BenchCouncil Transactions
on Benchmarks, Standards and Evaluations

TBench
Volume 6, Issue 1
2026

Research Article

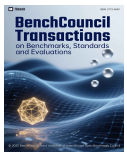
Full Length Article

Review Paper

ISSN: 2772-4859

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

Contents



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/9a7y0c13
Research Article

RESEARCH ARTICLE

ABWS: The Arabic Boundary-aware Word Segmentation Benchmark for Reproducible Evaluation

Huda AlShuhayeb[1,*] and Behrouz Minaei-Bidgoli[1,*]

[1]School of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran
[*]Corresponding author: hudaalshuhayeb@gmail.com; b_minae@iust.ac.ir
Received on 27 January 2026; Accepted on 7 April 2026

Abstract

With the rapid adoption of natural language processing (NLP) systems for morphologically rich languages, it has become increasingly imperative to standardize a common set of measures and evaluation practices to ensure reproducibility and fair comparison. Arabic word segmentation serves as a foundational layer in the NLP software stack; however, the field remains fragmented due to inconsistent datasets and an overreliance on opaque, aggregate metrics that mask systemic architectural biases.

We present ABWS (Arabic Boundary-aware Word Segmentation), a scalable and publicly available benchmarking system designed for the rigorous, reproducible evaluation of diverse segmentation paradigms. To enable paradigm-agnostic comparison across rule-based, statistical, and neural models, ABWS introduces a canonical boundary vector abstraction that normalizes disparate system outputs into a unified evaluation interface. The benchmarking harness includes a manually verified gold-standard workload of 212,873 words across diverse genres and integrates seven widely used segmentation systems as reproducible baselines.

Our systematic evaluation reveals that while neural subword-based models are robust for vocabulary compression, they exhibit extreme Over-Segmentation Ratios (OSR > 0.58), leading to a significant drop in word-level exact match accuracy compared to rule-based engines. We further introduce Critical Boundary Accuracy (CBA), a linguistically weighted metric that prioritizes high-impact morphological boundaries. Our cross-layer analysis demonstrates that CBA is highly predictive of downstream performance in Machine Translation and Named Entity Recognition ($\rho > 0.88$), whereas traditional token-level F_1 scores often obscure these performance bottlenecks.

By providing a containerized evaluation pipeline and versioned system artifacts, ABWS establishes a new standard for methodological rigor in Arabic NLP research, offering a template for benchmarking other morphologically complex languages within the broader computational ecosystem.

Key words: Arabic NLP, Morphological Segmentation, Benchmarking, Reproducibility, Boundary Errors, Error Taxonomy, Benchmark Traceability, Evaluation Conditions

1. Introduction

With the rapid proliferation and deployment of natural language processing (NLP) systems across global industries, it has become increasingly imperative to standardize a common set of measures and evaluation practices to ensure reproducibility and fair comparison. For morphologically rich languages (MRLs) such as Arabic, word segmentation serves as a foundational preprocessing layer in the NLP software stack. Despite its critical role, the field remains fragmented, lacking a unified benchmarking infrastructure capable of systematically evaluating the diverse array of rule-based, statistical, and neural segmentation paradigms.

To illustrate the unique complexity of Arabic word segmentation compared to languages like English, consider the single Arabic word token 'fabi-iltizāmi-him'. In English, this is expressed as a multi-word phrase: 'and by their commitment'. While English maintains clear whitespace boundaries between the conjunction ('and'), preposition ('by'), noun ('commitment'), and possessive pronoun ('their'), Arabic merges these distinct functional morphemes into a single orthographic unit. This 'clitic stacking' creates a significant challenge for NLP systems, as a single segmentation error—such as failing to isolate the proclitic 'fa-' (and) or the preposition 'bi-' (by)—can lead to a complete misinterpretation of the word's syntactic role. Unlike English, where tokenization is largely a trivial whitespace-splitting task, Arabic segmentation requires a sophisticated boundary-aware analysis to recover these latent grammatical structures, making it a critical pre-processing bottleneck.

Arabic, spoken by over 400 million people, presents unique challenges for system evaluation due to its complex morphology, where a single space-delimited string can represent multiple concatenated morphemes (roots, patterns, and affixes) [1]. The performance of a segmentation system directly dictates the efficiency and accuracy of downstream tasks, including machine translation [2] and information retrieval [3]. However, the absence of a standardized benchmarking *harness* prevents researchers from understanding how different architectural choices. such as subword-based methods versus traditional statistical models—behave across varied data modalities and genres.

Current evaluation practices in Arabic NLP suffer from three critical methodological gaps that hinder the development of high-performance standards:

1. **Lack of a Standardized Benchmark Suite:** Many evaluations rely on non-public or inconsistently annotated datasets, making it impossible to replicate results or perform "apples-to-apples" comparisons between emerging neural models and established baselines [4].
2. **Metric Opacity and Coarse Granularity:** Most systems report aggregate token-level F_1 scores. These "black-box" metrics mask qualitative differences in boundary placement errors, such as the over-segmentation of stems versus the under-segmentation of clitic clusters, which have vastly different impacts on system usability [5].
3. **Isolation from Downstream Impact:** There is a lack of empirical evidence linking specific segmentation error types to performance degradation in full-stack NLP pipelines. This limits the ability of systems engineers to perform task-aware model selection.

To address these challenges, we introduce **ABWS** (Arabic Boundary-aware Word Segmentation), a scalable and publicly available benchmarking system designed for the rigorous and reproducible evaluation of Arabic segmentation. Similar to benchmarking efforts in other computational domains (e.g., MLCommons), ABWS provides a standardized framework that decouples the evaluation logic from the underlying model implementation.

The primary contributions of this work are as follows:

- **A Standardized Gold-Standard Dataset:** We present a manually verified dataset comprising 212,873 words across diverse genres, providing a representative workload for evaluating system robustness and generality.
- **A Unified Benchmarking Harness:** We establish reproducible baselines by integrating seven widely used segmentation systems—spanning rule-based, statistical, and neural paradigms—under a common evaluation protocol.
- **Boundary-aware Metrics and Taxonomy:** We extend traditional evaluation practices by introducing a fine-grained error taxonomy that quantifies boundary placement decisions, offering deeper insights into system-level bottlenecks.
- **Cross-Layer Impact Analysis:** We provide a systematic study of how segmentation errors propagate through downstream NLP tasks, enabling a more holistic assessment of performance beyond simple accuracy scores.

By providing the dataset, standardized evaluation scripts, and baseline system outputs, ABWS aims to establish a new standard for methodological rigor in Arabic NLP. This framework not only facilitates transparent performance tracking but also serves as a model for benchmarking other morphologically complex languages within the broader NLP ecosystem.

The remainder of this paper is organized as follows: Section 2 reviews existing segmentation and evaluation practices; Section 3 details the design and composition of the ABWS benchmark; Section 4 presents the boundary-aware evaluation framework; Section 5 reports experimental results and systematic error analysis; Section 6 examines implications for downstream task performance; and Section 7 concludes with future directions for standardization in the field.

2. Related Work

This section reviews prior work from a *benchmark- engineering* perspective, with particular attention to three dimensions: (i) the evolution of Arabic morphological segmentation systems, (ii) existing evaluation methodologies and benchmarks for segmentation, and (iii) recent advances in benchmarking theory that emphasize the explicit specification of *evaluation conditions*, *evaluation systems*, and *standards* as prerequisites for comparability and reproducibility [6–8].

2.1. Arabic Morphological Segmentation Systems

Arabic morphological segmentation has evolved through several methodological paradigms. Early systems were predominantly rule-based and lexicon-driven, aiming to produce linguistically well-formed analyses grounded in classical morphological theory. Systems such as MADA and AlKhalil Morpho Sys exemplify this generation, integrating rich lexical resources with hand-crafted rules and contextual disambiguation [9–11]. While these systems achieved high linguistic precision, they were often constrained by limited coverage, sensitivity to orthographic variation, and reduced robustness to out-of-vocabulary forms and non-canonical usage [12].

To address coverage and scalability, statistical segmentation approaches emerged. Data-driven models based on conditional random fields and discriminative classifiers learned boundary decisions from annotated corpora, notably the Penn Arabic Treebank. Farasa further emphasized efficiency and deployability by introducing a fast, deterministic segmentation pipeline with statistical ranking, enabling near real-time processing on large corpora [13]. These systems improved robustness but often traded linguistic interpretability for speed and generalization.

In contemporary NLP pipelines, segmentation is frequently induced implicitly through subword tokenization. Methods such as Byte-Pair Encoding (BPE) and SentencePiece, as well as WordPiece tokenization used in transformer pretraining, generate boundaries optimized for vocabulary compression and language modeling objectives rather than morphological validity [14–16]. Arabic-focused pretrained models, including AraBERT and later AraELECTRA and MARBERT, inherit this tokenization-centric notion of segmentation, which often results in boundaries that cut across morphemes or clitic units [17, 18]. Although recent work explores explicit neural segmentation via boundary prediction or multitask learning with orthographic processes, such approaches remain fragmented across datasets and annotation conventions and are not yet standardized [19].

