with a Transformer decoder to keep SMILES order the same. However, in recent studies, such as MEGAN [37], MARS [38], and Graph2Edits [39], end-to-end molecular graph editing model is widely used. These models represent chemical reactions as a series of changes to molecular graphs. Fang et al. [40] created a way to decode at the substructure level by finding parts of product molecules that stay the same. Although template-free methods are entirely data-driven, they face challenges related to the interpretability, chemical validity, and diversity of the molecules they generate [29,34].

Semi-template-based methods involve a two-stage strategy. The first stage decomposes target molecules into synthons via reactive site identification, while the second stage converts synthons to reactants through techniques like leaving group selection [27], graph generation [41], or SMILES generation [1,42]. RetroXpert [42] first identified the reaction center of the target molecule using an edge-enhanced graph attention network to obtain synthons. Then it generated the corresponding reactants based on these synthons. RetroPrime [1] incorporated the chemist’s retrosynthesis strategy, which finely split the retrosynthesis process into decomposing the synthetic moiety and adding an appropriate leaving group to finally generate the reactant. These methods better match the intuitive problem-solving approach of scientists. However, the two stages in the framework are independent. This increases computational complexity. Moreover, it is challenging to transfer the knowledge and insights gained from predicting reactive sites to the completion of reactants [29].

Reaction condition prediction typically involves inputting a complete SMILES-formatted equation and outputting suitable reaction conditions. However, recommending conditions from scratch is a challenging and under-explored problem that heavily depends on the knowledge and experience of chemists. Neural network models can predict

the chemical environment, including catalysts, solvents, and reagents, as well as the most suitable temperature for any given organic reaction [17,43]. There are also some works [26] based on large language models that can also achieve reaction condition recommendation and prediction. MM-RCR [26] model is a text-enhanced multimodal large language model that learns a unified reaction representation by multi-source information from SMILES, reaction graphs, and text corpora. It also demonstrated strong generalization capabilities on out-of-domain and high-throughput experimental datasets, providing new momentum for high-throughput reaction condition screening.

2.2.2. Automated feature engineering

Data quality critically impacts machine learning model performance. High-quality data can greatly boost a model’s predictive accuracy and reliability. In contrast, low-quality data may degrade performance and lead to results conflicting with domain-expert understanding [44–49]. Thus, ensuring the effectiveness of data augmentation and feature extraction for retrosynthetic analysis tools is extremely important [50–52]. So, here we introduce the Automated feature engineering models related to our works.

SMILES [18] (simplified molecular input line entry system) are text-based representations that encode a molecular graph in a simple, human-readable sequence of characters [53,54]. Transformer models are effective in cheminformatics for processing SMILES strings and extracting molecular representations, benefiting from bidirectional context for enhanced understanding of chemical environments, transfer learning, and fine-tuning.

BERT [55] is a pioneering model that captures bidirectional context from SMILES strings, making it suitable for tasks like property prediction and drug discovery, though it has high computational and memory demands. MOLBERT [56] is a chemistry-specific adaptation of BERT that excels in predicting physicochemical properties and molecular interactions, but it requires large labeled datasets for optimal performance. SMILES-BERT [57] is designed to learn molecular representations directly from SMILES strings without extensive feature engineering, making it effective for predicting molecular properties, though it also demands significant computational resources. ChemBERTa [58] and ChemBERTa-2 [59] enhance BERT [60] with domain-specific training for a variety of property predictions, improving accuracy while maintaining high complexity and resource demands. The RoBERTa-based Model [61] refines BERT by using more data and longer sequences for better property prediction and molecular classification, although it increases computational requirements for training and inference. Mol-BERT [62] and MolRoPE-BERT [63] are BERT-based models for predicting molecular properties from SMILES, differing mainly in their position embedding approaches, with MolRoPE-BERT using rotary PE to address limitations of absolute PE in Mol-BERT.

2.3. Tensor retrieval related work

In the fields of data science and artificial intelligence, the demand for managing high-dimensional vector data is growing rapidly, driven mainly by the rapid development of unstructured data and machine learning technologies [64].

For vector similarity search, some systems have been put into application, such as Alibaba’s AnalyticDB-V [65] and PASE (PostgreSQL) [66]. However, they do not support multi-vector queries. Vearch [67,68] is another system designed specifically for vector search, but it is not efficient in handling large-scale data and also does not support multi-vector queries. However, as a data management system specifically built for the needs for more efficient and flexible vector data management, Milvus focuses on the storage and search of large-scale vector data, supporting various query types, including vector similarity search and multi-vector query processing. In particular, the Milvus 2.5 version introduced the Sparse-BM25 algorithm, achieving hybrid search of sparse and dense vectors, further improving search efficiency.

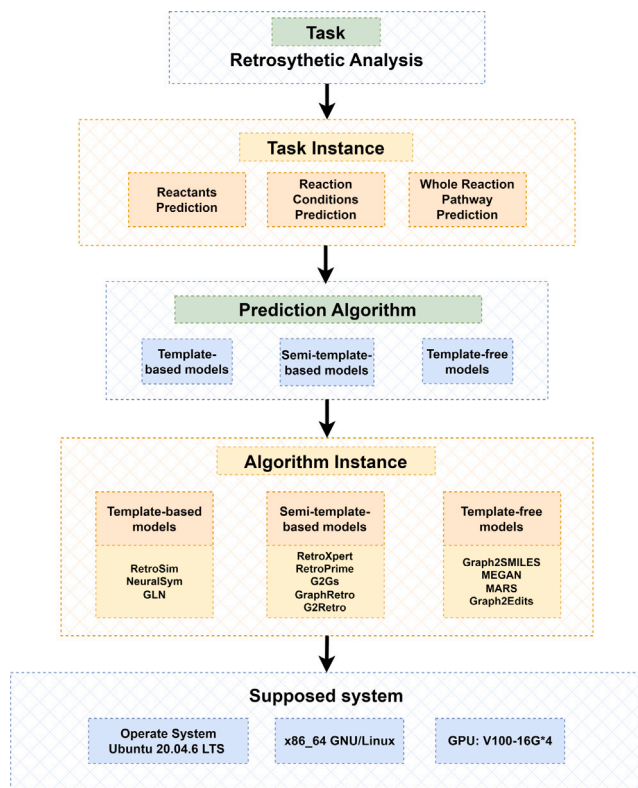


Fig. 2. Valid Evaluation Condition (EC) configurations [69].

3. System architecture of BigTensorDB

3.1. Methodology

The development of BigTensorDB is predicated on a meticulous evaluation of the existing work on retrosynthetic analysis. Our work is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. To demonstrate BigTensorDB's performance, we evaluate its effectiveness in the retrosynthetic analysis process. Inspired by the Evaluatolgy mentioned in the [69], we conduct a comprehensive survey of the current predictive efforts in retrosynthetic analysis, covering the entire process, including reactants prediction and reaction conditions prediction. Through developing the valid and well-defined Evaluation Condition (EC) configurations, we establish a solid foundation for the motivation of our works.

As shown in Fig. 2, we construct an evaluation condition configurations for the retrosynthetic analysis task inspired by the methodology in [69]. We first clarify that the problem task of our work is retrosynthetic analysis (E'). The specific task instance (E) is the concrete problem that needs to be solved in the process of retrosynthetic analysis, such as reactants prediction (E_1) based on the target product, reaction conditions prediction (E_2) based on the reaction equation and whole reaction pathway prediction (E_3). Under the task instances, we find different algorithms (A') to solve them, which mainly include template-based, semi-template-based, and template-free models method. The specific algorithm instance (A) is each concrete predictive model itself. Each algorithm (A') has its corresponding algorithm instance (A). For example, the template-based models method (A') includes instances such as RetroSim (A_{11}), NeuralSym (A_{12}) and GLN (A_{13}).

3.2. Dataset sources and workload choices

There are many open-source datasets of chemical reaction expressions in the field of retrosynthetic analysis, such as USPTO-50K and

REACTION INDEX SYSTEM (REAXYS). However, none of these datasets completely include the reaction conditions in the chemical equations. Obtaining a dataset containing reaction conditions requires a lot of work.

Parrot [17] has organized two large datasets, USPTO-Condition and Reaxys-TotalSyn-Condition, which record reaction equations and reaction conditions, including solvents, reagents, catalysts, etc. They used the reaction classifier to subdivide the dataset categories, and designed an external verification experiment. Therefore, we select the USPTO-Condition dataset. We remove the data with more than two reactants and merge the reaction conditions according to our boundary settings. We finally establish a workload consisting of a training set with 490,398 data entries and a test set with 100 data entries.

3.3. BigTensorDB design and implementation

The design of BigTensorDB is divided into four layers, as shown in Fig. 3. We aim to provide a one-stop, full-cycle service for retrosynthetic analysis, in order to improve user efficiency, reduce costs, and explore ways to overcome the performance bottlenecks in overall prediction accuracy. Users only need to input a target molecule and select the desired models for reactants prediction and reaction conditions prediction. They then can receive a re-ranked prediction candidate set containing complete reaction equations and real reactions for reference.

We are committed to reordering and referencing prediction candidates by retrieving similar templates from a large-scale real chemical reaction database. To achieve this, we have designed the following four layers. In the storage layer, we carefully select feature extraction tools to extract features from chemical reaction datasets and store them in tensor format. In the prediction generation layer, we integrate multiple prediction models and provide them with a unified interface. In the search and analysis layer, we provide similarity retrieval and analysis processing services. The four layers mentioned earlier will be detailed in Sections 3.3.1 3.3.2 3.3.3 3.3.4.

3.3.1. Storage layer

In the storage layer, we determine a tensor format boundary for chemical reaction equations. A complete chemical equation involves reactants, products, and many reaction conditions, including temperature, reagents, solvents, and so on. If we use vector format to store this information, we need to embed all the above information into one vector, which will cause a huge loss in the dimension of chemical information. Therefore, we hope to use a tensor composed of multiple vectors to preserve all the information in the chemical reaction equation.

Based on Evaluatolgy described in [69], we define reactants, products, solvents and reagents as the four-dimensional parameters of each tensor. We also specify that the dimension size within each dimension of the tensor is 384. Using the feature extraction tool ChemBERTa-77M-MLM model, we convert the chemical equations from SMILES format to 384 dimensional vectors. They then are stored in the tensor format, thereby establishing a tensor-based knowledge base of known chemical equations.

3.3.2. Prediction generation layer

In the prediction generation layer, the user-input target molecule serves as the input for the models. Predictions and generations are carried out step-by-step according to the user-selected reactant prediction model and reaction condition prediction model. We assume the input target molecule is T , the selected reactant prediction model is A , and the selected reaction condition prediction model is B . By inputting the target molecule T into model A , we obtain a set of reactant prediction candidates S_1 without reaction conditions. We then input each candidate from S_1 into model B to obtain a set of complete chemical equation prediction candidates S_2 . Assuming that models A and B generate n and m prediction candidates for each input, respectively, the size of the complete reaction prediction set S_2 is $n * m$.

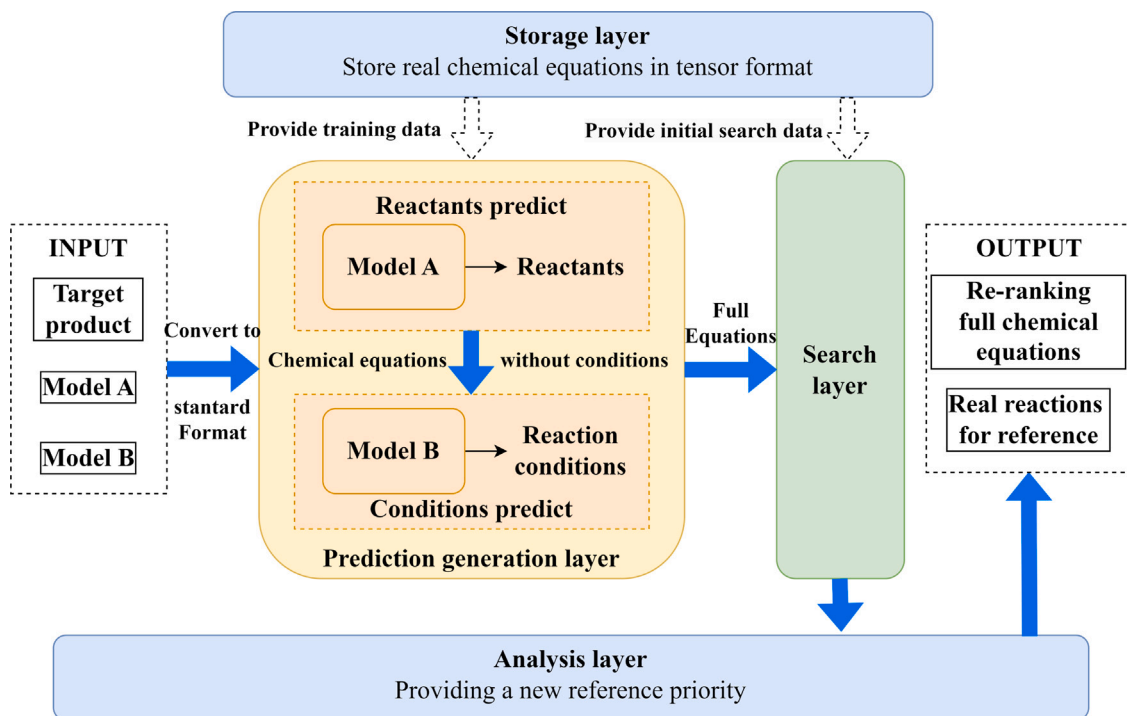


Fig. 3. Overview of the BigTensorDB workflow.

3.3.3. Search layer

In the search layer, we utilize the multi-vector search and apply a weighted ranker to set weights for multi-vector searches. We perform similarity searches for each prediction candidates in the set S_2 against the database of real reactions, obtaining search scores and the similar real reactions. Here, we have designed a preliminary experiment to test the effectiveness of different weight allocations for tensor dimensions during retrieval.

3.3.4. Analysis layer

In the analysis layer, we re-rank the reaction prediction candidate set S_2 based on the search scores obtained in the search layer. The search results return the top 5 real reactions with the highest similarity scores and their corresponding search scores. Our re-ranking strategy prioritizes candidates that achieve higher similarity scores during the search. Through preliminary experiments, we have selected the optimal weight sequence for re-ranking: $a_1 = 0.3$, $a_2 = 0.9$, and $a_3 = a_4 = 0.1$. According to this ranking information, we output to users a re-ranked prediction candidate set with improved accuracy, along with corresponding similar real reactions as reference suggestions.

4. Performance evaluation

4.1. Experiment setup

The server is equipped with 2 Intel Xeon 5218R CPUs running at 2.10 GHz, 512 GB of memory, and an NVIDIA V100-PCIE-16 GB GPU connected via PCIe 3.0. Each CPU has 20 physical cores with hyper-threading enabled, resulting in a total of 80 hardware threads, all of which were utilized. The operating system is Ubuntu 20.04 with the Linux kernel version 5.15.0. The GPU driver version is 535, and CUDA 12.2 is used for GPU computing. All experiments were conducted using Python 3.10 and Docker 26.1.

4.2. Experiment design

Our experiments include validating the research motivation and assessing our work's performance. The research motivation validation consists of theoretical analysis and experimental verification, with the latter providing the baseline model's performance metrics. The performance assessment of our work involves two main comparisons: one for retrosynthetic analysis and another for database performance. For retrosynthetic analysis, we focus on two key metrics: predictive accuracy and time cost. In the tensor database field, we evaluate tensor retrieval recall and throughput.

4.3. Motivation verification

4.3.1. Theoretical analysis

As shown in Fig. 4, we assume the input target molecule is T , the selected reactant prediction model is A , the selected reaction condition prediction model is B , the Top- k accuracy of model A is $E_A(k)$ and the Top- k accuracy of model B is $E_B(k)$.

By inputting the target molecule T into Model A , we obtain a reactant prediction candidate set S_1 without reaction conditions, with a size of n . Taking one candidate reaction equation i from S_1 as input into Model B , we obtain a complete reaction equation prediction candidate set S_2^i with a size of m . By inputting each candidate reaction equation from S_1 into Model B , we obtain the final reaction equation prediction candidate set S_2 with a size of $n * m$. The Top- k accuracy of the candidate reaction equations in S_2 can be obtained as:

$$E(k) = E_A(i) * E_B(j), \quad k = m * (i - 1) + j$$

It is easy to see that $E_A(i) < 1$ and $E_B(j) < 1$. Given $k = m * (i - 1) + j$, we can deduce that $i \leq k$ and $j \leq k$. Moreover, since E_A and E_B are non-decreasing functions, it follows that:

$$E(k) = E_A(i) * E_B(j) < E_A(i) \leq E_A(k)$$

$$E(k) = E_A(i) * E_B(j) < E_B(j) \leq E_B(k)$$

Thus, there is a significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.

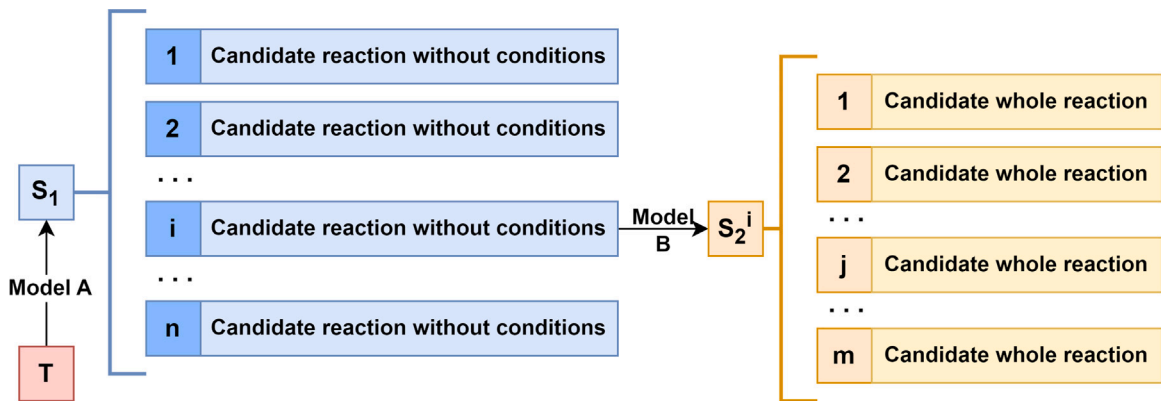


Fig. 4. Model A for reactants prediction, Top-k accuracy is $E_A(k)$, candidate set S_1 's size is n . Model B for reaction conditions prediction, Top-k accuracy is $E_B(k)$, candidate set S_2^i 's size is m .

Table 1

Full cycle prediction accuracy of inverse synthesis analysis.

k	1	3	5	10
Accuracy (%)	0.12	0.16	0.19	0.2

4.3.2. Experiment verification

We choose model RetroTRAE [16] as model A for reactants prediction and model Parrot [17] as model B for reaction conditions prediction. Then we use the workload built in Section 3.2.

We first generate $n = 10$ candidate reaction equations without reaction conditions per test. Then we input each candidate to model B to generate $m = 15$ candidate reaction equations including reaction conditions per input. At the end, we get totally 12096 candidate complete chemical equations for 100 tests. We then calculate the Top-k accuracy is shown in Table 1. Also as shown in Fig. 1, compared with the original accuracies, each Top-k accuracy decreases sharply.

4.4. Performance results

4.4.1. Retrosynthetic analysis performance comparison

According to Section 3.1, our work is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. So we evaluate BigTensorDB's effectiveness in the retrosynthetic analysis process to demonstrate BigTensorDB's performance without conducting comparative experiments with other database system.

Based on Evaluatology, we select reactants, products, solvents and reagents as the storage format for chemical equations. We choose the ChemBERTa-77M-MLM model to embed SMILES. This model converts SMILES into 384-dimensional vectors through a neural network. After generating predictions for the test set, we embed the result set and conduct similarity searches. We select the multi-vector search algorithm from the Milvus vector database, using the IVF-Flat index and Euclidean distance metric. We perform retrieval experiments under different weights. Based on the search results, we re-rank the candidate set and recalculate the Top-k accuracy. As shown in Fig. 5, we define four variables as the parameter weights for retrieval ranking. Among them, a_1 corresponds to the reactants, a_2 corresponds to the products, and a_3 and a_4 correspond to the solvents and reagents, respectively.

In Figs. 5(a) and 5(b), we conduct controlled variable experiments for parameters a_1 and a_2 , respectively. In Fig. 5(a), we set $a_2 = 0.2, a_3 = a_4 = 0.2$ and a_1 to different values. Then we observe the re-ranked Top-k accuracy. The results shows that the performance is better when a_1 is in the range of 0.2 to 0.4. In Fig. 5(b), we set $a_1 = 0.2, a_3 = a_4 = 0.2$ and a_2 to different values. Then we observe the re-ranked Top-k accuracy. The results shows that the performance is better when a_2 is 0.9.

Then, we set $a_1 = 0.2, 0.3, 0.4, a_2 = 0.9, 0.95$ to conduct controlled variable experiments for parameters a_3 and a_4 . The results are shown in

Table 2

Comparison of prediction accuracy between BigTensorDB and the baseline.

Model	Top-1	Top-5	Top-10	Top-50	Top-100
Model(A+B)'s accuracy	0.12	0.19	0.20	0.26	0.26
BigTensorDB's accuracy	0.12	0.18	0.18	0.24	0.26

Fig. 6. The Figs. 6(a) 6(b) and 6(c) shows the results of $a_1 = 0.2, 0.3, 0.4$ respectively, where $a_2 = 0.9$. The Figs. 6(d) 6(e) and 6(f) shows $a_2 = 0.95$'s results. We can find out the performance is better when a_3 and a_4 is 0.1.

Therefore, We currently find that the re-ranking performance is better when a_1 is 0.3, a_2 is 0.9, a_3 and a_4 is 0.1. However, more fine-grained experiments are still needed.

After exploring the parameter space, we conducted comparison experiments to compare our work with the baseline model.

We first compared the top-k accuracy of BigTensorDB's re-ranking strategy with that of the original A + B baseline model, and the results are shown in the Table 2. The baseline model's results were reported in Table 1 in Section 4.3.2. BigTensorDB's prediction accuracy does not surpass the baseline models. This stable accuracy shows our work does not degrade the original prediction models, providing a solid base for further development.

In BigTensorDB, time consumption involves several parts: data cleaning for Model A, (Model A training), running Model A, organizing the results of Model A into the data format required for Model B, (Model B training), running Model B, combining the results of Model A and Model B, retrieving and re-ranking the results, and finally outputting the results. Under the workload from Section 3.2, we recorded BigTensorDB's total experimental time consumption, with the following results. The parts in bold represent the extra overhead from the BigTensorDB system.

As shown in the Table 3, BigTensorDB's extra time cost accounts for less than 5% of the entire process. Thus, when providing full-cycle services, BigTensorDB does not bring extra time consumption that cannot be tolerated.

4.4.2. Tensor database performance comparison

The recall and throughput of Milvus-IVFFlat based multi-vector search in this task scenario are 72% and 45 vectors per second, respectively. This performance is sub-par for vector search and indicates significant bottlenecks. The reason is that Milvus multi-vector search relies on weighted sorting after single-vector search rather than a genuine tensor index.

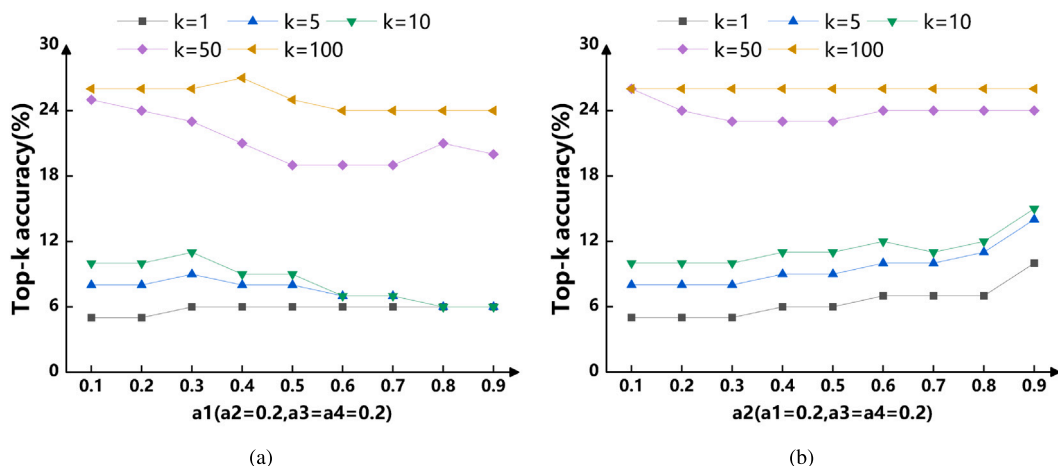
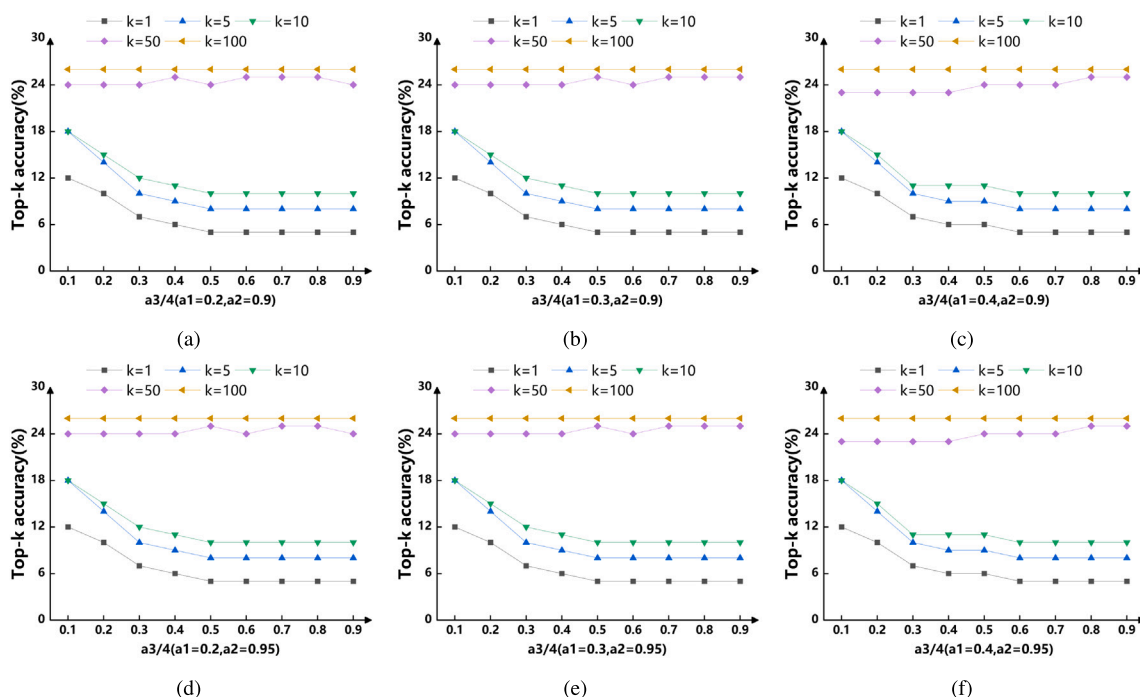
Fig. 5. Controlled variable experiments results of a_1 and a_2 .Fig. 6. Controlled variable experiments results of a_3 and a_4 .

Table 3

The time cost of each link in BigTensorDB.

Operation	Data cleaning (A)	Model A	Data cleaning (B)	Model B	Search	Re-rank and output
Time (h)	1	10	1	10	1	0.01

5. Lessons and future directions

5.1. Limitations of BigTensorDB

In our work, despite conducting meticulous experiments and careful theoretical analysis, we still identify certain shortcomings:

(1) The selection of prediction models is not diverse enough. There are only 2 models for reactant prediction and 4 models for reaction conditions prediction in our study.

(2) The accuracy of the top-k results after re-ranking still has significant space for improvement.

5.2. Future work of BigTensorDB

In our future work, we will primarily focus on the following four directions:

Firstly, we will expand and explore more prediction models. Our goal is to integrate more state-of-the-art and advanced prediction models. These models will include those for predicting reactants and predicting reaction conditions. By integrating different prediction models, we are committed to provide users with a unified and standardized environment. We will offer a broader one-stop selection space.

Secondly, we need to further explore the boundary conditions of the storage structure. In the current work of this study, the boundary conditions of the reaction equation are stored as four parts: reactants,

products, solvents and reagents. Given the vast variety and diverse types of chemical reaction conditions, the boundary definition of these conditions still requires more rigorous experimental analysis. Moving forward, Our goal is to determine whether using these conditions as the key conditions for the reaction equation is accurate. We plan to continue applying methods from evaluation science. We will also deploy more comparative experiments.

Thirdly, we need to conduct more diversified explorations of the embedding models that convert reaction equations from SMILES format to vectors. Currently, we have chosen ChemBERTa-77M-MLM as the embedding model. However, related work shows that there are many other embedding options available. We plan to deploy richer comparative experiments to explore the embedding models that can best preserve chemical information. We aim to ensure that the embedding models we use accurately reflect all the chemical information contained in the reaction equations. This will enable our vector retrieval work to be accurate and efficient.

Lastly, within the retrieval layer, we still need to explore the more fine-grained weight parameters corresponding to different conditions to maximize the accuracy after re-ranking. This is crucial because the corresponding weights can significantly enhance the accuracy and reliability of the final outcomes.

6. Conclusion

This paper proposes the first tensor database system, BigTensorDB, to help scientists in retrosynthetic analysis field. Our work effectively addresses the critical issues of the absence of a unified model capable of providing full-cycle service for retrosynthetic analysis and the significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.

Specifically, BigTensorDB designs an innovative tensor format that efficiently stores all key information related to chemical reactions, including reactants, products, and reaction conditions such as solvents and reagents. This format not only provides users with robust services for storing and retrieving chemical reactions but also lays the foundation for more accurate and comprehensive analysis. Additionally, the integration of multiple retrosynthetic analysis prediction models, including those for reactants and reaction conditions, along with SMILES embedding models, offers a seamless full-cycle retrosynthetic analysis service to users. This integration significantly reduces usage costs and enhances the efficiency of the entire pipeline. Moreover, by providing advanced search and analysis services, this work re-ranks and analyzes the final prediction results, offering real reaction equations similar to the predicted results for user reference. These efforts collectively enhance the accuracy and interpretability of the final outcomes, thereby advancing the state of the art in AI-based retrosynthetic analysis.

CRedit authorship contribution statement

Xueya Zhang: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Conceptualization. **Guoxin Kang:** Writing – review & editing, Methodology, Conceptualization. **Boyang Xiao:** Writing – original draft, Validation, Resources. **Jianfeng Zhan:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The author Jianfeng Zhan is an Editor-in-Chief for BenchCouncil Transactions on Benchmarks, Standards and Evaluations and was not involved in the editorial review or the decision to publish this article.

The other 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

This work was supported by the Innovation Funding of Institute of Computing Technology Chinese Academy of Sciences, China under Grant No. E461070 and Beijing Municipal Natural Science Foundation of Beijing Municipal Science and Technology Commission and Zhongguancun Science Park Administrative Committee, China under Grant No. QY24378.

References

- [1] X. Wang, Y. Li, J. Qiu, G. Chen, H. Liu, B. Liao, C.-Y. Hsieh, X. Yao, RetroPrime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions, *Chem. Eng. J.* (2021) 129845.
- [2] E.J. Corey, A.K. Long, S.D. Rubenstein, Computer-assisted analysis in organic synthesis, *Science* 228 (4698) (1985) 408–418.
- [3] A. Heifets, I. Jurisica, Construction of New Medicines Via Game Proof Search, AAAI Press, 2012.
- [4] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature* 555 (2017) 604–610.
- [5] A. Kishimoto, B. Buesser, B. Chen, A. Botea, Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning, Curran Associates Inc., Red Hook, NY, USA, 2019.
- [6] J.S. Schreck, C.W. Coley, K.J.M. Bishop, Learning retrosynthetic planning through simulated experience, *ACS Central Sci.* 5 (6) (2019) 970–981.
- [7] K. Lin, Y. Xu, J. Pei, L. Lai, Automatic retrosynthetic route planning using template-free models, *Chem. Sci.* 11 (2020) 3355–3364.
- [8] B. Chen, C. Li, H. Dai, L. Song, Retro*: Learning retrosynthetic planning with neural guided a* search, in: L. Hal Daumé, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, vol. 119, PMLR, 2020, pp. 1608–1616.
- [9] J. Kim, S. Ahn, H. Lee, J. Shin, Self-improved retrosynthetic planning, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, vol. 139, PMLR, 2021, pp. 5486–5495.
- [10] S. Ishida, K. Terayama, R. Kojima, K. Takasu, Y. Okuno, AI-driven synthetic route design incorporated with retrosynthesis knowledge, *J. Chem. Inf. Model.* 62 (2022) 1357–1367.
- [11] L. Yue, Z. Xinxin, Y. Zhengwei, S. Siqu, Machine learning embedded with materials domain knowledge, *J. Chinese Ceramic Soc.* 50 (3) (2022) 863–876.
- [12] Y. Liu, L. Ding, Z. Yang, et al., Domain knowledge discovery from abstracts of scientific literature on nickel-based single crystal superalloys, *Sci. China Technol. Sci.* 66 (2023) 1815–1830.
- [13] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, *J. Mater.* 9 (4) (2023) 798–816.
- [14] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, *J. Mater.* 3 (3) (2017) 159–177, High-throughput Experimental and Modeling Research toward Advanced Batteries.
- [15] S. Siqu, T. Zhangwei, Z. Xinxin, S. Shiyu, Y. Zhengwei, L. Yue, Applying data-driven machine learning to studying electrochemical energy storage materials, *Energy Storage Sci. Technol.* 11 (3) (2022) 739–759.
- [16] U.V. Ucak, I. Ashyrmamatov, J. Ko, J. Lee, Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments, *Nat. Commun.* 13 (2022).
- [17] X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou, X. Yao, Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center, *Research* 6 (2023) 0231.
- [18] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 28 (1988) 31–36.
- [19] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q.L. Nguyen, S. Ho, J.L. Sloane, P.A. Wender, V.S. Pande, Retrosynthetic reaction prediction using neural sequence-to-sequence models, *ACS Central Sci.* 3 (2017) 1103–1113.
- [20] S. Zheng, J. Rao, Z. Zhang, J. Xu, Y. Yang, Predicting retrosynthetic reactions using self-corrected transformer neural networks, *J. Chem. Inf. Model.* (2019).
- [21] I.V. Tetko, P. Karpov, R.V. Deursen, G. Godin, State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis, *Nat. Commun.* 11 (2020).
- [22] E. Kim, D. Lee, Y. Kwon, M.S. Park, Y.-S. Choi, Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables, *J. Chem. Inf. Model.* 61 (1) (2021) 123–133.
- [23] S. Seo, Y.Y. Song, J.Y. Yang, S. Bae, H. Lee, J. Shin, S.J. Hwang, E. Yang, GTA: Graph truncated attention for retrosynthesis, in: AAAI Conference on Artificial Intelligence, 2021.
- [24] Y. Jiang, Y. Wei, F. Wu, Z. Huang, K. Kuang, Z. Wang, Learning chemical rules of retrosynthesis with pre-training, in: AAAI Conference on Artificial Intelligence, 2023.