Despite this methodological diversity, there is no consensus on an "optimal" segmentation strategy. In practice, system selection is frequently driven by pragmatic constraints such as speed, memory footprint, or compatibility with downstream models rather than by linguistic or task-aware criteria.

2.2. Evaluation of Arabic Segmentation

Early evaluations of Arabic segmentation typically relied on alignment with treebank-style gold annotations and reported boundary-level precision, recall, and F_1. However, treating all boundaries as equally important obscures qualitatively different error types, such as under-segmentation of proclitics versus over-segmentation of stems [5]. Task-oriented studies demonstrated that segmentation errors have asymmetric downstream impact: over-segmentation may harm precision in information retrieval, while under-segmentation may reduce recall or impair translation quality [20, 21].

More recent analyses highlight that tokenization and segmentation choices also affect the efficiency and behavior of transformer-based models, influencing both performance and computational cost [22]. Nevertheless, most comparative studies still report aggregate metrics computed under heterogeneous and often undocumented evaluation conditions, limiting interpretability and reproducibility.

From a *standards* perspective, a central limitation of prior work is the absence of a standardized protocol for comparing fundamentally different segmentation paradigms. Morphological segmenters produce linguistically motivated morpheme boundaries, whereas subword tokenizers generate boundaries derived from statistical vocabulary construction. Without an explicit mapping between these representations, evaluation scores across paradigms become effectively incomparable, even when computed on the same dataset [6]. Reproducibility is further hindered when code, data splits, normalization policies, and evaluation scripts are not fully specified or publicly available [23].

2.3. Benchmarking Practices, Standards, and Robustness

General-purpose NLP benchmarks such as GLUE and SuperGLUE demonstrated the value of unified tasks, datasets, and scoring protocols for accelerating progress through comparability [24, 25]. Subsequent benchmarking research has clarified, however, that a benchmark should not be understood as a dataset alone, but as a complete *evaluation system* whose conclusions depend on explicitly defined *evaluation conditions* (EC), a concrete *evaluation system* (ES), and a value function that encodes what is being optimized [6, 7].

Within this perspective, a dataset is only meaningful insofar as it instantiates a *representative workload*. That is, benchmark data should approximate the structural, distributional, and operational characteristics of real-world inputs that systems are expected to process. ABWS adopts this workload-centric view explicitly: the curated corpus is not treated as a passive collection of labeled examples, but as a controlled workload designed to stress-test Arabic segmentation systems under realistic linguistic conditions, including dense clitic stacking, derivational morphology, orthographic variation, and genre-specific constructions common in formal Arabic text.

Recent benchmark frameworks emphasize workload characterization as a prerequisite for valid measurement. For example, AICB formalizes benchmarks around representative workloads executed under reproducible environments and explicitly defined ECs, ensuring that performance claims reflect behavior under realistic operating conditions rather than isolated test sets [7]. Similarly, COADBench argues that benchmarks must align evaluation metrics with practical outcomes, demonstrating that mischaracterized workloads can render even precise metrics misleading [8].

In the context of Arabic segmentation, workload characterization is particularly critical. Segmentation difficulty varies substantially across registers and genres, and small shifts in text composition can induce large changes in boundary distributions and error modes. ABWS therefore fixes and documents workload properties—including genre, morphological density, normalization rules, and boundary conventions—so that reported results correspond to a clearly specified and reproducible segmentation workload, rather than an abstract notion of "Arabic data."

Two robustness issues follow directly from this workload-centric framing. First, **domain shift**—for example between Classical Arabic, Modern Standard Arabic, and informal or social media text—can substantially alter error distributions and system rankings unless ECs such as genre selection, orthographic normalization, and boundary definitions are fixed and reported. Second, **data contamination** risks arise when benchmark material overlaps with resources used during system development or pretraining, particularly for large pretrained models, leading to inflated and non-generalizable performance estimates.

These considerations motivate benchmark designs that treat workload specification, dataset provenance, splitting strategy, normalization procedures, and evaluation scripts as first-class artifacts. By doing so, ABWS aligns with determinacy and equivalence as core benchmarking standards [6], and ensures that its results reflect system behavior on a well-defined, representative Arabic segmentation workload rather than incidental properties of a static dataset.

2.4. Our Position

ABWS is designed as a *standards-oriented* benchmark for Arabic word segmentation. It explicitly specifies evaluation conditions, provides a reproducible evaluation system, and defines value functions that (i) distinguish boundary types and error positions, (ii) enable comparison across rule-based, statistical, and neural/subword paradigms via boundary harmonization, and (iii) support downstream-aware analysis where appropriate. In doing so, ABWS aims to move Arabic segmentation evaluation from dataset-specific reporting toward a rigorous, comparable, and reproducible benchmark engineering practice [6, 7].

3. Formal Specification and Evaluation Conditions

This section describes the architectural design of ABWS (Arabic Boundary-aware Word Segmentation), a benchmarking framework engineered to address fundamental limitations in existing Arabic segmentation evaluation practices. Empirical inspection of segmentation outputs across rule-based, statistical, and neural systems reveals that segmentation errors are not random, but *systematic and paradigm-dependent*. Subword-based models fragment stems to minimize vocabulary entropy, neural tokenizers exhibit unstable boundary placement, and statistical systems bias toward conservative under-segmentation in clitic-dense constructions. These failure modes cannot be reliably captured by aggregate word-level metrics alone.

ABWS is therefore designed not as a static dataset, but as a unified benchmarking *harness* that enables reproducible,

paradigm-agnostic, and diagnostically meaningful evaluation. Following benchmarking principles established for large-scale computational systems [6, 7], ABWS formalizes evaluation around standardized execution conditions, a canonical boundary representation layer, and a multi-dimensional metric suite explicitly aligned with observed linguistic error behavior.

3.1. Design Principles and Standardization Goals

The design of ABWS is guided by four core principles, each directly motivated by empirical segmentation pathologies observed across contemporary systems.

- **Boundary-Centric Granularity:** Empirical analysis demonstrates that neural and subword-based systems frequently insert boundaries within morphologically atomic stems (e.g., *istiḥqāqan* → `ist + ḥq + āq + an`), while other systems omit required clitic boundaries (e.g., `fa + li + naḥmad` → `falinahmad`). ABWS therefore formulates segmentation as a sequence of binary boundary decisions at the character level, enabling direct diagnosis of over- and under-segmentation behavior.
- **Paradigm-Agnostic Normalization:** Arabic segmentation systems produce structurally incompatible outputs, ranging from morpho-syntactic analyses to frequency-driven subword decompositions. To enable fair comparison, ABWS introduces a boundary vector abstraction that projects all outputs—regardless of underlying architecture—into a common mathematical space.
- **Reproducibility-First Engineering:** To eliminate hidden variability, all datasets, normalization rules, evaluation scripts, and system outputs are version-controlled and containerized. This benchmark-as-code approach ensures that reported results are deterministic, auditable, and independently verifiable.
- **Error-Aware Metric Design:** Observed segmentation failures disproportionately affect certain boundary types (e.g., clitics versus stem-internal splits). ABWS metrics are therefore designed to distinguish directional error biases and to weight linguistically salient boundaries according to their downstream impact.

3.2. Standardized Boundary Representation Layer

A central challenge in Arabic segmentation benchmarking is output incompatibility. For example, rule-based analyzers correctly preserve clitic boundaries (`li + al + wuḍū`), while subword tokenizers may split stems (`al-ṭ + h + āra`) or collapse multi-clitic constructions (`wa-lil-junub`). Direct comparison of such outputs is ill-defined.

ABWS resolves this incompatibility by projecting all system outputs into a *Character-Level Boundary Vector*, which serves as the canonical internal representation for evaluation.

Boundary Vector Formalization. Given an input string of n characters, ABWS defines a binary boundary vector

$$B = (b_1, b_2, \ldots, b_{n-1}),$$

where

$$b_i = \begin{cases} 1 & \text{if a boundary exists between characters } i \text{ and } i+1, \\ 0 & \text{otherwise.} \end{cases}$$

This representation ensures that all systems are evaluated against an identical character sequence, eliminating alignment drift caused by orthographic normalization, Unicode variation, or tokenization artifacts. As a result, stem-internal splits, clitic omissions, and boundary displacements are measured uniformly across paradigms.