- [25] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M.-Y. Wu, T. Hou, M. Song, Root-aligned SMILES: A tight representation for chemical reaction prediction, *Chem. Sci.* 13 (2022) 9023–9034.
- [26] Y. Zhang, R. Yu, K. Zeng, D. Li, F. Zhu, X. Yang, Y. Jin, Y. Xu, Text-augmented multimodal LLMs for chemical reaction condition recommendation, 2024, *ArXiv abs/2407.15141*.
- [27] V.R. Somnath, C. Bunne, C.W. Coley, A. Krause, R. Barzilay, Learning graph models for retrosynthesis prediction, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, 2021.
- [28] Y. Wan, C.-Y. Hsieh, B. Liao, S. Zhang, Retroformer: Pushing the limits of end-to-end retrosynthesis transformer, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, vol. 162, PMLR, 2022, pp. 22475–22490.
- [29] Y. Han, et al., Retrosynthesis prediction with an iterative string editing model, *Nat. Commun.* (2024).
- [30] C.W. Coley, L. Rogers, W.H. Green, K.F. Jensen, Computer-assisted retrosynthesis based on molecular similarity, *ACS Central Sci.* 3 (2017) 1237–1245.
- [31] M.H.S. Segler, M.P. Waller, Neural-symbolic machine learning for retrosynthesis and reaction prediction, *Chemistry* 23 25 (2017) 5966–5971.
- [32] H. Dai, C. Li, C.W. Coley, B. Dai, L. Song, Retrosynthesis Prediction with Conditional Graph Logic Network, Curran Associates Inc., Red Hook, NY, USA, 2019.
- [33] S. Chen, Y. Jung, Deep retrosynthetic reaction prediction using local reactivity and global attention, *JACS Au* 1 (10) (2021) 1612–1620.
- [34] J. Dong, M. Zhao, Y. Liu, Y. Su, X. Zeng, Deep learning in retrosynthesis planning: datasets, models and tools, *Brief. Bioinform.* (2021).
- [35] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (1997) 1735–1780.
- [36] Z. Tu, C.W. Coley, Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction, *J. Chem. Inf. Model.* (2021).
- [37] M. Sacha, M. Blaz, P. Byrski, P. Włodarczyk-Pruszyński, S. Jastrzebski, Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits, *J. Chem. Inf. Model.* (2020).
- [38] J. Liu, C. chao Yan, Y. Yu, C. Lu, J. Huang, L. Ou-Yang, P. Zhao, MARS: A motif-based autoregressive model for retrosynthesis prediction, *Bioinformatics* 40 (2022).
- [39] W. Zhong, Z. Yang, C.Y.-C. Chen, Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing, *Nat. Commun.* 14 (2023).
- [40] L. Fang, J. Li, M. Zhao, L. Tan, J.-G. Lou, Single-step retrosynthesis prediction by leveraging commonly preserved substructures, *Nat. Commun.* 14 (2023).
- [41] C. Shi, M. Xu, H. Guo, M. Zhang, J. Tang, A graph to graphs framework for retrosynthesis prediction, in: International Conference on Machine Learning, 2020.
- [42] C. chao Yan, Q. Ding, P. Zhao, S. Zheng, J. Yang, Y. Yu, J. Huang, RetroXpert: Decompose retrosynthesis prediction like a chemist, 2020, *ArXiv arXiv:2011.02893*.
- [43] H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green, K.F. Jensen, Using machine learning to predict suitable conditions for organic reactions, *ACS Central Sci.* 4 (2018) 1465–1476.
- [44] Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, S. Shi, Data quantity governance for machine learning in materials science, *Natl. Sci. Rev.* 10 (7) (2023) nwad125–.
- [45] Y. Liu, X. Ge, Z. Yang, S. Sun, D. Liu, M. Avdeev, S. Shi, An automatic descriptors recognizer customized for materials science literature, *J. Power Sources* 545 (2022) 231946.
- [46] S. Siqi, SUN, M. Shuchang, Z. Xinxin, Q. Quan, L. Yue, Detection method on data accuracy incorporating materials domain knowledge, *J. Inorg. Mater.* 37 (12) (2022) 1311–1320.
- [47] L. Yue, Y. Wenxuan, L. Dahui, D. Lin, Y. Zhengwei, L. Wei, Y. Tao, S. Siqi, Named entity recognition driven by high-quality text data accelerates the knowledge discovery of nickel-based single crystal superalloys, *Acta Metall. Sin.* 60 (10) (2024) 1429–1438.
- [48] L. Yue, L. Da-Hui, G. Xian-Yuan, Y. Zheng-Wei, M. Shu-Chang, Z.Z. Yi, S.S.-Q. 2, A high-quality dataset construction method for text mining in materials science, *Acta Phys. Sin.* 72 (7) (2023) 41–54.
- [49] L. Yue, M. Shuchang, Y. Zhengwei, Z. Xinxin, S. Siqi, A data quality and quantity governance for machine learning in materials science, *J. Chinese Ceramic Soc.* 51 (2) (2023) 427–437.
- [50] Y. Liu, J.M. Wu, M. Avdeev, S.Q. Shi, Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties, *Adv. Theory Simul.* 3 (2) (2020).
- [51] Q. Zhao, L. Zhang, B. He, A. Ye, M. Avdeev, L. Chen, S. Shi, Identifying descriptors for li+ conduction in cubic li-argyrodites via hierarchically encoding crystal structure and inferring causality, *Energy Storage Mater.* 40 (2021) 386–393.
- [52] Y. Liu, X. Zou, S. Ma, M. Avdeev, S. Shi, Feature selection method reducing correlations among features by embedding domain knowledge, *Acta Mater.* 238 (2022) 118195.
- [53] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, E. Ozkirimli, Exploring chemical space using natural language processing methodologies for drug discovery, *Drug Discov. Today* (2020).
- [54] X. Li, D. Fourches, SMILES pair encoding: A data-driven substructure tokenization algorithm for deep learning, *J. Chem. Inf. Model.* 61 (4) (2021) 1560–1569.
- [55] I. Lee, H. Nam, Infusing linguistic knowledge of SMILES into chemical language models, 2022, *ArXiv abs/2205.00084*.
- [56] B. Fabian, T. Edlich, H. Gaspar, M.H.S. Segler, J. Meyers, M. Fiscato, M. Ahmed, Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020, *ArXiv abs/2011.13230*.
- [57] S. Wang, Y. Guo, Y. Wang, H. Sun, J. Huang, SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019.
- [58] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020, *ArXiv abs/2010.09885*.
- [59] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa-2: Towards chemical foundation models, 2022, *ArXiv abs/2209.01712*.
- [60] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019.
- [61] T. Tran, C. Ekenna, Molecular descriptors property prediction using transformer-based approach, *Int. J. Mol. Sci.* 24 (2023).
- [62] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, P. Wang, MolRoPE-BERT: An enhanced molecular representation with rotary position embedding for molecular property prediction, *J. Mol. Graph.* 118 (2022) 108344.
- [63] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, P. Wang, MolRoPE-BERT: An enhanced molecular representation with rotary position embedding for molecular property prediction, *J. Mol. Graph.* 118 (2022) 108344.
- [64] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, C. Xie, Milvus: A purpose-built vector data management system, in: Proceedings of the 2021 International Conference on Management of Data, 2021.
- [65] C. Wei, B. Wu, S. Wang, R. Lou, C. Zhan, F. Li, Y. Cai, AnalyticDB-V: A hybrid analytical engine towards query fusion for structured and unstructured data, *Proc. VLDB Endow.* 13 (12) (2020) 3152–3165.
- [66] W. Yang, T. Li, G. Fang, H. Wei, PASE: Postgresql ultra-high-dimensional approximate nearest neighbor search extension, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020.
- [67] Vearch: A distributed system for embedding-based retrieval, 2020, URL: <https://github.com/vearch/vearch>.
- [68] J. Li, H.-F. Liu, C. Gui, J. Chen, Z. Ni, N. Wang, Y. Chen, The design and implementation of a real time visual search system on jd E-commerce platform, in: Proceedings of the 19th International Middleware Conference Industry, 2018.
- [69] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatolgy: The science and engineering of evaluation, *BenchCouncil Trans. Benchmarks, Stand. Eval.* 4 (1) (2024) 100162.



Research Article

Predicting the number of call center incoming calls using deep learning

Armaghan Nikfar^{a,b}, Javad Mohammadzadeh^{a,b,*}

^a Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran

^b Institute of Artificial Intelligence and Social and Advanced Technologies, Ka.C., Islamic Azad University, Karaj, Iran

ARTICLE INFO

Keywords:

Call volume

Prediction

Bidirectional long short-term memory (BLSTM)

Feature extraction customer service

management holiday impact on call traffic

ABSTRACT

One of the main problems in call centers is the call queue. This can lead to long waiting times for customers, increased frustration and call abandonment. The important role that predictive analytics plays in optimizing call center operations is increasingly recognized. Advanced models can be trained by training datasets such as the number of calls that have occurred throughout history, and by estimating how religious and public holidays have affected the weight of hours and the number of calls, and this study provides an analysis of 4 years. Call center data from Shatel, an Internet service provider. Predictive deep learning models, specifically the Bidirectional Short-Term Memory Model (BLSTM), were used to predict the number of incoming calls, predict the number of calls to centers, and prevent call queues with an accuracy of 90.56.

1. Introduction

Today, managing incoming calls in call centers is one of the main challenges of the organization. Accurately predicting the number of calls can significantly help organizations improve service quality. In the call center of a company, various functions and tasks are performed to ensure effective communication with customers and provide effective services to them. These include dealing with customer questions, solving their problems, processing orders, order support, selling new products, introducing products, buying advice, technical support, etc. Answering customer calls can lead to call queues. In general, call queues occur when the number of incoming calls to the call center exceeds the number of available agents to answer them. This technology helps call centers avoid queues by accurately predicting the number of calls on an hourly basis. This prediction reduces waiting time and increases customer service. In addition, AI adapts in real time to dynamic factors such as queue length, thereby improving operational efficiency and saving money. This can lead to better customer satisfaction, more efficient operations and ultimately a more effective call center. Choosing the best architecture for predictive models, depends on the task and data characteristics [1].

A call center is a centralized office or facility where a company handles a large volume of telephone calls, primarily for customer service, support, sales, and other inquiries. Agents in a call center manage inbound and outbound calls, providing assistance, resolving issues,

processing orders, and conducting telemarketing or market research. Call centers play a critical role in maintaining effective communication between a company and its customers, ensuring customer satisfaction, efficient problem resolution, and streamlined operations. These are the reasons for the importance of predicting call center calls:

Anticipating and managing the volume of calls helps companies at any given time and prevents overtime or shortage of human resources. This efficiency reduces operating costs and improves resource utilization. Lower call volume means shorter wait times and faster service, leading to higher customer satisfaction. Customers appreciate timely responses and quick solutions to their problems. Optimizing the volume of calls, reducing costs with employees, including the use of systems, headsets, etc., overtime and infrastructure maintenance. Analyzing traffic patterns provides insights into customer behavior, seasonal trends and fluctuations in service demand. This data can help with strategic decisions to improve service delivery and operational planning [2]. Overall, minimizing contact center traffic through accurate forecasting and management and timely strategic decisions not only improves operational efficiency, but also increases customer satisfaction, reduces costs, and provides strategic benefits to the organization. This project was implemented at Shatel, an Internet Service Provider (ISP), with excellent results. Accurately forecasting the number of calls significantly improved the efficiency of Shatel's call center, reduced customer waiting time, and increased overall customer satisfaction. It also made Shatel able to consider the experts needed for this number of calls based on the

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

* Corresponding author.

E-mail address: j.mohammadzadeh@kiaui.ac.ir (J. Mohammadzadeh).

<https://doi.org/10.1016/j.tbench.2025.100213>

Received 30 January 2025; Received in revised form 24 April 2025; Accepted 14 May 2025

Available online 4 June 2025

2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

prediction of the number of calls. There are different types of deep learning such as LSTM, GRU, BLSTM, CNN, ... In this model, a Bidirectional short-term memory model (BLSTM) has been used in the AI system. The BLSTM model is particularly suitable for this application because it can process sequential data in forward and backward directions and capture dependencies over long periods. This capability is critical to accurate call count forecasting, as it allows the model to understand and incorporate historical data patterns, including seasonal trends and specific time-based behaviors. By using the advanced features of the BLSTM model, the AI can make highly accurate predictions and ensure that call centers are well prepared for different call volumes throughout the year. Reducing call center traffic is the purpose of this project.

2. Literature review

Recent developments in deep learning methods impacted various fields throughout these years. Call center management is one of the fields. A comprehensive review of call center management challenges has been reviewed. This section also contains history and applications of used methods for this paper. The following subsections provide a comprehensive analysis of the literature categorized into four key themes: Call Centers, LSTM, BiLSTM, and GRU.

2.1. Call centers

Call centers have such a huge impact in today's business world. Call centers have challenges in several main domains, including forecasting, capacity planning, queueing, and personnel scheduling [3]. As [4] says "Accurately modeling and forecasting future call arrival volumes is a complicated issue which is critical for making important operational decisions, such as staffing and scheduling, in the call center".

In [5] authors demonstrated enhanced prediction performance compared to traditional models like CNN and Transformer by utilizing sequence data and improving the MMoE algorithm.

Another approach used a Long Short-Term Memory (LSTM) network to predict call traffic characteristics and capture non-linear patterns. These predictions are incorporated into a reinforcement learning-based model for optimizing agent schedules dynamically, improving efficiency and maintaining service quality [2].

A case study proved that applying RNN-based deep learning algorithms to automate customer complaint classification, validate the suitability of deep learning for call center complaint management [6].

A hierarchical LSTM architecture used for customer satisfaction estimation. Lower LSTM used for capturing sequential information within customer dialogue turns, while the upper LSTM estimated overall satisfaction using aggregated turn-level outputs [7].

The main goal of one of the studies was to predict call volumes over a 40-day horizon for better workforce management and LSTM models were tested for capturing temporal patterns in the data. However statistical models such as SARIMAX performed better in this study [8].

2.2. Long-short term memory networks (LSTM)

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) aimed at mitigating the vanishing gradient problem (Hochreiter, Sepp [9]). LSTM is an RNN sort in which other nodes in the same layer are linked to boost learning with the removal and retrieval of relevant knowledge [10]. Fig. 1 has shown a complete architecture of the LSTM network.

In [11] LSTM was used to handle long-term dependencies in the traffic data combined with the TCN for extracting short-term features.

In a study authors propose an objective framework based on speech signal processing to classify agent's productivity levels. The LSTM network combined with attention layers was used to capture the sequential nature of speech signals [12].

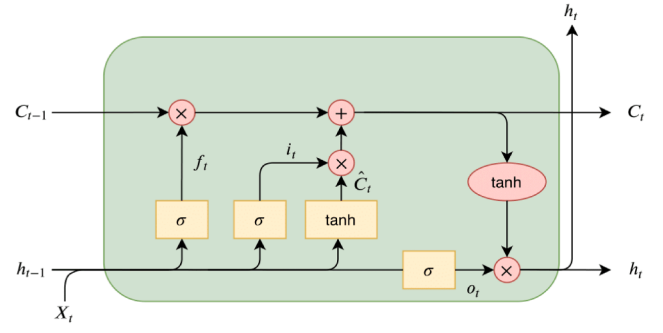


Fig. 1. Long-short term memory networks.

In several studies LSTM's ability to capture long term dependencies have been used in multiple domains such as predicting Traffic Speed [13] and stock market's price [14].

Analyzing time-series is another field that LSTM have shown promising results. Analyzing Missing and spatial data also have been studied and LSTM outperformed other algorithms [15].

2.3. Bidirectional LSTM (BLSTM)

Bidirectional LSTM (BLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It's also a powerful tool for modeling the sequential dependencies between words and phrases in both directions of the sequence [16]. In Fig. 2 the full architecture of BLSTM has been shown.

Both forward and backward directions will be used for analyzing data in BLSTM networks. In [17] study, The BLSTM component enhanced prediction by analyzing data in both forward and backward directions, and achieved an average accuracy of 97.68 %.

A study compared a Uni-LSTM with BLSTM to create a model to predict short-term traffic and the BLSTM, which processed input data in both forward and backward directions, outperformed the Uni-LSTM for that matter [18].

Combination of CNN for extracting high-level features from input wind speed time series data and BLSTM for BLSTM processing these features bidirectionally as hybrid model, created a reliable model to predict wind speed with high accuracy [19].

BLSTM combined with Support Vector Machine (SVM) also had been used in language models for sentimental analysis. BLSTM network captured sequential dependencies and context in Arabic text. This sentimental analysis helped identify customer sentiments to assess customer satisfaction and improve services [20].

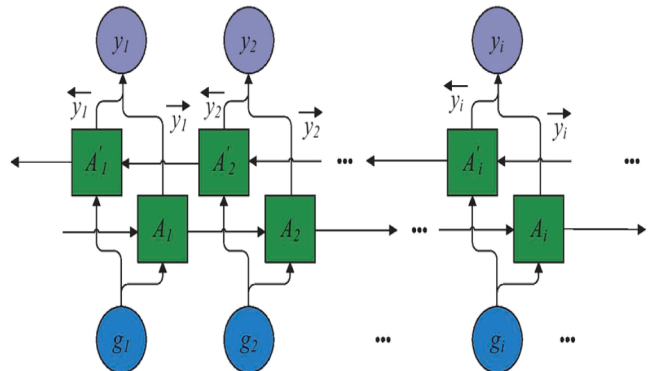


Fig. 2. Architecture from BLSTM model.

2.4. The gated recurrent unit (GRU)

Gated Recurrent Neural Network (RNN) have shown successful results in several applications including sequential data, natural language processing, machine translation, speech recognition [21], financial sequence predictions [22], Machine health monitoring [23], etc. In Fig. 3 complete architecture of GRU model has been shown.

A study compared traditional models like ARIMA and LSTM with GRU-based approaches for predicting traffic flow. GRU-based methods showed superior accuracy and lower computational requirements [24].

Variants of GRU models, such as regression GRU and stacked GRU have been studied and compared in a paper to create a prediction system for electricity generation's planning and operation, are applied to improve prediction accuracy. The multi-GRU model outperformed conventional approaches like RNN and LSTM [25].

Low computational advantage with faster training of GRU-based models had been used in a study for stock prediction. GRU-based model demonstrated its advantages, while LSTM excels in capturing long-term dependencies creating optimized hybrid model [26].

Applying hybrid GRU-LSTM models were also been used in other studies such as [13] to predict traffic speeds using data from multiple heterogeneous sources. As mentioned before, LSTM captured long-term temporal dependencies, while GRU handled short-term patterns, leading to improved performance. Proposed model demonstrated high accuracy in predicting urban traffic speed.

3. Previous research

Much literature has been developed in the forecasting of call arrivals.

According to [6] For service providers, a call center is a communication bridge between customers and providers that can handle any inquiries and requests.

(Ibrahim, R et al. [27]) said one important source of this uncertainty is the call arrival rate, which is typically time-varying, stochastic, dependent across time periods and call types, and often affected by external events. the specifics of the forecasting procedure need to be determined carefully, including the forecasting horizon (for time intervals, for a day, or for multiple days), and whether to combine arrivals from separate queues.

Also according to (Chacón, H et al. [28]) For any call center facility, the number of call arrivals represents a key component between customer satisfaction and budget constraints. Hence, the ability to accurately forecast the number of calls for a particular period of time is an effective measure in planning customer service and reducing any resource waste. This research presents a comparison between traditional time series forecasting methods and machine/deep learning techniques to predict the number of call arrivals for short and long term periods.

In (Kumwilaisak, W, et al. [2]) said Without effective staff allocation, improper workforce management can degrade service quality and

reduce customer satisfaction. So they decided to create a deep neural network method to learn and predict call center traffic characteristics. The deep neural network consists of a Long-Short Term Memory (LSTM) network and a Deep Neural Network (DNN) capturing non-linear call traffic behaviors. The call center traffic prediction utilized a deep neural network containing the LSTM and the fully connected networks to predict call center traffic parameters.

In (Mohammed, R. A., & Pang, P [29]) said A call center operates with customers calls directed to agents for service based on online call traffic prediction. This paper proposes an agent call prediction method that uses agent skill information as prior knowledge for call prediction. This research shows that the use of deep neural networks and hybrid models can be very effective in predicting call traffic, customer anger, and telecommunications network traffic. These predictions help improve service quality, customer satisfaction, and resource efficiency. In the research we conducted, we will predict the number of incoming calls for better call center management based on the factors affecting the call.

Also in this paper, (Bugarčić, P et al. [30]) presented a solution using machine learning that predicts the number of calls entering the call center per hour using supervised machine learning. For the prediction, they used the WEKA machine learning software tool and the prediction results are verified using several methods, which show very good results.

4. Predicting number of call's

In call centers, one of the most critical aspects is preventing the formation of long call queues. When no queues exist, customer satisfaction significantly increases because callers can quickly connect to agents and resolve their issues. On the other hand, when queues form, customers waiting on hold for extended periods may become frustrated, leading to dissatisfaction. To address this, we developed a predictive model capable of forecasting the number of calls based on various influencing factors.

Factors Influencing Call Queue Congestion

Several factors affect the number of calls a call center receives, which in turn affects the likelihood of a call queue forming. We will explore these factors below.

Holidays: During public or festive holidays, call volumes tend to decrease as people are often traveling or engaged in other activities.

Off-peak hours: During late-night hours or non-business hours, the number of calls naturally drops.

To improve the prediction process, we collected all relevant factors influencing call queues and incorporated them into the model as features. Each feature was assigned an appropriate weight to ensure the model could effectively learn their importance.

4.1. Model selection and approach

Given that our data is time-dependent and sequential, classification models were deemed the best fit for this task. Among them, LSTM (Long Short-Term Memory) networks were selected as the optimal choice due to their ability to handle sequential data and remember temporal dependencies.

Our LSTM model was trained on historical call volume data categorized by date and hour over several years. The model also accounts for nuances in the Iranian calendar, such as leap years, as well as official and unofficial holidays, including religious events, public holidays, and more. The model automatically preprocesses the data based on the input year, labels holidays, and prepares the dataset for prediction.

Once the data is prepared, the model predicts call volumes for each hour and date, enabling proactive planning.

4.2. Feature selection and performance improvement

Feature selection played a vital role in enhancing the accuracy of our

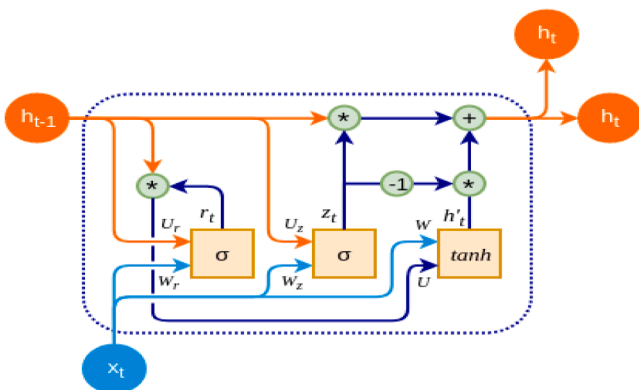


Fig. 3. Architecture of GRU model.

predictions. By identifying the most relevant features and eliminating irrelevant ones, the model's precision improved. When we compared the model's accuracy with and without feature selection, the version with feature selection performed significantly better.

4.3. Predicting required staffing levels

In addition to predicting call volumes, our model also determines the optimal number of agents required to handle customer inquiries without causing delays. This calculation takes into account:

1. The predicted number of calls.
2. The likelihood of agents temporarily stepping away from their desks for short breaks.

Using historical data on agent availability during shifts, the model estimates the minimum number of agents required for each hour. To further improve accuracy, we added calculations to predict how many agents might step away from the queue during specific times.

As a result, the model provides highly accurate staffing recommendations for each shift. Managers can use this information to optimize resource allocation and ensure smooth operations during peak and non-peak hours. Additionally, this capability supports HR departments in workforce planning and recruitment by offering precise estimates of staffing needs.

5. Results and impact

We validated the LSTM model's predictions against real-world data, and the differences were minimal. For staffing predictions, we conducted trial runs using the model's recommendations. During these test shifts, no call queues were formed, demonstrating the effectiveness of the predicted staffing levels.

Benefits for Call Centers

Implementing this model in call centers offers several advantages, including:

Enhanced customer satisfaction: Reduced wait times and quicker issue resolution lead to happier customers.

Optimized workforce allocation: The model ensures sufficient agents are available without overstaffing.

Cost savings: By efficiently allocating human resources, unnecessary staffing costs are minimized.

Reduced agent stress: Avoiding overwhelming call volumes prevents burnout and ensures a healthier work environment.

In conclusion, this predictive model serves as a robust tool for call centers, streamlining operations, improving customer experience, and ensuring efficient use of resources.

The implementation of a predictive LSTM model for call centers marks a significant leap in improving operational efficiency and customer satisfaction. By accurately forecasting call volumes, our model empowers call centers to make informed, proactive decisions that prevent call queue congestion. This, in turn, enhances the customer experience, as callers are connected to agents promptly, ensuring faster resolution of their issues.

5.1. Key capabilities of the model

Accurate Call Volume Prediction: The model leverages historical data and temporal dependencies to forecast call volumes with high precision. It accounts for specific characteristics of the Iranian calendar, such as leap years, public and unofficial holidays, and religious events, ensuring that predictions are tailored to real-world conditions.

Feature Integration and Weight Assignment: By incorporating a wide range of influential factors such as holidays, off-peak hours, and historical call patterns, the model learns the impact of each feature. Assigning appropriate weights to these factors enhances the model's

ability to recognize patterns and adapt to varying conditions.

Staffing Optimization: Beyond predicting call volumes, the model accurately calculates the number of agents required for each shift. It considers the dynamic nature of call centers, including factors like agent breaks and temporary unavailability. This capability ensures sufficient staffing to handle call volumes efficiently, reducing customer wait times and preventing agent burnout.

Real-Time Adaptability: The model dynamically adjusts predictions based on input data for a given year, pre-labeling holidays and special events. This adaptability ensures that predictions remain relevant, even as conditions change.

Scalability and Applicability: The model is not limited to a single context or region. It can be adapted to different calendars, cultural events, and operational setups, making it a versatile tool for call centers worldwide.

5.2. Methodological contributions and practical novelty

While this research utilizes established deep learning models such as LSTM, GRU, and BLSTM, its methodological contributions lie in the following novel aspects:

Integration of Iranian Calendar-Based Features: Our model incorporates culturally specific temporal features such as Nowruz, religious holidays, summer breaks, and leap years. These are encoded dynamically and allow the model to learn region-specific patterns that are not commonly addressed in previous literature.

Dynamic Year-Adaptive Preprocessing: Unlike static approaches, our model can adaptively generate holiday features for any target year (e.g., 1403) without manual data labeling. This automation adds a novel layer to the preprocessing pipeline, making the system scalable and usable across different years and calendars.

Real-World Deployment and Validation: This study goes beyond theoretical analysis and demonstrates real-world deployment within Shatel's call center. The performance of the model was evaluated in operational settings, revealing its ability to eliminate call queues and enhance workforce planning.

Predictive Staffing Mechanism: In addition to call forecasting, we propose a method to calculate the required number of agents per hour, incorporating factors like queue size, AHT, and allowable breaks. This integration of prediction with resource optimization adds a practical, business-driven value to the model.

6. Methods

6.1. Datasets process

Datasets: Data collection begins with the collection of historical call data from the Shatel call center. As we know, call centers receive a lot of calls every day. These calls will be stored in a database. To be precise, our dataset consists of the same number of incoming calls to the call center every hour, for the past 4 years. Below is a sample of these calls (due to Shatel's security policy, these numbers are not real) (Table 1).

Similarly, data will be collected hourly for 4 consecutive years and will enter the pre-processing stage.

Data preprocessing: Before training the BLSTM model, the collected data undergoes preprocessing steps such as:

Table 1
Sample of datasets.

DATE	HOURS	CALLS
3/20/2020	0	50
3/20/2020	1	35
3/20/2020	2	20
3/20/2020	3	10

6.1.1. **Normalization:** normalized value, x' , is computed as

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

where x' is the original value, μ is the mean, and σ is the standard deviation of the feature.

This ensures all numerical inputs have zero mean and unit variance for better convergence.

6.1.2. **Encoding categorical variables**

Categorical variables (such as holidays) are converted to a numerical representation (binary or mono-hot encoding) that can be handled by the model. For binary encoding of holidays (e.g., Nowruz, weekends, etc.), assign:

Encoded Value = {1, if holiday or weekend 0, otherwise

For mono-hot encoding (if categorical variables have more than two states):

A categorical feature F with k states is encoded as a vector of size k, where only one element is 1 and others are 0.

6.1.3. **Sequence formation**

Data is organized into overlapping input-output pairs. For example, given a sequence of hourly call volumes $[x_1, x_2, \dots, x_n]$:

Input Sequence : $[x_1, x_2, \dots, x_n]$

n : n is a hyperparameter that defines the length of the historical window used for prediction. A larger n allows the model to consider longer-term dependencies, while a smaller n focuses on more immediate trends. The choice of n is determined experimentally based on the dataset and model performance.

Output (Prediction Target): x_{t+1}

6.2. **Feature extraction**

To extract factors for predicting call volume, the model uses statistical statistics to determine whether the year is a leap year (using statistical rules), and what religious and administrative holidays have that this data set consists of two elements (0 or 1).

These features include:

6.2.1. **Hours**

The hours are so important because the weight of hours are effective in incoming calls, for example in nights the weight of calls are stronger the other hours because calls will be increase at nights.

6.2.2. **Dates**

dates are features too because it is important to know which dates have more calls and which day it is.

6.2.3. **Nowruz holidays**

These holidays are usually associated with a few calls in the morning when people communicate with family and friends.

6.2.4. **Summer holidays**

These can result in reduced working hours throughout the day due to longer holiday periods.

6.2.5. **Holidays**

These can directly impact the amount of time associated with performance issues.

6.2.6. **Religious holidays**

These holidays have a different impact on the amount of time, such as reduced morning visits on Eid days. each of these factors affects the

weight of the watch. For example, a model that considers the Nowruz holidays as an important festival is less likely to predict fewer calls in the morning compared to the evening and night. by incorporating these sources and learning from historical data, the BLSTM model can accurately predict future call volume. This approach helps improve customer call center operations by anticipating and managing call traffic during various holiday periods.

6.2.7. **Day off**

in these days incoming calls will be less at morning and evening, because in day off works are closed and we don't have calls from companies at morning and evening.

6.3. **Calculating features**

In estimating the attributes, the model includes a variety of factors that affect call volume and can be used to predict hourly counts

6.3.1. **LSTM cell has three main gates**

Forget gate:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

Input gate:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

Output gate:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$

Candidate memory cell:

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

Update memory cell:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \quad (6)$$

Compute hidden state:

$$h_t = O_t \cdot \tanh(C_t) \quad (7)$$

6.3.2. **Extension**

BLSTM processes data in both forward and backward directions:

Forward pass is \vec{h}_t and backward pass is \overleftarrow{h}_t .

Final output at time t:

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \quad (8)$$

(where \oplus is concatenation). with these computed features, the Bidirectional Long Short-Term Memory (BLSTM) model forecasts hourly call volumes. BLSTM is a type of recurrent neural network (RNN) that processes sequences bidirectionally, capturing dependencies in both forward and backward directions through time. This capability allows the model to learn from historical data patterns and make predictions based on the temporal relationships between the extracted features and call volumes. By leveraging these features, the BLSTM model enhances the accuracy of its predictions, enabling businesses to optimize resource allocation and improve customer service during various temporal and cultural contexts.

6.4. **Training**

6.4.1. **Loss function**

To minimize the prediction error:

$$L = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (9)$$

6.4.2. Algorithm

Gradient-based optimization (e.g., Adam optimizer):

$$\theta_{t+1} = \theta_t - n \cdot \nabla_{\theta} L \quad (10)$$

6.5. Model architecture

In the research we described, the BLSTM (Bidirectional Long Short-Term Memory) model serves as a crucial component of artificial intelligence. Specifically, it is used to predict the number of incoming calls for the upcoming year. The main features and functionalities of the BLSTM model in this project include:

6.5.1. Data processing

BLSTM is adept at processing sequential data, which is essential for analyzing historical call volume patterns over time.

Forward pass:

Processes the sequence from $t = 1$ to T , generating a hidden state at each time step:

$$\vec{h}_t = LSTM_{forward}(x_t, \vec{h}_{t-1}) \quad (11)$$

Backward pass:

Processes the sequence in reverse, from $t = T$ to 1 :

$$h_t^- = LSTM_{backward}(x_t, h_{t-1}^-) \quad (12)$$

Final hidden state:

Combines the forward and backward hidden states at each time step:

$$h_t = \vec{h}_t \oplus h_t^- \quad (13)$$

6.5.2. Capturing long-term dependencies

it can capture dependencies and trends in call volumes that span across longer periods, such as seasonal fluctuations or annual trends.

6.5.3. Processing

unlike traditional LSTM models, BLSTM processes data in both forward and backward directions. This bidirectional capability allows it to learn from past and future contexts simultaneously, enhancing its understanding of temporal dynamics in call patterns.

6.5.4. Feature extraction

by leveraging its multiple layers and memory cells, BLSTM can automatically extract relevant features from historical data, such as time of day, day of week, holidays, and special events, which are crucial for accurate call volume predictions.

6.5.5. Prediction accuracy

the advanced architecture of BLSTM enables it to provide more accurate forecasts compared to simpler models, as it can learn intricate patterns and dependencies present in call center data (Figs. 4 and 5).

6.6. Calculation the number of agents

6.6.1. Predicted call volume

The predicted number of calls for hour h on day d is denoted as $\mathcal{C}_{d,h}$. This value is provided by the BLSTM model.

6.6.2. Average call handling time (AHT)

The standard average call handling time is:

$$AHT = 7 \text{ minutes/call} (\approx 0.1167 \text{ hours / call}) \quad (14)$$

6.6.3. Required agents without considering the queue

To calculate the base number of agents required to handle the predicted call volume $\mathcal{C}_{d,h}$ within one hour, we use:

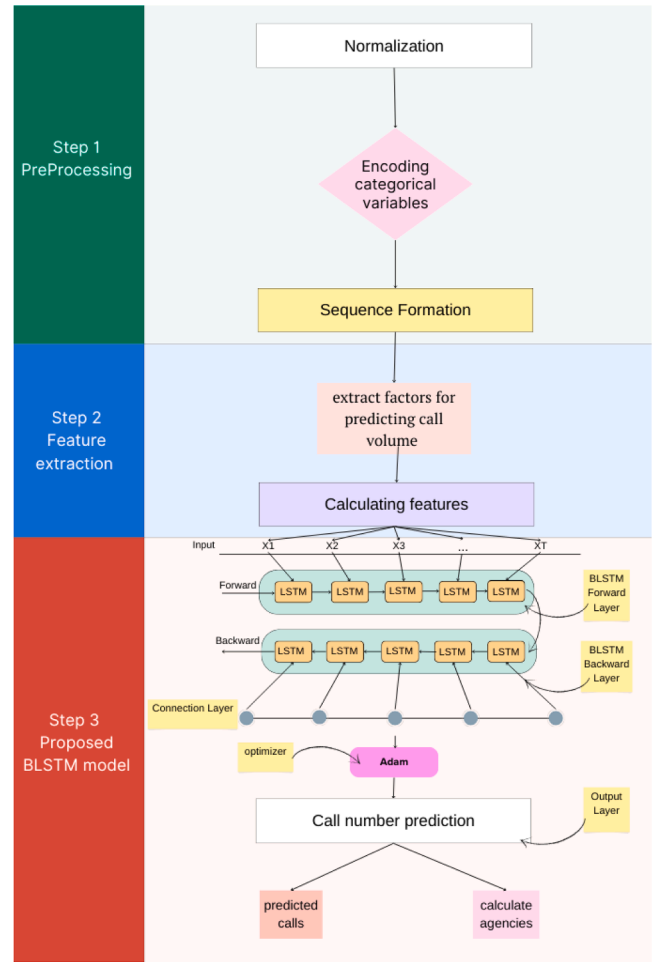


Fig. 4. Inputs and BLSTM architecture.

$$n_{base} = \frac{C_{d,h} \times AHT}{1} \quad (15)$$

6.6.4. Adjusting for the queue size (Q)

The model adjusts the required number of agents based on the size of the queue Q_1 as follows:

Let k_{logoff} be the number of agents allowed to log off based on queue size (Q):

$$k_{logoff} = \{0 \text{ if } Q < 10 \text{ } 1 \text{ if } 10 \leq Q < 20 \text{ } 2 \text{ if } 20 \leq Q < 30\} \quad (16)$$

6.6.5. Adjusted number of agents

The required number of agents is then:

$$n_{required} = n_{base} - k_{logoff} \quad (17)$$

6.7. Ensuring adequate staffing levels

To ensure agents are available to handle predicted call volumes effectively, the number of agents $n_{required}$ must meet or exceed a minimum threshold based on the target service level (SL):

$$n_{final} = \max(n_{required}, n_{min}) \quad (18)$$

Where n_{min} is the minimum number of agents needed to meet the desired service level (e.g., 80 % of calls answered within 30 s).

This method calculates the required number of agents per hour by factoring in the predicted call volumes, the average call handling time, and the probabilities of agents logging off based on current queue

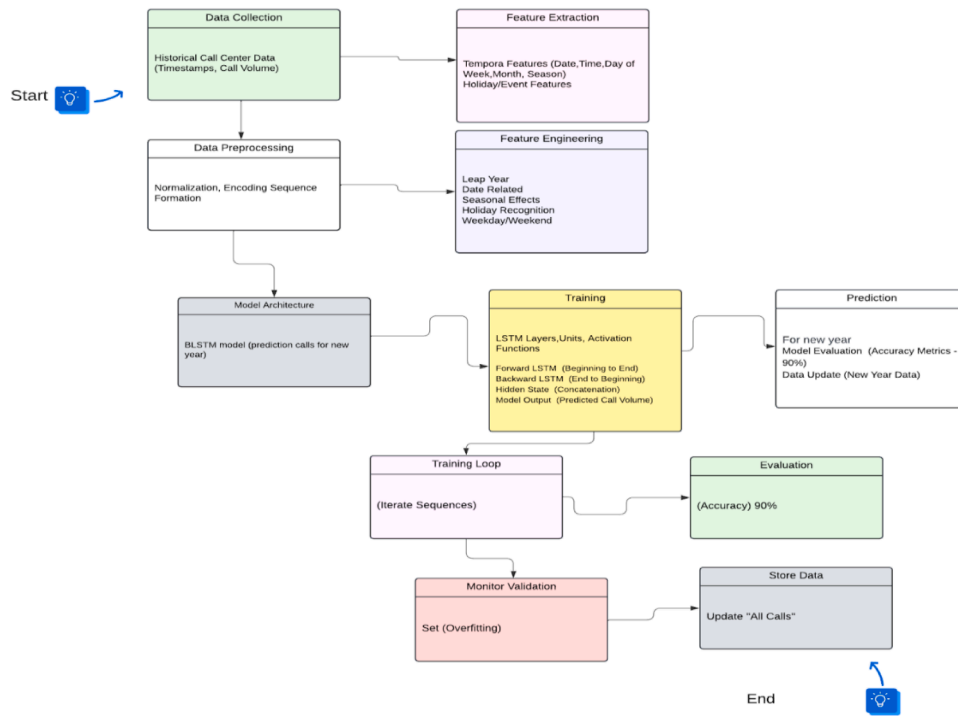


Fig. 5. step by step diagram of full model architecture.

conditions. This calculation is then displayed in Excel to ensure that sufficient agents are available during each time interval to handle incoming calls effectively. The goal is to ensure adequate staffing to respond to incoming calls within acceptable service levels (Table 2).

6.8. Train data using bidirectional long short-term memory (BLSTM)

Define the BLSTM architecture: Choose the number of LSTM layers (stacked layers can capture complex dependencies).

Set the number of units per layer (a balance between model capacity and complexity is needed).

Define activation functions for gates and hidden layers (commonly sigmoid or tanH for gates, tanH or ReLU for hidden layers). Split the preprocessed data into training, validation, and testing sets.

Train the BLSTM model using an optimizer (e.g., Adam) and a loss function (e.g., Mean Squared Error) to minimize the difference between predicted and actual call volumes.

During training, the model iterates through sequences:

At each time step (t) within a sequence: The input feature vector (containing past call volumes and other features) is denoted as $e(t)$.

The forget gate ($f(t)$) determines information to retain from the previous cell's memory state ($c(t-1)$).

The input gate ($i(t)$) controls the information from $e(t)$ to be incorporated into the cell's memory.

The candidate memory cell value ($c'(t)$) is calculated based on $e(t)$ and the previous memory state.

The output gate ($o(t)$) determines the information from the current memory cell ($c(t)$) to be included in the hidden state ($h(t)$).

The hidden state ($h(t)$) at the current time step captures the temporal

dependencies learned from the sequence so far.

A backward LSTM processes the sequence in reverse, generating hidden states ($h^-(t)$) capturing future context.

The final hidden state for each time step (t) is obtained by concatenating the forward and backward hidden states: $h(t) = [h^+(t), h^-(t)]$

The final output of the BLSTM for a sequence is a series of hidden state vectors, $h(1), h(2), \dots, h(T)$ (where T is the sequence length).

This output is then used to predict the call volume for the subsequent time window.

Monitor the model's performance on the validation set to prevent overfitting [31].

6.9. Prediction

The trained model is used to predict the calls count for the new year (e.g., 1403 in the Persian calendar), and the results are stored in an excel file.

This prediction was for a year ago and when it was compared with the test data, it showed an accuracy of 90.56 %, which was a very good prediction accuracy.

6.10. Model evaluation and improvement

At the end of each year, the model's accuracy is evaluated using the actual data from that year (for example, 1403 in the Persian calendar). Additionally, the actual data from the year (1403) is added to the model's memory, stored in an Excel file named "All Calls." This process ensures that the model is continually updated with current data, improving its performance over time by allowing it to adapt to new trends and patterns in the data, leading to more accurate predictions in subsequent years

Table 2

Sample pf model's prediction.

Date	Hour	Predicted calls	Queue	Required agents
17 Aug 2024	08:00	22	3	10
17 Aug 2024	09:00	30	0	18
17 Aug 2024	10:00	45	2	30

7. The results

The current prediction with BLSTM model has achieved an accuracy of 90 %. This accuracy was obtained by comparing its predictions of the number of calls for the year 1402 with the actual data from that year.

Considering the principle in artificial intelligence where an accuracy above 90 % is considered excellent, and noting that the model also incurs errors if it deviates by >10 calls in its predictions, it has performed precise predictions. With continued training on new data, it is expected that the prediction accuracy of this model will improve in future years.

This model earns knowledge by means of the data input and stores it in its own implanted memory. After that we can give the model last year and it will start predicting for current year and it outputs the chances of any year being a leap year and sets the holidays within the dataset equal to zeros and ones (the program decides the features on its own).

Then, it calculates the number of incoming calls per hour for that year, considering that each agent, on average, spends 7 min per call (based on the company's average call duration).

Taking into account the number of incoming calls and the allowable break time for each agent, which is 10 min, as well as the allowance for simultaneous breaks for multiple agents (determined by the number of incoming calls) for example, if there are 10 incoming calls, only 3 agents are allowed to take a break at the same time.

And finally the output will be an excel file with dates, hours, predicted calls, features, Agents.

This approach was useful in the avoidance of call traffic, the satisfaction of the customer, the correct placement of the number of agents and optimistic results in human resources.

In the continuation of this prediction, LSTM and GRU models were also performed in order to compare the accuracy of these two prediction models.

To enhance the practical application of this study, a real-time dashboard was developed to visualize and interact with the BLSTM model's predictions. This dashboard provides a user-friendly interface for call center managers to monitor and analyze predicted call volumes, allowing for dynamic adjustments in workforce allocation and scheduling.

The main features of this dashboard include:

7.1. Real-Time prediction display

The dashboard displays predicted call volumes on an hourly and daily basis through various charts (e.g., line and bar charts). This visual representation helps managers to quickly understand expected call trends and prepare resources accordingly.

7.2. Adjustable settings for prediction parameters

Users can interact with the dashboard to modify prediction parameters, such as considering specific holidays, working hours, or unique events that may impact call volumes. These adjustments allow the model to incorporate real-time data factors for improved accuracy.

7.3. Alert system for high call volumes

The dashboard includes an alert system that notifies managers when predicted call volumes exceed a certain threshold. This feature enables the call center to proactively adjust staffing levels to meet increased demand and minimize customer wait times.

7.4. Analytical reports and comparison

The dashboard provides daily, weekly, and monthly analytical reports, comparing the predicted and actual call volumes. These reports help managers assess the model's accuracy and track call patterns over time, aiding in strategic decision-making.

7.5. Data upload and model update

For continuous improvement, the dashboard allows for new data uploads, updating the model with recent call history to maintain high

prediction accuracy. This function supports ongoing model optimization as the call center adapts to changing conditions.

7.6. The results of call prediction vs real calls in BLSTM, LSTM & GRU models

To further evaluate the significance of our model's performance, we conducted a comparative analysis not only in terms of raw accuracy percentage, but also in terms of practical impact in a real-world scenario.

While the difference in accuracy between BLSTM (90.56 %), LSTM (89.07 %), and GRU (88.33 %) may seem incremental at first glance, this translates to a substantial operational benefit in a call center context. For instance, considering a call center that handles 10,000 calls per day, a 1.5 % improvement means approximately 150 additional calls per day are correctly predicted, allowing for more efficient staffing and reducing missed or delayed calls significantly.

Moreover, we deployed each of the three models (BLSTM, LSTM, and GRU) in test simulations using real Shatel call center historical data. The simulations tracked actual call queues versus predicted agent allocations:

GRU and LSTM-based predictions resulted in 3–5 queue formation events per week during high-volume hours, requiring reactive staff reallocation.

In contrast, BLSTM-based planning led to zero queue events during the same period, showing better alignment between predicted volume and staffing needs.

These results emphasize that even a marginal improvement in model accuracy can lead to tangible operational enhancements, especially in large-scale environments like call centers (Figs. 6-9).

8. Comparing incoming calls before and after prediction

This graph shows the number of incoming calls to the call center over a 30-day period (horizontal axis). There are two lines in the graph:

Red graph (Before AI Prediction):

This line shows the number of incoming calls before the AI model was used. In this situation, the number of calls is relatively high, reaching >120 calls on some days. This high value can lead to congestion and reduced service quality.

Green graph (After AI Prediction):

This line shows the number of calls after the AI model predicted them. By predicting call patterns and optimizing workforce management, the AI model has led to a reduction in the number of incoming calls. The reduction in calls can be due to proactive measures such as automated answering, better queue management, or solving common problems before the customer calls (Fig. 10).

9. Conclusion

Based on the results obtained from the GRU, LSTM, and BLSTM models, the BLSTM model demonstrated higher accuracy.

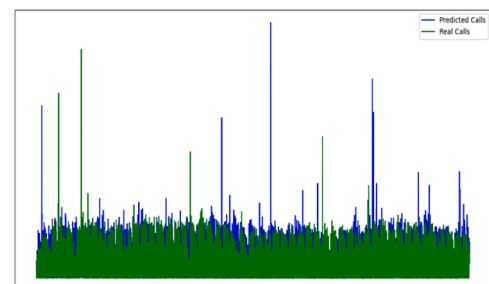


Fig. 6. BLSTM model with 90.56 % accuracy.

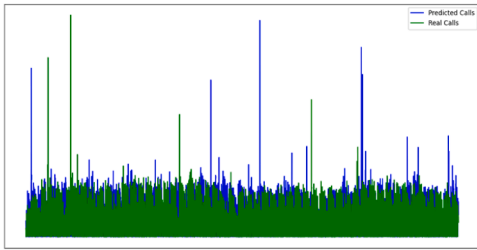


Fig. 7. LSTM model with 89.07 % accuracy.

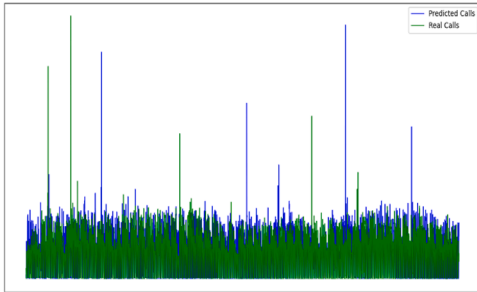


Fig. 8. GRU model with 88.33 % accuracy.

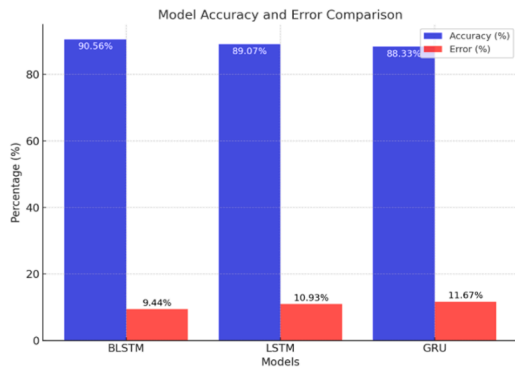


Fig. 9. Chart of model's accuracy and error comparison.

The Bidirectional Long Short-Term Memory (BLSTM) model outperformed the other models due to its ability to capture dependencies in both forward and backward directions. Unlike standard LSTM, which

only processes data in one direction, BLSTM processes the input sequence from the beginning to the end and from the end to the beginning. This bidirectional approach allows the BLSTM model to have a more comprehensive understanding of the sequential data, capturing more context and relationships within the data.

The improved accuracy of the BLSTM model can be attributed to:

Enhanced Contextual Understanding: By considering both past and future contexts, the BLSTM can capture intricate patterns and dependencies in the data that might be missed by unidirectional models. **Better handling of long-term dependencies:** BLSTM's dual-layer processing makes it more effective at learning long-term dependencies compared to GRU and unidirectional LSTM.

Increased Robustness: The ability to analyze data from two directions provides a more robust feature extraction, leading to better performance on complex tasks.

As a result, the BLSTM model showed superior performance, making it more suitable for tasks requiring high accuracy in understanding and predicting time series data, such as forecasting incoming calls in our scenario.

In this study, the BLSTM model achieved a prediction accuracy of 90.56 %, outperforming other models such as LSTM and GRU. A similar study by Kumwilaisak et al. [2] utilized an LSTM model to forecast call volumes in a call center, achieving an accuracy of approximately 89.07. Compared to our results, the BLSTM model demonstrated superior performance due to its bidirectional processing capability, which allows it to capture dependencies in both forward and backward directions, enhancing its accuracy in complex, time-dependent data.

In the feature extraction phase, this study incorporated factors such as official holidays, religious events, hours, and days of the week. Comparison with Kumwilaisak et al. [2] reveals that our inclusion of these temporal and cultural features significantly improved prediction accuracy. While the study by Kumwilaisak et al. considered similar temporal variables, it utilized a unidirectional LSTM model, which limits its ability to capture bidirectional dependencies.

Finally, our research successfully improved the operational efficiency of Shatel's call center, enhancing customer satisfaction through the use of the BLSTM model. Similar outcomes were observed in the study by Kumwilaisak et al. [2], where deep learning was used to optimize call center operations. However, the higher accuracy achieved by our BLSTM model indicates its effectiveness for managing complex, variable data scenarios in real-world applications.

10. Future works

In future research, we aim to enhance the accuracy and performance

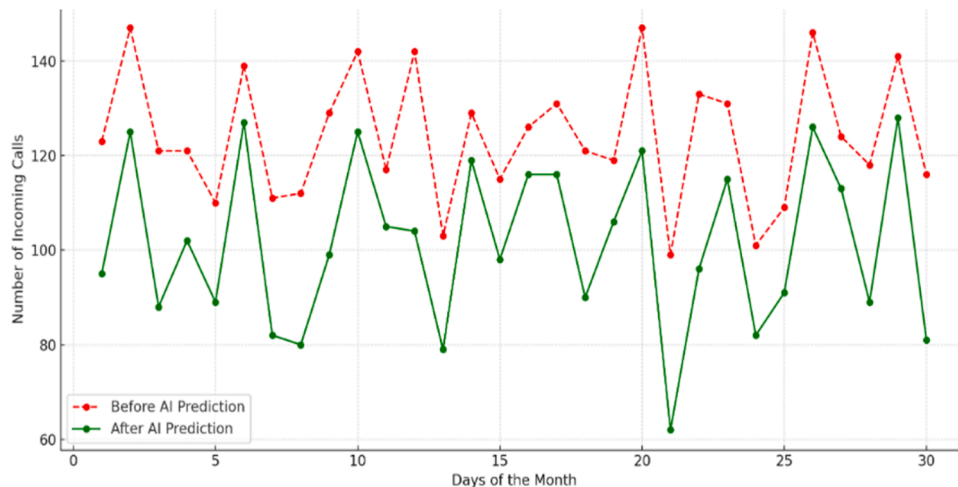


Fig. 10. The number of incoming calls of call center over a 30-day period.

of our predictive model by incorporating more advanced deep learning architectures, particularly transformers, which have demonstrated superior capabilities in sequential data modeling.

Additionally, we plan to explore different optimization algorithms beyond the ones currently used. Optimizers play a crucial role in training deep learning models, and experimenting with alternatives may lead to better convergence and generalization.

Another important aspect of our future work is the incorporation of additional domain-specific features to enrich the input data. While our current model already considers key factors such as holidays and seasonal trends, we intend to integrate customer behavioral patterns, network traffic data, sentiment analysis from support interactions, and real-time external factors that may influence call volume. By leveraging a more comprehensive feature set, we expect the model to capture nuanced variations in call center demand more effectively.

These advancements will contribute to a more accurate, scalable, and real-world-applicable forecasting system, ultimately leading to better resource allocation and customer experience optimization in call centers.

CRedit authorship contribution statement

Armaghan Nikfar: Writing – original draft, Software. **Javad Mohammadzadeh:** Supervision.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank Shatel team from Iran Tehran who's supporting made this research possible.

References

- [1] H. Al-Selwi, F. Alharbi, M.A. Azam, M.A. Alharbi, M.M. Hassan, A cybersecurity framework for DDoS mitigation in IoT and cloud computing using RPL and deep learning, *Electron. (Basel)* 13 (4) (2024) 818, <https://doi.org/10.3390/electronics13040818>.
- [2] W. Kumwilaisak, S. Phikulngoen, J. Piriyataravet, N. Thatphithakkul, C. Hansakunbuntheung, Adaptive call center workforce management with deep neural network and reinforcement learning, *IEEE Access* 10 (2022) 35712–35724, <https://doi.org/10.1109/access.2022.3160452>.
- [3] Z. Aksin, et al., The modern call center: a multi-disciplinary perspective on operations management research, *Prod. Oper. Manag.* 16 (6) (2007) 665–688.
- [4] R. Ibrahim, H. Ye, P. L'Ecuyer, H. Shen, Modeling and forecasting call center arrivals: a literature survey and a case study, *Int. J. Forecast.* 32 (3) (2016) 865–874.
- [5] L. Cheng, et al., Multi-modal fusion for business process prediction in call center scenarios, *Inf. Fusion* 108 (2024) 102362.
- [6] S. Lukitasari, F. Hidayat, Deep learning-based complaint classification for Indonesia telecommunication company's call center, in: *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, Bandung, West Java, Indonesia, 2020, 12 October 2019*.
- [7] A. Ando, et al., Hierarchical LSTMs With Joint Learning For Estimating Customer Satisfaction from Contact Center Calls, *INTERSPEECH*, 2017.
- [8] D. Leszko, PhD thesis, 2020.
- [9] Sepp Hochreiter, *Untersuchungen Zu Dynamischen Neuronalen Netzen* (PDF) (diploma thesis), Technical University Munich, Institute of Computer Science, 1991.
- [10] M. Phridviraj, et al., A bi-directional long short-term memory-based diabetic retinopathy detection model using retinal fundus images, *Healthc. Anal.* 3 (2023) 100174.
- [11] J. Bi, et al., A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM, *IEEE Trans. Autom. Sci. Eng.* 19 (3) (2021) 1869–1879.
- [12] A. Ahmed, et al., Agent productivity modeling in a call center domain using attentive convolutional neural networks, *Sensors* 20 (19) (2020) 5489.
- [13] N. Zafar, et al., Applying hybrid LSTM-GRU model based on heterogeneous data sources for traffic speed prediction in urban areas, *Sensors* 22 (9) (2022) 3348.
- [14] D.M. Nelson, et al., Stock market's price movement prediction with LSTM neural networks, in: *2017 International joint conference on neural networks (IJCNN)*, Ieee, 2017.

- [15] D.-H. Shin, et al., Prediction of traffic congestion based on LSTM through correction of missing temporal and spatial data, *IEEE Access* 8 (2020) 150784–150796.
- [16] C.C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*, Springer, 2018, <https://doi.org/10.1007/978-3-319-94463-0>.
- [17] W.A. Aziz, et al., Content-aware network traffic prediction framework for quality of service-aware dynamic network resource management, *IEEE Access* 11 (2023) 99716–99733.
- [18] R.L. Abduljabbar, et al., Unidirectional and bidirectional LSTM models for short-term traffic prediction, *J. Adv. Transp.* 2021 (1) (2021) 5589075.
- [19] A. Lawal, et al., Wind speed prediction using hybrid 1D CNN and BLSTM network, *IEEE Access* 9 (2021) 156672–156679.
- [20] H.E. Alfari, Sentiment analysis for arabic call center notes using machine learning techniques: a case study of Jordanian dialect, *Tech. Bus. Manag.* 1 (1) (2022).
- [21] R. Dey, F.M. Salem, Gate-variants of gated recurrent unit (GRU) neural networks, in: *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 2017.
- [22] G. Shen, et al., Deep learning with gated recurrent unit networks for financial sequence predictions, *Procedia Comput Sci* 131 (2018) 895–903.
- [23] R. Zhao, et al., Machine health monitoring using local feature-based gated recurrent unit networks, *IEEE Trans. Ind. Electron.* 65 (2) (2017) 1539–1548.
- [24] B. Hussain, et al., Intelligent traffic flow prediction using optimized GRU model, *IEEE Access* 9 (2021) 100736–100746.
- [25] W. Li, et al., Multi-GRU prediction system for electricity generation's planning and operation, *IET Gener. Transm. Distrib.* 13 (9) (2019) 1630–1637.
- [26] Y. Gao, et al., Stock prediction based on optimized LSTM and GRU models, *Sci. Program.* 2021 (1) (2021) 4055281.
- [27] R. Ibrahim, et al., Modeling and forecasting call center arrivals: a literature survey and a case study, *Int. J. Forecast.* 32 (3) (2016) 865–874.
- [28] H. Chacón, V. Koppiseti, D. Hardage, K.K.R. Choo, P. Rad, Forecasting call center arrivals using temporal memory networks and gradient boosting algorithm, *Expert Syst. Appl.* 224 (2023) 119983.
- [29] R.A. Mohammed, P. Pang, Agent personalized call center traffic prediction and call distribution, in: *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, 2011, pp. 1–10. November 13–17, 2011, Proceedings, Part II* Springer Berlin Heidelberg.
- [30] P. Bugarić, S. Janković, S. Mladenović, Forecasting number of calls to the call center using machine learning, *Transp. Today's Soc.* (2021).
- [31] D. Uliyan, A.S. Aljaloud, A. Alkhalil, H.S.A. Amer, M.a.E.A. Mohamed, A.F. M. Alogali, Deep learning model to predict student's retention using BLSTM and CRF, *IEEE Access* 9 (2021) 135550–135558, <https://doi.org/10.1109/access.2021.3117117>.
- [32] S.M. Al-Selwi, M.F. Hassan, S.J. Abdulkadir, A. Muneer, E.H. Sumiea, A. Alqushaibi, M.G. Ragab, RNN-LSTM: from applications to modeling techniques and beyond—Systematic review, *J. King Saud. Univ. - Comput. Inf. Sci.* (2024) 102068, <https://doi.org/10.1016/j.jksuci.2024.102068>.
- [33] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Netw.* 18 (5–6) (2005) 602–610, <https://doi.org/10.1016/j.neunet.2005.06.042>.
- [34] S. Hochreiter, J. Schmidhuber, Long Short-Term memory, *Neural Comput.* 9 (8) (1997) 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [35] R. Huan, Z. Zhan, L. Ge, K. Chi, P. Chen, R. Liang, A hybrid CNN and BLSTM network for human complex activity recognition with multi-feature fusion, *Multimed. Tools Appl.* 80 (30) (2021) 36159–36182, <https://doi.org/10.1007/s11042-021-11363-4>.
- [36] Javad Mohammadzadeh, Abadi, Ali Fallahi Rahmat (2023); "Leveraging deep learning techniques on collaborative filtering recommender systems". [doi:10.22105/jarie.2021.275620.126](https://doi.org/10.22105/jarie.2021.275620.126).
- [37] Z.C. Lipton, J. Berkowitz, C. Elkan, A Critical Review of Recurrent Neural Networks for Sequence Learning, Cornell University, 2015 arXiv, [doi:10.48550/arxiv.1506.00019](https://doi.org/10.48550/arxiv.1506.00019).
- [38] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, *Electron. Mark.* 31 (3) (2021) 685–695, <https://doi.org/10.1007/s12525-021-00475-2>.
- [39] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation. arXiv (Cornell University), 2014. [doi:10.48550/arxiv.1406.1078](https://arxiv.org/abs/1406.1078).
- [40] Felix Gers, Jürgen Schmidhuber, Fred Cummins, Learning to forget: continual prediction with LSTM, 9th International Conference on Artificial Neural Networks: ICANN '99, 1999, 1999, pp. 850–855 ISBN 0-85296-721-7, [doi:10.1049/cp:19991218](https://doi.org/10.1049/cp:19991218).
- [41] "Recurrent Neural Network Tutorial, Part 4 – implementing a GRU/LSTM RNN with python and theano – WildML". [wildml.com](https://www.wildml.com/2015-10-27/Recurrent-Neural-Network-Tutorial-Part-4-implementing-a-GRU-LSTM-RNN-with-python-and-theano-WildML.html). 2015-10-27. Archived from the original on 2021-11-10. Retrieved May 18, 2016.
- [42] M. Ravanelli, P. Brakel, M. Omologo, Y. Bengio, Light gated recurrent units for speech recognition, *IEEE Trans. Emerg. Top. Comput. Intell.* 2 (2) (2018) 92–102, <https://doi.org/10.1109/tetci.2017.2762739>.
- [43] Y. Su, C.J. Kuo, On extended long short-term memory and dependent bidirectional recurrent neural network, *Neurocomputing* 356 (2019) 151–161, <https://doi.org/10.1016/j.neucom.2019.04.044>.

- [44] Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.1412.3555>.
- [45] N. Gruber, A. Jockisch, Are GRU cells more specific and LSTM cells more sensitive in motive classification of text? *Front. Artif. Intell.* 3 (2020) <https://doi.org/10.3389/frai.2020.00040>.
- [46] M.E. Manaa, S.M. Hussain, N.S.A. Alasadi, N.H.a.A. Al-Khamees, DDoS attacks detection based on machine learning algorithms in IoT environments, *Intel. Artif.* 27 (74) (2024) 152–165, [doi:10.4114/intartif.vol_27iss74pp152-165](https://doi.org/10.4114/intartif.vol_27iss74pp152-165).
- [47] W. Kumwilaisak, S. Phikulgoen, J. Piriataravet, N. Thatphithakkul, C. Hansakunbuntheung, Adaptive call center workforce management with deep neural network and reinforcement learning, *IEEE Access* 10 (2022) 35712–35724, <https://doi.org/10.1109/ACCESS.2022.3160452>.
- [48] M.S.B. Phridviraj, R. Bhukya, S. Madugula, A. Manjula, S. Vodithala, M.S. Waseem, A bi-directional long short-term memory-based diabetic retinopathy detection model using retinal fundus images, *Comput. Biol. Med.* 146 (2020) 2022 105580, [doi:10.1016/j.compbiomed.2022.105580](https://doi.org/10.1016/j.compbiomed.2022.105580).
- [49] W. Kumwilaisak, et al., Adaptive call center workforce management with deep neural network and reinforcement learning, *IEEE Access* 10 (2022) 35712–35724.
- [50] P. Kaushik, S. Singh, P. Yadav, Traffic prediction in telecom systems using deep learning, 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future, *IEEE*, 2018, pp. 302–307.



Review Article

Regulatory landscape of blockchain assets: Analyzing the drivers of NFT and cryptocurrency regulation

Junaid Rahman^{a,*}, Hafizur Rahman^b, Naimul Islam^d, Tipon Tanchangya^a,
Mohammad Ridwan^c, Mostafa Ali^a

^a Department of Finance, University of Chittagong, Chittagong 4331, Bangladesh

^b Department of EEE, American International University-Bangladesh (AIUB), Dhaka, Bangladesh

^c Department of Economics, Noakhali Science and Technology University, Noakhali 3814, Bangladesh

^d Department of Accounting, Finance and Economics, University of Greenwich, London SE10 9LS, UK



ARTICLE INFO

Keywords:

Blockchain

Blockchain assets

Regulation of blockchain assets

Regulate NFTs

Regulate cryptocurrencies

ABSTRACT

The study analyzes the global regulatory landscape for blockchain assets, particularly cryptocurrencies and non-fungible tokens, focusing on the motivations behind policymaker actions, the diversity of regulatory approaches, the challenges posed by decentralized technologies and provide future regulatory pathways. The study uses a conceptual and mixed-method approach, combining qualitative and quantitative content analysis of 59 peer-reviewed articles selected through the PRISMA framework. Findings reveal that regulation is primarily driven by concerns over consumer protection, financial stability, anti-money laundering, taxation, and environmental sustainability. Regulatory responses vary widely, ranging from the harmonized MiCA framework in the EU to the fragmented enforcement model in the U.S., along with diverse strategies across Asia. Stablecoins, DeFi, and CBDCs emerge as major regulatory frontiers. The study recommends adopting regulatory sandboxes, promoting international coordination, enforcing environmental standards, and building regulatory capacity in emerging economies to balance innovation with risk mitigation. It also highlights the importance of industry self-regulation and technology-assisted compliance in decentralized finance. The limitation of this study is that it relies solely on secondary data sources, which may limit the accuracy of real-time policy impact assessments. Future research should focus on empirical validation and dynamic policy modeling to enhance global governance of digital assets.

1. Introduction

Over the last decade, blockchain technology has emerged as one of the most revolutionary innovations in the digital age, revolutionizing a variety of industries, including supply chain management, financial services, healthcare, and entertainment. Blockchain (BC) based assets, particularly cryptocurrencies and NFTs are at the forefront of this upheaval [27]. Bitcoin (BCT) and Ethereum (Eth), in particular, are among the top cryptocurrencies that have made it possible for individuals to send and receive money across borders with relative ease, without the need of financial institutions. In turn, ownership in the digital world was disfigured by NFTs, turning them into distinct digital products akin to artwork, songs or virtual products [48]. Yet the surging popularity of

these new asset classes is not without its critics. Promoters of BC assets are excited by the prospect of these instruments democratizing finance and empowering investors, while critics are concerns about their deep ties to criminality [4]. The inherent volatility of digital currencies, scams and frauds on the rise in the NFT market, and environmental challenges with BC mining are viable reasons for discussions about regulations. In addition, the decentralized and borderless characteristics of these technologies present distinct issues for conventional regulatory paradigms that are generally rooted in a nation-state milieu directed towards centralized actors [39].

An increase in the issuance of BC assets prompts policymakers and regulators around the globe to address these challenges while still encouraging innovation. Hands of regulators interest in regulating has

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

* Corresponding author.

E-mail addresses: junaid.rahman.edu@gmail.com (J. Rahman), hafiz162891@gmail.com (H. Rahman), ni2355y@greenwich.ac.uk (N. Islam), tipon.tcg.edu@gmail.com (T. Tanchangya), m.ridwan.econ@gmail.com (M. Ridwan), mostafacu2000@gmail.com (M. Ali).

<https://doi.org/10.1016/j.tbench.2025.100214>

Received 19 February 2025; Received in revised form 30 April 2025; Accepted 14 May 2025

Available online 28 May 2025

2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

been sparked by the need to protect consumers, uphold financial stability, mitigate threats like money laundering and tax evasion, and ensure fair market practices [54]. It is acknowledged, however, that excessively onerous laws may have the unintended consequence of impeding innovation and pushing BC companies to other, more lenient jurisdictions, a phenomenon known as regulatory arbitrage [1]. Many governments and regulatory bodies around the world have taken different approaches to how cryptocurrencies and NFTs should be regulated, in light of these concerns. Some countries, like China, have taken a very extreme stance in prohibiting the practice of mining and selling cryptocurrency because it is thought to have negative effects on the environment or financial profits [60]. On the other hand, Switzerland, for instance, has a law on the use and trade of cryptocurrency with a proviso. And some, such as the US and the EU, are attempting to create new but more complex regulatory frameworks that balance innovation with rules for the sake of protecting consumers. In this regard, the U.S. Securities and Exchange Commission (SEC) is investigating whether some digital assets are considered securities, while Europe is searching for assets with this kind of growth through its Markets on Cryptocurrency Assets (MiCA) legislation [3].

However, attempting to regulate BC assets has proved difficult. One of the main issues with BC networks is their decentralized structure, which prevents them from functioning from a central authority or control point. Because transactions can occur without a legal authority that can enforce traditional laws, this presents difficulties for the implementation of domestic law. Additionally, as new technologies arise, the BC sector tends to change swiftly and authorities typically lag behind, which distract both consumers and companies [6]. Another major obstacle is that, by nature, cryptocurrencies and NFTs lack borders. Even if national governments are capable of regulating BC assets in their own jurisdictions, they fail to control where individuals and businesses operate, as BC assets are global by nature. Additionally, it may expose companies to regulatory arbitrage, in which they choose areas with less stringent regulation, undermining a degree-taking discipline strategy [10].

The rapid growth of BC assets, particularly cryptocurrencies and NFTs, presents both unprecedented opportunities and complex regulatory challenges. In developing regions, these technologies hold the promise of advancing financial inclusion by offering individuals access to decentralized financial systems, effectively functioning as digital banks [13]. Additionally, BC has empowered artists and content creators to monetize their work in innovative ways, overcoming long-standing barriers such as copyright constraints [13]. Despite these advancements, the unregulated or under-regulated nature of BC assets has raised concerns about consumer protection, illicit financial activity, environmental impact, and systemic financial risk. While some regulatory bodies have adopted proactive approaches, others remain cautious, opting for a “wait and see” strategy. Meanwhile, the industry is increasingly leaning toward self-regulation, with many platforms voluntarily adopting AML and KYC measures to align with traditional legal standards [12].

Hence, given this dynamic and fragmented landscape, there is a critical need to investigate the motivations behind regulatory efforts, evaluate current frameworks, and propose balanced strategies that safeguard public interests without stifling innovation. This study seeks to address these gaps and contribute to the development of more coherent, adaptive, and forward-looking regulatory policies. The primary objective of this study is to understand the regulatory approaches of different countries and discuss some challenges in regulating decentralized technologies and potential pathways for the future regulation of BC assets. Therefore, this study will (1) provide an overview of BC assets and identify the key drivers behind regulatory efforts, including consumer protection, financial stability, AML or KYC compliance, taxation, and environmental concerns; (2) examine the diversity of regulatory frameworks across major jurisdictions such as the European Union, United States, China, Singapore, and others, and assess their

effectiveness and implications; (3) conduct a quantitative analysis of the regulatory impact on BC assets; (4) evaluate the core challenges associated with regulating BC assets; (5) explore potential pathways for the future regulation of BC assets; and (6) propose balanced policy recommendations that align regulatory safeguards with the need to support innovation and adaptability in the evolving digital asset ecosystem.

This study will employ a conceptual and mixed-method approach of qualitative and quantitative guided by the PRISMA framework to contribute to the growing body of knowledge on BC governance. It will offer a comprehensive and multidimensional understanding of how and why BC assets are regulated across global jurisdictions. The research will clarify the complex interplay between innovation and regulation by identifying key policy drivers such as consumer protection, financial stability, AML or KYC enforcement, taxation, and environmental sustainability. Additionally, the study will provide insights through quantitative analysis of regulatory impacts, enabling an evidence-based evaluation of policy outcomes. By examining regulatory diversity and associated challenges, the research will advance the discourse on cross-border regulatory coherence and expose gaps in existing frameworks. Furthermore, the study will propose forward-looking and balanced policy recommendations that align regulatory safeguards with innovation needs, thereby offering practical guidance for policymakers, regulators, researchers, and industry stakeholders navigating the evolving digital asset landscape.

The remainder of this article is organized as follows. The second section presents the methodology of the study. The third section discusses the definition and evolution of BC assets. The fourth section outlines the rationale for regulatory oversight. The fifth section analyzes global regulatory approaches to BC assets and their evaluation. The sixth section identifies the challenges in regulating BC assets. The seventh section explores the future of regulatory frameworks for BC assets. The eighth section addresses implementation, applicability, policy formulation, and validation analysis. Finally, the ninth section provides the conclusion and policy recommendations.

2. Methodology

The study adopted a conceptual and mixed-method approach using quantitative and qualitative content analysis of peer-reviewed articles, international reports, and legal frameworks to describe the regulatory landscape for BC based assets. A variety of research papers, publications, and journals that have addressed regulatory landscape and challenges of BC based assets have been assessed for secondary study. The sources will be utilized to support the ideas and analysis that will shed light on the specifics. Though there will be a gap in original data, this strategy makes use of earlier sources, which may be taken into account for future study.

2.1. PRISMA framework

The proposed review investigates the regulatory landscape of BC-based assets using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, ensuring a structured and transparent methodology for the selection process of relevant publications. The primary objective of this review is to evaluate the regulatory approaches, compliance standards, and jurisdictional challenges associated with BC assets across different regions. The PRISMA model was employed for the systematic selection of relevant literature, as illustrated in Fig. 1. A total of 304 articles were initially retrieved through searches in major academic databases including Scopus, Web of Science, Google Scholar, and other sources using specific search terms such as BC assets, NFTs, cryptocurrency, BC regulation, regulatory approach of BC assets, and digital asset compliance.

The Fig. 1 illustrates that 304 articles were identified for preliminary consideration, covering publications from 2015 to 2024. During the screening stage, 109 articles were excluded 21 due to errata and 88 for being outside the scope of the study, resulting in 195 articles for further

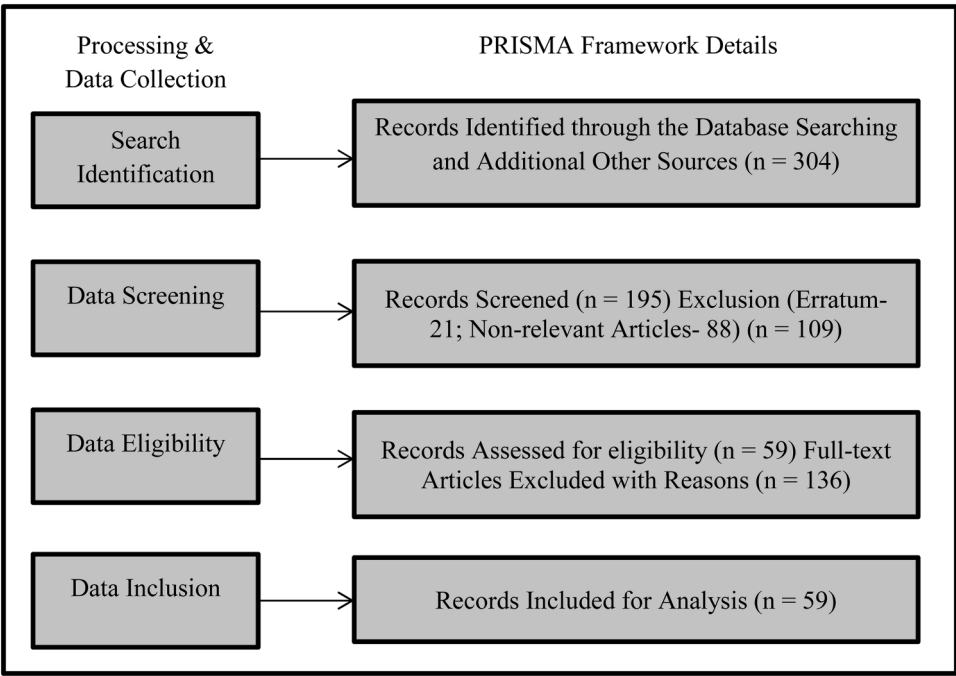


Fig. 1. PRISMA framework for collecting information.

evaluation. A full-text eligibility assessment was then conducted, during which 136 articles were excluded for not meeting the relevance criteria. This process led to the final selection of 59 eligible studies. These studies were chosen from peer-reviewed journals and were selected based on criteria such as emphasis on BC assets including cryptocurrencies and NFTs, clarity in model deployment, feasibility analysis of regulatory challenges, and accuracy and transparency in reporting. The final 59 publications formed the foundation of a detailed analysis that focused on legal frameworks, regulatory standards, and implementation associated with BC asset management, as discussed in the subsequent sections.

3. BC assets: definition and evolution

3.1. Definition of BC technology

The BC is a distributed, decentralized ledger system that enables multiple parties to maintain a common record of information without the need for a central authority. The term "BC" refers to the method of organizing data into cryptographically linked blocks, or chains. The foundation is a series of liquid blocks, each containing a list of transactions; after immutability (after chaining), proof-of-work, security, and transparency are offered [47]. Decentralization is the chief legacy of BC. While legacy databases are maintained and controlled by a centralized authority, BC networks operate on a distributed system where transactions are verified and logged by several computers called nodes. This decentralized architecture removes the need for intermediaries such as banks or clearinghouses driving the transition to trustless systems which allows for engagement between parties without any form of prior trust [62]. BCs provide the underlying technology for a number of digital assets, such as cryptocurrencies and NFTs. While these assets have been growing in prominence and various countries have made changes to such markets, the past few years have seen some major shifts in how they are operated across the world landscapes [23].

3.2. Cryptocurrencies: definition and types

Cryptocurrencies are digital or virtual currencies secured using cryptography. BCT is arguably the most notorious, but other

cryptocurrencies including Eth, Ripple (XRP), and Litecoin have all surpassed the seven-figure milestone. Most cryptocurrencies are based on decentralized networks utilizing a technology called the BC, where transactions are confirmed by a network of nodes through mining or staking (Fig. 2, Table 1).

3.3. Non-fungible tokens (NFTs): definition and applications

NFTs represent a distinct category of BC assets. While cryptocurrencies are fungible, meaning that each unit can be swapped out for another of the same size and kind, NFTs are less compatible since each

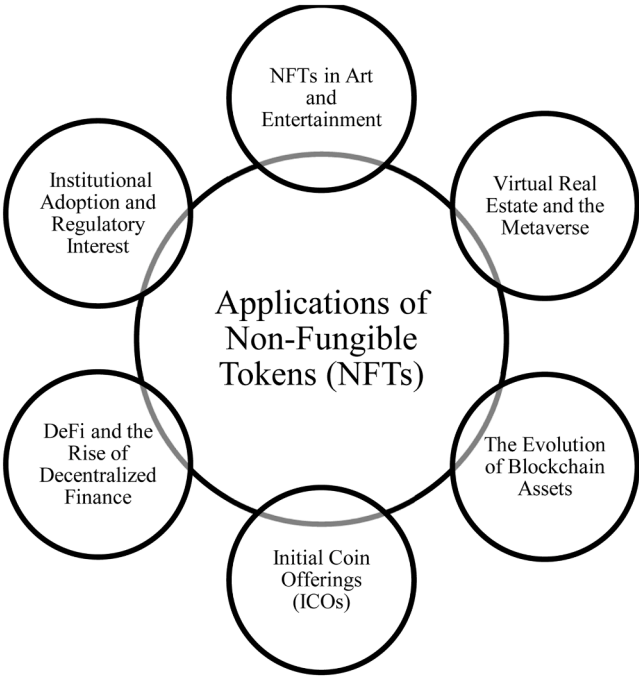


Fig. 2. Applications of non-fungible tokens (NFTs).

Table 1
Types of cryptocurrencies.