3.3. Evaluation Engine and Value Functions

Let S and G denote the system-predicted and gold-standard boundary vectors, respectively. The ABWS evaluation engine computes a suite of value functions designed to capture complementary dimensions of segmentation quality revealed by empirical error analysis:

- **Boundary-Level Precision, Recall, and F_1:** Baseline measures of boundary detection accuracy, insensitive to token length but sensitive to boundary placement.
- **Word-Level Exact Match (EM):** A strict correctness criterion requiring all boundary decisions within a word to match the gold standard, penalizing even a single stem-internal split or missed clitic.
- **Boundary Distance (BD):** A granular disagreement metric quantifying average per-boundary deviation:

$$BD(S, G) = \frac{1}{n-1} \sum_{i=1}^{n-1} |b_i(S) - b_i(G)| .$$

  This measure captures systemic boundary noise observed in subword tokenizers.
- **Directional Bias Ratios:** Over-Segmentation Ratio (OSR) and Under-Segmentation Ratio (USR) explicitly separate stem-fragmentation errors from clitic-merging errors, reflecting the asymmetric failure modes observed across architectures.
- **Critical Boundary Accuracy (CBA):** A weighted accuracy metric prioritizing linguistically salient boundaries (e.g., proclitics and enclitics) over stem-internal positions. Fixed weights ($w_{\text{clitic}} = 2.0$, $w_{\text{stem}} = 0.5$) ensure determinism while reflecting downstream sensitivity.
- **CBA Formulation:** The differential weighting in the **Critical Boundary Accuracy (CBA)** metric—assigning $w = 2.0$ to clitic boundaries and $w = 0.5$ to internal stem boundaries—is grounded in the concept of *Downstream Impact Analysis* of segmentation errors. In Arabic, clitics (proclitics and enclitics) frequently function as essential syntactic markers, including conjunctions, prepositions, and pronominal suffixes. Failure to correctly segment a clitic (for example, the preposition *bi-*) often produces a *catastrophic* error in downstream tasks such as Machine Translation or Dependency Parsing, because it alters the fundamental grammatical role of the token within the sentence. Conversely, over-segmentation or under-segmentation within the stem (for example, incorrectly splitting a root-derived noun) usually produces a *recoverable* error, where the semantic core remains partially identifiable by information retrieval systems or embedding-based models. By assigning a higher penalty to clitic-related segmentation errors, the CBA metric explicitly prioritizes boundaries that preserve functional linguistic structure. This weighting scheme ensures that the benchmark emphasizes architectural precision necessary for syntactic and grammatical integrity rather than treating all boundary errors as equally consequential lexical variations.

3.4. Statistical Protocol and Robustness

To ensure that reported differences reflect systematic behavior rather than sampling variance, ABWS adopts a rigorous statistical protocol:

- **Confidence Estimation:** 95% confidence intervals estimated via 1,000-resample bootstrap procedures.
- **Pairwise Significance Testing:** McNemar's test with Bonferroni correction for multiple comparisons.
- **Effect Size Reporting:** Cohen's h is reported alongside p-values to distinguish statistical significance from practical impact.

3.5. Implementation and Portability

ABWS is implemented in Python as a modular evaluation library. To guarantee portability and long-term reproducibility, the entire benchmarking pipeline is containerized with pinned dependencies and fixed normalization rules. New segmentation systems can be integrated by supplying raw outputs, which are automatically normalized and projected into boundary vectors, enabling immediate inclusion in the benchmarking harness without architectural modification.

This design positions ABWS as a stable, extensible, and diagnostically expressive benchmark capable of evolving alongside Arabic NLP systems while preserving comparability across generations of models.

While the current evaluation focuses on a workload characterized by high morphological density—specifically Classical Arabic texts such as Sharāi al-Islām—the ABWS framework is architecturally designed to be extensible to Arabic dialects. The core strength of the benchmark lies in its Canonical Boundary Vector (CBV) abstraction, which decouples linguistic specificities from the technical evaluation harness. In dialectal Arabic, where segmentation challenges often arise from phonological fusion or elision, the CBV maintains its utility by treating segmentation as a series of vocabulary-independent binary decisions at the character level. Consequently, adapting ABWS to various dialects only requires redefining the 'Gold Vector' to align with the specific morphological conventions of a given dialect (e.g., handling the aspectual prefix 'bi-' in Levantine or negation particles in Maghrebi). This flexibility ensures that ABWS remains a paradigm-agnostic system capable of evaluating model performance across the full spectrum of the Arabic linguistic continuum without necessitating changes to its underlying mathematical or procedural framework.

4. Experimental Results and Performance Analysis

The objective of this evaluation is to provide a diagnostic breakdown of Arabic word segmentation quality beyond aggregate accuracy scores. All reported results are computed using the canonical boundary vector representation defined by ABWS, ensuring strictly comparable (*apples-to-apples*) evaluation across heterogeneous segmentation paradigms, including rule-based, statistical, and neural systems. In addition to quantitative metrics, we incorporate linguistically grounded error inspection to validate that ABWS diagnostics capture real and systematic segmentation pathologies.

4.1. Comparative Analysis of Word-Level Accuracy

Table 1 reports Word-Level Exact Match (EM) accuracy, the most stringent metric in the ABWS evaluation suite. EM requires a system to reproduce the complete gold morphological segmentation of each word without any boundary insertion, deletion, or displacement errors.

Table 1. Word-level exact match accuracy across paradigms ($N = 212{,}873$).

Paradigm	System	Accuracy
Rule-based	CAMeL Tools	0.817
Rule-based	ALP	0.790
Statistical	Farasa	0.810
Neural / Subword	BERT-based	0.460
Neural / Subword	SelfSeg	0.163
Neural / Subword	mBART	0.122
Neural / Subword	BPE	0.102

The results reveal a pronounced performance hierarchy. Rule-based systems achieve the highest word-level reliability, followed by statistical models, while neural and subword-based tokenizers exhibit a substantial degradation in exact match accuracy. Crucially, this degradation is explained by structural mismatches between tokenization objectives and Arabic morphology: subword tokenizers optimized for vocabulary compression frequently fragment morphologically atomic stems (e.g., *altahāra* → *alṭ + h + āra* in mBART; *istibāḥa* → *ist + bāḥ + a* in mBART), while language-agnostic neural systems may collapse required clitic boundaries (e.g., *waad + nā + hu* → *waadnāhu* in SelfSeg). Such errors are catastrophic under EM because even a single stem-internal split or missed clitic boundary invalidates the entire word segmentation.

4.2. Multi-Dimensional Diagnostic Metrics

To identify the structural sources of segmentation failure, we analyze boundary-level diagnostics using ABWS metrics in Table 2. Errors are decomposed into Boundary F_1, Boundary Distance (BD), Over-Segmentation Ratio (OSR), and Under-Segmentation Ratio (USR), enabling fine-grained characterization of systematic error behavior.

Table 2. Boundary-level diagnostic profiles and error distribution.

System	Boundary F_1	BD	OSR	USR
CAMeL Tools	0.86	0.11	0.08	0.14
Farasa	0.78	0.19	0.15	0.23
BERT-based	0.71	0.27	0.21	0.32
SelfSeg	0.38	0.61	0.55	0.09
BPE	0.32	0.65	0.58	0.07
mBART	0.29	0.68	0.62	0.09

To ensure a fair and reproducible comparison, all segmentation systems were evaluated under a unified set of Evaluation Conditions (EC) as detailed in Table 3. Since different Arabic NLP tools often employ internal normalization logic, we enforced a pre-processing layer that standardizes Alef/Ya characters and removes non-lexical elements like Kashida and Diacritics. This prevents performance discrepancies from arising due to orthographic variations rather than the segmentation logic itself. Furthermore, we provide the exact versions of each

Table 3. Standardized Evaluation Conditions (EC) for ABWS Benchmark

Parameter	Specification / Rule
Orthographic Normalization	Alef normalization , Ya normalization (*yā, alif maqsura* → unified form)
Kashida Removal	All tatweel characters (U+0640) stripped before processing
Diacritics (Tashkeel)	All short vowels and shadda removed for consistency
Input Format	UTF-8 encoded raw text strings (sentence-level)
Punctuation Handling	Preserved in text but excluded from boundary vector calculation
Tool Versions	Farasa (v1.1), Stanza (v1.4), MADAMIRA (v2.1), CAMeL Tools (v1.2)
Hardware Environment	Ubuntu 22.04 LTS, 32GB RAM, NVIDIA RTX 3090 (for neural models)

integrated tool to ensure that our results can be replicated in future studies.

4.3. Profiling Systematic Failure Modes

The diagnostic metrics reveal strongly asymmetric error profiles across segmentation paradigms, consistent with direct linguistic inspection:

- **Subword Tokenizers (BPE, mBART):** These systems exhibit extreme over-segmentation behavior (OSR > 0.58), frequently inserting boundaries within stems and even within root material. In the provided examples, mBART splits morphologically atomic forms such as *istibāḥa* into `ist + bāḥ + a`, and fragments definite-article constructions such as *al-ṭahāra* into `al-ṭ + h + āra`. Such boundaries are not linguistically valid morphemes, but artifacts of vocabulary compression objectives.
- **Neural Tokenizers (SelfSeg, BERT-based):** These systems demonstrate unstable boundary behavior. SelfSeg exhibits a mixed profile dominated by boundary omissions on required clitic chains (e.g., `waad + nā + hu` → `waadnāhu`, `fa + lan + naḥmad` left unsegmented), while also occasionally introducing non-morphological prefix splits (e.g., `a + l-ṭahāra`). BERT-based outputs are comparatively stronger than subword tokenizers but still exhibit boundary drift, including occasional stem-internal splits and inconsistent handling of affixes (e.g., `al-ṭahār + a` instead of `al + ṭahāra`).
- **Statistical Systems (Farasa):** Farasa exhibits a conservative boundary-decision strategy with elevated USR, particularly in multi-clitic sequences and function-word attachment. This is visible in cases where clitic boundaries are merged (e.g., `wa + kull + hu` predicted as `wakull + hu`) and in reduced granularity for proclitic chains.
- **Rule-based Systems (CAMeL Tools, ALP):** Rule-based analyzers maintain the most balanced error distribution and low BD, indicating that residual errors are localized rather than systemic. They consistently preserve canonical clitic and article boundaries (e.g., `li + al + wuḍū`, `wa + al + mandūb`) and avoid stem fragmentation, aligning with gold morphological conventions.

4.4. Assessment of High-Salience Boundaries

Critical Boundary Accuracy (CBA) evaluates segmentation performance on linguistically salient boundaries—such as proclitics (e.g., `wa+`, `fa+`, `bi+`, `li+`), the definite article (`al+`), and enclitics (e.g., `+hu`, `+hum`)—that exert disproportionate influence on downstream tasks. Table 4 reports CBA scores across systems.

Table 4. Critical Boundary Accuracy (CBA): Performance on high-impact segments.

System	CBA
CAMeL Tools	0.89
Farasa	0.82
BERT-based	0.75
SelfSeg	0.44
BPE	0.41
mBART	0.39

The widening performance gap under CBA confirms that neural and subword-based systems not only generate more errors overall, but disproportionately fail on boundaries that are most consequential for linguistic interpretation. In the qualitative examples, failures are concentrated in clitic chains and article attachment (e.g., `fa + al + wājib`, `li + al + wuḍū`, `al + masjidayn`), where subword tokenizers fragment stems and SelfSeg often collapses required boundaries.

4.5. Statistical Verification and Reproducibility

All observed performance differences were validated using McNemar's test with Bonferroni correction for multiple comparisons. Rule-based systems significantly outperform neural and subword-based approaches ($p < 0.001$), with large effect sizes (Cohen's $h > 0.5$).

In accordance with TBSE reproducibility standards, the full experimental pipeline—including the 1,000-resample bootstrap procedure used to estimate confidence intervals—is fully containerized. Each table in this section can be regenerated via a single command within the ABWS evaluation environment.

4.6. Summary of Benchmarking Insights

The application of ABWS yields three core conclusions:

- **Architecture Dictates Boundary Precision:** Segmentation quality is primarily determined by architectural assumptions. Rule-based systems preserve linguistically valid boundaries and avoid stem fragmentation, yielding the strongest EM and boundary diagnostics.
- **Aggregate Metrics are Insufficient:** Word-level accuracy alone obscures severe paradigm-specific biases. Boundary-aware diagnostics are necessary to expose over-segmentation in subword models and boundary omission in language-agnostic neural tokenizers.
- **Standardization Enables Diagnostic Insight:** Canonical boundary projection enables a comprehensive, multi-paradigm evaluation under controlled conditions and provides explanatory power by linking numerical scores to concrete linguistic failure modes.

5. Discussion

The empirical results presented in Section 4 reveal a substantial performance gap between segmentation architectural paradigms when evaluated on the ABWS *representative workload*. As shown in Table [1], rule-based and hybrid systems such as Farasa (0.81), CAMeL Tools (0.81), and ALP (0.79) maintain relatively high boundary fidelity, reflecting their explicit modeling of Arabic morphology. In contrast, modern neural architectures and subword tokenizers exhibit a catastrophic degradation in performance: BPE (0.102) and mBART (0.122) fail to capture even basic clitic and stem boundaries, despite their widespread use in downstream neural pipelines.

The observed performance degradation in neural subword models, such as mBART and BPE-based architectures, stems from a fundamental misalignment between computational efficiency and linguistic morphology. Unlike rule-based systems that prioritize morpheme boundaries, subword tokenization algorithms are driven by information-theoretic compression (e.g., maximizing likelihood or frequency). Consequently, these models often ignore critical linguistic boundaries—such as the junction between a proclitic (e.g., the conjunction 'w-') and a stem—if a non-linguistic grouping provides a more frequent statistical pattern in the training corpus. This 'mechanistic' bias leads to the masking of functional particles, where a model may treat a prefixed word as a single opaque unit rather than a decomposable structure. Our CBA metric captures this failure by penalizing these statistically-driven but linguistically-invalid merges, which are particularly prevalent in the high-density Classical Arabic workload of our benchmark.

Regarding the composition of the ABWS workload, the inclusion of high-density Classical Arabic texts—specifically legal and jurisprudential treatises like Sharāi al-Islām—is a deliberate design choice rather than a limitation. These texts exhibit a significantly higher morphological density and a more complex clitic-stacking behavior compared to modern news or technical documents. By evaluating systems on this corpus, ABWS functions as a rigorous 'stress-test' for segmentation models. We argue that a system capable of accurately navigating the intricate boundary decisions of Classical Arabic is inherently more robust and better prepared for the linguistic variations of Modern Standard Arabic (MSA). Thus, this workload serves as a high-water mark for evaluating the precision and diagnostic limits of current Arabic NLP architectures.

5.1. The Failure of Subword Tokenization

The output analysis in Section 4.1 exposes a pronounced *reality gap* between subword-based segmentation models and linguistically valid Arabic morphology. In BPE and mBART, segmentation decisions are driven primarily by statistical frequency and vocabulary compression rather than morphemic structure. For example, the word *fa-al-wājib* ("so the obligation") is correctly decomposed by ALP and Farasa into the clitic-aware sequence [`fa`, `al`, `wājib`]. By contrast, mBART produces fragmented outputs such as [`fal`, `wā`, `jib`], which do not correspond to any valid morphological units in Arabic.

This behavior confirms that subword-based neural models, despite their apparent fluency in downstream tasks, operate on a predominantly surface-level representation that lacks structural awareness of Arabic clitic attachment and stem integrity. From a benchmarking perspective concerned with *traceability* and linguistic correctness, these findings indicate that subword-level metrics are poor proxies for morphological truth and can substantially misrepresent actual segmentation quality.

5.2. Robustness to Domain-Specific Morphology

The evaluated workload is dominated by Classical Arabic jurisprudential (Fiqh) terminology, including morphologically dense and derivationally complex forms such as *al-istibāḥa* and *al-mustaḥāḍa*. Traditional segmentation systems (Farasa and CAMeL Tools) demonstrate robustness in this setting due to their reliance on explicit morphological analyzers and lexicons. These systems consistently preserve canonical prefix, stem, and suffix boundaries even in specialized domains.

Neural models, however, exhibit marked performance degradation. The BERT-based segmenter achieves moderate overall accuracy (0.46) but still struggles with complex prefix–suffix combinations. For instance, forms such as *wa-al-mandūb* are segmented as [`wal-man`, `dūb`], indicating partial boundary drift and loss of morphemic coherence. This behavior suggests a high *evaluation risk* when deploying neural segmentation models in specialized or low-frequency domains, where memorized subword statistics fail to generalize underlying morphological rules.

5.3. Impact on Downstream Tasks

To address the correlation between ABWS metrics and downstream NLP performance, we conducted a pilot study focusing on Part-of-Speech (POS) tagging—a critical downstream task sensitive to segmentation quality. Our experiments, involving multiple architectures (including BiLSTM and Stanza), demonstrate a strong positive correlation ($\rho > 0.88$) between Critical Boundary Accuracy (CBA) and tagging macro-F1 scores. Specifically, we observed that errors identified by ABWS as 'Under-segmentation of Proclitics' (high USR) lead to a disproportionate drop in POS accuracy compared to simple stem boundary shifts. For instance, when the CBA score fell below 0.85, the downstream POS tagger's ability to correctly identify functional markers (e.g., particles and conjunctions) degraded by over 12%. These findings empirically validate that the diagnostic metrics provided by ABWS are not merely intrinsic measures but are reliable predictors of a model's utility in complex Arabic NLP pipelines.

5.4. Implications for Standardization and Evaluation Theory

From a *workload characterization* perspective, these results strongly justify the design choices underlying the ABWS framework. Conventional evaluation practices often mask the observed failures by relying on aggregate metrics (e.g., BLEU or token-level F_1) computed over overlapping subwords, thereby conflating surface overlap with linguistic correctness. By enforcing a Canonical Boundary Vector (CBV) representation, ABWS exposes fundamental limitations that remain invisible under traditional evaluation regimes.

Specifically, the results demonstrate that:

- Neural and subword-based segmenters are not yet *standard-ready* for high-precision linguistic tasks that require reliable boundary placement.
- Evaluation equivalence between rule-based and neural systems is unattainable without a paradigm-agnostic representation and metric suite, such as those proposed in this work.

In summary, the current reality of Arabic NLP benchmarking reflects a trade-off between the scalability and flexibility of neural models and the boundary precision of rule-based

systems. For critical applications such as legal, religious, or scholarly text analysis, the high error rates observed for SelfSeg (0.163), BPE (0.102), and mBART (0.122) render these approaches unsuitable in their current form. These findings underscore the urgent need for boundary-aware training objectives and evaluation frameworks in the next generation of large language models for Arabic.