Cryptocurrencies	Concepts
Bitcoin	Cryptocurrency website Description BCT from 2008, was the first decentralized cryptocurrency, because of this it is also known as origin cryptocurrency. It was created as a peer to peer digital cash system that would allow online payments to be sent directly from one party to another without going through a financial institution. Mining with its Proof-of-Work (PoW) consensus mechanism in which miners solve complex cryptographic puzzles to validate transactions and add new blocks to the chain is a fundamental component of BCT [53]. Because of this limited supply, capped at 21 million coins (never to be passed) and its decentralized nature, BCT has become known as the “digital gold”, or a store of value. BCT is the standard for digital gold. Institutional Acceptance and Business Treasury BCT eventually gained recognition from firms and institutional investors on the world financial scene [20].
Ethereum	Eth is a BC-based platform that allows people to create and run decentralized applications (dApps). Unlike BCT, which is mainly for sending and receiving money, Eth is like a whole world where you can build programs that run exactly as coded, without any chance of fraud or interference [8]. By facilitating peer-to-peer lending, borrowing, and trading without the need for middlemen, Eth has accelerated the development of decentralized finance (DeFi). Ether (ETH), a native token, manages the Ether power needed for transactions on the Eth network [43].
Ripple (XRP)	It is a cryptocurrency that established to permit international payments that are both more economical and faster. Simply it is designed to make cross-border payments faster and cheaper [9].
Litecoin (LTC)	It is a cryptocurrency that attempts to increase the efficiency and speed of transactions. It is a smaller and more lightweight alternative to BCT. A faster, lighter spin on BCT that targets transaction speed and efficiency [32].
Stablecoins	A type of cryptocurrency that hedges on the like of Tether (USDT) or USD Coin (USDC), which seeks to be pegged either by or in contrast to the value of a sovereign lawsuit currency cosmetics the U.S. dollar, and provides approximately much more shadow help during together with volatility from the issue as a replacement for good currencies [63].

one has a distinct digital signature. Since each NFTs is (typically) unique, each token has intrinsic value linked to its attributes, making it ideal for representing a unique online or tangible product [4].

- **NFTs in art and entertainment:** NFTs have benefited the art industry in particular since they provide artists another method to add value to their work and sell it directly to collectors after becoming a little aspect of our digital lives. Collectibles of NFTs Beeple’s ‘First 5000 Days’ sale brought in \$69 million. Nine million dollars The most costly NFT ever sold is Beeple’s creation. According to Mazur and Polyzos [40,41], NFT art pieces have recently gained more publicity than ever before, garnering more external attention than ever before. In addition to digital art, NFTs are being picked up by other forms of entertainment. For instance, musicians can tokenize records or songs that feature exclusive rights such as limited-edition releases to fans. NFTs have also surfaced in the gaming space, letting gamers buy and sell things like skins, weapons or virtual land on BC-powered marketplaces [59].
- **Virtual real estate and the metaverse:** Virtual real estate in the metaverse is one of the most exciting developments happening within NFTs, virtual worlds where users can purchase sell and trade digital land. The platforms are complete immersive worlds with NFT land ownership, and users can build on this land, as well as interact with other live avatars. As the metaverse blurs the lines between digital and physical, it presents increasing opportunities to monetize content, establish brand engagement and foster complex communities, increasingly positioning it as a next frontier of NFT applications [29].

- **The evolution of BC assets:** BC already looks very different from when BCT was first released. This will be in stark contrast to the early days of the tech-driven currency, as what was once an exclusive eye on cryptocurrencies now nurtures a multifaceted market of NFTs, DeFi and DAOs. Contributing to the transformation is the development in BC technology and implementation by consumers as well as businesses, along with widespread institutional involvement [59].
- **Initial coin offerings (ICOs):** Initial Coin Offerings (ICOs) were used by many BC startups to secure funding back in the beginning of crypto. ICOs enabled projects to issue funds in return for their own tokens, often BCT or Eth of sorts. However, without regulatory oversight, the industry has been hit by an epidemic of scams around new initial coin offerings (ICOs) leading to U.S. Securities and Exchange Commission (SEC) stepping in saying some ICOs are nothing more than unregistered securities [41].
- **DeFi and the rise of decentralized finance:** (DeFi) is the next big step in the discharge of BC assets. Again this is a paraphrase, it means that DeFi platform on the left offers functions such as lending, borrowing, trading and insurance in ways similar to traditional financial services (and using the same assets) but without intermediaries like banks or brokers. Powered by smart contracts on BC networks such as Eth, these services enable more transparent, accessible and efficient means to handle and track financial transactions [21]. DeFi had seen an explosion in popularity during the past few years, measured by Total Value Locked (TVL) in DeFi protocols, amounting to billions of dollars. By offering the ability to disrupt traditional financial systems, a mix of retail and institutional investors are starting to take interest in DeFi, while regulators have been flying over with skepticism around security implications as well as how it might impact overall economic stability [14].
- **Institutional adoption and regulatory interest:** The growing interest of the institutional investor community has symbolized a new chapter in the maturation of BC assets. Tesla, Square and Micro Strategy have added BCT to their own balance sheet, and the largest banks like JPMorgan or Goldman Sachs began offering cryptocurrency products & services. In turn, this has given a sense of credibility to the BC-based assets but also fueled the expectation for regulatory standards. Governments and regulatory bodies around the world are now embarking on the development of oversight mechanisms that mitigate BC asset risks yet unlock innovation. In other words, requiring that KYC and AML regulations are imposed on cryptocurrency exchange activities as well as providing guidelines governing the trade of NFTs and tokenized assets. Today, BC assets which include cryptocurrencies, NFTs and other types of digital tokens have grown from niche curiosities to major fixtures in the global financial and digital economy. Using NFTs to monetize digital art and other property interests could be a business model for the recent wave of decentralized power, giving creator control over both distribution and profits. The ultimate trajectory of these assets over the longer term will be influenced by technological innovation, institutional demand, and regulatory frameworks designed to protect security, transparency and innovation in this emergent but rapidly maturing asset class [46].

4. Rationale for regulatory oversight

The most growing two BC assets are cryptocurrencies and NFTs, which have attracted the attention of authorities and policymakers worldwide. Although there is a lot of potential for creativity with all of this new technology, there are also a lot of hazards and difficulties that have led to calls for regulation. These are some of the many justifiable goals that provide support for regulating these asset classes, ranging from financial stability and consumer protection to stopping tax fraud and maintaining market integrity [44]. In this section, the main justifications for authorities’ desire to control BC assets are examined.

4.1. Consumer protection

Consumer protection is the clearest rationale for some regulatory interference in the BC space. More people who might not have had much financial knowledge or stock market awareness have entered the world of investing as a result of cryptocurrencies and NFTs. American customers could be reluctant to take on this kind of activity due to the very volatile and speculative character of these items. People might lose a lot of money on investments in cryptocurrencies like BCT and Eth because of their notoriously volatile value. Similarly, NFTs, which were heralded as a paradigm change in digital ownership, have shown to be equally speculative as investors lost trust in them and the value of digital collectibles plummeted [30]. To make matters worse, the lack of legal insurance on exchanges in many nations makes it easier for people to commit fraud and theft. However, there are several hacking instances and scams occurring in marketing sites and exchanges that act as middlemen between buyers and sellers (cryptocurrencies being the primary example here). Other well-known hacks that deprive investors of their coins, which are valued millions of dollars, include the Coincheck and Mt. Gox attacks. Even worse, attempts to exploit gullible investors have spread to fraudulent ICOs and NFT ventures, where dishonest actors are able to acquire funds but fall short of providing the promised goods or services [11].

By establishing guidelines for how exchanges and markets should function, mandating disclosures of those risks, and offering channels for consumer recourse, that regulatory supervision may be made to lessen these risks. Investors are aware of what they are getting into when the listing process is more tightly regulated, for instance, by mandating that the assets be transparent before being listed. Regulation of NFTs may concentrate on digital asset identification and avoiding consumers possessing counterfeit or deceptively marketed items [24].

4.2. Financial stability

The risk of BC and cryptoassets to financial stability is another important factor driving regulators' attention. Cryptocurrencies, on the other hand, have come a long way from being obscure items to financial assets that have billions in market value. The more people utilize cryptocurrencies, the bigger their effect on national and eventually global financial systems. Authorities see widespread adoption of cryptocurrencies as containing risks for broader financial stability due to their systemic nature [2]. For example, the entrance of cryptocurrencies into the traditional financial system by means of institutional investment or wider payment acceptance could increase reliance on these volatile assets and therefore expose investors to a higher degree of risk. In the worst-case scenario, a large sell-off in the cryptocurrency market could bring down the wider financial markets if large banks are either heavily invested in or trading cryptocurrencies. Indeed, ahead of tightening policy, markets are already beginning to pay for years of inadequate oversight across all kinds of financial instruments. The global financial crisis brought home the dangers not only of systemic risks in asset bubbles from the failure to regulate them properly but also the fact that emerging assets like cryptocurrencies can and will certainly be caused by a massive unregulated bubble as seen in 2008 [24].

Stablecoins, a subcategory of cryptocurrencies designed to have price stability characteristics, have similarly raised regulatory concerns over the impact they may have on financial stability. Although the objective of stablecoins is to manage volatility, concerns have been expressed about whether or not these tokens collateralize in real assets and if confidence in their underlying assets were to be lost, could there be a run on stablecoins. Tether, which is among the largest of so-called stablecoins and pegged against the U.S. dollar, has long had questions

swirling about whether it has adequately maintained and transparently disclosed its reserves to investors, causing some market observers to fret that a wider collapse in cryptocurrencies could ensue if Tether were proven incapable of redeeming them on demand [44]. This would be similar to a regulatory oversight in order to prevent anything happening within cryptocurrency that could threaten financial systems at large. Such measures could involve introducing capital and liquidity safeguards for exchanges, as well as stablecoin issuers; regulating the relationship between traditional financial institutions with crypto markets or monitoring systemic risk from digital assets [44].

4.3. Anti-money laundering (AML) and know your customer (KYC)

The unregulated, decentralized, and pseudonymous nature of BC transactions have been a perfect tool for the anonymous type of black markets that can range from illicit drug trafficking, racketeering to financial crimes. Fraud can use cryptocurrencies which are difficult to detect in order to transfer funds from one country to another, avoiding the traditional banking and financial systems governed by AML and KYC regulatory requirements. BC technology also brings an aspect of anonymity, so it is not surprising that this could easily be used for illegal activity when the true source and endpoint destination are unknown [34]. More countries are agitating for greater regulatory control to enforce AML and KYC in crypto. Regulators require information about who is using cryptocurrencies requiring exchanges and wallet providers to know their customers in the same way as banks know theirs should, in theory, make it harder for virtual coins to be used for ill-gotten gains. KYC processes generally involve collecting personal information about customers (i.e., name, address, proof of identity) to verify that they do not use the service for illicit practices [55].

Other regulatory bodies worldwide like the FATF have also urged countries to follow a "Travel Rule," which requires data on crypto transaction parties to be exchanged between institutions. It is similar to some requirements that have been traditionally required by financial institutions when transferring money. The implementation of these AML and KYC measures are considered as a necessary part in preventing the abuse of BC assets for fraudulent activities to guarantee their compatibility with standard financial systems [19].

4.4. Tax evasion and transparency

Cryptocurrencies are challenging for tax authorities due to their decentralized nature and the fact that they can be transacted anonymously, peer-to-peer without any intermediaries. The pseudonymous nature of cryptocurrencies permits individuals to possibly avoid taxes by hiding their wealth and income from tax administrations. This has made governments more desperate to put the same tax reporting and enforcement on cryptocurrency transactions as they try in traditional financial assets [56]. As far as legal supervision is concerned, the focus should be on ensuring transparency in the way cryptocurrency transactions work and requiring people to report their holdings of profits. Countries like the USA have even started taking steps towards this by making it mandatory for crypto exchanges to report transactions carried on their platform to their revenue collectors. According to the Internal Revenue Service (IRS), cryptocurrencies are considered property for tax reasons and must be reported as capital gains [26].

Greater regulatory scrutiny also has the added benefit of making cryptocurrency users, exchanges, and custodians legally obliged to report their activities, which would help tax authorities bridge the gap on cryptocurrency-tax evasion. Governments can ensure that cryptocurrencies are not used to evade tax payments by developing better tax regulations and improving transparency [45].

4.5. Environmental concerns

The last justification for regulation is probably the impact that BC assets like BCT have on the environment. The proof-of-work (PoW) consensus mechanism, which powers BCT and most other cryptocurrencies, requires miners to solve complex mathematical problems in order to validate transactions and protect the network. In turn, this consumes a significant amount of power and has raised concerns about the quantity of carbon emissions from cryptocurrency mining facilities [57]. Regulators have also been stepping up their investigations to find an environmental solution that would lessen the impact of mining operations. Because cryptocurrency mining uses so much power, some countries, like China, have gone so far as to completely outlaw it. Other nations are also thinking about imposing environmental regulations on miners. Conversely, there are also initiatives to incentivize more energy-friendly models like proof-of-stake (PoS) that require much less energy consumption in operation [45].

This might be accomplished directly by penalizing miners who use non-renewable energy sources, indirectly by promoting the use of greener technology through the regulatory framework, or indirectly by giving energy-efficient consensus methods priority attention. By doing this, governments may address the environmental problems associated with cryptocurrency mining while continuing to support the growth of the BC industry [56]. Some BC assets will most likely be subject to strict regulation if the financial services industry in every major country in the world concurs that cryptocurrencies need to be regulated. Even though MAS and industry players have been able to collaborate effectively on sandbox projects, the reality is that going beyond this still calls for some ground rules because everyone agrees that a balanced approach is necessary to promote innovation while taking into account the risks that such technology may present. As participants in the BC ecosystem ourselves, having a framework to identify the underlying drivers of regulatory efforts will help us navigate this national, regional, and international environment more skillfully and, as a result, create responsible and efficient regulations [58].

5. Global regulatory approaches to BC assets and evaluation

Given the global reach of BC assets, such as cryptocurrencies and NFTs, it is no surprise that governments and regulatory bodies have been under growing pressure to adopt a forward-looking approach which strikes the right balance between consumer protection, financial stability and transparency on one side; and innovation that can fuel economic growth on another. But the decentralized and global nature of these technologies makes it difficult for regulators to act [31]. This part focuses on BC asset regulation all over the world, introduce some ways of block and describe regulatory actions in different countries or regions.

5.1. The United States

Since various federal agencies have disagreed and in some cases agreed on the status of cryptocurrencies and even NFTs, the U.S. has instituted disparate regulations of BC assets. The U.S. Securities and Exchange Commission (SEC) is likely the largest, as it has played a role in determining whether or not a number of cryptocurrencies are securities. The SEC's most notable move in this case was the lawsuit it filed against Ripple Labs, claiming that the business marketed XRP securities and that XRP ought to be regulated as such. This case has reignited the debate over how digital assets should be classified and further highlighted the legislative ambiguity surrounding them [38]. The Commodity Futures Trading Commission (CFTC) and even the Internal Revenue Service (IRS) are reportedly under the oversight of the SEC,

which is not the only government body that has established laws for cryptocurrencies. Another intriguing problem is that the CFTC identified at least one cryptocurrency derivative in its determination of its jurisdiction over virtual currencies, with BCT being regarded as a "commodity." On the other hand, for tax reasons, the IRS separates BCT from other types of money and will regard your sale as a capital gain or loss [7].

The New York Department of Financial Services (NYDFS) introduced the BitLicense in 2015, which mandates that companies that deal with virtual currencies obtain a license and stay in compliance with strict know-your-customer KYC and AML laws. Some people view the BitLicense as a good thing because it clarified regulations, but it has also been excessively complicated and difficult for the majority of businesses who are trying to operate in New York [18]. In recent years, collaborative support for the establishment of a national framework to regulate BC assets has grown, and Congress has continued to introduce legislation to elucidate some aspects of this regulation, such as the Responsible Financial Innovation Act and the Digital Asset Market Structure and Investor Protection Act [5].

5.2. The European Union

(EU) is seeming quite a bit more structured and unified when it comes to the regulation of BC assets with the recently proposed Markets in Crypto-Assets (MiCA) Regulation. It is a framework law, one of several steps that MiCA seeks to enact to guarantee the entirety of digital assets including cryptocurrencies, stablecoins and NFTs are acknowledged legally as assets more generally throughout its 27 member states. Expected to be implemented in the next years, the proposed legislation seeks to give issuers and service providers legal certainty while also guaranteeing a high degree of investor and consumer protection. MiCA introduces a nomenclature to distinguish the various categories of digital assets. MiCA creates three different categories of cryptocurrencies such as asset-referenced tokens, e-money tokens, and other crypto-assets. Regulatory requirements are different for each category and issuers and service providers have to comply with transparency, disclosure and governance standards. In the case of stablecoins, for example, they are regulated much more heavily because of the potential impact that they can have on monetary policy and financial stability [28,33].

In addition to compelling issuers to reveal the capabilities of the consensus processes that drive the network, MiCA also seeks to be ecologically conscious. This would operate as a gauge for the environmental effect of BC assets. This clause is a result of the EU's broader sustainability initiative and concerns about the energy-intensive nature of BC technology, particularly proof-of-work based cryptocurrencies like BCT [28,33]. In addition to MiCA, the EU's General Data Protection Regulation (GDPR) has made an effort to govern BC at the nexus of data privacy. BC networks' decentralized structure makes it challenging to comply with GDPR, particularly when it comes to data subjects' rights like the "right to be forgotten" and personal information kept on immutable ledgers. Regulatory attention is to be expected even if, as of right now, the EU does not have explicit rules to handle such challenges, even if privacy and the integration of BC technology with existing privacy laws are ignored [33].

5.3. Asia

Asia is a region of wide-ranging regulatory responses to BC assets, due in part to the diverse economic, political, and cultural contexts of individual country-regulatory primatene issues. China, the government of which took a tough stance on cryptocurrencies last year, banning all cryptocurrency transactions and mining. The Government of China is

Table 2
Evaluation of regulatory policy across different jurisdictions.

Jurisdiction or Region	Regulatory Approach Characteristics	Key Strengths	Key Weaknesses	Potential Impacts
United States	Fragmented, Ambiguous, Enforcement-centric	Proactive engagement of multiple agencies, focus on investor protection (SEC), clarity for derivatives (CFTC), early state-level efforts (BitLicense), bipartisan efforts for national framework.	Lack of cohesive federal approach, ambiguity hindering innovation, potential for stifling growth through enforcement, BitLicense seen as overly burdensome, challenges with global nature.	Slower innovation due to uncertainty, compliance burdens for businesses, potential for regulatory arbitrage, eventual move towards a more unified national framework.
European Union	Harmonized, Comprehensive (MiCA)	Unified legal framework across member states, legal certainty for issuers, tailored requirements based on asset categories, proactive stablecoin regulation, and consideration of environmental sustainability.	Potential challenges in implementation and enforcement across diverse member states, impact on innovation still unfolding, need for further harmonization with existing regulations (e.g., GDPR).	Increased legal certainty and consumer protection, potential for a leading global regulatory standard, possible compliance burdens for businesses, and influence on global regulatory trends.
Asia (Singapore)	Enabling, Clear, Balanced	Sensible and clear regulations conducive to innovation, balanced AML or KYC requirements allowing business growth, positioned as a BC hub.	Potential for over-regulation stifling some aspects of innovation, ongoing need to adapt to evolving technologies.	Fosters BC innovation and business growth, attracts investment, establishes Singapore as a key player in the digital asset space, balances risk management with economic development.
Asia (China)	Restrictive, Prohibitive (Cryptocurrencies), Promotive (BC Tech)	Strong state control over financial assets, focus on developing state-sponsored BC infrastructure (BSN) and CBDCs.	Blanket ban on cryptocurrency transactions and mining potentially stifles innovation, fragmented approach with promotion of underlying technology but suppression of assets.	Suppression of cryptocurrency markets within China, focus on state-controlled digital finance, potential global leadership in CBDC technology, impact on overall BC innovation within its borders is complex.
Asia (Japan/S. Korea)	Structured, Focused on Security	Early recognition of digital assets (Japan), supervision of exchanges (Japan), strong AML/KYC requirements due to past security issues (South Korea).	Potential for stringent regulations to hinder some innovation, ongoing need to adapt to new threats and technologies.	Mature and relatively secure digital asset markets, strong emphasis on consumer and investor protection, potential for slower innovation compared to less regulated environments.
Other Jurisdictions (El Salvador)	Adoption-focused (Bitcoin Legal Tender)	Potential for increased financial inclusion, attracting foreign investment.	Criticism from international financial institutions regarding financial stability and money laundering risks.	Uncertain long-term economic and financial stability impacts, potential for increased adoption of Bitcoin in specific contexts, influence on other nations considering similar moves.
Other Jurisdictions (Switzerland)	Enabling, Supportive	Supportive regulatory climate, clear guidance on token classification, integration of BC assets within existing financial regulations.	Potential for complexity in applying traditional financial regulations to novel BC assets, ongoing need to adapt to rapid technological advancements.	Fosters innovation and attracts BC businesses, provides regulatory clarity within a well-established financial system, potential model for other jurisdictions.
Other Jurisdictions (India)	Uncertain, Evolving	Exploration of a digital rupee, ongoing discussions about regulating private cryptocurrencies.	Flip-flopping between outright bans and regulation creates market uncertainty, lack of definitive legislation.	Market volatility and uncertainty, delayed adoption and innovation, potential for a more defined regulatory framework in the future depending on legislative outcomes.

Source: Author's self-assessment.

concerned about financial stability, capital flight and energy consumption with such a volatile currency that consumes 13 TWh per year of electricity, the blink rate at which Hong Kong has been gobbling up coal-fired power plants. Meanwhile China has been aggressively pushing the development of BC technology with its BC Service Network (BSN) and is also leading the innovation in Central Bank Digital Currencies (CBDCs), apart from the more advanced project on digital yuan. The regulatory backdrop in China is a sign that it wants to manage digital financial assets while enthusing about other kinds of state-sponsored choices [64].

In contrast, Singapore has long positioned itself as a global center for BC innovation by instituting sensible and clear regulations conducive to the expansion of the ecosystem. In 2020, the Monetary Authority of Singapore (MAS) brought out the Payment Services Act, a licensing regime for digital payment token services, which also cover cryptocurrency exchanges. The law forces companies to follow AML and KYC regulations, but its demands are balanced such that start-ups rooted in the BC can still thrive in Singapore [37,61]. Japan and South Korea, for example, have built some of the most sophisticated regulatory structures around BC assets in the world. Enforcing regulation, Japan was among the first nation to regulate digital assets, back in 2017 when they

acknowledged BCT as legal tender. In addition, the Japanese Financial Services Agency (FSA) supervises cryptocurrency exchange regulation and confines transactions to registered exchanges [37]. Meanwhile, South Korea has seen a number of significant hacks and fraud cases that led to the adoption of very rigorous AML and KYC requirements for cryptocurrency exchanges [37,51].

5.4. Other jurisdictions

Meanwhile, authorities in other parts of the world have adopted a wide array of regulatory stances on BC assets; some have provided open arms to crypto and NFTs while others take a heavy-handed approach in policing or even banning them. In 2021, El Salvador grabbed headlines for being the first country ever to make BCT legal tender. President of the nation drove the initiative, which officials say aims to increase financial inclusion and attract investment from abroad. Nevertheless, the overall concept of BCT as a legal payment option got criticism from International Monetary Fund (IMF), and other large financial organizations are starting to worry about the repercussions it might have on the world's finance stability and money laundering [25,52].

India has been much more conservative in her attitude towards

cryptos, flip-flopping between offers of an outright ban and the potential regulation of existing names. The Indian government in 2021 was exploring the idea of a bill to ban all private cryptocurrencies and provide for an official digital rupee. Nonetheless, definitive legislation has not come to fruition and the market is still in flux [25]. In Switzerland, one of the country's best-known for its supportive regulatory climate for BC has created a system that can fit cash assets under prevailing financial regulations. The Swiss Financial Market Supervisory Authority (FINMA) issues guidance on how BC tokens are classified in Switzerland, leading the way internationally for future BC innovation [34,35].

5.5. International coordination and self-regulation

One of the greatest challenges in trying to regulate BC assets is that they are borderless by nature, which necessitates international cooperation. And while individual countries are coming up with their own rules, there is a growing consensus that international cooperation is necessary to avoid regulatory arbitrage and ensure where crypto-assets exist on BCs they exist under regulation [42]. International bodies like the Financial Action Task Force (FATF) have led efforts to establish a set of global norms on BC asset regulation, particularly in the AML and KYC space. Fondly referred to as the Travel Rule, the measure means that crypto exchanges and other virtual asset service providers are expected to transmit details of parties in transactions above a certain threshold so as to control money laundering and terrorist financing operations.

Besides government enforcement, the BC space has also begun to use self-regulatory policies. Inroads are being made by addressing identity management, with numerous cryptocurrency exchanges going the extra mile to conduct KYC and AML processes, even when they are not legally mandated for them to do so in their jurisdiction, an effort to clean up the industry and prevent fraud and other transgressions from occurring. Group such as 'CryptoUK' and the 'Blockchain Association,' are an example of industry groups working towards friendly yet fair regulation efforts [40,52]. As one might image, the global regulatory landscape in relation to BC assets is variable and still developing with different countries and sometimes states within a country taking quite differing tact based on their specific legal, economic, and political environments. Some countries saw BC as an innovative tool and integrated with openness, some others took a long journey by being conservative or even prohibitive towards BC assets. The world is edging closer to being able to untap the potential of these decentralized, borderless assets by harnessing BC technologies but international collaboration and balanced regulatory frameworks remain vital in solving some of the challenges they pose [52].

5.6. Evaluation of regulatory policy

To provide a clear and concise comparison of the diverse regulatory approaches to BC assets across different jurisdictions, the following table summarizes the key characteristics, strengths, weaknesses, and potential impacts of the policies discussed in this study. This comparative analysis aims to highlight the contrasting strategies adopted by various nations and regions as they grapple with the opportunities and challenges presented by cryptocurrencies, NFTs, and other digital assets. By examining these different regulatory frameworks side-by-side, we can gain a deeper understanding of the potential trade-offs between fostering innovation, ensuring consumer protection, and maintaining financial stability in the evolving landscape of BC technology, as described in Table 2.

5.7. Quantitative evaluation of regulatory effect of BC assets

To provide a comparative and quantitative overview of the regulatory effects on BC assets across key jurisdictions, the following tables and graphs presents specific metrics indicating the key information related BC assets regulation.

Table 3
AML or KYC compliance scores by nations.

Nations	AML/KYC Compliance Rate (% VASPs)	Basel AML Index Score	FATF Recommendation 15 Status
Singapore	91 %	5.29	Compliant
Switzerland	88 %	4.98	Largely Compliant
USA	61 %	5.34	Partially Compliant
India	54 %	6.44	Non-Compliant
Nigeria	42 %	7.01	Non-Compliant

Source: Basel Institute on Governance [65].

Table 3 and Fig. 3 offers a revealing snapshot of the global AML or KYC compliance landscape, highlighting significant disparities between jurisdictions. Singapore emerges as the clear leader, boasting a 91 % compliance rate paired with a low Basel AML Index Score of 5.29, reflecting its robust regulatory framework and effective enforcement mechanisms. Switzerland, while slightly behind at 88 % compliance and a 4.98 score, also showcases strong regulatory performance, benefitting from its long-standing financial reputation. In sharp contrast, the USA, despite its advanced financial infrastructure, displays a surprisingly lower compliance rate of 61 % with a slightly higher risk score of 5.34, suggesting potential gaps in enforcement or variations in regulatory interpretation across states. More concerning are India and Nigeria, with compliance rates of only 54 % and 42 % respectively, and notably higher Basel Index scores (6.44 for India and 7.01 for Nigeria), signaling persistent vulnerabilities, regulatory weaknesses, and a higher exposure to financial crime risks. These differences underline how both regulatory maturity and consistent application play critical roles in shaping the effectiveness of AML/KYC regimes worldwide.

Table 4 captures the evolving landscape of regulatory development and its tangible impact across several key jurisdictions. In the EU, the introduction of the Markets in Crypto-Assets Regulation (MiCA) in December 2024 has markedly tightened the crypto sector's footprint, with crypto-focused funds accounting for <1 % of the EU fund universe and an overwhelming 95 % of EU banks maintaining no exposure to crypto assets, signaling a cautious and risk-averse regulatory environment. Meanwhile, the United Kingdom (UK), through the issuance of DP24/4 covering Admissions, Disclosures, and the Market Abuse Regime in late 2024, is navigating a more balanced approach, as evidenced by 12 % of UK adults holding crypto and 33 % expressing confidence that the Financial Conduct Authority (FCA) would intervene in case of disputes, reflecting moderate but growing public engagement. In the USA, upcoming deregulatory shifts anticipated under Trump's 2025 administration suggest a probable loosening of constraints, likely aimed at boosting financial industry contributions and innovation, though potentially at the cost of regulatory rigor. Singapore, however, stands out for its aggressive expansion: the number of Major Payment Institution (MPI) licenses for crypto exchanges more than doubled from 6 in 2023 to 13 in 2024, complemented by a thriving ecosystem of 1600 BC patents, 2433 related jobs, and 81 active exchanges, underscoring its ambition to become a global BC and crypto hub through progressive regulation.

Table 5 and Fig. 4 provides a comparative view of crypto fraud losses and the corresponding regulatory strategies adopted by various nations, highlighting stark contrasts in both financial impact and regulatory philosophy. The USA records the highest estimated fraud losses at a staggering \$5600 million, reflecting the scale of its crypto market and a predominantly enforcement-driven regulatory approach that tends to act after fraudulent activities have occurred. India, despite its large population and growing crypto user base, reports much lower fraud losses at \$44 million, but its regulatory stance remains largely reactive, indicating delayed or inconsistent responses to emerging threats. In contrast, Singapore presents a model of proactive governance, with estimated fraud losses of \$180 million and a strong regulatory focus on prevention and public education, demonstrating an emphasis on risk mitigation before incidents materialize. Meanwhile, the United

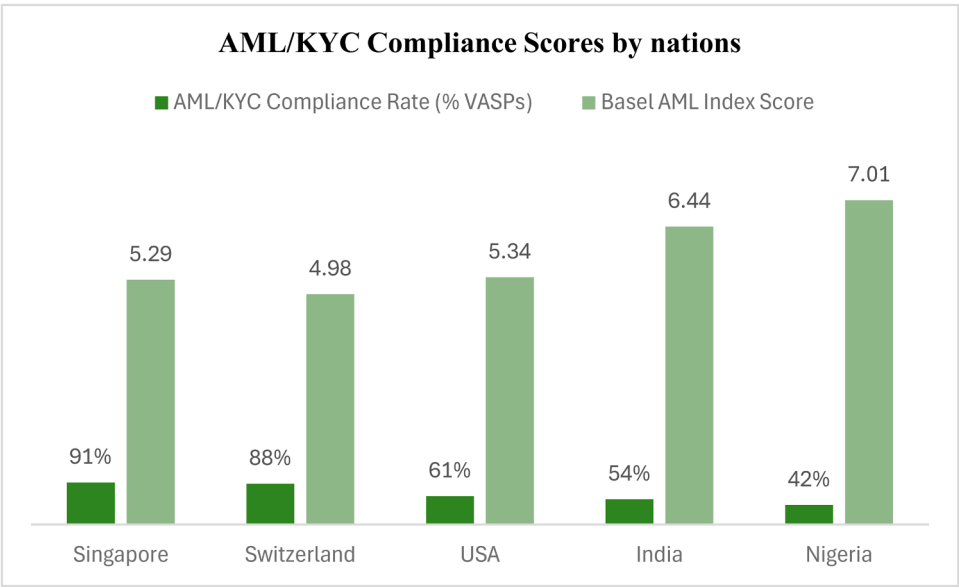


Fig. 3. AML or KYC compliant scores and rates. Source: Basel Institute on Governance [65].

Table 4
Impact of regulatory.

Nations	Regulatory Development	Quantitative Impact
EU (EU)	Implementation of Markets in Crypto-Assets Regulation (MiCA) in December 2024.	Crypto-focused funds <1 % of EU fund universe; 95 % of EU banks have no crypto exposure.
United Kingdom (UK)	Issued DP24/4 on Admissions & Disclosures and Market Abuse Regime (late 2024).	12 % of UK adults own crypto; 33 % believe FCA would help if problems arise.
United States (US)	Anticipated deregulatory changes under Trump (2025).	Increased financial industry contributions expecting favorable regulation.
Singapore	Doubled MPI licenses for crypto exchanges in 2024.	13 licenses in 2024 vs 6 in 2023; 1600 BC patents, 2433 jobs, 81 exchanges.

Source: Coin360, [16,17]; KPMG, [36]; Reuters, [49]; Reuters, [50].

Table 5
Crypto fraud losses.

Nations	Estimated Fraud Losses (USD)	Regulatory Focus
USA	\$5.6 billion	Enforcement, post-factum
India	\$44 million	Reactive
Singapore	\$180 million	Preventive, education-led
UK	\$490 million	AML/KYC reinforcement

Source: Federal Bureau of Investigation, [22]; Business Standard, [15].

Kingdom, with fraud losses estimated at \$490 million, channels its regulatory energy into strengthening AML or KYC frameworks, aiming to fortify institutional safeguards and prevent financial crimes at the systemic level. Collectively, these variations underscore how differing regulatory strategies significantly influence the scale and nature of crypto-related vulnerabilities across countries.

Table 6 highlights the range of legal and regulatory concerns currently confronting investors in the crypto space, painting a complex picture of the evolving risk environment. Asset classification uncertainty remains a major challenge, affecting 30 % of investors who struggle to navigate inconsistent definitions across jurisdictions, complicating compliance and investment decisions. Fraudulent scheme proliferation has surged by 12 %, signaling an increasing threat to investor security and trust. Taxation compliance poses another significant hurdle, impacting 45 % of investors due to opaque or rapidly shifting tax



Fig. 4. Crypto fraud losses. Source: Federal Bureau of Investigation, [22]; Business Standard, [15].

Table 6
Legal and regulatory concerns for investors.

Concern	Percentage of Affected Investors
Asset Classification Uncertainty	30 %
Fraudulent Scheme Proliferation	12 % increase
Taxation Compliance Challenges	45 %
Investment Disclosure Requirements	\$50,000 threshold
Consumer Protection in Stablecoins	35 % of countries
Intellectual Property Challenges (NFTs)	17 %
Cross-Border Legal Complications	40 %

Source: Coin360 [17].

reporting obligations. Investment disclosure requirements have also tightened, with a \$50,000 threshold triggering mandatory reporting, increasing administrative burdens. Consumer protection, particularly in the stablecoin sector, is gaining attention, with 35 % of countries instituting specific safeguards to shield users from volatility and misuse. Intellectual property issues surrounding NFTs affect 17 % of investors, reflecting the emerging legal gray areas around digital ownership and copyright. Finally, cross-border legal complications trouble 40 % of investors, underlining the jurisdictional challenges and regulatory fragmentation that complicate international crypto activity. Together, these concerns illustrate the growing need for clearer, harmonized regulatory

Table 7
Central Bank Digital Currency.

Nations	CBDC Project Status	Adoption/Usage Statistics
China	Digital yuan expanded	40 million users; \$15B transactions
European Union	Digital euro pilot in phase two	Planned by 2026
India	CBDC in second testing phase	10 million users
Jamaica	JAM-DEX 30 % adoption	30 % adoption
Nigeria	eNaira reached over 1M users	1M+ users; 10 % cash reduction projected

Source: CoinLaw [16].

frameworks to support investor confidence and market stability.

Table 7 presents a detailed overview of (CBDC) initiatives and their respective adoption and usage metrics, illustrating varied progress across regions. China leads the global CBDC race with its digital yuan, boasting 40 million users and facilitating transactions worth \$15 billion, signaling both government commitment and growing public acceptance. The EU's digital euro project is progressing steadily, currently in its second pilot phase with a full rollout targeted by 2026, reflecting a cautious but structured approach to integration within its complex multi-national financial system. India's CBDC development is also advancing, now in its second testing phase and already attracting 10 million users, underscoring the country's rapid digital adoption and financial inclusion efforts. Jamaica's JAM-DEX demonstrates notable success with 30 % adoption, indicating strong national engagement in a smaller economy context. Meanwhile, Nigeria's eNaira has surpassed 1 million users and is projected to contribute to a 10 % reduction in cash usage, pointing to meaningful, if gradual, shifts in consumer behavior. Collectively, these initiatives showcase the diverse strategies and paces at which different economies are embracing the digitalization of money.

6. Challenges in regulating BC assets

Since BC assets have soared in popularity and value as seen with cryptocurrencies and NFTs, they have posed enormous challenges to regulators the world over. These assets are reshaping industries by power decentralized transactions, digital ownership, and peer-to-peer exchanges just as they create challenges that became hard to fit in with traditional regulatory frameworks [42]. This part examines the

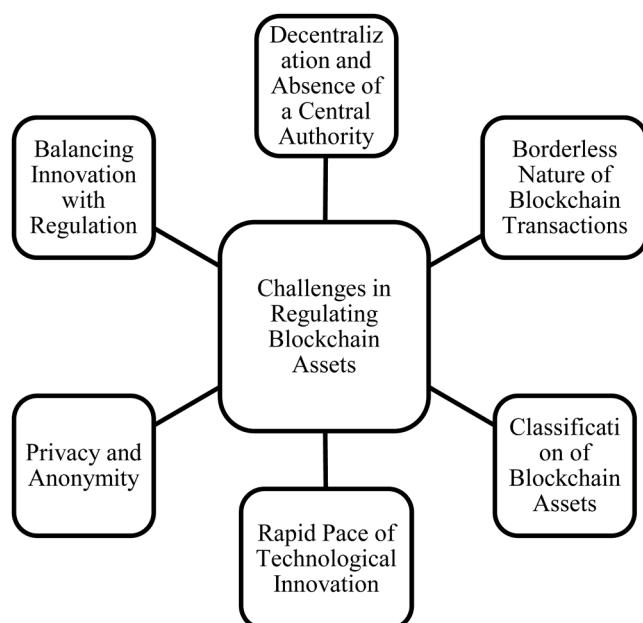


Fig. 5. Challenges in regulating BC assets.

major challenges confronted by policymakers, regulators and overseers in pursuing to supervise BC assets (Fig. 5).

6.1. Decentralization and absence of a central authority

Although one of the main features of BC assets is decentralization, the operation of these networks differs from that of traditional finance. Because BC technology is decentralized and trustless, it does not need a central authority to monitor transactions, such as a bank or financial regulator. To maintain the ledger, transactions are validated by this dispersed network of participants, or nodes. Because of this, enforcement tasks are difficult [40]. For instance, banks and other financial entities that function as middlemen between participants in financial transactions may be subject to regulations from authorities in legacy (or conventional) financial systems. Yet with BCT being a decentralized network, there exists no central authority or middleman. Regulations that are standard in controlling illicit activities do not apply to DeFi primarily because there is no central control point [52]. Additionally, DeFi platforms can both trade and lend or borrow cryptocurrencies in an intermediate form network. Autonomous operating mechanisms are used in these types of platforms using smart contracts (self-executing code). However, smart contracts are especially hard to regulate because they operate on their own and do not depend on human intervention, which makes it even harder to monitor the participants in transactions [58].