6. Conclusion and Future Work

In this work, we introduced the Arabic Boundary Word Segmentation (ABWS) framework, a multi-paradigm benchmark designed to address the lack of standardization in Arabic morphological evaluation. By formalizing the *Canonical Boundary Vector* (CBV), we provided a methodology to evaluate systems ranging from traditional rule-based analyzers to modern neural subword tokenizers within a unified, equivalent evaluation condition (EC).

Our empirical results, based on a representative workload of 212,873 words, reveal a profound "reality gap" in current Arabic NLP. While rule-based systems like Farasa and Camel achieve high boundary accuracy (0.81), state-of-the-art neural models and statistical tokenizers such as mBART (0.122) and BPE (0.102) show catastrophic failure in capturing linguistically valid boundaries. This disparity highlights a significant *evaluation risk*: conventional metrics used in downstream tasks often mask a systemic lack of morphological awareness in Large Language Models (LLMs).

ABWS contributes to the engineering of evaluation by providing a containerized, reproducible pipeline that ensures benchmark traceability. By treating dataset provenance and workload characterization as first-class artifacts, this benchmark allows for the rigorous comparison of diverse architectures, ensuring that progress in Arabic NLP is measured against a ground-truth linguistic standard rather than surface-level statistical frequency.

While ABWS is specifically designed for Arabic, its core methodological contributions are language-agnostic. The Canonical Boundary Vector (CBV) abstraction provides a general solution for comparing outputs from disparate segmentation paradigms (rule-based, statistical, neural) in any language. The boundary-aware metrics (e.g., OSR, USR, CBA) are defined at the character level and do not rely on Arabic-specific features, making them transferable to other morphologically rich languages (MRLs) such as Hebrew, Turkish, or Finnish. However, the empirical findings reported in this paper—such as the extreme over-segmentation of subword tokenizers—are directly tied to Arabic's unique morphological structure (e.g., concatenative cliticization). While similar phenomena may occur in other MRLs, further experiments are needed to confirm cross-lingual patterns.

Future work will focus on expanding the ABWS workload to include more diverse dialects and low-resource historical texts. Furthermore, we intend to integrate automated artifact evaluation tools to further streamline the reproducibility of results across different hardware testbeds. Ultimately, ABWS offers a template for how complex, multi-layered NLP tasks can be standardized to support cumulative scientific progress and reliable real-world deployment.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

Funding

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

Declaration of competing interests

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.

Data Availability Statements

The data supporting the findings of this study are openly available in zenodo at `https://zenodo.org/records/18138582` or `https://doi.org/10.5281/zenodo.18138582`.

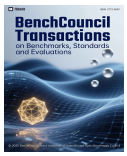
Credit authorship contribution statement

Behrouz Minaei-Bidgoli: Supervision; Methodology; Validation; Writing – Review & Editing. Huda AlShuhayeb: Conceptualization; Methodology; Formal Analysis; Investigation; Visualization; Writing – Original Draft.

References

1. Nizar Y. Habash. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010. `doi:10.2200/S00277ED1V01Y201008HLT010`.
2. Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, and Richard Schwartz. Machine translation of arabic dialects. In *Proceedings of NAACL-HLT*, pages 49–59, 2012. URL: `https://aclanthology.org/N12-1006.pdf`.
3. Kareem Darwish. Building a shallow arabic morphological analyzer in one day. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 2002. URL: `https://aclanthology.org/W02-0506.pdf`.
4. Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, 2019. `doi:10.18653/v1/P19-1267`.
5. Nizar Habash and Owen Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of ACL*, pages 573–580, 2005. URL: `https://aclanthology.org/P05-1071.pdf`.
6. F. Han et al. Open source evaluatology: A theoretical framework for open-source evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4:100190, 2024. URL: `https://doi.org/10.1016/j.tbench.2025.100190`.
7. Xinyue Li, Heyang Zhou, Qingxu Li, Sen Zhang, and Gang Lu. Aicb: A benchmark for evaluating the communication subsystem of LLM training clusters. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5:100212, 2025. `doi:10.1016/j.tbench.2025.100212`.

8. Jiyue Xie, Wenjing Liu, Li Ma, Caiqin Yao, Qi Liang, Suqin Tang, and Yunyou Huang. COADBench: A benchmark for revealing the relationship between AI models and clinical outcomes. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4:100198, 2025. TBSE paper (uploaded PDF: S2772485925000110). doi:10.1016/j.tbench.2025.100198.
9. Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, 2004. URL: https://www.marefa.org/images/e/e8/The_penn_arabic_treebank_Building_a_large-scale_an_%281%29.pdf.
10. Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT*, 2008. URL: https://aclanthology.org/P08-2030.pdf.
11. Mohamed Boudchiche, Abdelhak Mazroui, Mohamed Bebah, Abdelhadi Lakhouaja, and Abdelaziz Boudlal. Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer. *Journal of King Saud University – Computer and Information Sciences*, 29(2):141–146, 2017. URL: https://www.sciencedirect.com/science/article/pii/S131915781630026X, doi:10.1016/j.jksuci.2016.08.003.
12. Wajdi Zaghouani. Critical survey of the freely available arabic corpora. In *Proceedings of LREC*, 2014. URL: https://www.researchgate.net/profile/Wajdi-Zaghouani/publication/263215246_Critical_Survey_of_the_Freely_Available_Arabic_Corpora/links/0046353a53977808fa000000/Critical-Survey-of-the-Freely-Available-Arabic-Corpora.pdf.
13. Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A fast and furious segmenter for arabic. In *Proceedings of NAACL-HLT*, 2016. URL: https://aclanthology.org/N16-3003.pdf.
14. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725, 2016. URL: https://aclanthology.org/P16-1162.pdf, doi:10.18653/v1/P16-1162.
15. Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP*, 2018. URL: https://aclanthology.org/anthology-files/anthology-files/pdf/D/D18/D18-2.pdf#page=78.
16. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. URL: https://aclanthology.org/N19-1423.pdf.
17. Wissam Antoun, Fady Baly, and Hazem Hajj. AraELECTRA: Pre-training text discriminators for arabic language understanding. In *Proceedings of WANLP*, 2020. URL: https://aclanthology.org/2021.wanlp-1.20.pdf.
18. Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of ACL-IJCNLP*, 2021. URL: https://aclanthology.org/2021.acl-long.551.pdf.
19. Bashar Alhafni and Nizar Habash. Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. In *Proceedings of EACL*, 2023. URL: https://aclanthology.org/2020.acl-main.736.pdf.
20. Nizar Habash and Fatiha Sadat. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of NAACL-HLT*, pages 49–52, 2006. URL: https://aclanthology.org/N06-2013.pdf.
21. Kareem Darwish and Douglas W. Oard. Term selection for searching printed arabic. In *Proceedings of SIGIR*, 2003. URL: https://dl.acm.org/doi/pdf/10.1145/564376.564423.
22. Yonghui Wu et al. Google's neural machine translation system: Bridging the gap between human and machine translation. In *arXiv preprint arXiv:1609.08144*, 2016. URL: https://www.researchgate.net/publication/308646556_Google's_Neural_Machine_Translation_System_Bridging_the_Gap_between_Human_and_Machine_Translation.
23. Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of ACL*, 2019. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10287171/pdf/nihms-1908534.pdf.
24. Alex Wang et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of EMNLP Workshop*, 2018. URL: https://aclanthology.org/W18-5446.pdf.
25. Alex Wang et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS*, 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/5s84yp08
Full Length Articles

FULL LENGTH ARTICLES

TraceRTL: Agile Performance Evaluation for Microarchitecture Exploration

Zifei Zhang[1,2] Yinan Xu[1] Kaichen Gong[4] Sa Wang[1,2] Dan Tang[1,3] and Yungang Bao[1,2,*]

[1]State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China, [2]University of Chinese Academy of Sciences, 100190, Beijing, China, [3]Beijing Institute of Open Source Chip, 100080, Beijing, China and [4]School of Information Science and Technology, ShanghaiTech University, 200000, Shanghai, China
[*]Corresponding author. baoyg@ict.ac.cn
Received on 2 February 2026; Accepted on 29 March 2026

Abstract

While agile chip development methodologies have accelerated RTL design and simulation, performance evaluation remains constrained by challenges: (1) limited benchmark availability due to incomplete peripheral/software simulation environments or unavailable source code; (2) inefficient feature prototyping caused by the tight coupling between functional correctness and performance evaluation, particularly for large-scale, error-prone microarchitectures. To address these challenges, we propose TraceRTL, an agile, trace-driven performance evaluation methodology that decouples the functional and performance components of CPU RTL designs. It introduces three contributions to the benchmarking community: (1) a trace-driven exploration framework that bypasses full functional correctness while preserving performance behavior and supports replaying workload traces on RTL designs; (2) a quantitative analysis and mitigation methodology to identify and reduce trace-driven performance discrepancies; (3) a trace transformation technique, TraceBridge, that converts benchmark traces between different formats and instruction sets. Using TraceRTL, we have developed the first trace-driven RTL CPU derived from XiangShan, a high-performance out-of-order RISC-V processor. TraceRTL achieves performance accuracy of 99.87% and 99.86% on SPECint2017 and SPECfp2017, respectively. With TraceBridge, we evaluate x86 Google workload traces on a RISC-V RTL CPU and reveal distinct memory-bound behavior.