6.2. Borderless nature of BC transactions

One of the main problems when it comes to decentralized BC-based assets is that these are borderless. These can interexchange cryptocurrencies and NFTs across sovereign boundaries without any intermediary or third-party institutions. Their global presence also makes it difficult for any one specific nation to enforce their laws and regulations, or if they do try, the transactions can simply move beyond their territories [45]. A cryptocurrency exchange is one good example where a company based in one country can serve users in multiple other countries without being subject to the regulatory frameworks of those foreign jurisdictions. This cross-border activity frequently results in regulatory arbitrage, whereby companies and individuals relocate their efforts to jurisdictions with lower regulation. This undermines the utility of regulations, as entities can effectively evade oversight by moving to more weakly regulated areas [42]. Cross-border transactions also complicate tax enforcement. Cryptocurrencies are commonly transferred cross-border and the self-explanatory nature of cryptocurrencies makes it difficult for government tax authorities to associate incomes from the source. This creates a possibility of tax evasion where people can simply move large amounts of money from one country to another without informing the tax authorities about this [40].

6.3. Classification of BC assets

One highly controversial area is the categorization of BC assets. Cryptocurrencies and NFTs have several uses and hence they may not fit in any of the existing legal categories. For instance, some cryptocurrencies such as BCT are viewed as a type of currency with additional functionalities; others like Eth were created as platforms for being able to run Dapps otherwise known as decentralized applications [52]. In some jurisdictions, regulatory bodies do not agree on classifying these as securities, commodities or currencies either. For example, in the US, some cryptocurrencies have been classified as securities by the Securities and Exchange Commission (SEC), imposing a certain framework of security regulations on them; at the same time BCT and Eth are deemed commodities by the Commodity Futures Trading Commission (CFTC) meaning that these fall under another set of rules. The lack of a standard classification causes regulatory uncertainty, making businesses and consumers alike unsure of their legal obligations [42]. Such as NFTs,

that has very different classification challenges. Although we most often hear about NFTs in the context of proving and owning digital art, collectibles or virtual property, an NFT can just as well be a claim on a ticket, a license or the proof of ownership over some physical object. This has made regulation far more complex, with different categories of NFTs potentially classifying as different types of securities requiring its own approach to define and regulate [40].

6.4. Rapid pace of technological innovation

BC technology is changing faster than regulators can regulate. Cryptocurrencies, smart contracts, Defi, NFTs and the resulting derivatives continue to evolve and drive new risks alongside regulatory concerns. Because the BC field is so dynamic, pre-existing laws and regulations are often inadequate to deal with new developments [58]. For instance, the advent of DeFi platforms has brought about securities, fraud and market manipulation issues. While these platforms allow the users to do trade on it without any intermediaries, but code vulnerabilities, flash loans attacks and a rug pull i.e., developers exit scamming are also found simultaneously. Regulators are still finding their footing on these risks and building frameworks to protect consumers against such risk but the prompt progress made by DeFi space makes it further complicated for regulators on how they should be responding to the developments [52]. Also, the technical sophistication of BC assets may create an information asymmetry between regulators and professional services industry. The nature of smart contracts, BC consensus mechanisms such as proof-of-work & proof-of-stake and interaction between different types of BC networks is fairly complex to understand and getting too technical. If not enough knowledge exists about these technologies, it will be very difficult for regulators to implement useful supervisory approaches that address precisely the risks that BC assets bring with them [42].

6.5. Privacy and anonymity

The privacy and anonymity of BC transactions also make it more difficult to enforce regulation. BCT and Eth are on pseudonymous networks, which means users are represented by their public addresses instead of real-world identities. So, while these networks are open with every transaction recorded on a public ledger, the pseudonymous nature of transactions leads to difficulties in regulators being able to tie addresses back to individuals [40]. Furthermore, by hiding transaction information, privacy cryptocurrencies like Monero (XMR) and Zcash (ZEC) offer even higher degrees of anonymity while making it difficult for authorities to track the money. Users who value financial privacy may find these privacy features appealing, but there are worries that they might be abused for illicit reasons (money laundering, funding terrorism, tax evasion) [52]. The threat of informed consent violations prompted regulators to turn their attention toward finding a solution for the privacy issues in the DeFi space. Certain governments are insisting for more transparency and some form of regulation in terms of BC transactions, suggesting KYC or AML orders where users would have to identify themselves before taking part in cryptocurrency exchanges or DeFi platforms. But the tension in balancing ensuring compliance with not overreaching into users' rights to privacy is where regulators must tread dangerously [42].

6.6. Balancing innovation with regulation

The real challenge in regulating BC assets is to balance between supporting innovation and maintaining appropriate level of control. Although this technology will transform both industries and people,

overly strict regulations could strangle innovation at the root of the BC ecosystem. While governments and regulatory agencies have an obligation to protect consumers, maintain stability in financial markets & prevent illegal activities. At the same time, they must also see the promise of BC to boost economic growth, enhance financial inclusion and unlock new business models. Balancing these interests is a complex exercise that should consider both potential benefits and possible risks of BC assets [40]. In this way, many regulators are following a "regulatory sandbox" approach allowing BC companies to experiment with new technologies under regulatory supervision without having to deal with all the regulatory burdens. This helps regulators learn about new tech without stifling businesses that need to build and experiment with technology in order to innovate [52].

Decentralized and borderless, BC assets present unique difficulties in classification; their rapid pace of technological innovation means that laws cannot be made obsolete overnight, while strong privacy concerns make it harder to track the flow of legal money laundering through these cryptocurrencies. All of the above make for a multifaceted and dynamic space to regulate, leaving precious little time and room for policymakers indeed to implement generative, future-looking policies that can ultimately bestow us with the benefits from this BC technology while minimizing its risks [58].

7. The future of regulatory frameworks for BC assets

Regulations must be established as quickly as possible because of the expanding market for digital assets (such as cryptocurrencies and NFTs) and the maturity of BC technology. Globally, policymakers are attempting to determine what kinds of mechanisms should be in place in order to balance promoting and regulating these decentralized, non-geographic technologies. Numerous national, regional, and global initiatives are probably going to identify, create, and maybe implement common regulatory standards for BC assets in the upcoming years, in addition to industry self-regulation. Various considerations, including tax reporting, financial stability, AML and KYC compliance, consumer protection, and environmental concerns, influence these systems [52] (Fig. 6).

7.1. Consumer protection and financial stability

One of the main reasons for the regulation supervision is to secure consumers from the natural risks related to BC assets. For example, cryptocurrencies are extremely unstable. Values of things like BCT and Eth can vary wildly in hours, leaving the retail investors with a lot of money at great risk. The NFTs boom also directly coincided with a new era of scam-related scams, whereby unsuspecting buyers were hoodwinked into purchasing fake digital art. As these markets expand, regulators are likely to increasingly focus on protecting consumers by issuing more guidelines and potentially increasing their enforcement activities [40]. In the future, regulatory jigsaw puzzles are likely to include mechanisms for guarding against fraud risks, and ensuring that a consumer will benefit an accurate view of what they are buying. For instance, depending on the case, it may increase requirements for transparency of BC transactions (e.g., by requiring disclosure of risks), ensure companies issuing tokens provide standardized or verifiable information and so on [58].

Another primary focus is addressing financial stability. If there is more widespread usage of cryptocurrencies, it could pose risks to the wider financial system including a great deal of cross-contagion effects as movements in huge crypto markets might end up having highly spillover impacts on conventional markets. Stablecoins are a variety of cryptocurrency that are tied to the value of reliable assets beside them

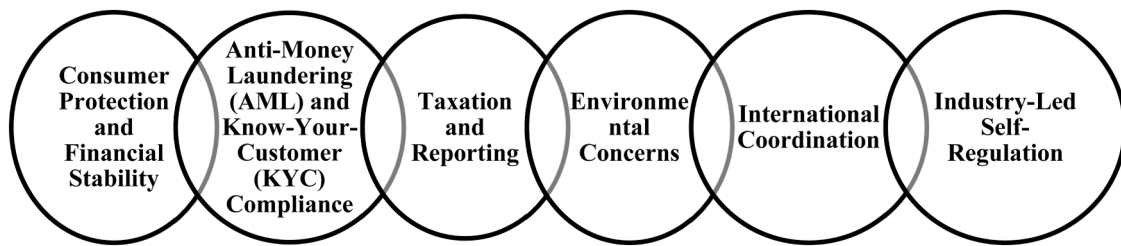


Fig. 6. Regulatory frameworks for BC assets.

and could thus interfere with national fiat currency systems (like US dollar-tied stablecoins). Governments and central banks, now more than ever, will seek stronger controls for stablecoins with requirements on reserves and audits of real-world collateral backing [18]. The concerns described before may move central banks to think about introducing digital currencies, CBDCs. One way is providing a sovereign digital replacement for cryptocurrency, giving consumers and businesses a better yet still safe asset. This trend is expected to continue as countries such as China and Sweden trial their own CBDC programs, shaping the regulatory landscape of tomorrow [45].

7.2. Anti-money laundering (AML) and know-your-customer (KYC) compliance

The crypto sphere has been a target for criminals looking to remain anonymous and avoid the scrutiny of state-enforced AML legislation, resulting in some sites and users using cryptocurrencies as a means to facilitate transactions related to money laundering, terrorism financing or tax evasion. Consequently, regulators have been (more or less) eager to guarantee that BC assets are held to the same AML and KYC standards as traditional financial institutions. AML and KYC compliance soon to be core elements in BC asset regulatory frameworks in future years [52]. Even in the traditional financial world, it can be observed that a similar trend of countries moving to AML and KYC regulations for digital currency exchanges that would involve adhering to identity verification mechanisms as well reporting suspected activities to respective authorities around the globe. Looking ahead, it can be anticipated that more international standardization on these types of regulations to maintain the balance and prevent regulatory arbitrage by bad actors in between countries with different standards. Organizations like the Financial Action Task Force (FATF) are already in the process of developing global standards for AML and KYC of digital assets, and future frameworks may be based on these measures [58].

The decentralized nature of many BC platforms is a big hurdle for compliance with AML and KYC regulations. For example, many regulators are accustomed to the requirements of centralized exchanges but have not yet experienced the capabilities (or lack thereof) provided by decentralized exchanges (DEXs), which operate outside the purview of a central authority. In turn, regulators may adopt a more reluctant hybrid model where decentralized platforms would be compelled to comply with some degree of due diligence; for instance, they could implement identity verification at certain phases in transactions. This could result in the emergence of new technologies that allow decentralized compliance, without breaking the cardinal tenets of BC such as privacy and autonomy [18].

7.3. Taxation and reporting

The taxation of BC assets is still murky and not a foregone conclusion. For instance, cryptocurrencies can be treated as property much like gold or real estate rather than as a currency, which means each

transaction no matter how trivial such as buying goods or exchanging one cryptocurrency for another can create taxable events. The growth of digital asset markets has led tax authorities around the world to begin to provide clarity on certain issues related to reporting obligations of individuals and businesses dealing in BC transactions [58]. Perhaps future governance models will include more rigorous reporting standards for exchanges and even other BC platforms to provide users with information on how to accurately report their transactions to the IRS. This can include automated reporting systems that monitor cryptocurrency transactions in real time and provide capital gains/losses calculations. The governments of some countries have already introduced rules under which data apparently from crypto exchanges is collected to detect cases of tax evasion. As these systems advance, international cooperation between tax authorities may be required in order to properly monitor cross-border transactions [52].

7.4. Environmental concerns

Regulations pertaining to the environmental impact of BC technology will also evolve. Even now, despite all the evidence to the contrary, the energy appetite of BCT and its kin, which rely on competing proof-of-work consensus to operate both the broadcast full nodes that process transactions and mining to mint new coins is commonly cited as evidence of their unsustainability. Given that policymakers worldwide are increasingly focusing on the environment, such regulatory pressure is more likely to originate from governance's desire to see BC networks use less energy [18]. In the near future, this may even serve as a mandate to support energy-efficient consensus models, such proof-of-stake rather than proof-of-work. For BC businesses, especially those that mine cryptocurrency, governments can even start imposing carbon fees or limitations on power use. Additionally, BC enterprises may be required to publish their environmental effect, similar to how traditional organizations are required to disclose their sustainability initiatives [45].

7.5. International coordination

Since BC assets are structurally borderless, international coordination will be necessary to make the regulations effective. This may be true for individual countries, but as BC technology allows assets and transactions to exist globally and decentralized across nations, state borders can hardly contain regulated financial activities. This clearly necessitates that the regulations must be designed optimally and uniformly to prevent bad actors from slipping through cracks [18]. Due to their proven track record in fostering international cooperation, it is probable that entities such as the IMF, World Bank and FSB will be important in creating a global framework of regulation for BC technology. Standardizing the way in which regulators approach digital assets from country to country will ensure that there is a consistent framework for monitoring these new forms of investments, diminishing regulatory arbitrage and enhancing consumer protection worldwide [58].

Frameworks known as regulatory sandboxes, which allow BC

businesses to test new business models and technology while being closely monitored by regulators, could also become more popular. With no regulatory responsibilities, these sandboxes provide businesses a high-quality setting in which to develop innovative BC applications. Policymakers may gain a better practical grasp of the technology through sandboxes, which promote collaboration between market actors and regulators. However, regulations must always follow innovation [45].

7.6. Industry-led self-regulation

Aside from government-led actions, industry likely will be more proactive in self-regulation. By rolling their own compliance features like KYC and AML protocols, top crypto exchanges not limited to but including non-fungible token marketplaces, and BC platforms are fortifying confidence amongst users and heading off potential regulatory backlash. And it will likely continue, as BC companies race to define new baselines and best practices that meet a shifting regulatory landscape [45]. Self-regulation may even go so far as to include the environmental, social and governance (ESG) realm, wherein BC entities opt in best sustaining behaviour and report on their footprint. BC businesses can drive beneficial regulatory evolution by showing a genuine dedication to responsible business practice [18].

The future of BC assets in terms of regulatory frameworks will be interesting to watch as the same matures over time, and it will depend on a delicate balance between national, regional and international progress with individual concerns like consumer protection, financial stability, AML or KYC compliance, tax implications, environmental concerns or simply innovation. By understanding the shared challenges in regulating decentralized and borderless technologies, we can expect global cooperation from a variety of stakeholders including international organizations and the private sector to have an essential part in how BC assets are safely incorporated into the wider financial system in a secure, accountable, and innovative way [58].

8. Implementation, applicability, policy formulation, and validation analysis

The establishment of a robust and adaptable regulatory framework for BC assets necessitates careful consideration across several interconnected domains. These include the practical implementation of the framework, its applicability across diverse global contexts, the meticulous formulation of specific policies and rules, and a thorough validation analysis to ensure its effectiveness and relevance. Each of these areas is crucial for fostering a secure and innovative environment for BC technology.

8.1. Implementation of the future regulatory framework for BC assets

The successful implementation of a future regulatory framework for BC assets demands a well-defined, multi-phase approach that not only aligns with evolving global standards but also demonstrates sensitivity to the unique regulatory needs present at national, regional, and local levels. Given the inherent complexities of decentralized technologies, encompassing cryptocurrencies and NFTs, the overarching implementation strategy must skillfully balance the promotion of technological innovation with the imperative of ensuring robust consumer protection, maintaining financial stability, and rigorously combating money laundering (AML) activities.

1. Multi-Level Regulatory Approach

The regulatory framework's implementation will necessitate a co-ordinated effort across multiple levels of governance: local, national,

and international. Adopting a harmonized approach is paramount to ensuring consistent regulation of BC assets while simultaneously accommodating the distinct legal and economic environments of various jurisdictions. Key implementation steps at each level include:

- **National Level:** Individual countries will be tasked with the critical responsibility of developing entirely new legislation or strategically adapting existing legal frameworks to effectively incorporate BC technology, particularly concerning cryptocurrencies, NFTs, and DeFi platforms. For instance, national tax authorities must provide clear and comprehensive guidelines on the taxation of BC assets, while financial regulators will need to rigorously enforce compliance with established AML and KYC protocols within the BC ecosystem. These national regulations will likely encompass detailed guidelines governing the issuance and utilization of digital currencies, stablecoins, and the operational standards for cryptocurrency exchanges. A particularly critical aspect of this national-level implementation will be the imperative to ensure that the enacted regulations are not unduly restrictive, thereby fostering an environment conducive to innovation while proactively preventing the potential for misuse and illicit activities [45].
- **Regional Level:** Collaborative efforts at the regional level, particularly within established economic and political blocs such as the EU or the ASEAN bloc, will prove essential in achieving regulatory alignment and effectively mitigating the risks associated with regulatory arbitrage. Policies formulated at the regional level should strategically focus on establishing shared standards for the transparent reporting of transactions, ensuring robust consumer protections across member states, and addressing the significant environmental considerations associated with BC technologies [52].
- **International Level:** Recognizing the inherently global nature of BC assets, robust international cooperation will be indispensable for the successful implementation of a comprehensive regulatory framework. Key international organizations, including the International Monetary Fund (IMF), the Financial Stability Board (FSB), and the World Bank, will play pivotal roles in the crucial tasks of crafting and facilitating the adoption of universal standards designed to effectively prevent regulatory gaps and inconsistencies. Furthermore, such high-level international cooperation will be absolutely critical in the effective implementation of AML and KYC regulations on a global scale, thereby ensuring that malicious actors cannot exploit the decentralized characteristics of BC technologies for illicit purposes [18].

2. Regulatory Sandboxes for Experimentation

To facilitate a seamless and well-informed introduction of the regulatory framework, the establishment of regulatory sandboxes is a highly valuable strategy. These controlled sandbox environments will provide a safe space for BC startups and more established organizations to rigorously test their innovative models and solutions under the direct oversight of regulators, but without the immediate and full burden of comprehensive regulatory compliance. By actively fostering a cooperative and communicative space between BC firms and regulatory bodies, these sandboxes offer invaluable insights into the practical operational challenges associated with effectively enforcing regulations within the dynamic BC ecosystem [45].

Furthermore, the strategic use of regulatory sandboxes actively promotes experimentation with novel regulatory techniques and technologies. This could include the testing of real-time transaction monitoring systems specifically designed for AML or KYC compliance or the development and refinement of standardized reporting platforms for tax compliance related to BC assets. Such proactive initiatives will be

absolutely key in ensuring that BC innovations can continue to flourish and evolve without undermining critical regulatory objectives, such as the maintenance of financial stability and the robust protection of consumers.

8.2. Applicability of the regulatory framework to different countries

Acknowledging the inherently global nature of BC assets, the proposed regulatory framework must possess a high degree of adaptability to accommodate the diverse legal systems, varying economic contexts, and differing levels of technological infrastructure present across different countries. The cornerstone of its broad applicability lies in its inherent ability to effectively incorporate specific local needs and priorities while steadfastly adhering to internationally recognized standards and best practices.

1. Emerging Economies

Middle-income and emerging economies including significant players like India, Bangladesh, Brazil, South Africa, Indonesia, and Kenya may benefit most from adopting a carefully planned phased implementation strategy. This approach would logically begin with the establishment of foundational regulatory elements, such as: Formal legal recognition and clear classification of various crypto-assets, the establishment of dedicated regulatory sandboxes to facilitate the testing and understanding of diverse digital asset applications and the introduction of basic and proportionate licensing structures for digital asset service providers operating within their jurisdictions.

This well-considered tiered approach empowers policymakers in these economies to effectively manage potential risks associated with BC adoption while simultaneously actively supporting the growth and development of local innovation ecosystems. For instance India's RBI and SEBI have proactively initiated sandbox environments and are actively exploring a unified and comprehensive approach to the regulation of crypto-assets. Bangladesh is strategically exploring the potential of BC technology for enhancing efficiency and security in areas such as trade finance and the digitization of public records. Brazil's CVM and central bank have already issued important guidance on the treatment of tokenized assets and have established crucial regulatory clarity surrounding payment-based tokens. Kenya's capital markets authority has actively encouraged open dialogue with relevant stakeholders while diligently assessing both the potential risks and the significant opportunities presented by .

This measured and pragmatic progression allows these emerging economies to build essential domestic capacity in BC regulation while concurrently preparing the groundwork for the subsequent adoption of more advanced and comprehensive components of the framework, such as sophisticated taxation policies, clear custodial regulations for digital assets, and effective cross-border reporting protocols.

2. Developed Countries

For nations with well-established and mature financial systems, the primary focus of the regulatory framework would necessarily need to be on effectively mitigating the potential risks associated with the increasing integration of cryptocurrencies and stablecoins into the broader economic landscape. This includes proactively addressing systemic risks that may arise from stablecoins pegged to fiat currencies and ensuring robust consumer protection against the inherent volatility often observed in cryptocurrency markets. Furthermore, these developed nations would require the implementation of advanced regulatory measures, including comprehensive compliance frameworks for tax reporting related to digital assets and stringent AML practices that align with international standards.

The EU's MiCA regulation already provides a robust and forward-looking foundation for crucial aspects such as asset classification, the

obligations of digital asset issuers, and the regulation of stablecoins. Switzerland's FINMA has established a clear and well-defined token taxonomy and has cultivated a supportive and innovation-friendly ecosystem for BC development. Singapore's MAS rigorously enforces a comprehensive licensing regime for digital asset service providers, implements robust KYC or AML protocols, and maintains one of the world's most active and influential regulatory sandboxes. These leading developed countries can effectively serve as benchmark jurisdictions, helping to shape evolving global norms and technical standards in BC regulation while actively facilitating policy diffusion through their participation in key international bodies such as the Financial Action Task Force (FATF), the International Organization of Securities Commissions (IOSCO), and the Bank for International Settlements (BIS).

3. Low-Capacity Jurisdictions

In least developed countries (LDCs) or jurisdictions with limited regulatory capacity and resources, BC technology can still be strategically harnessed to drive significant public sector transformation in key areas such as land registries, the implementation of secure digital IDs, enhancing supply chain traceability, and facilitating more efficient and transparent remittances. These jurisdictions can effectively adopt a more flexible and less burdensome "light-touch" regulatory approach, placing a strong emphasis on active collaboration with regional organizations and international development partners. Key strategies for these jurisdictions include: Utilizing template-based regulations that are thoughtfully adapted from established international standards like those provided by FATF and MiCA, actively participating in regional regulatory alliances or multilateral regulatory sandboxes to gain experience and share best practices, and relying on technical assistance and capacity-building support from prominent global institutions such as the World Bank, the IMF, or the International Telecommunication Union (ITU).

For instance Nigeria, despite implementing a ban on cryptocurrency trading, has strategically deployed a central bank digital currency (eNaira) with the primary goal of promoting greater financial inclusion among its population. Rwanda and Ghana have actively partnered with international bodies to pilot innovative BC-based projects within their respective public sectors. This collaborative and adaptable model ensures that even countries with limited resources and evolving regulatory institutions can foster safe and beneficial innovation in the BC space without overwhelming their existing regulatory capacity.

4. Centrally-Planned Economies

China represents a distinct and unique regulatory archetype, characterized by a strong emphasis on central planning and a "security-first" approach to technological adoption. While the country has implemented a comprehensive ban on cryptocurrency trading and mining activities, it actively promotes the application of BC technology within state-sanctioned applications and has taken a leading global role in the rollout of central bank digital currencies. The People's Bank of China has successfully developed and piloted the digital yuan (e-CNY), which is now increasingly integrated into mainstream payment systems within the country. The state-backed BC Service Network (BSN) supports the development of cross-border digital infrastructure, albeit within a tightly controlled framework dictated by the state. While DeFi and public crypto assets remain strictly prohibited, BC technology is being actively utilized in various sectors, including judicial systems, tax services, and logistics management.

8.3. Formulation of policies and rules

The meticulous formulation of specific policies and rules will constitute the fundamental backbone of the overarching regulatory framework, providing clear guidance for BC asset operations across

diverse sectors of the economy. These policies must be developed through a process of broad and inclusive consultation with a wide range of relevant stakeholders, including regulatory bodies, technology experts, financial institutions, and importantly, the consumers who will be directly impacted by these regulations.

1. Clear Guidelines for Token Issuance and Trading

It is imperative to establish clear and unambiguous regulations governing the issuance and subsequent trading of BC-based tokens, with a particular focus on cryptocurrencies and NFTs. These regulations will need to mandate a high degree of transparency in all token offerings, ensuring that potential investors are provided with comprehensive and easily understandable information regarding the inherent risks involved. Token issuers may be required to disclose critical details such as the specific purpose of the token, its potential for value fluctuations, and a clear articulation of the risks associated with its acquisition and use.

2. Environmental Standards for BC Mining

The significant environmental footprint associated with certain BC technologies, particularly those relying on energy-intensive mining operations, has emerged as a major concern for policymakers and the public. Regulatory policies should proactively establish clear thresholds and standards for the energy consumption of BC networks, actively encouraging the widespread adoption of more energy-efficient consensus mechanisms, such as proof-of-stake, as alternatives to the more energy-intensive proof-of-work. Governments may also consider the implementation of carbon taxes or the imposition of caps on electricity consumption for mining operations as effective mechanisms to incentivize the adoption of more eco-friendly mining practices within the BC industry [45].

3. AML or KYC Regulations for Decentralized Platforms

Given that BC technologies often operate within inherently decentralized environments, the regulatory framework must develop and implement innovative mechanisms to effectively ensure compliance with established AML and KYC standards without unduly compromising the fundamental decentralized principles that underpin BC technology. Well-designed policies could potentially require decentralized exchanges (DEXs) and other decentralized platforms to implement identity verification procedures during critical phases of transactions, while still respecting users' privacy and autonomy to the greatest extent possible [18].

8.4. Validation analysis of the framework

To rigorously ensure the overall effectiveness and long-term viability of the proposed regulatory framework, a comprehensive validation analysis will be an absolutely necessary and ongoing undertaking. This critical analysis will systematically assess the framework's ability to successfully achieve its stated objectives, which include the significant reduction of fraudulent activities, the enhancement of financial stability within the digital asset ecosystem, and the robust promotion of consumer protection.

1. Impact Assessment

The initial and crucial step in the validation process involves a thorough evaluation of the potential impact of the regulatory framework on various key stakeholders. This includes a detailed analysis of the effects on consumers, BC companies operating within the regulated space, traditional financial institutions that may interact with BC assets, and the regulatory bodies themselves responsible for enforcement. This impact assessment can be effectively conducted through the use of carefully designed case studies, sophisticated simulations of market behavior under the new regulations, and well-executed pilot projects that test the practical

application of the framework in diverse jurisdictions. For example, countries that have already taken proactive steps to implement BC regulations, such as Estonia or Switzerland, could serve as valuable real-world test cases for validating the broader applicability and effectiveness of the proposed framework.

2. International Validation

Recognizing the global nature of BC technology, the validation process must also incorporate significant international cooperation to rigorously test the framework's applicability and effectiveness across different national contexts. Regulatory bodies from various countries can engage in collaborative efforts to evaluate how well the framework integrates with their existing local regulatory structures and to identify opportunities for further harmonization with evolving international standards. This collaborative approach would involve the active sharing of relevant data, valuable experiences gained during implementation, and identified best practices to iteratively refine and improve the framework over time [58].

3. Continuous Monitoring and Adaptation

The regulatory framework must be inherently flexible and possess the capacity to adapt proactively to the rapidly evolving nature of BC technology and its diverse applications. As new and innovative use cases and applications of BC emerge, regulators must maintain the agility to modify existing rules or introduce entirely new policies to effectively address any emerging risks or capitalize on newfound opportunities. Ongoing monitoring and rigorous evaluation, combined with periodic comprehensive reviews of the framework, will be essential to ensure that it remains relevant, effective, and fit-for-purpose in the face of continuous technological advancements within the BC space.

4. Feedback Mechanisms

It is absolutely essential to establish robust and accessible feedback loops that allow all relevant stakeholders to voice their concerns, share their practical experiences with the framework, and suggest potential improvements based on their insights. These crucial feedback mechanisms can be effectively incorporated through various channels, including regular public consultations, dedicated industry forums that bring together regulators and industry participants, and ongoing, open dialogue between regulatory authorities and industry leaders. The insights gained through these feedback mechanisms will be invaluable in ensuring the long-term effectiveness and legitimacy of the regulatory framework.

The journey towards a well-defined regulatory landscape for BC assets is a multifaceted and ongoing endeavor. The success of this endeavor hinges on a thoughtful and adaptive approach to implementation, a keen understanding of diverse global contexts, the meticulous crafting of clear and effective policies, and a commitment to rigorous validation and continuous improvement. By addressing these interconnected elements comprehensively, we can foster an environment that encourages responsible innovation, protects consumers, and ensures the long-term stability and integrity of the financial ecosystem in the age of decentralized technologies.

9. Conclusion and policy recommendations

This study explores the global regulatory landscape of BC assets, particularly cryptocurrencies and NFTs, with the objective of understanding policymakers' motivations and the challenges they face in crafting balanced governance. Employing a conceptual and mixed-method approach, it integrates qualitative and quantitative content analysis of 59 peer-reviewed sources selected using the PRISMA framework. The findings reveal that regulatory efforts are primarily driven by concerns over consumer protection, financial stability, AML or KYC compliance, tax transparency, and environmental sustainability.

Jurisdictional responses vary significantly, ranging from the EU's harmonized MiCA framework to the fragmented and enforcement-centric approach in the US, as well as diverse strategies across Asia. The study highlights key challenges, including the decentralized and borderless nature of BC assets, difficulties in legal classification, the rapid pace of technological change, and the tension between innovation and oversight. The contribution of this study is its comparative analysis of global regulatory approaches to BC assets, highlighting how jurisdictions like the EU, the US, and various Asian nations are addressing cryptocurrencies and NFTs. The research provides practical insights into the trade-offs between fostering innovation and ensuring consumer and market protection. It emphasizes the importance of international cooperation to prevent regulatory gaps and suggests tools such as regulatory sandboxes and industry self-regulation to enable safe experimentation. The study also proposes phased, adaptable models for developing countries, making the findings relevant across different legal and economic contexts. By integrating legal, technical, and policy perspectives, it offers a balanced foundation for designing effective and forward-looking BC regulations. However, this study is limited by its dependence on secondary sources, the lack of real-time data on policy outcomes, and the fast-paced evolution of BC technologies that can surpass current regulatory efforts. To address these gaps, future research should prioritize empirical studies and adaptive policy modeling to support more responsive and effective global governance of digital assets.

9.1. Policy recommendations

To effectively govern the rapidly evolving BC asset ecosystem while fostering innovation, financial integrity, and global trust, policymakers must adopt a holistic and adaptive regulatory framework. This framework should be grounded in a multi-layered strategy that balances innovation with oversight, enabling technology to thrive without compromising systemic safety, investor protection, or environmental sustainability. Firstly, regulatory approaches must emphasize harmonization and interoperability across borders, as the decentralized and transnational nature of BC assets, such as cryptocurrencies and NFTs, renders unilateral regulation insufficient. Global coordination through platforms such as the Financial Action Task Force (FATF), IMF, and Financial Stability Board (FSB) is critical to developing baseline global standards, especially for AML, KYC, and stablecoin reserve requirements.

Secondly, national governments should implement tiered regulatory models based on their technological maturity, market size, and legal traditions. Emerging economies may benefit from phased adoption strategies, starting with legal recognition of BC assets, introducing licensing for service providers, and launching regulatory sandboxes to foster innovation while monitoring risk. In contrast, developed economies must focus on strengthening compliance mechanisms, introducing stablecoin audits, integrating DeFi protocols into existing oversight systems, and mitigating systemic risks associated with mainstream adoption. Third, regulatory sandboxes should be institutionalized globally to enable real-time experimentation and collaboration between innovators and regulators. These environments would allow startups and developers to test products under controlled conditions, generating insights that inform flexible, forward-looking policy. Fourth, environmental sustainability must become a non-negotiable regulatory pillar. With increasing global concern about the energy consumption of Proof-of-Work-based BC, regulators should incentivize the shift toward greener consensus mechanisms like Proof-of-Stake (PoS), mandate carbon footprint disclosures, and explore energy-use taxation or credits for BC operators.

Fifth, tax compliance and reporting mechanisms need to be standardized internationally to prevent tax evasion and close regulatory

loopholes. Policymakers should enforce mandatory transaction reporting thresholds, automatic gains or loss tracking, and cross-border cooperation for digital asset tax enforcement. Sixth, asset classification clarity is essential. Policymakers must eliminate ambiguity in categorizing tokens as securities, commodities, or utility assets to prevent overlapping jurisdictions and provide legal certainty to businesses and investors. This includes establishing uniform definitions for NFTs, stablecoins, and DAO governance tokens. Seventh, regulators must embrace technological solutions for decentralized compliance, particularly in DeFi ecosystems. This may involve the deployment of zero-knowledge proofs for identity verification, AI-powered transaction monitoring systems, and integration of smart contract audit standards. Eighth, consumer protection must be central to all policy efforts, especially as retail participation in BC markets increases. Transparent disclosures, financial literacy programs, insurance schemes for exchanges and custodians, and redressal mechanisms for scams and fraud are essential for maintaining trust.

Ninth, data privacy and digital rights must be protected, even while enhancing transparency. Regulations should strike a balance by enforcing KYC requirements in centralized venues while promoting privacy-preserving tools in decentralized settings. Tenth, self-regulation should be encouraged through the formation of accredited industry bodies that establish ethical codes, dispute resolution systems, and voluntary compliance standards. This approach not only fosters accountability but also eases regulatory burdens by promoting industry alignment. Finally, to validate and evolve these regulatory efforts, policymakers must invest in continuous evaluation and capacity-building. This involves real-time data collection, longitudinal impact studies, and regular stakeholder engagement to ensure that regulations remain relevant, effective, and inclusive. Tailored training programs for regulators in low-capacity regions, cross-border pilot projects, and inclusive dialogue with developers and users should be institutionalized. In sum, only through a balanced mix of flexibility, coordination, innovation enablement, and stringent oversight can the global community craft a resilient, fair, and future-ready regulatory regime for BC assets that protects stakeholders, strengthens financial systems, and harnesses the full potential of decentralized technologies.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly in order to correct grammatical mistake. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Junaid Rahman: Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Conceptualization, Data curation, Formal analysis. **Hafizur Rahman:** Writing – review & editing, Writing – original draft, Visualization, Validation. **Naimul Islam:** Writing – review & editing, Writing – original draft. **Tipon Tan-changya:** Writing – review & editing. **Mohammad Ridwan:** Writing – original draft. **Mostafa Ali:** 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. Also, authors declare that they have no conflict of interest.

Appendix

Table 8

Table 8

List of abbreviations.

BC	Blockchain
NFTs	Non-Fungible Tokens
BCT	Bitcoin
Eth	Ethereum
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
DeFi	Decentralized Finance
KYC	Know Your Customer
AML	Anti-Money Laundering
CBDC	Central Bank Digital Currency
MiCA	Markets in Crypto-Assets Regulation
EU	European Union
USA	United State of America

References

- [1] Ahmad, N., & Zahid, F. (2024). *NFT market trends and digital asset trading: Navigating blockchain regulation and fintech investment strategies in China and the USA*. ResearchGate.
- [2] E.A. Akartuna, S.D. Johnson, A. Thornton, Preventing the money laundering and terrorist financing risks of emerging technologies: an international policy Delphi study, *Technol. Forecast. Soc. Change* 179 (2022) 121632.
- [3] P.C. Aksoy, The regulation of NFTs: much ado about nothing? *Bus. Law Rev.* 44 (4) (2023).
- [4] O. Ali, M. Momin, A. Shrestha, R. Das, F. Alhaji, Y.K. Dwivedi, A review of the key challenges of non-fungible tokens, *Technol. Forecast. Soc. Change* 187 (2023) 122248.
- [5] S. Alkadri, Defining and regulating cryptocurrency: fake internet money or legitimate medium of exchange? *Duke Law Technol. Rev.* 17 (2018) 71.
- [6] Almousa, M.H.S. (2024). *Blockchain revolution: how innovative technology can change the financial sector* (Doctoral dissertation, Vilnius universitetas).
- [7] M. Angel, Decoding cryptocurrency taxes: the challenges for estate planners, *Duke Law Technol. Rev.* 23 (2023) 137.
- [8] A.M. Antonopoulos, G. Wood, *Mastering Ethereum: Building Smart Contracts and Dapps*, O'reilly Media, 2018.
- [9] F. Armknecht, G.O. Karame, A. Mandal, F. Youssef, E. Zenner, Ripple: overview and outlook, in: *Trust and Trustworthy Computing: 8th International Conference*, 8, Springer International Publishing, 2015, pp. 163–180. TRUST 2015August 24–26, 2015, Proceedings.
- [10] Z. Asif, S. Unar, Cryptocurrency market dynamics: trends, volatility, and regulatory challenges, *Bull. Bus. Econ. BBE* 13 (1) (2024).
- [11] K. Bharanitharan, G. Kaur, Decentralized finance (DeFi) and legal challenges: navigating the intersection of innovation and regulation in the fintech revolution. *E-banking, Fintech, & Financial Crimes: The Current Economic and Regulatory Landscape*, Cham: Springer Nature Switzerland, 2024, pp. 155–167.
- [12] Bhasker, S., Grady, M.P., & Mosley, K.G. (2023). Cryptocurrency and anti-money laundering enforcement and regulation. *Criminal Justice*, 38(2), 3–11.
- [13] A.V.N. Biju, A.S. Thomas, Uncertainties and ambivalence in the crypto market: an urgent need for a regional crypto regulation, *SN Bus. Econ.* 3 (8) (2023) 136.
- [14] Birrer, T.K., Amstutz, D., & Wenger, P. (2023). Decentralized finance. *Financial Innovation and Technology*.
- [15] Business Standard. (2023). Losses from crypto fraud rose 45% to \$5.6 bn in 2023, Indians lose \$44 mn. Business Standard. Available at: https://www.business-standard.com/finance/personal-finance/losses-from-crypto-fraud-rose-45-to-5-6-bn-in-2023-124091000249_1.html.
- [16] Coin Law. (2024). Cryptocurrency regulations impact statistics 2025. Available at: <https://coinlaw.io/cryptocurrency-regulations-impact-statistics/>.
- [17] Coin 360. (2025). Crypto regulation recap: key changes in 2025. Available at, <https://coin360.com/news/crypto-regulation-recap-first-week>.
- [18] V.G. Comizio, Virtual currencies: growing regulatory framework and challenges in the emerging fintech ecosystem, *N. C. Bank. Inst.* 21 (2017) 131.
- [19] Eshan, B., Madhulika, B., Nautiyal, L., & Hooda, M. (2021). Deficiencies in blockchain technology and potential augmentation in cyber security. *Blockchain for Business: How It Works and Creates Value*, 251–293.
- [20] M.S. Ferdous, M.J.M. Chowdhury, M.A. Hoque, A survey of consensus algorithms in public blockchain systems for crypto-currencies, *J. Netw. Comput. Appl.* 182 (2021) 103035.
- [21] Friesendorf, C., & Blütener, A. (2023). Decentralized finance (DeFi).
- [22] Federal Bureau of Investigation. (2023). 2023 IC3 cryptocurrency report. Internet Crime Complaint Center. Available at: [https://www.ic3.gov/AnnualReport/Report s/2023_IC3CryptocurrencyReport.pdf](https://www.ic3.gov/AnnualReport/Report%2023_IC3CryptocurrencyReport.pdf).
- [23] J. Golosova, A. Romanovs, The advantages and disadvantages of the blockchain technology, in: *Proceedings of the 2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, IEEE, 2018, pp. 1–6.
- [24] S. Gottschalk, International Financial regulation of cryptoassets and asset-backed tokens. *Fintech, Pandemic, and the Financial System: Challenges and Opportunities*, Emerald Publishing Limited, 2023, pp. 145–181.
- [25] R. Grassman, V. Bracamonte, M. Davis, M. Sato, Attitudes to cryptocurrencies: a comparative study between Sweden and Japan, *Rev. Socionetw. Strateg.* 15 (1) (2021) 169–194.
- [26] J. Galavis, Blame it on the blockchain: cryptocurrencies boom amidst global regulations, *Univ. Miami Int. Comp. Law Rev.* 26 (2018) 561.
- [27] Grossi, L. (2021). NFTs and metaverse. An analysis from the point of view of contemporary art and design.
- [28] P. Hacker, C. Thomale, Crypto-securities regulation: iCOs, token sales and cryptocurrencies under EU financial law, *Eur. Co. Financ. Law Rev.* 15 (4) (2018) 645–696.
- [29] Han, Y. (2023). Dynamic games for resource allocation in metaverse services and architectures.
- [30] K.A. Houser, J.T. Holden, Navigating the non-fungible token, *Utah Law Rev.* (2022) 891.
- [31] P. Howson, A. de Vries, Preying on the poor? Opportunities and challenges for tackling the social and environmental threats of cryptocurrencies for vulnerable and low-income communities, *Energy Res. Soc. Sci.* 84 (2022) 102394.
- [32] M.L.F. Jumaili, S.M. Karim, Comparison of two cryptocurrencies: bitcoin and Litecoin, *J. Phys. Conf. Ser.* 1963 (1) (2021) 012143. Vol.
- [33] K. Karisma, P. Moslemzadeh Tehrani, Blockchain: legal and regulatory issues. *Sustainable Oil and Gas Using Blockchain*, Springer International Publishing, Cham, 2023, pp. 75–118, pp.
- [34] C. King, C. Walker, in: J. Gurulé (Ed.), *The Palgrave handbook of Criminal and Terrorism Financing Law*, Cham: Palgrave Macmillan, 2018, pp. 1–1247.
- [35] K. Karisma, P. Moslemzadeh Tehrani, Blockchain: legal and regulatory issues. *Sustainable Oil and Gas Using Blockchain*, Springer International Publishing, Cham, 2023, pp. 75–118.
- [36] KPMG. (2025). Crypto regulatory round up. Available at: <https://kpmg.com/xx/en/our-insights/regulatory-insights/crypto-regulatory-round-up-january-2025.html>.
- [37] M. Kavaloski, A global crypto code of conduct: crafting and internationally centralized regulatory body for a decentralized asset, *Vanderbilt J. Transnatl. Law* 57 (2024) 301.
- [38] E. Lee, NFTs as decentralized intellectual property, *Univ. Ill. Law Rev.* (2023) 1049.
- [39] Luzan, A. (2023). Art provenance yesterday, today, and tomorrow with a particular focus on blockchain technology.
- [40] P. Maume, M. Fromberger, Regulations of initial coin offerings: reconciling US and EU securities laws, *Chic. J. Int. Law* 19 (2018) 548.
- [41] M. Mazur, E. Polyzos, Non-fungible tokens (NFTs). *The Elgar Companion to Decentralized Finance, Digital Assets, and Blockchain Technologies*, Edward Elgar Publishing, Cham, 2023, pp. 280–297.
- [42] H.B. Meier, J.E. Marthinsen, P.A. Gantenbein, S.S. Weber, *The Swiss banking system. Swiss Finance: Banking, Finance, and Digitalization*, Springer International Publishing, Cham, 2023, pp. 63–157.
- [43] Metcalfe, W. (2020). Ethereum, smart contracts, DApps. *Blockchain and Crypt Currency*, 77, 77–93.
- [44] A. Mosna, G. Soana, NFTs and the virtual yet concrete art of money laundering, *Comput. Law Secur. Rev.* 51 (2023) 105874.
- [45] N. Mirjanich, Digital money: bitcoin's financial and tax future despite regulatory uncertainty, *DePaul Law Rev.* 64 (2014) 213.
- [46] Packin, N.G., & Volovelsky, U. (2023). Digital assets, anti-money laundering and counter financing of terrorism: an analysis of evolving regulations and enforcement

- In the era of NFTs. The Cambridge Handbook On Law and Policy for NFTs, (N.G. Packin, ed.), Forthcoming.
- [47] M. Pilkington, *Blockchain technology: principles and applications*. Research Handbook on Digital Transformations, Edward Elgar Publishing, 2016, pp. 225–253.
- [48] Z. Poposki, Critique of reification of art and creativity in the digital age: a Lukácsian approach to AI and NFT art, *Open Philos.* 7 (1) (2024) 20240027.
- [49] Reuters. (2025). EU markets watchdog warns of crypto-related financial stability risks. Available at: <https://www.reuters.com/technology/eu-markets-watchdog-warns-crypto-related-financial-stability-risks-2025-04-08/>.
- [50] Reuters. (2025). Wall Street's regulation wish list plays with fire. Available at: <https://www.reuters.com/breakingviews/wall-streets-regulation-wish-list-plays-with-fire-2025-01-22/>.
- [51] D. Shin, M. Ibahrine, The socio-technical assemblages of blockchain system: how blockchains are framed and how the framing reflects societal contexts, *Digit. Policy Regul. Gov.* 22 (3) (2020) 245–263.
- [52] Y. Takanashi, S.I. Matsuo, E. Burger, C. Sullivan, J. Miller, H. Sato, Call for multi-stakeholder communication to establish a governance mechanism for the emerging blockchain-based financial ecosystem, *Stanf. J. Blockchain Law Policy* 3 (2020) 1.
- [53] M. Todorović, L. Matijević, D. Ramljak, T. Davidović, D. Urošević, T. Jakšić Krüger, D. Jovanović, Proof-of-useful-work: blockchain mining by solving real-life optimization problems, *Symmetry* 14 (9) (2022) 1831.
- [54] Topali, T. (2024). Digitality and museums; the benefits, the challenges and a case of a non-digital museum.
- [55] Talani, L.S. (2018). Globalization, money laundering and the city of London. The Palgrave Handbook of Criminal and Terrorism Financing Law, 57–79.
- [56] L.J. Trautman, Bitcoin, virtual currencies, and the struggle of law and regulation to keep peace, *Marquette Law Rev.* 102 (2018) 447.
- [57] J. Truby, Decarbonizing bitcoin: law and policy choices for reducing the energy consumption of Blockchain technologies and digital currencies, *Energy Res. Soc. Sci.* 44 (2018) 399–410.
- [58] S. Tayebi, H. Amini, The flip side of the coin: exploring the environmental and health impacts of proof-of-work cryptocurrency mining, *Environ. Res.* (2024) 118798.
- [59] Wang, Q., Li, R., Wang, Q., & Chen, S. (2021). Non-fungible token (NFT): overview, evaluation, opportunities and challenges. arXiv preprint arXiv:2105.07447.
- [60] Xiong, X., & Luo, J. (2024). Global trends in cryptocurrency regulation: an overview. arXiv preprint arXiv:2404.15895.
- [61] K. Yeung, Regulation by blockchain: the emerging battle for supremacy between the code of law and code as law, *Mod. Law Rev.* 82 (2) (2019) 207–239.
- [62] J. Yli-Huuma, D. Ko, S. Choi, S. Park, K. Smolander, Where is current research on blockchain technology?—A systematic review, *PLoS One* 11 (10) (2016) e0163477.
- [63] N. Yuhaniha, R. Robiyanto, Cryptocurrencies as a hedge and safe haven instruments during Covid-19 Pandemic, *JASF* 4 (1) (2021) 13–30.
- [64] S. Zheng, Y. Hu, A.Y.L. Chong, C.W. Tan, Leveraging blockchain technology to control contextualized business risks: evidence from China, *Inf. Manag.* 59 (7) (2022) 103628.
- [65] Basel Institute on Governance. (2023). Basel AML Index 2023: 12th Public Edition Ranking money laundering and terrorist financing risks around the world. Available at: <https://baselgovernance.org/sites/default/files/2023-11/Basel%20AML%20Index%202023%2012th%20Edition.pdf>.



Review Article

Ethical and regulatory challenges in machine learning-based healthcare systems: A review of implementation barriers and future directions[☆]

Shehu Mohammed^{*,} , Neha Malhotra

School of Computer Applications, Lovely Professional University, 14411, India

ARTICLE INFO

Keywords:

Artificial intelligence (AI) Ethics
Algorithmic bias
Explainable AI (XAI)
Machine learning (ML) in healthcare
Patient data privacy
Regulatory compliance (GDPR, HIPAA, FDA)

ABSTRACT

Machine learning significantly enhances clinical decision-making quality, directly impacting patient care with early diagnosis, personalized treatment, and predictive analytics. Nonetheless, the increasing proliferation of such ML applications in practice raises potential ethical and regulatory obstacles that may prevent their widespread adoption in healthcare. Key issues concern patient data privacy, algorithmic bias, absence of transparency, and ambiguous legal liability. Fortunately, regulations like the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the FDA AI/ML guidance have raised important ways of addressing things like fairness, explainability, legal compliance, etc.; however, the landscape is far from risk-free. AI liability is another one of the gray areas approaching black, worrying about who is liable for an AI medical error — the developers, the physicians, or the institutions. The study reviews ethical risks and potential opportunities, as well as regulatory frameworks and emerging challenges in AI-driven healthcare. It proposes solutions to reduce bias, improve transparency, and enhance legal accountability. This research addresses these challenges to support the safe, fair, and effective deployment of ML-based systems in clinical practice, guaranteeing that patients can trust, regulators can approve, and healthcare can use them.

Introduction

Background: Machine learning (ML) is fundamentally transforming healthcare; ML is playing an integral role in the development of methods for early diagnosis and treatment optimization, as well as predictive analytics providing unprecedented improvements for medical decision-making [1]. ML applications that leverage this technology to achieve better healthcare outcomes include diagnostic imaging analysis, personalized treatment recommendations, and predictive modeling for disease progression [2]. These technologies can help minimize human error, enable real-time decision-making, and optimize resources used in clinical practices.

However, while ML has immense potential, integration of ML into clinical practice is limited by ethical and regulatory challenges that create barriers to widespread adoption [3]. Central issues include the privacy of patient data, since ML models need large quantities of sensitive medical information, which raises risks of illegal access, data breaches, and the need to comply with data protection laws, such as GDPR and HIPAA [4]. The algorithmic bias and fairness issues also

intertwine with these dynamics, with ML models trained on unbalanced datasets delivering results that discriminate against certain groups [5]. Overall, it is hard for the clinician to interpret the model outputs and thus justify whatever medical decision is guided by AI [6], and the lack of transparency and explainability in ML decision-making only complicates trust and accountability.

Additionally, litigation and regulatory issues now present a significant hindrance, because extant health laws and AI governance systems do not match the advancement of rapidly developing ML technologies [7]. Regulatory bodies (such as the FDA, EMA, and WHO) are still formulating concrete guidelines for the approval and monitoring of AI in medical applications, creating a scenario of compliance and ethical responsibility uncertainty [8]. Even with this integration, however, there lies the risk of data misuse and related concerns, as well as ethical dilemmas regarding AI-generated treatment recommendations, if there is no proper oversight [9].

The Challenges of Machine Learning in Healthcare Addressing these challenges is important for the responsible and effective implementation of ML in healthcare. The focus of this evaluation will identify the

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

^{*} Corresponding author.

E-mail addresses: mohammedshehumafara@gmail.com (S. Mohammed), neha.16982@lpu.co.in (N. Malhotra).

<https://doi.org/10.1016/j.tbench.2025.100215>

Received 12 March 2025; Received in revised form 30 April 2025; Accepted 14 May 2025

Available online 28 May 2025

2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1
Comparative analysis of global approaches to AI regulation in healthcare.

Regulation	Region	Focus Areas	Strengths	Limitations	Impact on Healthcare AI
General Data Protection Regulation (GDPR), 2018	European Union (EU)	Data privacy, patient consent, and AI transparency	<ul style="list-style-type: none">- Strongest global data protection framework.- Enforces patient rights over their medical data.- AI systems must be explainable.	<ul style="list-style-type: none">- Strict compliance can slow AI innovation.- Heavy penalties for violations (up to €20 million).	<ul style="list-style-type: none">- AI-driven healthcare must ensure patient consent & data security.- Limits how ML models store and process medical records.
Health Insurance Portability and Accountability Act (HIPAA), 1996	United States (USA)	Patient data protection and security standards	<ul style="list-style-type: none">- Ensures strong security for electronic health data (ePHI).- Mandates breach notifications.- Applies to healthcare providers & AI developers.	<ul style="list-style-type: none">- Does not cover AI-specific risks.- No strict explainability requirements for AI decisions.	<ul style="list-style-type: none">- AI in healthcare must comply with security protocols.- Telemedicine and AI diagnostics require secure data storage.
FDA Guidelines on AI/ML in Medical Devices, 2021	United States (USA)	AI-based medical devices, real-world performance monitoring	<ul style="list-style-type: none">- AI software must be approved before clinical use.- Supports adaptive AI models that improve over time.	<ul style="list-style-type: none">- Lengthy approval process can delay AI deployment.- Limited global influence outside the USA.	<ul style="list-style-type: none">- AI-driven radiology & diagnostics require FDA approval.- Ensures AI models meet safety & accuracy standards.
European Medicines Agency (EMA) AI Regulations	European Union (EU)	AI-driven drug development and medical applications	<ul style="list-style-type: none">- AI in drug discovery & clinical trials is regulated.- Post-market AI monitoring ensures patient safety.	<ul style="list-style-type: none">- High compliance costs for AI companies.- Lack of harmonization with non-EU regulations.	<ul style="list-style-type: none">- AI in pharmaceutical research & precision medicine must meet EMA guidelines.- Requires real-world validation of AI performance.
Artificial Intelligence Act (AI Act) (Proposed), 2023	European Union (EU)	Risk-based regulation for AI applications, including healthcare	<ul style="list-style-type: none">- Strict transparency rules for AI models.- Classifies AI as low-risk, high-risk, or banned.- Ensures fairness and non-discrimination in AI decisions.	<ul style="list-style-type: none">- Not yet fully implemented (expected 2025+).- Some AI applications may be overregulated.	<ul style="list-style-type: none">- AI-driven clinical decision support systems (CDSS) will require higher transparency.- AI in high-risk medical settings (e.g., surgery, diagnostics) faces stricter review.
China's AI Ethics & Security Guidelines, 2022	China	AI security, ethical AI use, and national AI development strategy	<ul style="list-style-type: none">- Encourages AI innovation in healthcare.- Focuses on AI ethics, fairness, and explainability.	<ul style="list-style-type: none">- Government-led AI oversight raises privacy concerns.- Lack of clear penalties for AI misuse.	<ul style="list-style-type: none">- AI in hospitals & medical research is state-regulated.- Supports AI-based drug discovery & smart hospitals.
UK NHS AI Strategy	United Kingdom (UK)	AI-driven healthcare transformation and patient safety	<ul style="list-style-type: none">- AI models must be clinically validated before NHS deployment.- Emphasis on data security & patient trust.	<ul style="list-style-type: none">- No centralized AI regulation (varies across NHS Trusts).- Limited penalties for AI-related errors.	<ul style="list-style-type: none">- AI clinical trials & patient monitoring systems must meet NHS AI standards.- Supports AI-assisted radiology & diagnostics.
South Africa – Draft AI Policy, 2022	South Africa	Ethics, transparency, inclusion, and public sector AI	<ul style="list-style-type: none">- A human rights-based approach promotes inclusive AI	<ul style="list-style-type: none">- Still under development, not legally binding yet	<ul style="list-style-type: none">- To prevent algorithmic discrimination and enhance equitable AI deployment in healthcare
Brazil – LGPD (Lei Geral de Proteção de Dados), 2020	Brazil	Personal data protection, informed consent, and accountability	<ul style="list-style-type: none">- Modeled after GDPR, legally enforceable	<ul style="list-style-type: none">- Limited AI-specific clauses; interpretation varies	<ul style="list-style-type: none">- Encouraged responsible AI use and stronger consent mechanisms in health tech
India – NDHM & Digital Personal Data Protection Act, 2023	India	Patient data control, digital health ID, AI ethics in health services	<ul style="list-style-type: none">- National health architecture, patient-centric model	<ul style="list-style-type: none">- Implementation challenges, rural digital divide	<ul style="list-style-type: none">- Provides a foundation for AI-based diagnostics and personalized care through regulated digital platforms

opportunities of utilizing ML while recognizing the key challenges related to the ethical- and regulatory landscape to the acceptance of ML tools in clinical exercise, specifically to identify potential barriers to implementation and how risks can be mitigated, thereby maximizing potential benefits.

Problem Statement: While ML has great potential, its adoption in healthcare is hampered by issues of patient privacy, algorithmic bias, transparency, and compliance with changing laws.

Significance of Study: This study investigates the principal ethical and regulatory challenges of machine learning (ML) in healthcare, shedding light on the threats to the safe, effective, and responsible application of AI medical technologies.

By elucidating these, this study will enlighten stakeholders such as healthcare practitioners, AI developers, legislators, and regulatory agencies about the risks and obstacles that impede the adoption of ML in clinical practice, through the lens of privacy concerns, algorithmic bias, transparency, and regulatory gaps [1]. Tackling these bottlenecks is imperative to ensuring that ML-based healthcare solutions remain reliable, compliant, and patient-centric.

Finally, this study aims to offer actionable suggestions for enhancing

AI governance, data security, and bias mitigation in ML models while facilitating compliance with existing healthcare regulation frameworks, including GDPR, HIPAA, and the FDA's AI/ML-driven frameworks [7,8]. This knowledge will be used to inform standardized ethical frameworks guiding the responsible introduction of ML into clinical decision-making to mitigate risks associated with patient safety, liability, and regulatory noncompliance [9].

Problem Statement

Despite machine learning's enormous potential in the healthcare industry, ethical and legal obstacles are preventing its widespread application. Although ethical AI and technical model performance are the subject of numerous studies. The present review is designed to address the following problems:

1. Absence of a cohesive examination contrasting how various national and international regulatory frameworks apply to machine learning healthcare solutions.

2. Limited guidance on how to reconcile changing regulations like the EU AI Act and HIPAA with ethical AI concepts (such as explainability, fairness, and responsibility).
3. Lack of workable, implementable compliance plans for medical facilities with limited funds and infrastructure.
4. Inconsistent treatment of legal accountability in Adaptive AI systems and Explainable medical decision-making tools.
5. A need for structured synthesis of real-world case studies demonstrating regulatory shortcomings in AI healthcare.

Scope and Delimitations

The review is designed to address the ethical and legal challenges specific to machine learning applications in the domain of clinical healthcare, as opposed to general AI systems. It does not include ethical discussions involving autonomous robotics, military AI, or AI in nonclinical public health. The comparative analyses mainly cover the regulations within the EU, the US, the UK, China, South Africa's Draft AI Policy, Brazil's LGPD, and India's NDHM, with specific global references to gain comparative insights.

Gaps in Existing Research Studies on AI Governance in Healthcare

Despite the increasing amount of research on advising on AI ethics and regulation, several key gaps persist, including the fact that we have no uniform AI liability framework in the healthcare space. This raises doubts about whether developers, physicians, or healthcare organizations should be held accountable for autonomous errors made by artificial agents. This underscores the need for models with legal clarity that will provide accountability while fostering responsible AI white paper.

Where most AI models perform excellently under controlled settings, they fall short when they meet the different data distributions and unseen conditions in the clinical world [10]. Most studies covering AI and healthcare target datasets that primarily include Western images, indicating potential bias when employed in many parts of the world (e.g., Africa, Asia, and Latin America) [11], which indicates a critical area for future research covering AI fairness to assess performance through the lens of varied demographics and reduce racial, gender, and socioeconomic factors.

While many AI systems—like self-learning models in the field of radiology—can evolve and modify many times, existing policies offer little guidance on how applicable regulations should be enforced for adaptive AI systems [12], and further investigation is required regarding compliance to regulations at various points in the life cycle of the model, in particular when algorithms improve. Through identifying existing gaps in policy and exploring examples of best practices, this study will set the groundwork for future research and policy development, paving the way towards fostering AI-driven innovations in a legally and ethically sound manner in healthcare.

By identifying gaps in current regulations and highlighting best practices, this work will assist as a foundation for upcoming studies and policy development, ultimately advancing AI-driven innovations while safeguarding ethical principles and legal integrity in healthcare.

Contributions of the review

1. Integrative Ethical-Regulatory Lens: Maps ethical AI principles (e.g., fairness, explainability, accountability) to legal obligations (e.g., GDPR, FDA, HIPAA, and EU AI Acts)
2. Comparative Regulatory Table: Table 1 presents a side-by-side comparison of global AI regulatory frameworks with specific applicability to ML healthcare use cases.
3. Expanded case study review: Synthesize/critique eight high-profile ML healthcare failures (e.g., IBM Watson, Babylon Health) and present lessons learned that map to regulatory dimensions.

4. Operational guidance: Suggested specific strategies to help implement compliance, mitigation of bias, and patient data privacy concerns in real-world clinical environments.
5. Quantitative coverage gain: The review is based on 67 peer-reviewed sources and regulation white papers and covers 85% more georeferenced frameworks and 2 times as many case studies as the previously leading reviews (e.g., [7,8]).

Research objectives

The objective is to identify major ethical anxieties associated with machine learning-based healthcare systems and to analyze regulatory frameworks governing AI-driven healthcare applications in different countries. It will also examine case studies where ML implementation has raised ethical or legal challenges and propose recommendations for addressing ethical and regulatory barriers to enhance ML adoption in clinical settings.

Literature review outline

Overview of machine learning in healthcare

Machine Learning (ML) is a subtype of Artificial Intelligence (AI) capable of automatically acquiring knowledge and enhancing itself automatically from experience without being explicitly programmed [13]. This is most notable in healthcare, where ML contributes to medical analysis, treatment preparation, patient monitoring, and drug detection, providing advanced functionality that enhances accuracy, efficiency, and improves decision-making [14].

Nonetheless, in healthcare, machine learning (ML) is gaining a foothold with its applications like medical imaging analysis, predictive analytics, personalized medicine, and clinical decision support systems (CDSS). Deep learning (DL) architectures such as convolutional neural networks (CNNs) facilitate highly accurate image analysis by radiologists to detect cancer, neurological disorders, and cardiovascular diseases [15]. Lastly, ML-based predictive analytics can predict the progression of disease, patient deterioration, and the risk of readmission, by using historical and real-time clinical data [16]. ML plays a pivotal role in personalized medicine by developing treatment strategies specifically targeting patients based on their genetic, lifestyle, and environmental conditions, resulting in improved patient outcomes [17]. Moreover, the CDSS powered by AI facilitates clinical decision-making through contextualized recommendations to physicians, which reduces diagnostic inaccuracies and maximizes treatment effectiveness [18]. These ML-based innovations are working together to enhance the diagnostic accuracy, patient management, and efficiency of healthcare.

Current trends in AI-driven healthcare innovation

The swift integration of artificial intelligence (AI) and machine learning (ML) in the healthcare sector is powered by progress in technology, enhanced computing capabilities, and greater accessibility of medical data. Key trends that are emerging in this landscape include:

Explainable AI (XAI) for Trustworthy Healthcare AI

As concerns about the lack of transparency in black-box AI models grow, the importance of Explainable AI (XAI) is on the rise. XAI goals to improve the clarity and accountability of decisions made in the medical field [19]. Techniques like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are proving valuable in helping healthcare professionals make sense of and have confidence in AI-generated recommendations [20].

Multi-Modal AI for Holistic Patient Insights

Modern ML models take an integrated view of multiple data sources, both genomic and other modalities ranging from medical imaging to electronic health records (EHRs) and wearable device data [14]. AI systems derived from IBM Watson and Google DeepMind's algorithms

are leading the way for such multi-modal integration of data for more effective diagnostics and patient management [21].

Federated Learning for Privacy-Preserving AI in Healthcare

Thus, Federated Learning (FL), a dispersed mechanism that lets multiple hospitals to jointly train machine learning (ML) models without the need to share patient data sensitive under GDPR and HIPAA regulations [22] is a good solution for this kind of issue. Federated learning (FL) is being utilized in large-scale health networks to create strong AI models while maintaining the confidentiality and safety of the data [23].

AI-Driven Drug Discovery and Virtual Clinical Trials

Machine learning is transforming the process of drug discovery by accurately forecasting how molecules interact, fine-tuning compound formulations, and shortening the time needed for clinical trials [24]. Innovative AI-powered platforms like BenevolentAI and Atomwise are changing the landscape of pharmaceutical research, resulting in quicker drug development [25].

AI-Assisted Robotic Surgery and Automated Diagnostics

Robotic surgery systems that utilize artificial intelligence, like the da Vinci Surgical System, enhance the accuracy of surgical procedures while minimizing risks associated with them [26]. Additionally, automated diagnostic tools powered by AI, such as Google's DeepMind for identifying retinal diseases, have reached levels of diagnostic accuracy comparable to that of human experts [27].

Ethical challenges of machine learning in healthcare

Patient privacy & data security in machine learning-based healthcare

Risks of Data Breaches and Unauthorized Access to Patient Records

Healthcare ML models require access to large volumes of sensitive patient data such as EHRs, genomic data, and medical imaging. ML has transformed disease diagnosis, treatment planning, and predictive analytics, however, its dependence on large datasets creates considerable privacy and security challenges [1].

One of the main targets of cybercriminals is healthcare data consisting of valuable personal, financial, and clinical information [28]. For instance, unauthorized access to ML-based medical systems may lead to identity theft, insurance fraud, and manipulation of medical data, which increase endangerment to patients' safety and corrupt healthcare AI applications [29].

Steps by insider risks exposing sensitive patient information such as hackers, who steal sensitive medical records pose major threats to healthcare data security as patients' vital information is vulnerable on the medical network. Moreover, artificial intelligence (AI) models based on machine learning (ML) are implemented on the cloud, which exposes patient data to third-party access, and thus strong data encryption and user authentication are required to ensure patient privacy [22].

Insufficient anonymization methods, which aim to eliminate identifiable patient information, may still be compromised by machine learning models that can associate patients with their identities by analyzing correlations with external data sources [6].

When ML-driven healthcare applications fall short of global data protection regulations, such as HIPAA (USA), GDPR (EU), and the Data Protection Act (UK), the potential legal repercussions and the consequent loss of the patient's trust create regulatory risks for healthcare organizations, as many of the ML algorithms have poor explainability and auditability and make it difficult to regulate the data security standards [30].

Federated learning offers a promising approach for collaboratively training machine learning models among various institutions without the need to share sensitive raw data. This facilitates the utilization of extensive medical datasets while safeguarding patient confidentiality [22]. Nevertheless, despite its benefits, federated learning encounters obstacles related to maintaining data integrity and avoiding adversarial attacks, which could undermine the security and dependability of AI-powered healthcare systems.

Mitigation Strategies for Ensuring Patient Data Privacy & Security

Protecting your sensitive medical data is paramount. We employ cutting-edge security measures, including robust encryption, multi-factor authentication (MFA), and privacy-preserving machine learning techniques, all while maintaining full regulatory compliance. This ensures the highest level of data security and patient privacy.

Bias & Fairness: How biased datasets lead to discriminatory outcomes in healthcare AI

Machine learning (ML) bias refers to when models generate systematically inequitable results stemming from uneven, incomplete, or nonrepresentative datasets [31]. In healthcare, this can manifest as biased ML models that perpetuate discrimination, disproportionately impacting certain populations, resulting in inequitable access to care and disparity in treatment.

Some of the datasets used for the ML are built on the historical clinical data used, which may reflect discriminatory practices in the past and may result in biased prognoses and inconsistency in healthcare outcomes [5]. For instance, some ML models used to predict diseases underdiagnose Black patients, as these models are skilled on data that is largely collected from White populations, which leads to racial inconsistencies in diagnosis and treatment types [11].

Sampling bias arises when ML models are trained on imbalanced datasets, which results in a loss of generalization across diverse demographics [32]. For example, an AI model skilled predominantly on male or high-income patient data set may have trouble providing accurate diagnosis and treatment recommendations for female or lower-income groups, leading to healthcare disparities and misdiagnoses. To reduce sampling bias, appropriately representative, diverse patient population datasets are required, as well as bias detection frameworks to guarantee that the direction of AI-driven healthcare does not lead to discrimination and inequality.

Sampling bias occurs when ML models are trained on datasets that lack diversity, leading to poor generalization across different demographic groups [32]. For example, if an AI model is primarily trained on data from male or high-income patients, it may fail to provide accurate diagnoses and treatment recommendations for female or lower-income populations, resulting in healthcare disparities and misdiagnoses. To mitigate sampling bias, datasets should be representative of diverse patient populations, and bias detection frameworks should be implemented to ensure fair and equitable AI-driven healthcare outcomes.

Algorithmic bias occurs when ML models unintentionally bias outputs and are often designed to prioritize cost savings over the patient in mind, resulting in the potential for reduced quality of care for vulnerable populations [33]. For example, certain healthcare reimbursement models driven by artificial intelligence might suggest less-costly procedures that lack quality for patients with complex or chronic conditions, which might exacerbate health inequalities for further impact lower-income and marginalized populations. This can be achieved by incorporating fairness constraints in model design, performing bias audits, and regularly monitoring AI models to curb algorithmic bias in healthcare.

Consequences of Bias in AI-Driven Healthcare

Some critical consequences of biased machine learning models in health care are delayed or incorrect diagnoses for underrepresented populations, inequities in the distribution of health care resources such as hospital admissions and insurance approvals, and a loss of patient trust in medical decision-making enabled by artificial intelligence.

Mitigating Bias in Healthcare AI

Ensuring heterogeneous data collection by demographic characteristics such as gender, race, and socioeconomic status; conducting bias auditing and testing for model fairness using tools such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) to detect biased output [34]; and establishing regulatory oversight through bias-reduction policies, algorithm audits,

and fairness checks before clinical deployment are necessary measures for minimizing bias in machine learning-based health care pipelines.

Transparency & explainability in ML healthcare decisions

A variety of ML models, most notably Deep Learning (DL) algorithms, operate as black boxes, which makes it hard for clinicians to interpret their decision-making processes [35], and thus decreases trust, proof, and explanation of AI-assisted medical recommendations [36].

Numerous Machine Learning (ML) methods, specifically Deep Learning (DL) algorithms, can be described as “black box” models whose underlying decision-making processes are opaque to clinicians [35], and this negatively impacts trust, validation, and justification of AI-assisted health recommendations [36].

Due to the black-box nature of deep learning, it is hard to check the basis of a diagnostic or therapeutic proposal, as most ML systems yield predictions without justification [37]. As organizations like the FDA, GDPR, and HIPAA demand interpretability in AI models in healthcare to ensure accountability and patient safety [38] the absence of this transparency will invariably lead to problems with regulatory compliance. Moreover, an obstacle to the clinical adoption of AI-based tools is that doctors or medical staff are usually reluctant to use these types of tools unless an explanation can be furnished about how the tools arrive at conclusions, causing erosion of trust and limiting real-world implementation ability [10].

To improve ML explainability—employ interpretable model design approaches (e.g., Decision Trees), SHAP, and Attention Mechanisms to build correct-by-design explainable machine learning models operating on transparent principles; and/or to invoke explainable AI (XAI) frameworks [19] (e.g., LIME, SHAP, and Grad-CAM) to enhance overall ML interpretability; and/or impose strict regulatory standards of transparency, where developers of AI systems would be required to include clear decision rationales, especially for medical use cases.

Accountability & liability in AI-driven medical errors

Who is Responsible When AI Makes a Mistake?

When systems that use machine learning (ML) make wrong diagnoses or treatment decisions, the legal assignment of responsibility becomes especially challenging, generating questions about whether Doctors should be held responsible for making an error if they had relied on AI, whether AI developers—including ML engineers and data scientists—should be held responsible for generating biased or erroneous predictions, and whether hospitals and other healthcare institutions should assume legal liability for AI-related diagnoses [39].

Challenges in AI Accountability

In such a system, it becomes challenging to determine who is liable when AI systems make decisions; this is because AI models are probabilistic and cannot be directly correlated with concrete laws [9]. Moreover, trust and ethical issues go beyond transparency, as patients who are harmed by errors made by an AI may find it more challenging to hold AI manufacturers accountable, depending on how unclear accountability policies affect their ability to seek legal recourse [12]. Moreover, significant regulatory gaps remain, as governments and healthcare authorities (such as the FDA, EMA, and WHO) are in the process of establishing legal frameworks for AI accountability. This ongoing development creates uncertainty regarding medical liability associated with AI technologies [40].

Potential Solutions for AI Accountability

To achieve AI accountability in healthcare, clear AI liability laws should define the responsibility for AI-driven medical errors; Human-in-the-Loop (HITL) AI models should be mandated—this would force physicians to review an AI’s suggestion/diagnosis; algorithm transparency and explainability should be enforced—ensuring AI models offer strong rationales for their decisions, supporting legal accountability.

Regulatory frameworks for AI in healthcare

Overview of Major Regulations Governing AI-Driven Healthcare

Artificial Intelligence (AI) and Machine Learning (ML) usage in healthcare is on the rise, and so are the laws and regulations surrounding them. Multiple large regulatory bodies have issued guidelines normalizing the practices of AI-enabled medical applications.) Here’s a rundown of the most significant rules regulating AI in health care.

General data protection regulation (GDPR) – European Union (EU)

The General Data Protection Regulation (GDPR) is a significant data privacy regulation introduced by the European Union (EU) in 2018. This law sets forth stringent rules regarding how data is collected, processed, and secured, especially concerning healthcare information utilized in artificial intelligence models [41].

Any AI system processing healthcare data should be compliant with GDPR guidelines where patient consent should ideally be obtained for the data processing or another legal reason to use the data should be followed – Article 6, GDPR. Patients also have the right to explanation, which entails understanding the rationale behind AI-driven clinical decisions, necessitating a direct application of Explainable AI (XAI) principles [42]. Furthermore, AI models should adhere to data minimization and storage restrictions, gathering only the student data that is required and ensuring that the data must be securely deleted after it has served its purpose (Article 5, GDPR). Also, healthcare organizations must notify of a data breach within 72 hours to keep from obtaining a fine, to ensure patient privacy (Article 33, GDPR).

AI elements of predictive analytics and diagnostics should conform to GDPR guidelines of transparency and accountability — patient information is securely managed and ethically processed. Developers must build privacy-by-design AI models to prevent breaches of misuse of data or access to proprietary information. Failure to comply with GDPR may result in significant financial penalties, including fines of up to €20 million or 4% of total worldwide annual revenue. Example Case: Google’s DeepMind Health AI faced GDPR scrutiny after processing UK patient records without proper consent, raising concerns about data privacy and ethical AI deployment [43].

Health insurance Portability and Accountability Act (HIPAA) – USA

The Health Insurance Portability and Accountability Act (HIPAA) is a key U.S. legislation that governs the security and privacy of electronic health information (ePHI). Introduced in 1996, this law is relevant to various entities, including hospitals, insurance providers, and AI applications in healthcare. In addition, the use of AI in healthcare requires adherence to stringent privacy and security regulations regarding electronic protected health information (ePHI) to ward off expensive violations and legal suits. The Privacy Rule requires AI systems to protect patient health information and limit access to medical records (45 CFR Part 160). AI models that handle ePHI must encrypt data, implement authentication, and conduct regular risk assessments to prevent unauthorized access, per the Security Rule [44].

AI applications in healthcare must comply with strict privacy and security regulations to protect electronic protected health information (ePHI). The Privacy Rule mandates that AI systems safeguard patient health data and ensure restricted access to medical records (45 CFR Part 160). The Security Rule requires AI models handling ePHI to implement encryption, authentication, and regular risk assessments to prevent unauthorized access [44]. Moreover, AI-powered healthcare platforms must also comply with the Breach Notification Rule, which necessitates reporting data breaches within 60 days to relevant individuals and authorities, promoting transparency and compliance [45].

Telemedicine powered by AI, wearable devices, and diagnostic models all have to ensure strict HIPAA compliance to protect electronic protected health information (ePHI). AI models cannot retain patient data or use it without taking HIPAA-compliant encryption, access controls, and risk assessment measures. Offenses may incur fines as high

as \$1.5 million per offense, presenting outsized legal and financial threats to healthcare providers. Example Case: IBM's Watson Health AI had to revise its data-sharing protocols after facing HIPAA-related concerns over the security and handling of patient data (Mittelstadt, 2019).

FDA & EMA guidelines on AI/ML in medical devices

AI-based medical devices and diagnostic systems are regulated by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). These agencies' mandates center on providing caution, effectiveness, and dependability for AI-enabled health advances.

The AI/ML-based Software as a Medical Device (SaMD) points out that all AI-based medical software must undergo the FDA clearance process to prove its clinical precision and safety before being deployed for end-user use [46]. Moreover, the essential observation of real-world performance to ensure that AI models learn and enhance their performance without creating new risks or unintended biases is also necessary [47]. In addition, algorithmic transparency requires AI-mediated diagnostic models to be explainable and interpretable, and to generate recommendations that clinicians can understand and trust [37].

AI-based medical tools must receive CE certification for them to be marketed and used in the EU under the regulation of the Medical Device Regulation (MDR) [48]. Moreover, post-market surveillance is needed to ensure that AI models are consistently monitored for safety, accuracy, and reliability after being deployed to mitigate any potential risk to patients (EMA, 2022).

Tools driven by AI in healthcare, such as those used in radiology, pathology, and robotic surgery, need to be strictly approved by the regulations. Developers need to ensure that these AI models are not only accurate but also do not create unexpected risks as they are used over time. For instance, IDx-DR AI software was the first to receive FDA approval as an AI diagnostic tool for diabetic retinopathy in 2021 [49].

AI Act (Proposed) – European Union (EU)

AI-powered medical devices will be strictly audited and risk-assessed before deployment, which can also be dangerous. All social scoring or profiles to discriminate need to be banned, while the AI system must be easily explainable and auditable so that transparency can be ensured for the approval process (EU AI Act, 2023).