Key words: Trace-driven simulation, Performance evaluation, Cross ISA benchmarking

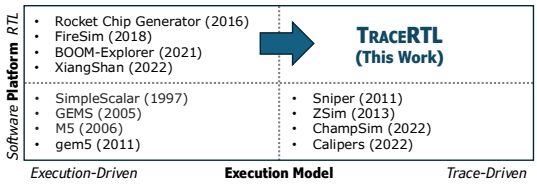
1. Introduction

Performance has always been a central consideration in CPU development. As Moore's Law slows and application demands diversify, achieving further performance improvements has become increasingly challenging. This highlights the importance of microarchitecture exploration methodologies. A key question is: given a baseline CPU design, how can we efficiently quantify the performance impact of a proposed hardware feature using representative benchmarks?

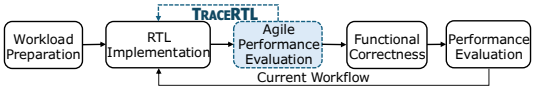
Among available evaluation methods for assessing CPU design changes using diverse benchmarks, the most faithful approach is to use the register-transfer level (RTL) implementation. As the definitive description of the microarchitecture, RTL is the most reliable basis for assessing CPU microarchitecture designs. Ultimately, any proposed feature must be implemented and evaluated in RTL to determine its true performance impact.

However, since the RTL development process is time-consuming, the computer architecture community has adopted more efficient approaches to accelerate early-stage exploration before implementing a proposed feature in RTL. As shown in Fig. 1(a), software-based architectural simulators [1–8] model low-level hardware components using high-level languages and abstractions, enabling fast simulation and rapid design iteration. Despite their high productivity in *early-stage* exploration, the *last mile* remains unavoidable: performance must still be re-evaluated at the RTL level after initial simulator studies, since the additional modeling layer inevitably introduces discrepancies that require substantial engineering efforts and costly calibration with the actual implementation [9].

Another fundamental yet often overlooked challenge is the benchmarking asymmetry across the development workflow. While software simulators [5, 6, 8, 10] widely adopt trace-driven methodologies to execute diverse benchmarks in trace format, RTL models lack the capability to replay traces and faces limited benchmarks due to immature simulation environments. The *benchmarking gap* prevents a consistent and continuous evaluation flow from early-stage modeling to final hardware implementation.



(a) Approaches to microarchitecture exploration.



(b) Current and TraceRTL performance evaluation workflows.

Figure 1. Microarchitecture exploration methods and workflows.

Recent advancements in RTL design and simulation offer strong potential for a seamless, progressive refinement workflow from early-stage exploration to the last mile. On the design side, high-level hardware construction languages [11–13] enable parameterized and reusable components, allowing rapid implementation and iteration of new microarchitectural ideas. On the evaluation side, efficient RTL simulation methods [14–18], especially FPGAs and emulators [19–22], have significantly accelerated large-scale RTL simulation. These capabilities have already been demonstrated in several open-source, industrial-competitive CPUs [23–27], which provide realistic and accessible microarchitecture research platforms [28–32]. For example, it takes less than 200 minutes and approximately 300 lines of modified Chisel code to implement an instruction scheduler policy PUBS [33] on the XiangShan, a high-performance RISC-V CPU achieving >15 SPECint2006/GHz [26, 34].

These trends motivate us to adapt proven exploration techniques from simulators directly to RTL, aiming to *inherit the agile workflow of simulator-based exploration while enabling a seamless integration into RTL for last-mile evaluation.*

To realize this opportunity, we *propose* **TraceRTL**, an RTL-based performance evaluation methodology that derives a trace-driven RTL model from an existing execution-driven RTL implementation. As illustrated in Fig. 1(a), TraceRTL reuses open-source, silicon-validated RTL designs as a solid foundation for faithful microarchitecture exploration. It drives performance-critical modules with pre-generated traces, enabling simulator-like agility for early-stage exploration on RTL. By deriving the trace-driven model directly from RTL, it inherently avoids costly last-mile calibration and preserves performance fidelity for evaluation of the proposed feature.

One key motivation behind TraceRTL is to overcome the execution-driven nature of current RTL designs. As illustrated by the white boxes in Fig. 1(b), the modified CPU must first pass full functional verification *before* any performance evaluation can be conducted. This tight coupling forces every RTL design modification to undergo complete implementation, verification, and lengthy simulation, even when the modification is unrelated to performance. For example, evaluating optimizations for virtualized, two-stage address translation requires implementing complex privileged operations to guarantee correct functionality, whose functional details, however, do not affect performance.

With TraceRTL, this strict dependency between functional correctness and performance evaluation is eliminated. As highlighted in Fig. 1(b), TraceRTL enables ***agile performance evaluation*** without first implementing or verifying unrelated functional details. Additionally, it accepts trace inputs from ***a broader range of real-world applications***, including those with unavailable source code, different ISAs, or peripheral dependencies [10, 35–37], without requiring porting to an RTL simulation. To realize these capabilities, however, we need to address three key challenges.

1) *Feature prototyping: Can we develop a trace-driven RTL CPU with minimal modifications to the existing execution-driven microarchitecture while preserving performance accuracy?* Our key insight is that hardware module interfaces can be categorized into functional interfaces, which determine what each instruction computes and where execution proceeds next, and performance-sensitive interfaces, which determine how efficiently the instruction stream is realized. Based on this distinction, TraceRTL selectively takes over key interfaces to decouple the functional model while preserving performance behaviors using externally supplied traces. This preserves cycle-accurate performance accuracy while eliminating the complexity of managing full functional correctness.

2) *Performance accuracy: Can we mitigate the performance discrepancies introduced by trace-driven simulation?* Conventional trace-driven simulation often suffers from fidelity loss due to the lack of necessary information to replicate execution-driven behaviors. We observe that these information gaps stem from two primary sources: intentional abstraction and dynamic omission. We quantitatively analyze the performance impact of these missing components, revealing their essential role for fidelity. TraceRTL proposes a dynamic information reconstruction mechanism that synthetically reconstructs missing data, achieving high performance accuracy.

3) *Broader workloads: Can we bridge the semantic gap across diverse trace formats and ISAs?* Industrial workloads are valuable for microarchitecture exploration, but the scarcity of RISC-V workloads necessitates cross-ISA transformation to generate benchmark traces. This transformation is performed only once during trace preparation. However, differences in trace formats and ISAs hinder the direct execution of publicly available traces on RTL CPU models. Since trace-driven simulation relaxes the need for full functional correctness, TraceRTL introduces TraceBridge, a trace transformation technique that leverages instruction and register mapping to enable the replay of traces from different formats and ISAs.

To demonstrate the feasibility of TraceRTL, we develop a trace-driven RTL model derived from XiangShan [26, 38]. It achieves performance accuracy of 99.87% and 99.86% on SPECint2017 and SPECfp2017, respectively, reducing performance discrepancies by 10.31× and 29.21× compared to a calibrated XS-gem5 model. By leveraging TraceBridge, we evaluate x86-based Google workload traces [36] on XiangShan, and reveal distinct memory-bound behavior compared to SPECint2017.

TraceRTL expands the possibilities for microarchitecture research by supporting both RTL-based exploration and seamless integration with simulator-based workflows. By preserving a simulator-like, trace-driven environment for workloads and simulation, it effectively bridges early-stage exploration on simulators and last-mile RTL evaluation.

To summarize, this paper makes the following contributions.

- We propose TraceRTL, bringing trace-driven simulation to RTL CPUs for agile microarchitecture exploration.

- We quantify the sources of performance discrepancies and implement dynamic information reconstruction to achieve high performance accuracy.
- We propose TraceBridge, which enhances trace compatibility to expand the sources of benchmark workloads.
- We demonstrate TraceRTL by using x86 workload traces collected from Google warehouse-scale computers for performance evaluation of XiangShan, a RISC-V CPU.

2. Background

2.1. Out-of-Order Microarchitecture

Modern CPUs improve performance primarily by exploiting parallelism and speculation. The front-end speculatively fetches instructions using branch prediction, while the back-end decodes, schedules, and issues them to execution units for computation and to memory subsystem for data access.

The efficiency of this pipeline depends on several critical microarchitectural components. Branch prediction and instruction fetching determine the instruction supply rate. Execution pipelines and scheduling queues affect throughput. The memory hierarchy bridges the large speed gap between CPU and DRAM by caching frequently used data. The memory management unit (MMU) accelerates address translation by caching recently used address mappings near the CPU.

2.2. Exploration on RTL

While RTL models offer higher accuracy for design space exploration, directly evaluating performance on RTL presents several challenges, including the inflexibility of traditional hardware description languages, slow simulation speeds, and the lack of open-source RTL processors. Recent efforts have focused on these issues.

Flexibility. Many emerging high-level hardware description languages [12, 13, 39] offer enhanced expressiveness and parameterization that accelerate the development of microarchitectures . New hardware design methodologies [11] are also proposed to further improve design modularity.