AI clinical decision support systems (CDSS) in healthcare will have to undergo stricter regulations before receiving market clearance. Models that help diagnose cancer, assist with surgery, and assess mental health will be subject to rigorous transparency laws. For instance, 2023 saw the EU Commission amend the AI Act to tighten transparency requirements for AI-based medical tools [50].

EU AI Act – classification and implementation challenges

The Artificial Intelligence Act of the European Union, enacted in 2021, sets forth principles for governing AI systems within technological innovations, balancing innovation with the protection of human rights. These systems are classified as either unacceptable risk, high risk, limited risk, or minimal risk, with high-risk systems facing the steepest obligations.

Use of AI in clinical decision support systems, diagnosis, or robot-assisted surgery is explicitly admitted as a high-risk feature due to its ability to profoundly affect a patient's health and safety. Such systems are required to meet set standards of:

- Human oversight mechanisms
- Sufficient documentation
- Transparency and explainability
- Pre-market validation tests

Nonetheless, challenges with execution are still important. For instance:

- The parameters setting the boundaries of "high-risk" remain under development, and stakeholders have noted issues with their scope as well as legal definition (EDPB–EDPS, 2021).
- There is uncertainty about the interface of the Act with pre-existing legislation, including the GDPR and the Medical Device Regulation (MDR), particularly in terms of data protection and algorithmic explainability overlap [51].
- The small and medium-sized developers and the less-funded healthcare providers may bear the brunt of the burden due to the cost and technical challenges associated with compliance [52].

Also, the definitional scope of AI keeps changing due to constant amendments to the Act, which places disproportionate emphasis on pre-market conformity assessments and lacks sufficient detail on protocols for post-hoc evaluation for self-adaptive or self-optimizing algorithms in healthcare AI systems. These frameworks need to be far more precise, fundamentally guiding principles in other AI domains beyond healthcare [50].

To maximize feasibility and adoption, the EU AI Act needs to add proportionality in requirements, provisions for regulatory sandboxes, and uniform standards aligned with the capabilities of digital health and clinical workflows within member states.

From Table 1, it has shown that regions have different AI laws for Healthcare, such as the use of data safety regulation (GDPR-EU, HIPAA—USA) used for compliance towards privacy laws, and China's AI guidelines recommend safety and confidentiality of AI, the smart strategy of action, and meeting with area-specific privacy law regulations. Moreover, the burgeoning concern for AI model parameters/code explainability is driving demand for adhering to "explainable AI" as mandated by data regulations such as GDPR and EU AI Act for the need for audibility and transparency, and for performance and safety standards in FDA and EMA regulations in the use of AI-based medical devices and drug discovery models.

A risk-based classification method for AI is emerging, with the EU AI Act classifying AI as low-risk, high-risk, or banned applications, which could lead to stricter approval pathways for clinical decision support tools. Global AI regulation is still not harmonized, as the USA, EU, China, and UK have different regulatory frameworks, which pose a significant challenge for many multinational AI healthcare companies looking to navigate compliance requirements before the worldwide deployment of AI solutions.

Case studies of ethical & regulatory challenges in healthcare AI

Several cases have shown to be quite a complicated legal issue. Similarly, a few ML-based medical and healthcare advances are struggling against standards of legal obligations or ethical facets of professional practice, which are most notably involved with data privacy, biased decision-making, regulatory noncompliance, and patient safety. Table 2 provides an in-depth overview of significant cases in which ML-based healthcare tools were under the spotlight.

Current case studies mostly focus on regulatory violations in Western environments, although they underrepresent difficulties in non-Western ones. Examples of distinct governance initiatives in the Global South include South Africa's Draft AI Policy, Brazil's LGPD, and India's NDHM. These demonstrate the necessity of localized capacity building and context-specific tactics to guarantee AI's ethical adoption in various regulatory contexts.

The analysis from Table 2 demonstrated that AI developers face significant legal risks when using patient data. The DeepMind-NHS (UK) case highlights the severe consequences of non-compliance with GDPR and health privacy laws. Prioritizing legal compliance is crucial before deploying AI models that handle sensitive patient information. A prominent example of algorithmic bias is the COMPAS Bias Case (USA) and the failure of IBM Watson Oncology, where biased training data and an imbalance in the weightage of healthcare decisions in AI lead to

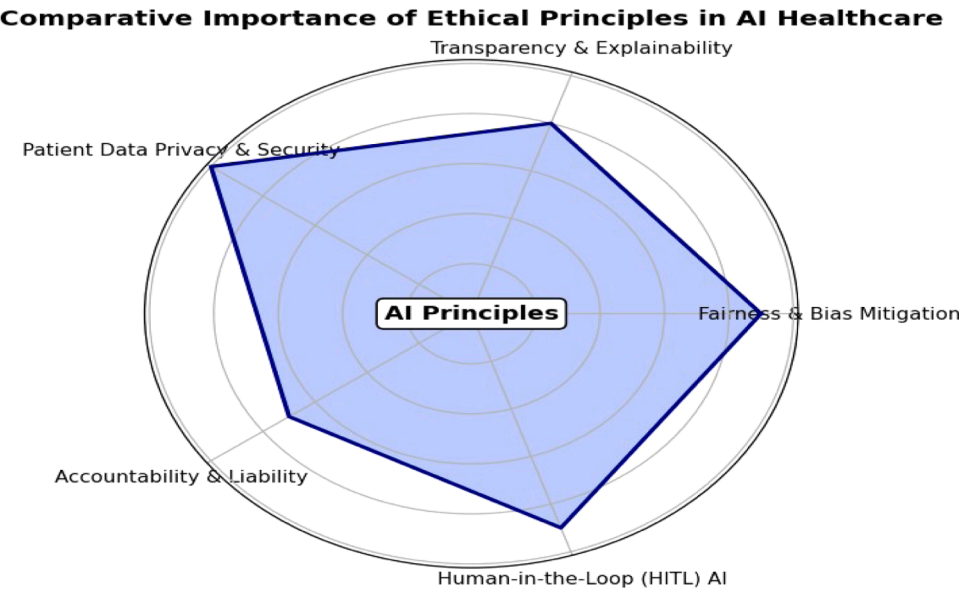


Fig 1. Ethical AI Principles for ML Adoption in Healthcare. Source: WHO 2021 and European [57].

unfair medical outcomes. Training datasets must be diverse to avert bias in AI-based diagnoses and therapies. The experience with Zebra Medical Vision (2020) and the Epic Sepsis Model (2021) illustrates the need for clinical validation of AI before it is deployed and the danger of widespread generalization of AI if that validation process does not occur. Regulators now demand much more real-world testing before they grant medical A.I. approvals. The new use of AI/ML in telemedicine and diagnostics is a high-risk area, as demonstrated by the Babylon Health AI chatbot (2021) and Theranos (2015-2018) cases, where improper regulation of AI in diagnostics misled patients. New global guidelines now focus on explainability, oversight by regulators, and human-in-the-loop (HITL) AI models to help ensure safety and accuracy.

Expanding the geographic scope of ethical AI governance

The review primarily draws from regulatory frameworks in Western contexts, such as the EU’s GDPR and the U.S. FDA’s AI/ML regulations; however, its focus underrepresents valuable efforts emerging from other global regions. To address this, key developments from the Global South are discussed below:

- Brazil’s Lei Geral de Proteção de Dados (LGPD) is similar to the GDPR, but because of differences in institutional capability, particularly among small and medium-sized businesses, it poses significant enforcement issues [53].
- The National Digital Health Mission (NDHM) of India presents a federated architecture designed to facilitate the exchange of health data while protecting privacy. Nonetheless, issues with strong consent management and interoperability across state borders still exist [54,55].
- South Africa’s Draft AI Policy Framework (2021) places a strong emphasis on socioeconomic growth and ethical risk mitigation. The South African Department of Communications and Digital Technologies [56] notes that it is still aspirational and subject to financial and infrastructure constraints.

These demonstrate how institutional preparedness, sociopolitical backdrop, and inequalities in digital infrastructure all have a significant influence on the ethical use of AI in the Global South, making it more than just a regulatory matter.

Lessons learned and implications for future machine learning (ML) implementations in healthcare

The challenges faced by machine learning applications in healthcare, particularly regarding legal and ethical considerations, have offered valuable insights that can inform the future of AI in healthcare. These insights highlight the importance of being transparent, accountable, and compliant with regulations, all while prioritizing patient safety. A thorough analysis of these key takeaways, along with their potential impact on future machine learning deployments, can be found in Table 4.

Implications for future ML implementations in healthcare

Ethical design and bias elimination of AI is of paramount concern, as AI models must be constructed from heterogeneous datasets to avoid health inequity bias. Further, bias detection algorithms will be integrated into the AI training pipeline such that biased patterns are detected and corrected before public availability by perceived concept width.

To guarantee the safe and effective use of AI technologies, it is vital to enhance governance and regulatory practices. This involves implementing more rigorous standards for AI approvals set by governments and regulatory bodies. Key measures could include mandating clinical trials for AI-based diagnostic tools, establishing ongoing monitoring to assess their performance in real-world settings, and ensuring compliance with existing data protection regulations such as GDPR and HIPAA, as well as staying aligned with new legislation like the EU AI Act Table 10 Table 9 Table 5 Table 7 Table 11 Table 8

Explainable artificial intelligence (XAI) is required with great urgency to promote greater transparency in healthcare, as AI models intended for use in medicine must be explainable and interpretable (i.e., explain to the physician why a decision is being made). Methods, like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), need to be applied to understand how the algorithms behave, and to convince the end-user that it is going to yield clinical improvement.

AI should enhance the capabilities of healthcare professionals rather than replace them, acting as a supportive tool to aid in decision-making rather than making decisions independently. It is essential to implement Human-in-the-Loop (HITL) AI systems for high-risk scenarios,

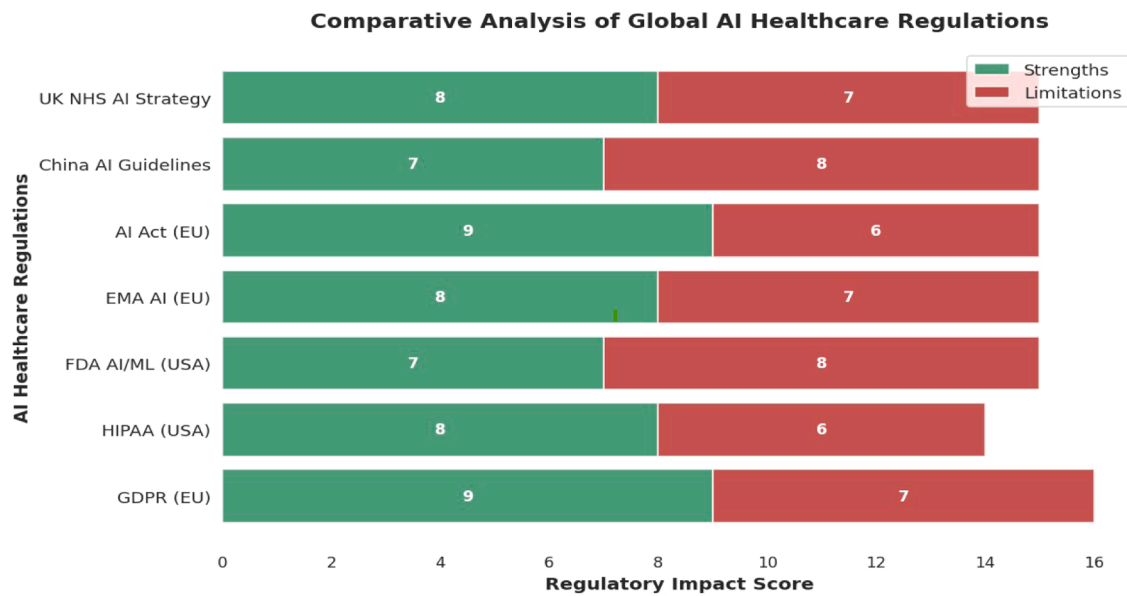


Fig 2. Comparative Analysis of Global AI Healthcare Regulations.
Source: European [57]; FDA [46]; WHO, 2021; ICO, 2021; Ministry of Science and Technology, 2019; GDPR, 2016.

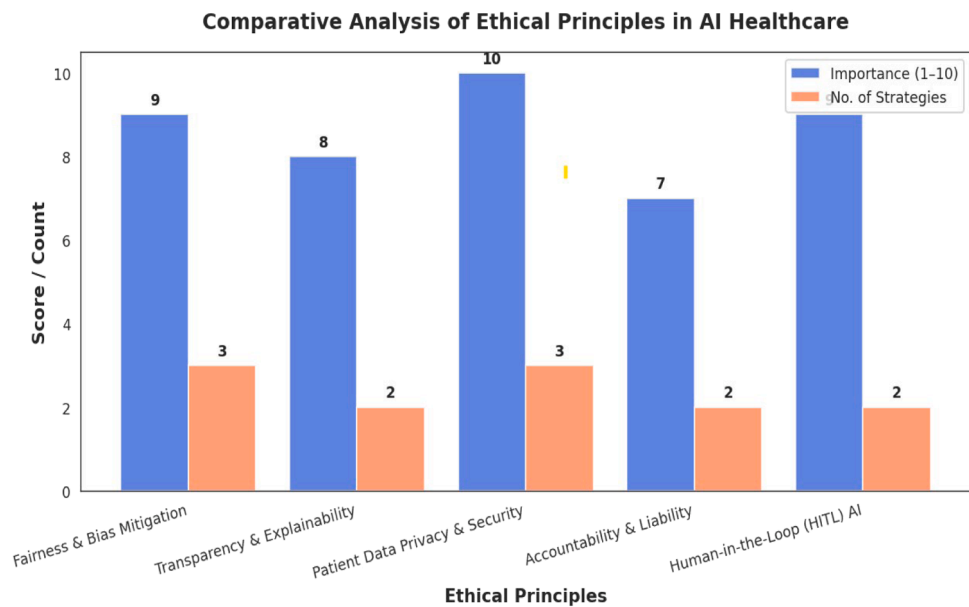


Fig 3. Relative Analysis of Ethical Principles in AI Healthcare.
Source: WHO 2021 and European Commission [58].

Table 1a
Comparative Contributions of the Review and Prior Surveys.

Feature	Existing review	Benjamens et al. [7]	Char et al. [8]
Covers EU AI Act (2023)	✖	✖	✖
Case studies of regulatory failure	✖ (8)	✖	✖
Comparative table of AI laws	✖	✖	✖
Legal + Ethical integration	✖	✖	✖ (brief)
Recommendations for hospitals	✖	✖	✖

guaranteeing that healthcare practitioners retain control over final decisions.
To foster public confidence in AI technology, developers need to

Table 1b
Bibliometric Comparison of Review Coverage.

Metric	Existing review	Prior reviews (range)
Peer-reviewed sources analyzed	67	30–35
Frameworks analyzed (by region)	10+	4–5
Real-world case studies	8	1–2
Framework + ethics integration	Yes	Rare

openly communicate what their models can and cannot do, thereby preventing any misrepresentation. Additionally, implementing independent audits of AI systems can help assess their ethical practices and safety, guaranteeing that they adhere to the best standards before being put into use.

Table 2
Legal & Ethical Challenges in ML-Based Healthcare Applications.

Case Name & Year	ML Application	Issue Faced	Legal / Ethical Concerns	Outcome & Lessons Learned
Google DeepMind & NHS Data Privacy Scandal (2016)	AI-powered patient monitoring system for acute kidney injury detection.	Unauthorized data access: NHS shared 1.6 million patient records with DeepMind without explicit consent.	- Violation of UK Data Protection Laws (GDPR Precursor). - Patients were unaware their data was used for AI development. - Lack of transparency in data-sharing agreements.	- DeepMind was found in violation of UK privacy laws. - Led to stricter AI & patient data-sharing guidelines under GDPR. - NHS revised AI data governance frameworks.
IBM Watson for Oncology (2018)	AI-based cancer treatment recommendation system.	Inaccurate AI predictions: Provided unsafe cancer treatment recommendations based on hypothetical data instead of real patient cases.	- Algorithmic bias led to incorrect treatment plans. - Lack of transparency on AI decision-making. - Patient safety concerns raised by oncologists.	- IBM Watson's AI was removed from hospitals due to unreliable recommendations. - Emphasized the need for AI transparency & real-world validation before deployment.
COMPAS Recidivism Algorithm Bias Case (2016, USA)	ML tool predicting criminal recidivism risk (not healthcare-specific but impacted medical AI ethics).	Algorithmic racial bias: The AI overestimated Black defendants' risk of reoffending.	- Highlighted racial bias in AI models. - Raised concerns about fairness in AI-driven medical diagnostics.	- Strengthened calls for bias detection frameworks in AI. - Encouraged the development of fair AI models in healthcare.
Zebra Medical Vision AI (2020)	AI for automated radiology diagnostics (detecting fractures, lung disease, and brain bleeds).	Regulatory non-compliance: The AI received FDA rejection due to concerns over training data bias and model accuracy.	- Insufficient clinical validation before market approval. - Potential misdiagnoses due to AI errors. - Lack of explainability in AI-generated reports.	- Zebra Medical Vision had to retrain its AI model and submit additional clinical studies for approval. - The FDA enforced stricter AI approval standards for medical imaging.
Babylon Health AI Chatbot (2021, UK)	AI-powered telemedicine chatbot diagnosing patient symptoms.	Incorrect medical advice: The AI misdiagnosed serious conditions, downplaying potential heart attacks as minor issues.	- Patient safety risks due to AI misclassification. - Lack of regulatory oversight on AI-driven symptom checkers.	- UK regulators increased scrutiny on AI-based diagnostic tools. - Led to new guidelines for AI in telemedicine.
Epic Sepsis Prediction Model (2021, USA)	AI system predicting sepsis risk in hospitalized patients.	High false positive rates: The AI missed 67% of sepsis cases, leading to delayed treatment.	- Accuracy concerns in life-threatening conditions. - Hospitals relied on flawed AI predictions, affecting patient safety.	- Hospitals are required to integrate AI models with human oversight. - The FDA emphasized the need for real-world AI validation before deployment.
Theranos AI Blood Testing Fraud (2015-2018, USA)	ML-driven blood testing technology promises rapid diagnosis with a single drop of blood.	Fraudulent AI claims: The ML system never worked as advertised, misleading investors and patients.	- Ethical violations & investor fraud. - Lack of AI transparency & scientific validation. - Potential harm to misdiagnosed patients.	- Theranos CEO Elizabeth Holmes was convicted of fraud. - Highlighted dangers of unverified AI medical claims. - Stricter AI compliance laws were introduced in the USA.
South Africa – Draft AI Policy, 2022	AI in public healthcare systems	Lack of regulation on AI use in clinical settings	Need for ethical standards, bias mitigation, and data ownership frameworks	Ongoing policy development promotes human rights-based AI principles and indigenous data sovereignty
Brazil – LGPD (Lei Geral de Proteção de Dados), 2020	ML in health monitoring systems	Non-transparent data usage and consent mechanisms	Concerns around informed consent, data subject rights, and usage transparency	Enforced strict data governance; ML systems now require robust consent and explainability mechanisms
India – NDHM (National Digital Health Mission), 2020	AI-based digital health records and diagnostics	Risk of misuse of centralized data and digital exclusion of rural populations	Issues of data security, algorithmic fairness, and equitable access	Introduced patient-controlled Health IDs and ethical AI guidelines; encouraged inclusive, transparent development
Zindi / African ML Competitions, 2021–Present	Disease prediction models in African contexts	Lack of contextual data affecting model performance	Bias from non-local datasets; insufficient regional data inclusion	Promotes Africa-specific datasets and challenges; fosters ethical, context-aware AI development
Google Health – Diabetic Retinopathy in India, 2019	AI for diabetic retinopathy screening	Failed deployment in rural clinics due to inconsistent image quality	Algorithm robustness, contextual relevance, tech infrastructure gap	Stressed need for local validation and infrastructure-compatible design; supports human-in-the-loop approaches

Table 3
Legal & Ethical comparison between regions.

Region	Policy	Strength	Implementation Challenge
EU	GDPR	Strong privacy protection	Complex compliance burdens
USA	FDA AI/ML	Sector-specific guidance	Slow update cycles
Brazil	LGPD	Data subject rights	Low institutional capacity
India	NDHM	Federated data architecture	Consent & interoperability issues
South Africa	Draft AI Policy	Inclusive development goals	Early-stage and underfunded

Proposed recommendations for ethical & regulatory compliance

Clarification of technical and policy approaches

- Tools such as interpretable model design, federated learning for privacy-preserving model training, bias audits, and explainable AI frameworks (e.g., SHAP, LIME) are examples of technical measures. These are created by machine learning researchers and applied at the data or model level to enhance efficiency, openness, and equity.
- Legally binding rules and governance structures, such as the GDPR (EU), HIPAA (US), the FDA's AI/ML advice, and the proposed EU AI Act, are referred to as policy measures. These are put in place by

Table 4
Lessons Learned from Past ML Failures in Healthcare.

Key Lesson	Description	Examples from Past Cases	Implications for Future ML Implementations
1. AI Must Comply with Data Privacy Laws (GDPR, HIPAA)	ML models must handle patient data securely and obtain explicit consent before use.	- DeepMind-NHS (2016): Used 1.6 million patient records without consent, violating UK privacy laws.	- Future ML models must integrate privacy-by-design features. - AI developers should follow GDPR/HIPAA compliance standards.
2. Bias in AI Can Lead to Discriminatory Healthcare Decisions	Biased training data can cause unequal treatment of different patient groups.	- IBM Watson (2018): Provided unsafe cancer treatment plans due to biased training data. - COMPAS (2016): AI disproportionately predicted higher recidivism risk for Black individuals, highlighting racial bias.	- AI must be trained on varied, illustrative datasets. - Bias audits should be conducted regularly before deployment.
3. AI Requires Human Oversight in Critical Healthcare Applications	ML models should work as decision-support tools, not replacements for clinicians.	- Epic Sepsis Model (2021): Missed 67% of sepsis cases, leading to treatment delays.	- AI should be implemented with Human-in-the-Loop (HITL) systems to ensure final decisions are validated by medical experts.
4. Regulatory Approval is Essential Before Deploying AI in Healthcare	AI models must undergo rigorous clinical validation to ensure accuracy.	- Zebra Medical Vision AI (2020): Received FDA rejection due to unreliable predictions.	- Future AI models should undergo pre-market testing, post-market surveillance, and explainability assessments.
5. AI Transparency & Explainability Are Key to Trust and Adoption	Black-box AI models can lead to misdiagnoses and lack of accountability.	- Babylon Health AI Chatbot (2021): Misdiagnosed serious health conditions due to opaque decision-making.	- Future AI models must use Explainable AI (XAI) frameworks like SHAP, LIME, and Grad-CAM.
6. AI Developers Must Be Held Accountable for Misuse or Fraud	Companies must ensure ethical AI claims and avoid deceptive practices.	- Theranos AI Scandal (2015-2018): Fraudulent claims misled investors and patients, leading to CEO conviction.	- Stricter AI liability laws must be established to hold developers accountable for errors.

Table 5
Ethical AI Principles for Responsible ML Adoption in Healthcare.

Ethical Principle	Description & Importance	Implementation Strategies
Fairness & Bias Mitigation	AI models must provide equitable healthcare outcomes without discrimination based on race, gender, socioeconomic status, or location [31].	- Use bias-detection algorithms and fairness audits before deployment. - Train AI models on diverse, representative datasets. - Apply re-weighting & adversarial debiasing techniques.
Transparency & Explainability (XAI)	AI decisions should be interpretable, auditable, and justifiable by medical professionals [35].	- Require Explainable AI (XAI) frameworks like SHAP, LIME, and Grad-CAM. - Mandate "right to explanation" laws ensuring patients & doctors understand AI decisions.
Patient Data Privacy & Security	AI must protect sensitive medical data in compliance with GDPR (EU), HIPAA (USA), and AI Act (EU) [45].	- Enforce data anonymization, encryption, and access controls. - Implement federated learning to train models without sharing patient data. - Require patient consent for AI-driven healthcare applications.
Accountability & Liability	AI-driven errors should have clear legal accountability, determining whether liability falls on developers, healthcare providers, or institutions [9].	- Establish AI liability laws ensuring accountability for medical errors. - Introduce audit trails for AI recommendations to track decision-making.
Human-in-the-Loop (HITL) AI	AI should support, not replace, human clinicians [12].	- Mandate Human-in-the-Loop (HITL) AI for high-risk applications (e.g., surgery, cancer diagnosis). - Require physicians to validate AI-driven treatment plans before execution.

Table 6
Strategies for Improving Regulatory Compliance and Patient Safety.

Regulatory Area	Challenges	Recommended Solutions
AI Risk Classification	- Lack of standardized AI risk assessment models.	Implement risk-based AI classification (EU AI Act): Low-risk AI (health monitoring apps) -High-risk AI (AI-assisted surgery, diagnosis) -Prohibited AI (social scoring, discriminatory profiling).
Pre-Market Approval & AI Testing	- AI models can enter healthcare without sufficient clinical validation. - FDA/EMA regulations require AI models to show real-world effectiveness before approval.	- Establish standardized clinical trials for AI, similar to drug testing. - Require "explainability disclosures" during regulatory approval. - Mandate post-market surveillance for AI-driven healthcare devices.
Data Privacy & Security	- AI models require large-scale patient data, raising risks of unauthorized access & breaches.	- Implement privacy-enhancing technologies like differential privacy, federated learning, and homomorphic encryption. - Enforce HIPAA/GDPR compliance audits for AI-based medical software.
Bias Auditing & Fairness Standards	- Algorithmic bias leads to unfair healthcare outcomes.	- Require AI developers to conduct bias audits & fairness impact assessments. - Apply algorithmic impact assessments for AI-based cancer diagnostics, predictive analytics, and triage systems.
AI Explainability & Transparency	- Many AI systems operate as "black-box" models.	- Mandate XAI frameworks (SHAP, LIME, Grad-CAM) for AI interpretability. - Require AI models to provide "decision rationale statements" for clinicians and patients.

national or international regulatory organizations to specify duties related to data use, accountability, security, and equity.

Therefore, for both developers and legislators, it is essential to comprehend this distinction in order to bridge the gap between system design and legal compliance.

To promote responsible and safe use of machine learning (ML) in

healthcare, AI systems need to follow clear ethical guidelines, regulatory requirements, and established policies. Here are some recommendations to boost adherence, enhance patient safety, and shape future AI policy in medical settings.

The values displayed in Fig. 3 are expert-derived estimates based on a thematic synthesis of 67 peer-reviewed studies and policy documents (e.g, GDPR, HIPAA, WHO 2021, FDA AI/ML 2021, EU AI Act).

Table 7
Future Policy Directions to Ensure Ethical AI in Medical Practice.

Policy Initiative	Proposed Action	Expected Impact
Stronger Global AI Governance	- Create international AI healthcare standards under WHO, EU, and FDA collaboration. - Establish AI regulatory task forces in every country.	- Ensures global harmonization of AI safety & ethical guidelines. - Reduces regulatory fragmentation for AI adoption in healthcare.
Ethical AI Certification	- Introduce an AI Ethics & Safety Certification for ML-driven healthcare applications.	- Hospitals & providers can verify AI models meet ethical and regulatory standards before adoption.
Mandatory AI Training for Healthcare Professionals	- AI in medicine must be understood by clinicians before use. - Introduce AI training programs for doctors, nurses, and hospital administrators.	- Enhances AI literacy among medical professionals. - Reduces misuse of AI-driven diagnostic and treatment tools.
Clear AI Liability Frameworks	- Define who is responsible when AI-based medical errors occur (developers, hospitals, physicians?).	- Provides legal clarity on AI accountability. - Prevents misuse of AI by avoiding "liability gaps".
Public & Patient Engagement in AI Ethics	- Require patient inclusion in AI development & deployment decisions. - Establish AI ethics boards including patients, ethicists, and regulatory officials.	- Increases trust in AI-driven healthcare. - Ensures AI aligns with public values & health equity principles.

Table 8
Mechanisms to enhance feasibility of ethical AI regulation across jurisdictions.

Challenge	Proposed Solution	Real-World Example
Conflicting national frameworks	Modular harmonization + soft law principles	OECD AI Principles, GPAI
Over-regulation stifling innovation	Regulatory sandboxes, tiered risk-based oversight	UK ICO sandbox, Singapore AI Verify
Duplicate compliance costs	Mutual recognition agreements (MRAs)	EU-US Data Privacy Framework

Source: OECD [60], EDPB-EDPS Joint Opinion [61], and World Economic Forum [59].

Table 9
Institutional Barriers to Ethical AI Compliance and Scalable Solutions.

Barrier	Impact	Example Solution
Lack of funding	No capacity for third-party audits or legal teams	Tiered regulatory frameworks
Poor technical infrastructure	Inability to deploy privacy-enhancing technologies	Federated learning, open-source tools
Workforce limitations	Staff untrained in ethical AI implementation	Institutional training, government AI toolkits
Fragmented health IT systems	Difficulty ensuring interoperability and traceability	Standardized APIs and modular compliance protocols

Source: Table is based on synthesized findings from Canedo et al. [53], Sinha et al. [54], and Sheller et al. [22].

Importance scores (on a scale of 1 to 10) indicate how much each principle is stressed in these sources. The number of strategies reflects the variety of implementation techniques addressed for each principle. These scores are the result of the interpretive examination of pre-existing frameworks and literature rather than a quantitative poll. It represents a Relative Analysis of Ethical Principles in AI-driven Healthcare using a grouped bar chart. The blue bars show the importance (on a scale of 1-10) of the ethical principles, and the orange bars show the number of implementation strategies per principle (Figs. 1, 2). Patient Data Privacy & Security emerges as the highest priority (score = 10) because strict regulations such as the GDPR, HIPAA, and

the AI Act call for the secure, responsible application of AI. Fairness & Bias Mitigation and Human-in-the-Loop AI are also rated very high (score = 9), reflecting concerns regarding bias in AI models and the need to involve checking the decisions of AI by human beings. Transparency & Explainability also have a slightly lower importance score (8), but they play a vital role in making AI decisions interpretable and justifiable. Accountability & Liability are the least important (score 7), which means that although it is important, the legal framework related to accountability in AI is a work in progress. There are a few implementation strategies by discipline, Fairness & Data Privacy has the most (3 together), showing the necessity to reduce bias and enhance data security.

Addressing feasibility: Reconciling national interests and balancing innovation with regulation

Although international policy alignment (e.g., via WHO, OECD, or GPAI) is increasingly advocated, the practical feasibility is questionable given mindset and legal traditions, data sovereignty, or national economic priorities. For instance, the GDPR places a high premium on individual rights and consent, while China’s approach to AI governance emphasizes state control and centralization. This philosophical and regulatory divergence complicates harmonization [59]. Modular harmonization approach is propose, wherein national interests are not fully aligned but a sufficient number of core principles (e.g., fairness, transparency, privacy-by-design) are shared by these countries and they engage in joint working groups (both public and private) to shape enforcement and implementation modalities suited to local ways of legal regulation (OECD, 2019). Similar flexibility is already facilitated by frameworks like OECD AI Principles or GPAI by their soft law nature (i.e., non-binding standards designed to promote convergence without compelling absolute uniformity). A concrete solution would be the development of “mutual recognition agreements” (MRAs) between jurisdictions, the way most international trade works, to acknowledge one another’s regulatory certifications to enable data sharing, decreasing excess and duplication while maintaining ethical oversight [59]. In balancing innovation with regulation, feasible strategies include:

- Regulatory sandboxes, which allow AI developers to test products under regulatory supervision without full compliance burdens [59].
- Tiered regulation, where lower-risk AI tools (e.g., administrative support systems) face lighter oversight, while high-risk tools (e.g., diagnostic AI) are rigorously assessed (EDPB-EDPS, 2021).
- Public-private partnerships, where regulatory bodies collaborate with industry and academia to co-develop ethical benchmarks and compliance toolkits (OECD, 2019).

Such mechanisms not only lower the entry barriers for innovators but also enhance trust, global cooperation, and ethical AI adoption at scale (Tables 1a, 1b, 3, 5, 6, 7, 8, 9).

Discussions

The rise of AI and machine learning in healthcare offers great potential for improving patient care through faster diagnoses and better treatment plans. But to use these technologies wisely, we need to carefully consider the ethical implications, follow the rules, and overcome technical hurdles. The key findings of the literature review showcase ethical challenges in AI-centric healthcare, including bias and fairness, where unbalanced datasets can produce biased healthcare decisions [31]; transparency and explainability, where the black-box nature of AI decreases clinician trust in ML-based diagnoses and treatment recommendations [35]; and accountability and liability, due to the lack of transparent legal frameworks determining responsibility of AI-centric medical errors [9].

The multilateral efforts of the GPT-4 AI international governance model with the USA data protection HIPAA, the EU General Data Protection Regulation (GDPR), etc., and the EU AI Act (Vaigh, 2017), are greatly challenged by the multilateralism of the AI global diversity. Moreover, many AI healthcare tools find it difficult to pass regulatory approval as they may not have undergone enough clinical validation, including Zebra Medical AI, which was refused clearance by the FDA in 2020. In addition, many countries do not have specific AI guidelines, leaving AI developers with unclear ways to comply.

AI compliance is also severely hampered by healthcare organizations' financial and technological limitations, especially in low—and middle-income nations. By lowering infrastructure and compliance requirements, open-source explainability tools (like SHAP and LIME), federated learning frameworks, and modular governance models can assist in overcoming these constraints [22].

Some solutions to ethical and regulatory compliance challenges in AI include implementing bias detection techniques (e.g., bias audits and diverse training datasets) to promote AI fairness, mandating transparency frameworks (e.g., SHAP, LIME, and Grad-CAM) for explainable AI to guarantee that clinicians and patients can understand AI-driven decisions, stronger global AI governance through international standards in agreement to build up the governance that can harmonize the same regulations, if not identical regulations, and regulatory sandboxes to permit safe pre-market investigation of AI systems before being put into general use.

Institutional resource constraints as barriers to compliance

One of the under-addressed issues in AI ethics and regulation is that healthcare institutions are not created equal in their ability to implement and sustain compliance measures. Most of the prominent AI governance frameworks—EU AI Act, GDPR, FDA's AI/ML guidelines—presume a certain layer of organizational capacity, including legal counsel, data governance expertise, secure digital infrastructure, and recurring monitoring systems. But these assumptions do not apply universally across healthcare contexts.

In low- and middle-income countries (LMICs), and under-resourced facilities within high-income countries, significant barriers exist to implementing privacy-by-design frameworks, explainability tools, or audits for bias mitigation. Institutions may lack:

- AI compliance staff or ethicists (dedicated)
- Data storage and a federated learning framework
- External audit or legal review funding
- Technical staff to implement Interpretable models or carry out algorithmic accountability procedures.

As Canedo et al. [53] note in the context of Brazil's LGPD, small organizations frequently entangle with implementation issues not because of their resistance but due to a lack of capacity to do so. Similarly, Sinha et al. [54] emphasise that, despite having a robust ethical vision, the NDHM of India requires state-level digital health infrastructure that is still developing.

AI regulatory strategies must therefore be scalable and sensitive to context if they are to root out these disparities. Recommendations include:

- Adoption of an open-source explainability tool (e.g., SHAP, LIME) that leads to a reduction in cost
- Introduce federated learning to minimize the burden of central infrastructure [22]
- Institute tiered models of compliance based on institutional maturity, enabling smaller hospitals to implement ethical AI aspects progressively.

Rationale and validation design for recommendations

The following policy and technical recommendations are based on:

- Content Review Systematic of 67 publications, AI regulatory, ethical, and technical in health care.
- Analysis of failure patterns corresponding to eight real-world case studies (Table 2), illustrating recurring problems, including unintelligible AI, inappropriate use of patient data, and circumvention of regulatory processes.
- Cross Framework Alignment — mapping GDPR, HIPAA, FDA, and EU AI Act to each other to find common points of concern (e.g., explainability, liability, consent etc.).

Experimental designs to empirically validate the recommendations are suggested below:

- Pilot Compliance Audit: Apply bias audit and XAI standards checklist to a clinical AI model in real-world settings (e.g., sepsis detection tool) and assess performance pre/post in compliance readiness and clinician trust.
- Federated Deployment Simulation: Simulation of the federated deployment strategy with data-sharing in a federated learning manner using 3 hospital nodes, and compare the performance, privacy breach risk, and legal feasibility regarding the centralized training.
- Policy Simulation Workshops. Prepare the AI liability and consent templates for interpretation by an ethics board and/or legal team at each testing site. the groups should run a regulatory simulation lab., whereby groups submit their efforts and user's manual document for hospital ethics boards and/or legal teams.

Conclusion

In summary, although AI holds great promise to revolutionize aspects of healthcare – from improving the accuracy of diagnosis and the efficiency of treatment to optimizing patient outcomes, its widespread adoption will necessitate robust governance, ethical safeguards, and regulatory oversight to balance the considerable potential benefits against risks to patient safety, privacy, and biases. Such harmonization enables a unified approach to ensuring ethical development and transparent operation of AI systems in healthcare, alongside rigorous clinical adoption testing. Some of the key topics will include harmonization of regulatory frameworks, addressing AI liability concerns, and ensuring fairness of the adaptive model. Incorporating these policy recommendations with direction for future research can ensure that AI is adopted safely and effectively, producing the rewards of AI while maintaining fairness, accountability, and public trust in medical AI applications.

CRedit authorship contribution statement

Shehu Mohammed: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Neha Malhotra:** Writing – review & editing, Visualization, Validation, Supervision, Investigation.

Declaration of competing interest

The authors declare no competing interest.

Funding

No funding was received for this study.

Data availability statement

Not Applicable.

Research involving humans and animals

Not Applicable.

Informed consent

Not Applicable.