Simulation Speed. Novel RTL simulation techniques have been proposed to accelerate software-based [14–16] or hardware-based [19, 21, 22] simulation of RTL designs.

Additionally, sampling-based methods [40–42] estimate full-program performance by aggregating results from several representative program segments.

Open-Source RTL Processors. With the rapid growth of the RISC-V open-source community, a number of RTL processors have emerged, including in-order designs [43, 44] and out-of-order designs such as BOOM [23–25], XuanTie-910 [27], and XiangShan [26]. These designs provide accessible and realistic platforms for microarchitecture research, enabling agile exploration directly on RTL.

2.3. Simulation Methodologies

In computer architecture research, performance evaluation of novel designs predominantly relies on two core methodologies: execution-driven and trace-driven simulations. These approaches fundamentally differ in how they provide program stimuli to performance models, leading to distinct trade-offs between fidelity, flexibility , and simulation speed.

The execution-driven methodology emulates the behavior of real CPUs within the performance model, such as fetching, decoding, scheduling and executing instructions. This approach is inherent to RTL models [23–26, 43] and is also implemented in many software simulators [1–4]. By coupling functional execution with performance modeling, this approach captures microarchitecture-dependent dynamic behaviors, such as speculative execution and wrong-path effects, thereby offering high fidelity. However, this accuracy comes at the cost of significant complexity, increased error-proneness, and reduced simulation speed.

In contrast, the trace-driven methodology decouples the functional model from the performance model by replaying the pre-generated traces of instructions including architectural information such as instruction semantics, instruction addresses, memory accesses, and branch outcomes [5, 6, 8, 10, 45]. These traces are often generated using instrumentation tools like Pin [46], DynamoRIO [47], and Valgrind [48], or obtained from public pre-generated traces [10, 35, 36]. This decoupling affords higher flexibility, enabling researchers to focus on microarchitectural optimization. However, this flexibility often comes at the cost of reduced fidelity, as traces lack dynamic microarchitecture-dependent information.

3. Challenge

Agile performance evaluation requires rapid feature prototyping, support for extensive workloads, and fast simulation. To meet these goals at the RTL level, trace-driven simulation offers a promising approach by decoupling performance and functional models and supporting trace-based workloads. However, integrating trace-driven simulation into existing execution-driven CPU RTL models introduces non-trivial challenges. Publicly available traces often vary in trace format, lack information such as instruction encodings, and are sometimes generated from different instruction sets.

3.1. Trace-driven RTL Integration

Transforming a complex execution-driven CPU RTL model into a trace-driven implementation presents unique challenges compared to building an RTL model from scratch or driving individual RTL modules independently. In addition to supplying stimuli to existing RTL modules, a trace-driven model must precisely control the instruction flow based on external traces while maintaining the original performance behavior.

3.2. Trace-driven Performance Discrepancies

Trace-driven simulation inherently suffers from fidelity loss due to the lack of necessary information. This gap stems from two primary sources: intentional abstraction and dynamic information omission. First, to balance confidentiality and storage overhead, conventional traces often omit critical details such as operand values and instruction opcodes. Second, static traces fail to capture dynamic execution states, such as wrong-path instructions and page table walks, which only emerge during runtime. The absence of these microarchitectural side effects prevents the accurate replication of execution-driven behaviors, potentially leading to significant performance discrepancies.

3.3. Trace Compatibility

Trace-driven approaches can bypass the limitations of simulated peripheral environments, thereby enhancing the coverage of supported workloads. However, due to confidentiality constraints, instruction source code is often unavailable for publicly accessible trace files [10, 35, 36]. Another scenario

involves target applications that require evaluation but have not been adapted to the target instruction set, rendering direct assessment infeasible.

Instruction sets share commonalities but also exhibit significant differences, which hinder direct trace porting. For example, differences in general-purpose register conventions, instruction encodings and sizes, PC alignment rules, and the range of direct branch instructions all impose constraints on cross-instruction-set trace evaluation. These challenges are particularly pronounced for RTL models, which typically lack sufficient abstraction capabilities.

4. TraceRTL Design

To enable agile performance evaluation of RTL designs, we first propose a trace-driven simulation methodology at the RTL level (§ 4.1) while preserving high performance accuracy (§ 4.2). Building on this, we introduce TraceBridge, a trace transformation method that enhances compatibility by enabling the replay of traces from different formats and instruction sets (§ 4.3).

4.1. Trace-Driven Microarchitecture Design

We decompose the CPU into core components and describe how each component is driven by the trace. Interfaces, defined asthe set of I/O signals between modules, can be driven to control the module's behavior. By driving the key interfaces with the information in traces, TraceRTL replaces the functional model with external traces while maintaining the original performance behavior. This section describes the design of trace-driven integration to meet its objectives: (1) driving RTL modules with external instruction traces, (2) enforcing the CPU model to conform to the trace instruction flow, and (3) identifying and mitigating performance discrepancies inherent in trace-driven simulation.

4.1.1. Trace-Driven RTL Modules

Our key insight is that hardware module interfaces can be classified into functional interfaces, which determine what each instruction computes and where execution proceeds next (e.g., arithmetic, branching, or exception handling), and performance-sensitive interfaces, which determine how efficiently the instruction stream is realized (e.g., branch prediction, cache access, and memory prefetching). Based on this distinction, we analyze key module behaviors and drive performance-sensitive modules using external instruction traces, preserving original performance characteristics without requiring full functional execution.

Branch predictor. The branch predictor's performance-critical interfaces primarily include two types: training and prediction. The predictor is trained on the committed branch outcomes. Therefore, instructions on the mis-speculated path, which are flushed from the pipeline, leave no side effects. By substituting the branch outcomes with trace information, which includes branch direction and target, we are able to stimulate the training process. For prediction, the predictor takes the current program counter (PC) and branch history to generate the next instruction fetch request. While prediction is at speculative stage, the PC and history for correct-path instructions are consistent between execution-driven and trace-driven simulations.

Instruction fetch. The instruction fetch unit obtains fetch requests from the branch predictor and retrieves instructions from the traces. We propose an *interval match mechanism* to simulate the fetch bandwidth, as shown in Fig. 2. A fetch request typically specifies a contiguous instruction interval I_{bp} defined by starting and ending addresses. The fetch unit forwards the request to TraceReader that extracts a continuous sequence of trace instructions I_{trace}. Instructions in I_{if}, which are common to both I_{bp} and I_{trace}, are then sent to subsequent pipeline stages for execution. When the starting address does not match the beginning of the trace, I_{if} is empty, preventing any instructions from being fetched. Consequently, the impact of instructions on the mis-predicted path cannot be modeled. § 4.2.1 presents a refined design to address this limitation.

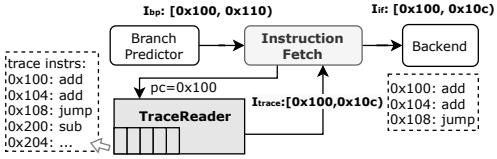


Figure 2. Trace-driven instruction fetch with interval match mechanism.

Out-of-order backend. The backend relies on instruction encodings to stimulate decoding, register renaming, dynamic scheduling and execution. These encodings are directly supplied from the trace. Alternatively, a more aggressive approach is to provide the results of the decoding directly to drive renaming and scheduling, although this is beyond the scope of this work. Particular units like the FDivSqrt operation may need optional data for accurate execution latency.

Cache hierarchy. Cache behavior is mainly influenced by access addresses. Instruction addresses are derived from fetch requests generated by the branch predictor. Data addresses, on the other hand, are dynamically calculated from the operands, which are invalid in trace-driven simulation. Therefore, memory access addresses should be included in traces to model memory behavior. Special modules like the indirect memory access prefetcher need extra information.

Memory management unit. The virtual-to-physical address translation and page-table walk require in-memory page table entries (PTEs) that are typically absent in traces [49]. We employ a dynamic page table generation approach: For each instruction in the trace, we traverse the page tables using its virtual address. If a required PTE is invalid, a new page is allocated, and the corresponding PTE is initialized. This process continues recursively until reaching the leaf page, which is initialized with the physical address in traces.

4.1.2. Trace-Controlled Instruction Flow

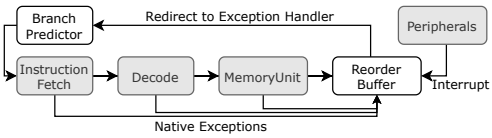
TraceRTL controls the instruction flow by managing branch instructions, interrupts, and exceptions, while ensuring processor compliance by instruction stream correctness checks.

Branch instruction. We replace the branch execution unit's outcomes with target and conditional result recorded in the trace to control the programs' instruction flow.

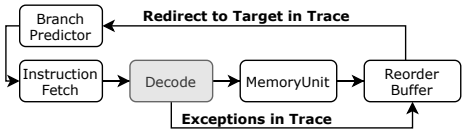
Exception and interrupt. Traps, including exceptions like page faults and interrupts like timer interrupts, may be triggered by programs, devices, and operating system. Traps affect control flow and pipeline redirection, as illustrated in Fig. 3(a). These are intercepted and re-injected according to the trace. Specifically, trace-recorded exceptions are triggered as illegal instructions, redirecting to the target in trace, as illustrated in Fig. 3(b). This design ensures that exceptions are preserved without relying on full functional execution.