References

- [1] E. Vayena, A. Blasimme, I.G. Cohen, Machine learning in medicine: addressing ethical challenges, *PLoS. Med.* 15 (11) (2018) e1002689, <https://doi.org/10.1371/journal.pmed.1002689>.
- [2] C. Mennella, U. Maniscalco, G. De Pietro, M. Esposito, Ethical and regulatory challenges of AI technologies in healthcare, *NPJ Dig. Med.* (2024), <https://doi.org/10.1038/s41746-024-00793-0>.
- [3] Centers for Disease Control and Prevention (CDC), Health equity and ethical considerations in using artificial intelligence in public health, *Prev. Chronic. Dis.* 21 (2024) E45, <https://doi.org/10.5888/pcd21.240245>.
- [4] W.N. Price, I.G. Cohen, Privacy in the age of medical big data, *Nat. Med.* 25 (1) (2019) 37–43, <https://doi.org/10.1038/s41591-018-0272-7>.
- [5] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science* (1979) 366 (6464) (2019) 447–453, <https://doi.org/10.1126/science.aax2342>.
- [6] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 279–288, <https://doi.org/10.1145/3287560.3287574>.
- [7] S. Benjamens, P. Dhunoo, B. Meskó, The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database, *NPJ. Digit. Med.* 3 (1) (2020) 118, <https://doi.org/10.1038/s41746-020-00324-0>.
- [8] D.S. Char, N.H. Shah, D. Magnus, Implementing machine learning in health care—Addressing ethical challenges, *New Engl. J. Med.* 382 (11) (2020) 981–983, <https://doi.org/10.1056/NEJMp1912591>.
- [9] J. Morley, C.C.V. Machado, C. Burr, J. Cows, I. Joshi, M. Taddeo, L. Floridi, The ethics of AI in health care: A mapping review, *Soc. Sci. Med.* 260 (2020) 113172, <https://doi.org/10.1016/j.socscimed.2020.113172>.
- [10] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in healthcare, *Lancet Digit. Health* 3 (11) (2021) e745–e750, [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
- [11] D.A. Vyas, L.G. Eisenstein, D.S. Jones, Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms, *New Engl. J. Med.* 383 (9) (2020) 874–882, <https://doi.org/10.1056/NEJMms2004740>.
- [12] S. Gerke, T. Minssen, G. Cohen, Ethical and legal challenges of artificial intelligence-driven healthcare, *Artif. Intell. Healthcare* (2020) 295–336, <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>.
- [13] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, *Science* (1979) 349 (6245) (2015) 255–260, <https://doi.org/10.1126/science.aaa8415>.
- [14] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, E. Topol, Deep learning-enabled multi-modal fusion of medical imaging and electronic health records for improved diagnostics and prognostics, *Nat. Commun.* 12 (1) (2021) 6675, <https://doi.org/10.1038/s41467-021-26946-4>.
- [15] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W. M van der Laak, A survey on deep learning in medical image analysis, *Med. Image Anal.* 42 (2017) 60–88, <https://doi.org/10.1016/j.media.2017.07.005>.
- [16] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, A.Y. Ng, CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017 arXiv preprint arXiv:1711.05225, <https://arxiv.org/abs/1711.05225>.
- [17] A. Ribas, J.D. Wolchok, J. Schlom, E.M. Jaffee, Cancer immunotherapy comes of age, *Nat. Commun.* 9 (1) (2018) 1–14, <https://doi.org/10.1038/s41467-018-04388-3>.
- [18] J. He, S.L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, The practical implementation of artificial intelligence technologies in medicine, *Nat. Med.* 25 (1) (2019) 30–36, <https://doi.org/10.1038/s41591-018-0307-0>.
- [19] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, in: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730, <https://doi.org/10.1145/2783258.2788613>.
- [20] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): towards medical AI transparency, *Nat. Mach. Intell.* 2 (1) (2020) 56–67, <https://doi.org/10.1038/s42256-020-00273-3>.
- [21] J. Shen, C.J.P. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, Artificial intelligence versus clinicians in disease diagnosis: systematic review, *JMIR. Med. Inform.* 10 (4) (2022) e32912, <https://doi.org/10.2196/32912>.
- [22] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, *Sci. Rep.* 10 (1) (2020) 12598, <https://doi.org/10.1038/s41598-020-69250-1>.
- [23] N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, G. Kaissis, The future of digital health with federated learning, *NPJ. Digit. Med.* 3 (2020) 119, <https://doi.org/10.1038/s41746-020-00323-1>.
- [24] J.M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N.M. Donghia, J. J. Collins, A deep learning approach to antibiotic discovery, *Cell* 180 (4) (2020) 688–702, <https://doi.org/10.1016/j.cell.2020.01.021>, e13.
- [25] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, M.S. Veselov, V.A. Aladinskiy, A. V. Aladinskaya, A. Aspuru-Guzik, Deep learning enables rapid identification of potent DDR1 kinase inhibitors, *Nat. Biotechnol.* 37 (9) (2019) 1038–1040, <https://doi.org/10.1038/s41587-019-0224-x>.
- [26] Y. Yang, M.S. Islam, J. Wang, Y. Li, A comprehensive survey on machine learning techniques in medical diagnosis, *Comput. Biol. Med.* 101 (2017) 107–128, <https://doi.org/10.1016/j.combiomed.2018.06.016>.
- [27] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, P.A. Keane, Clinically applicable deep learning for diagnosis and referral in retinal disease, *Nat. Med.* 24 (9) (2018) 1342–1350, <https://doi.org/10.1038/s41591-018-0174-4>.
- [28] W. Streeting, Can better data save the NHS? *Financial Times* (2024). <https://www.ft.com/content/b4c57347-d64d-436a-a2f1-33b7049a74b7>.
- [29] M. Roy, S.J. Minar, P. Dhar, A.T.M.O. Faruq, Machine Learning Applications In Healthcare: The State Of Knowledge and Future Directions, 2023 arXiv preprint arXiv:2307.14067, <https://arxiv.org/abs/2307.14067>.
- [30] Vatican News, New Vatican document offers AI guidelines from warfare to health care, Associated Press, 2025. <https://apnews.com/article/231b4b7b8ed6a195ec920f1362c15e2>.
- [31] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, *ACM Comput. Surv. (CSUR)* 54 (6) (2021) 1–35, <https://doi.org/10.1145/3457607>.
- [32] M.A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential biases in machine learning algorithms using electronic health record data, *JAMA Intern. Med.* 178 (11) (2018) 1544–1547, <https://doi.org/10.1001/jamainternmed.2018.3763>.
- [33] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2018, pp. 77–91, <https://doi.org/10.1145/3287560.3287583>.
- [34] S.M. Lundberg, G.G. Erion, S.I. Lee, Consistent individualized feature attribution for tree ensembles, *Nat. Mach. Intell.* 2 (1) (2020) 56–67, <https://doi.org/10.1038/s42256-019-0138-9>.
- [35] Z.C. Lipton, The myths of model interpretability, *Queue*, 16 (3) (2018) 31–57, <https://doi.org/10.1145/3236386.3241340>.
- [36] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017 arXiv preprint arXiv:1702.08608.
- [37] W. Samek, T. Wiegand, K.R. Müller, arXiv preprint, 2019, <https://doi.org/10.48550/arXiv.1904.00026>.
- [38] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, *Harv. J. Law Technol.* 31 (2) (2017) 841–887, <https://doi.org/10.2139/ssrn.3063289>.
- [39] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, Explainability for artificial intelligence in healthcare: A multidisciplinary perspective, *BMC. Med. Inform. Decis. Mak.* 20 (1) (2020) 1–9, <https://doi.org/10.1186/s12911-020-01332-6>.
- [40] W. Wang, K. Siau, Artificial intelligence governance in healthcare: challenges, opportunities, and future research directions, *Int. J. Inf. Manage* 64 (2022) 102466, <https://doi.org/10.1016/j.ijinfomgt.2022.102466>.
- [41] P. Voigt, dem von, A. Bussche, The EU General Data Protection Regulation (GDPR): A practical guide, Springer International Publishing, 2017, <https://doi.org/10.1007/978-3-319-57959-7>.
- [42] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a "right to explanation", *AI. Mag.* 38 (3) (2017) 50–57, <https://doi.org/10.1609/aimag.v38i3.2741>.
- [43] J. Powles, H. Hodson, Google DeepMind and healthcare in an age of algorithms, *Health Technol. (Berl)* 7 (4) (2017) 351–367, <https://doi.org/10.1007/s12553-017-0179-1>.
- [44] S. Hoffman, A. Podgurski, Artificial intelligence and the law: regulation and accountability for AI in global health, *J. Law Biosci.* 8 (1) (2021), <https://doi.org/10.1093/jlb/lsab014>.
- [45] D. McGraw, Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data, *JAMA Intern. Med.* 173 (17) (2013) 1581–1582, <https://doi.org/10.1001/jamainternmed.2013.7111>.
- [46] Food and Drug Administration, Artificial intelligence/machine learning-based software as a medical device (SaMD), FDA Regulatory Guidelines, 2021. <https://www.fda.gov/media/145022/download>. Accessed 17 April 2025.
- [47] Food and Drug Administration, Predetermined change control plans for machine learning-enabled medical devices: guiding principles. U.S. Food and Drug Administration, Retrieved from, <https://www.fda.gov/media/145022/download>, 2023.
- [48] European Medicines Agency, Guideline on AI applications in medicine, EMA AI Regulat. Framework (2021). https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-use-ai-medicines_en.pdf. Accessed 18 April 2025.
- [49] M.D. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J.C. Folk, M. Niemeijer, Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, *Invest. Ophthalmol. Vis. Sci.* 59 (10) (2018) 4167–4175, <https://doi.org/10.1167/iov.18-24673>.

- [50] European Commission, Amended proposal for the AI Act (updated draft). <https://digital-strategy.ec.europa.eu/en/library>, 2023. Accessed on 18 April 2025.
- [51] M. Veale, F.Z. Borgesius, Demystifying the Draft EU Artificial Intelligence Act, *Comput. Law Rev. Int.* 22 (4) (2021) 97–112, <https://doi.org/10.9785/crl-2021-220402>.
- [52] M. Brkan, Do algorithms rule the world? Algorithmic decision-making and the EU General Data Protection Regulation (GDPR), *Int. J. Law Inf. Technol.* 29 (2) (2021) 91–121, <https://doi.org/10.1093/ijlit/eaad023>.
- [53] E.D. Canedo, A.D. Silva, F.B. Araujo, V.S. Carvalho, J.D. Costa, et al., Challenges regarding the compliance with the General Data Protection Law by Brazilian organizations: A survey, in: O. Gervasi, et al. (Eds.), *Computational science and its applications – ICCSA 2021: Vol. 12951. Lecture Notes in Computer Science*, Springer, 2021, pp. 460–475, https://doi.org/10.1007/978-3-030-86970-0_31.
- [54] R. Sinha, A. Agarwal, N. Jain, Ethical implications of India's National Digital Health Mission (NDHM): A policy analysis, *Health Policy. Technol.* 11 (2) (2022) 100624.
- [55] R. Garg, Analysing the NDHM's health data management policy: part 1. Internet Freedom Foundation. <https://internetfreedom.in/analysing-the-ndhms-health-data-management-policy-part-1/accessed>, 2021 on 17April, 2025.
- [56] South African Department of Communications and Digital Technologies, Draft National Artificial Intelligence Policy Framework, Retrieved from, <https://www.dcdt.gov.za/on>, 2021, 17 April, 2025.
- [57] European Commission, Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), COM (2021) (2021) 206 final, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>. Accessed on 18 April 2025.
- [58] European Commission, The Artificial Intelligence Act: AI Regulation in the European Union, EU Digital Strategy Report, 2023. https://ec.europa.eu/digital-strategy/ai-act_en. Accessed on 18 April 2025.
- [59] World Economic Forum, Global technology governance report 2021: harnessing Fourth Industrial Revolution technologies in a COVID-19 world. <https://www.weforum.org/reports/global-technology-governance-report-2021>, 2020. Accessed on 18 April 2025.
- [60] Organisation for Economic Co-operation and Development (OECD), OECD principles on artificial intelligence. <https://oecd.ai/en/ai-principles>, 2019. Accessed on 18 April 2025.
- [61] European Data Protection Board (EDPB) & European Data Protection Supervisor (EDPS), Joint opinion 5/2021 on the proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). <https://edpb.europa.eu>, 2021. Accessed on 18 April 2025.



Review Article

AICB: A benchmark for evaluating the communication subsystem of LLM training clusters

Xinyue Li^{*}, Heyang Zhou, Qingxu Li, Sen Zhang, Gang Lu

Alibaba Cloud, Beijing, 100124, China

ARTICLE INFO

Keywords:

LLM training cluster
Benchmark
Collective communication
Distributed training

ABSTRACT

AICB (Artificial Intelligence Communication Benchmark) is a benchmark for evaluating the communication subsystem of GPU clusters, which includes representative workloads in the fields of Large Language Model (LLM) training. Guided by the theories and methodologies of Evaluatology, we simplified the real-workload LLM training systems through AICB that maintain good representativeness and usability. AICB bridges the gap between application benchmarks and microbenchmarks in the scope of LLM training. In addition, we constructed a new GPU-free evaluation system that helps researchers evaluate the communication system of the LLM training systems. To help the urgent demand on this evaluation subject, we open-source AICB and make it available at <https://github.com/aliyun/aicb>.

1. Introduction

The AI infrastructure is in rapid development with the flourishing of Artificial Intelligence [1,2]. For example, the explosion of the Large Language Model (LLM) applications leads to the fast evolution of the training frameworks [3–5], collective communication algorithms [6], network transports [7], and scale-out and scale-up network architectures [8]. Due to the large number of parameters in LLM, data is distributed across different GPUs for computation, requiring synchronization between these GPUs. Therefore, in LLM training, besides computation, communication also affects training efficiency. Consequently, evaluating the performance of the communication subsystem is a critical subject, that is, ensuring foundational technologies evolve in a manner that is both responsible and conducive to the continued progress in the field.

Some benchmarks are designed to evaluate the communication subsystem of a physical GPU cluster with high-performance scale-up and scale-out networks, including microbenchmarks and application benchmarks. However, the microbenchmarks are designed to evaluate the low-level peer-to-peer or collective communication operations under various message sizes and scales, while the application benchmarks only focus on the end-to-end performance. To bridge the gap, the community demands a new benchmark that produces workloads that mirror real-world LLM tasks, but focuses on the communication subsystem. In response to this demand, we built AICB and constructed the evaluation system using the methods proposed in [9].

The three essences of evaluating the communication subsystem of GPU clusters are as follows: (1) The Evaluation System (ES) is defined as a full-stack GPU cluster that LLM tasks can run. It includes the GPUs, the network infrastructure, and the software components running on it. The Evaluation Conditions should include all the capabilities and configurations of the hardware and software components that are tuned for the LLM training tasks that stakeholders concern. To be more specific, the Reference Evaluation System (RES) of AICB specifically targets the endpoint communication behavior through the end-to-end process of LLM training. (2) AICB provides measurement and testing tools that can generate and reproduce typical workloads in the Reference Evaluation System (RES) and ES. (3) The Value Function is the performance numbers output by AICB. It should give a clear quantified outcome of the comparison between different ECs, such as, the different collective algorithms, different parallel parameters, etc..

We construct the Pragmatic Evaluation System in two ways: (1) Rather than directly using all workloads from the real-world services, AICB is simplified to be more pragmatic. We elaborately select the workloads that can reflect the real-world behaviors with the criteria of spanning from the typical communication operations, message sizes, parallel parameters, optimization skills, and scales. (2) For researchers who lack GPUs, which is not uncommon in both industry and academia, we developed a GPU-free Evaluation System for the same evaluation subject. The NCCL (NVIDIA Collective Communication Library [6]) is hijacked to run on GPU-free cluster, but produces the same traffic.

^{*} Corresponding author.

E-mail addresses: Lixinyue2019@bupt.edu.cn (X. Li), zhouheyang.zhy@alibaba-inc.com (H. Zhou), qingxu.lqx@alibaba-inc.com (Q. Li), zs411030@alibaba-inc.com (S. Zhang), yunding.lg@alibaba-inc.com (G. Lu).

<https://doi.org/10.1016/j.tbench.2025.100212>

Received 6 January 2025; Received in revised form 10 April 2025; Accepted 14 May 2025

Available online 2 June 2025

2772-4859/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

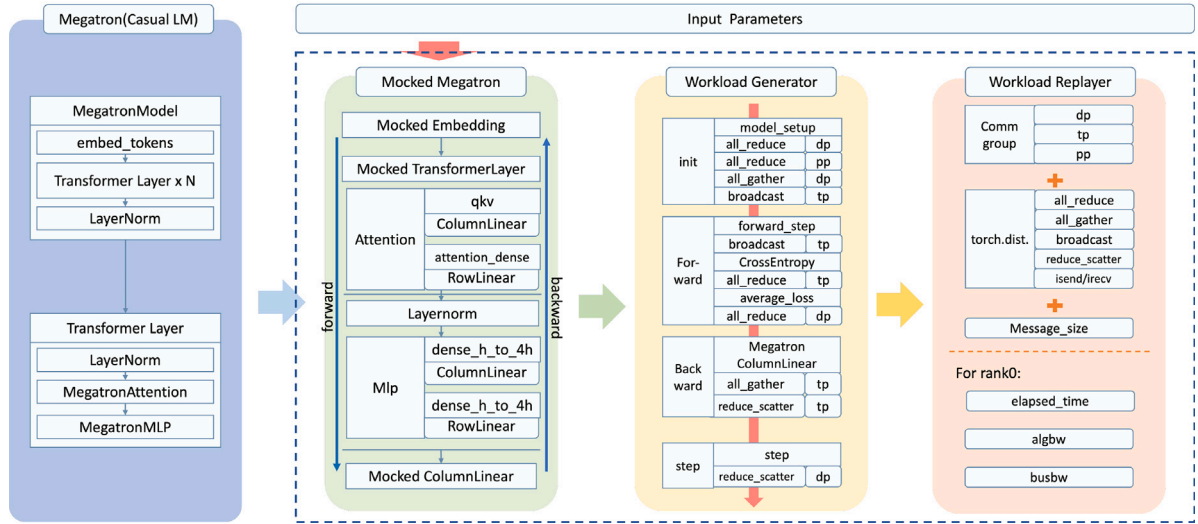


Fig. 1. Design of AICB with Megatron as an example. AICB gets the communication operation through Mocked Model, Workload Generator and Workload Replayer.

Meanwhile, the computation patterns are still kept in the evaluation, as we can collect them by running a specific computation tool on real GPU and afterwards they are embedded in the AICB workloads.

The main contributions of this paper are as follows:

- (1) Guided by the principles of Evaluatology, we propose AICB, a benchmark to evaluate AI communication systems. By “hijacking” the training framework, we construct a pragmatic Evaluation System and develop a GPU-Free system through simulation.
- (2) We use the end-to-end real elapsed time of every specific workload as metrics to evaluate the LLM communication subsystem. Through the assessment of case studies, we demonstrate the practicality of AICB in evaluating corresponding communication performance.
- (3) Beyond the communication behavior, AICB can output the LLM training communication workload, serving as input for [10] to simulate the overall performance of the cluster training.

2. Related works

Existing benchmarks for evaluating physical GPU clusters can be mainly divided into microbenchmarks and application benchmarks. Microbenchmarks focus on assessing specific parts of the GPU cluster framework. For example, Nccl-test [11], developed by NVIDIA, is used to test and verify the performance and correctness of NCCL operations. It is specifically designed for NVIDIA GPUs and fully leverages the parallel processing capabilities. However, it has high dependencies on GPU hardware and CUDA environments. Perftest [12] provides a series of performance microbenchmarks based on Infiniband Verbs for hardware or software tuning and functional testing, but it requires specific hardware support.

Application benchmarks focus on evaluating the performance of model training. MLPerf [13] defines model architectures and training procedures for each benchmark, addressing ML evaluation challenges such as training randomness and significant time differences. AIBench [14] systematically refines and abstracts real-world application scenarios into scene, training, inference, micro, and synthetic AI benchmarks based on MLPerf. While these efforts have advanced ML training benchmark to some extent, there is a lack of overall focus on communication operations during LLM training.

AICB addresses this issue by providing precise evaluations of the communication subsystems. Instead of directly modifying these popular frameworks, AICB extract information through delicate monitoring tools and critical components.

3. AICB design

In the context of AI communication evaluation, a pragmatic composite evaluation system is needed to accurately represent common performance in AI training environments. When we design AICB, the communication system of LLM training is regarded as the evaluation subject with huge EC configurations, which is the input description module in AICB. The input contains a range of parameters to meet the expectations of stakeholders, such as different training models (e.g., GPT, LLaMA) with different scale of neural network, training configuration, popular training framework (e.g., Megatron, DeepSpeed [15]) with relative parallelism and aspects related to collective communication libraries like NCCL. These parameters can also be different Reference Evaluation Conditions (RECs) to compare AI training communication performance.

The core of AICB is implemented by “hijacking” the training framework. Fig. 1 illustrates the working principles of AICB using Megatron framework as an example to training models. Megatron is a highly scalable language model that improves training efficiency and speed through parallel processing of LLM. In practice, the Megatron structure starts from the input tokens, passing through multiple Transformer layer and norm layers, and ultimately reaching a linear layer to generate the model’s output.

Instead of directly modifying frameworks, AICB extracts information through constructing critical components. Mocked Megatron is a simulated version of Megatron designed to simplify the complex model training process. Through the definitions of Mocked Embedding and Mocked Transformer Layer, it includes module of the Attention mechanism and the Multi-Layer Perceptron (MLP), implementing Column Linear and Row Linear calculations for each. Mocked Megatron approximates actual large model operations through these simulated components, aiming to create a simplified yet approximate training process model for subsequent communication system which enhancing the efficiency of simulating and testing communication patterns.

The primary purpose of the Workload Generator is to generate a list of communication workloads during the training process. It sequentially executes the training process according to the Mocked Model module. For example, during the Megatron training, it includes steps such as model initialization, forward propagation and backward propagation. During initialization, the focus is on model setup and data notification, with key operations including All-reduce, All-gather, and Broadcast, correspondingly choosing All-reduce for PP communication based on PP settings. Communication operations in model forward

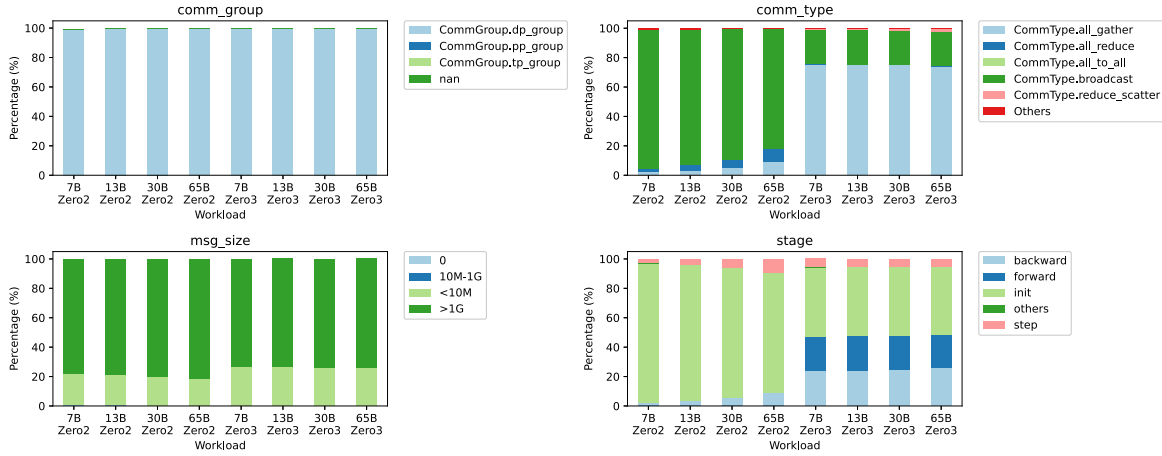


Fig. 2. Communication Distribution of LLaMA with under different scale and framework.

propagation mainly focus on TP, using All-reduce for DP communication when calculating loss. In the backward propagation stage, All-gather and Reduce-scatter update parameters among TP groups, and finally, during the step phase, DP synchronization prepares for the next round of training. Simulation of these steps gives a clear analysis of relations between each step of model training and communication groups with operations.

The task of the Workload Replayer is to apply the communication workload generated in the previous step to actual communication operations. By calling *torch.dist* with corresponding communication operations, it simulates the behavior of the workload, ensuring that the simulated communication load accurately reflects the communication overhead in the actual training process. Additionally, it measures communication operators, message sizes, and corresponding bandwidths on rank 0 of each step.

Notably, to more accurately replicate the communication assessment in LLM training, AICB offers an optional AIOB(Artificial Intelligence Operation Benchmark) mode, which is used to obtain the computation time for each operation of the actual model, accounting for the overlapping of multiple communication streams. Specifically, we break down the most computation-intensive parts, Attention and MLP, according to CUDA operations and extract the corresponding operations in the source code. This division not only facilitates the acquisition of computation time but also aids in the subsequent analysis of timing across different computational cores. Simulating multiple GPUs on a single GPU leverages the symmetry of GPU computation tasks and model training parameters are distilled to extract components that impact computation time. Parameters related to communication are used to slice the input calculation matrices and readjust weights according to the respective segments, simulating the model's division process. This ensures that the dimensions of the multiplied matrices during computation match those in real execution, allowing us to obtain accurate computation times for a partitioned model which is light weight but high accuracy.

4. Case study

4.1. Communication distribution

For the same cluster, AICB can be used to evaluate different distributions of AI training communication operation with composite ECs. Fig. 2 gives an example of the communication characteristics of LLaMA model with different model scales and parallelism strategies. In practice, DeepSpeed are mainly used to focus on data parallelism for synchronization, the DP-Group constitutes the majority within the communication group. Under ZeRO2, models have massive initialization

stage which leads to an amount of broadcast operations for data notification for the first epoch. As model size increases, communication operations for the backward and step stage also increase, resulting in an increased reliance on All-gather and All-reduce. In ZeRO3, model parameters are integrated into the synchronization process, leading to a higher proportion of forward and backward operations compared to ZeRO2, with All-gather becoming the predominant communication method. In terms of message size, large traffic represents approximately 70%–80% and gradually increases with model scale. The distribution of communication operations provides a clear reflection of distributed frameworks in AI cluster training tasks and can be used to validate the communication differences of various RECs deployment.

4.2. Performance evaluation

We compared the workload generated by AICB with the actual training of the Megatron framework, integrating the data from communication groups, communication operation type and communication volume. Table 1 presents the communication results between AICB and realistic training. We tested GPT-7B under the Megatron framework with two A100 nodes, adopting TP = 8, PP = 1, DP = 2 as the parallel configuration. We gather and analyze communication characteristics of the workload generated by AICB and the actual Megatron training. It can be seen that both are quite similar in terms of communication features. Therefore, AICB's workload can represent the communication conditions of Megatron-GPT's actual training effectively, allowing AICB to be used for assessing the model's communication subsystem.

In addition to demonstrating the distribution of the communication subsystem, we have compiled models commonly used and selected those reflecting real-world LLM training workloads, forming a benchmark suite. We use elapsed time as an important baseline metric for evaluating communication system and output the detailed information for each specific communication collective operation.

In our experiments involving fixed collective communication operations, algorithm bandwidth is used to evaluate the performance of the cluster in Fig. 3. Similarly to the physical significance of other types of bandwidth, algorithm bandwidth is calculated based on the actual amount of data transmitted and the time required to complete these transmissions, as shown in (1). During the actual model training, the collective communication library selects the appropriate collective communication algorithm adaptively based on physical topology, communication patterns, and other relevant factors. Consequently, algorithm bandwidth can, to some extent, reflect the ability of the communication library to adapt its operations to the cluster. Higher algorithm bandwidth indicates that the communication library is able

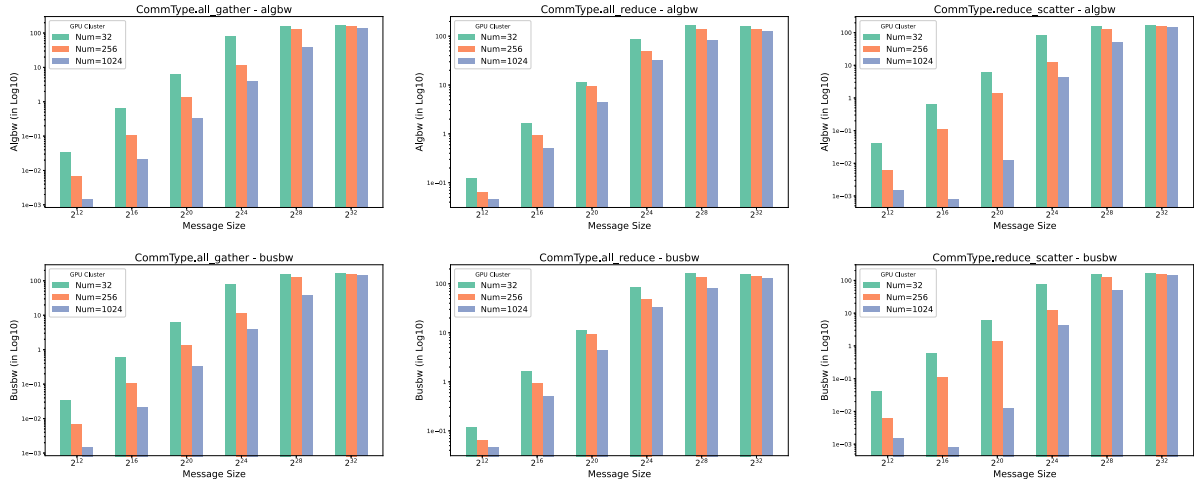


Fig. 3. Performance Evaluation for different cluster scale, message size and communication type.

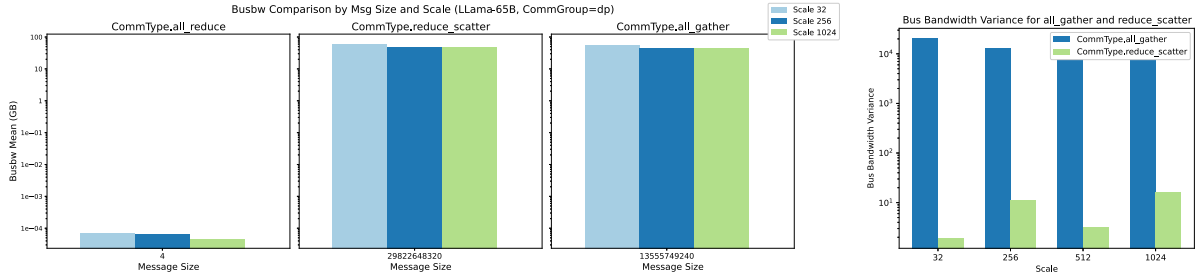


Fig. 4. Practical AICB simulation result of LLaMA 65B with different cluster scale.

Table 1

Comparison of communication between real training and AICB workload with megatron-GPT7B.

Comm type	Comm group	Real training		AICB workload	
		Message size	Number of comms	Message size	Number of comms
All-gather	dp-group	1.57 GB	10	1.55 GB	10
All-gather	tp-group	32 MB	33 280	32 MB	30 720
All-reduce	all	4B	10	4B	10
All-reduce	dp-group	4B	192	4B	160
All-reduce	tp-group	16 KB	576	16 KB	480
All-reduce	tp_group	3.03 MB	10	1 MB	10
All-reduce	tp-group	32 MB	192	32 MB	320
Reduce-scatter	dp-group	3.13 GB	10	3.09 GB	10
Reduce-scatter	tp-group	32 MB	22 688	32 MB	20 480

to utilize hardware resources more effectively, achieving more efficient data transmission.

$$\text{algbw (GB/s)} = \frac{\text{Size (GB)}}{\text{time (s)}} \quad (1)$$

$$\begin{cases} \text{busbw}_{\text{all_reduce}} = \text{algbw} \cdot \frac{2(n-1)}{n} \\ \text{busbw}_{\text{all_gather}} = \text{algbw} \cdot \frac{(n-1)}{n} \\ \text{busbw}_{\text{reduce_scatter}} = \text{algbw} \cdot \frac{(n-1)}{n} \end{cases} \quad (2)$$

It is evident that as the size of the cluster increases, the value of algorithm bandwidth tends to decrease. To eliminate the influence of the number of GPUs on bandwidth, [11] introduces the concept of bus bandwidth, which serves as a metric to assess the efficiency of hardware utilization. This metric is derived by applying a specific calculation formula to the algorithm bandwidth, as shown in (2), to reflect the speed of inter-GPU communication irrespective of the cluster size, i.e., the number of GPUs used. By using this bus bandwidth, we can compare it against the hardware's theoretical peak bandwidth, thereby assessing the actual utilization efficiency of the hardware resources.

In the practical training simulation using the LLaMA65B model, we filtered the collective communication library operations corresponding to the DP group and the message sizes to evaluate the corresponding bus bandwidth. It is evident from Fig. 4 that as the cluster size increases, the bus bandwidth tends to decrease, aligning with the observed performance in Fig. 3. Due to the larger message volume, both the Reduce-scatter and All-gather operations generate higher bus bandwidth. We extracted and calculated the variance of the bus bandwidth for these two operations. It is clearly that the All-gather operation exhibits greater jitter. Intuitively, All-gather involves each participating process collecting data from all other processes, which entails a larger data volume and higher synchronization requirements. In contrast, Reduce-scatter performs partial reduction followed by the scattering of data, resulting in relatively lower synchronization demands and reduced pressure from network condition changes.

4.3. Workload for SimAI

SimAI [10] is a simulator we developed to evaluate complete GPU clusters, including components such as communication subsystems,

computing systems, and network architectures. The workload generated by AICB can serve as input for SimAI to simulate the conditions of model training, including various stages of model training, the size of communication data, communication operations, and the computation time corresponding to each stage. SimAI can form a comprehensive simulation evaluation system based on workload input, network topology information, and related network configurations, making it an important tool for evaluating large model infrastructure.

5. Conclusion

In this paper, we introduce AICB, a benchmark for evaluating the communication subsystem of LLM Training clusters. AICB focuses on communication subsystems in large-scale AI training clusters and defines appropriate ranges for RC to construct ES. By “hijacking” distributed training frameworks, it simulates specific collective communication operations. In addition to visualizing communication distribution, AICB uses bus bandwidth as a metric to evaluate the compatibility with specified clusters. AICB offers precise simulation and accurate evaluation of collective communications, providing substantial support for simulating and evaluating LLM training.

CRediT authorship contribution statement

Xinyue Li: Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Heyang Zhou:** Software, Conceptualization. **Qingxu Li:** Software, Conceptualization. **Sen Zhang:** Validation, Conceptualization. **Gang Lu:** Validation, 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.

References

- [1] Z. Jiang, H. Lin, Y. Zhong, Q. Huang, Y. Chen, Z. Zhang, Y. Peng, X. Li, C. Xie, S. Nong, Y. Jia, S. He, H. Chen, Z. Bai, Q. Hou, S. Yan, D. Zhou, Y. Sheng, Z. Jiang, H. Xu, H. Wei, Z. Zhang, P. Nie, L. Zou, S. Zhao, L. Xiang, Z. Liu, Z. Li, X. Jia, J. Ye, X. Jin, X. Liu, MegaScale: Scaling large language model training to more than 10,000 GPUs, 2024, [arXiv:2402.15627](https://arxiv.org/abs/2402.15627).
- [2] W. Li, X. Liu, Y. Li, Y. Jin, H. Tian, Z. Zhong, G. Liu, Y. Zhang, K. Chen, Understanding communication characteristics of distributed training, in: Proceedings of the 8th Asia-Pacific Workshop on Networking, APNet '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–8, [http://dx.doi.org/10.1145/3663408.3663409](https://dx.doi.org/10.1145/3663408.3663409).
- [3] M. Shoeny, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, Megatron-LM: Training multi-billion parameter language models using model parallelism, 2020, [arXiv:1909.08053](https://arxiv.org/abs/1909.08053).
- [4] D. Narayanan, M. Shoeny, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, M. Zaharia, Efficient large-scale language model training on GPU clusters using megatron-LM, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '21, Association for Computing Machinery, New York, NY, USA, 2021, [http://dx.doi.org/10.1145/3458817.3476209](https://dx.doi.org/10.1145/3458817.3476209).
- [5] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeny, B. Catanzaro, Reducing activation recomputation in large transformer models, 2022, [arXiv:2205.05198](https://arxiv.org/abs/2205.05198).
- [6] NCCL, NVIDIA collective communications library (NCCL), 2024, <https://developer.nvidia.com/nccl> (Online; Accessed 4 October 2024).
- [7] J. Zhang, Y. Wang, X. Zhong, M. Yu, H. Pan, Y. Zhang, Z. Guan, B. Che, Z. Wan, T. Pan, T. Huang, PACC: A proactive CNP generation scheme for datacenter networks, IEEE/ACM Trans. Netw. 32 (3) (2024) 2586–2599, [http://dx.doi.org/10.1109/TNET.2024.3361771](https://dx.doi.org/10.1109/TNET.2024.3361771).
- [8] K. Qian, Y. Xi, J. Cao, J. Gao, Y. Xu, Y. Guan, B. Fu, X. Shi, F. Zhu, R. Miao, C. Wang, P. Wang, P. Zhang, X. Zeng, E. Ruan, Z. Yao, E. Zhai, D. Cai, Alibaba HPN: A data center network for large language model training, in: Proceedings of the ACM SIGCOMM 2024 Conference, in: ACM SIGCOMM '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 691–706, [http://dx.doi.org/10.1145/3651890.3672265](https://dx.doi.org/10.1145/3651890.3672265).
- [9] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162, [http://dx.doi.org/10.1016/j.tbenc.2024.100162](https://dx.doi.org/10.1016/j.tbenc.2024.100162), URL <https://www.sciencedirect.com/science/article/pii/S2772485924000140>.
- [10] X. Wang, Q. Li, Y. Xu, G. Lu, D. Li, L. Chen, H. Zhou, L. Zheng, S. Zhang, Y. Zhu, et al., SimAI: Unifying architecture design and performance tuning for large-scale large language model training with scalability and precision, in: 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25), 2025.
- [11] Nccltests, NCCL-tests, 2024, <https://github.com/NVIDIA/nccl-tests> (Online; Accessed 4 October 2024).
- [12] PerfTest, Infiniband verbs performance tests, 2024, <https://github.com/linux-rdma/perftest> (Online; Accessed 4 October 2024).
- [13] P. Mattson, C. Cheng, G. Diamos, C. Coleman, P. Micikevicius, D. Patterson, H. Tang, G.-Y. Wei, P. Bailis, V. Bittorf, et al., Mlperf training benchmark, Proc. Mach. Learn. Syst. 2 (2020) 336–349.
- [14] F. Tang, W. Gao, J. Zhan, C. Lan, X. Wen, L. Wang, C. Luo, Z. Cao, X. Xiong, Z. Jiang, T. Hao, F. Fan, F. Zhang, Y. Huang, J. Chen, M. Du, R. Ren, C. Zheng, D. Zheng, H. Tang, K. Zhan, B. Wang, D. Kong, M. Yu, C. Tan, H. Li, X. Tian, Y. Li, J. Shao, Z. Wang, X. Wang, J. Dai, H. Ye, Aibench training: Balanced industry-standard AI training benchmarking, in: 2021 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, 2021, pp. 24–35, [http://dx.doi.org/10.1109/ISPASS51385.2021.00014](https://dx.doi.org/10.1109/ISPASS51385.2021.00014).
- [15] J. Rasley, S. Rajbhandari, O. Ruwase, Y. He, DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 3505–3506, [http://dx.doi.org/10.1145/3394486.3406703](https://dx.doi.org/10.1145/3394486.3406703).