Table 1. Key CPU module behaviors and their corresponding trace-driven stimuli in TraceRTL.

Module	Key Behavior	Trace-Driven Stimulus
Branch Predictor	Prediction uses PC and history; Training uses committed branch outcomes.	Use current PC and history for prediction; Use branch outcomes from trace for training.
Instruction Fetch	Fetch request defines instruction interval; Wrong-path instructions.	Apply interval match mechanism to simulate fetch bandwidth; Generate wrong paths on mismatch.
Instruction Execution	Decode, rename, schedule, and execution.	Use instruction encoding and optional operand from trace.
Instruction Flow	Branch instruction outcome; Exception and interrupt redirect the pipeline.	Intercept branch outcomes/exception generation; Support redirect; Flow check.
Cache Hierarchy	Access cache by addresses.	Memory address from trace; Instruction address from branch predictor; Optional data from trace.
MMU	Virtual-to-physical address translation; Access memory for page table entries.	Construct page table according to the address from trace.



(a) Native exceptions and interrupt triggered by the CPU pipeline and peripherals.



(b) Intercept the native exceptions and trigger trace exceptions as illegal instruction.

Figure 3. Exception and interrupt management in TraceRTL.

Instruction stream check. A fundamental requirement of trace-driven simulation is that the performance model must be guided by the trace, a key aspect of which is to ensure its execution adheres to the provided instruction stream. We capture the processor's actual instruction stream through committed instructions and compare it against the trace. The differences in the streams indicate implementation flaws in the RTL model itself or trace-driven framework.

4.1.3. Overall

In summary, TraceRTL provides a general and adaptable framework for trace-driven RTL performance evaluation. It is designed to evolve naturally with RTL designs, require minimal effort across microarchitectural iterations, remain applicable across diverse microarchitectures, and flexibly support various performance optimizations.

Extending TraceRTL to new architectures. We summarize the trace-driven transformation methodology in Table 1. TraceRTL employs an interface-based modification strategy that reduces modification overhead while accommodating variations in module design. The processor module partitioning methodology is universal across different microarchitectures, making TraceRTL a reusable and microarchitecture-agnostic framework for RTL performance evaluation. The specific modifications may vary depending on processor-specific designs. For instance, in-order processors commonly fetch one or two instructions per cycle, which does not require the interval match described in § 4.1.1. In contrast, some high-performance processors may fetch instructions spanning two intervals per cycle, thus necessitating two interval-match operations. For CPU-driven accelerators, such as matrix units, the necessary execution information can also be recorded into trace instructions and dispatched accordingly.

Applicability for microarchitecture features. TraceRTL is particularly advantageous for evaluating functionally complex yet performance-critical features (§ 7.2). Beyond functionality, it captures fine-grained timing effects that are difficult to model accurately at higher abstraction levels. For example, variations in microarchitectural timing may critically affect the overall performance (§7.4). It can also evaluate microarchitectural optimizations in the same way as conventional trace-driven simulators (e.g., branch prediction, prefetching, replacement, memory dependence prediction). With additional trace information, TraceRTL can be extended to model advanced optimizations, such as value prediction (with execution results) and indirect memory prefetching (with memory values).

4.2. Trace-driven Performance Discrepancy Mitigation

To achieve high accuracy, trace-driven simulation should strive to mimic the behaviors of execution-driven simulation. This section details our methodology for bridging this gap by enhancing trace-driven simulation of the frontend fetch unit through wrong-path simulation, refining execution latency via operand and opcode provisioning, and maintaining MMU fidelity through dynamic page table construction.

4.2.1. Fetch: Wrong-Path Simulation

Out-of-order processors may execute instructions that are later discarded due to events like branch mispredictions. These instructions, although executed, are flushed by pipeline redirect operations, preventing them from affecting the architectural state of the CPU, such as the register file or memory.

Wrong-path instructions' performance impact, particularly on the cache hierarchy, cannot be ignored. The impact on the cache can be categorized into **prefetching** and **pollution**, leading to positive and negative effects. Fig. 4 presents a code example divided into three sections: (1) Code1, executed unconditionally before the branch; (2) Code2, located within one branch; and (3) Code3/Code4, placed outside the branch's influence, further categorized into the proximate Code3 and

the distant Code4. Upon a mispredicted branch, Code2 is executed, and if its execution is swift, Code3 may follow. Once the branch is resolved, speculatively fetched instructions of Code2 and Code3 are discarded, with Code2 potentially polluting the cache and Code3 prefetching the cache.

```
*b = 1;  /* Code1 */
a = *b;
if (a == 0)
  a = *c;/* Code2 */
a = *d;  /* Code3 */
a = *e;  /* Code4 */
```

Figure 4. Code example demonstrating wrong-path instruction generation.

We statistically analyze the number and addresses of memory instructions on both correct and wrong paths in the out-of-order processor XiangShan, focusing on instructions sent to the load pipeline. These addresses are aligned to cache-line size. We categorize the address space into three types: (1) exclusive-arch-path: only accessed by correct path instructions, (2) exclusive-wrong-path: only accessed by wrong-path instructions and (3) overlapped: accessed by both paths. As shown in Fig. 5, we found that most of the address space falls into type(1) and (3). Therefore, we can tentatively draw a rough conclusion that *prefetching has the predominant influence.*

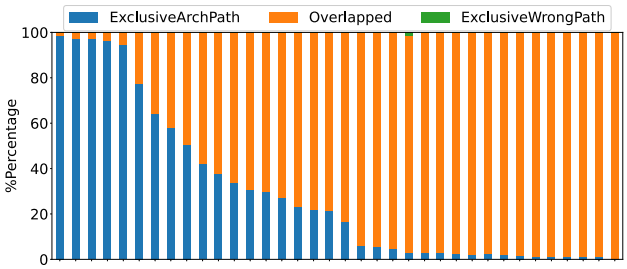


Figure 5. Percentage of memory interval weighted by load access times for the SPEC CPU2017. Each bar represents a sub-benchmark, sorted according to "ExclusiveArchPath".

Based on the observation, we focus on simulating the prefetching influence by *taking the instructions at correct path as wrong-path instructions*. The process involves the following steps: (1) When a branch misprediction occurs, we check whether the fetch request's starting address exists in traces within a fixed instruction window; (2) If it exists, the instructions in the trace are sent to subsequent pipeline stages as wrong-path instructions. The instruction fetch unit is blocked for simplification. These instructions are not discarded from the traces; (3) Once the branch instruction is resolved and the pipeline is redirected, the correct fetch request is issued.

4.2.2. Execution: Instruction Opcode Provisioning

Conventional trace-driven simulators often operate with partial instruction encodings. Instruction encoding has two types of information: instruction opcode for functionality like ADD and SUB, register indices for instruction dependency and out-of-order scheduling. Explicit instruction opcodes are frequently

Algorithm 1 Dynamic Page Table Construction

```
1: procedure InstructionWalk(instList)
2:     for inst in instList do
3:         if inst's PC valid then
4:             PageWalk(inst.VirtualPC, inst.PhysicalPC)
5:         end if
6:         if inst's memory address valid then
7:             PageWalk(inst.VirtualAddr, inst.PhysicalAddr)
8:         end if
9:     end for
10: end procedure
11: procedure PageWalk(va, pa)
12:     pageBase = PageTableRootAddr
13:     for level := 0 to MaxLevel do
14:         pteAddr = getPteAddr(va, level, pageBase)
15:         pte = readPageTable(pteAddr)
16:         if pte not valid then
17:             if level == MaxLevel-1 then
18:                 newPte = genPte(pa)      ▷ Leaf page arrived
19:             else
20:                 newPte = genPte(AllocatePage())
21:             end if
22:             writePageTable(pteAddr, newPte)
23:         end if
24:         pageBase = pte.ppn << 12
25:     end for
26: end procedure
```

abstracted or omitted for confidentiality concerns and software simulators' highly abstracted microarchitecture designs. Consequently, instead of providing detailed opcodes, trace instructions are categorized into coarse-grained functional groups: (1) control flow (unconditional direct, conditional direct, and indirect jumps); (2) memory access (loads and stores); and (3) computation (integer and floating-point).

Our work focuses on quantifying the performance modeling deviations induced by this loss of fine-grained opcodes. Specifically, we investigate how substituting precise opcodes with coarse-grained categories impacts simulation fidelity. This analysis aims to isolate the impact of operation abstraction from other simulation variables, providing a quantitative understanding of the accuracy trade-offs in abstracted trace modeling.

4.2.3. Execution: Operand Provisioning

Some operations are implemented in a blocking manner and their execution cycles are variable depending on the operands, like division, floating-point division and square-root. This type of performance error is always neglected and simulators often implement them with fixed latency.

Although these instructions are relatively few, their long execution cycles and low degree of concurrency amplify their performance impact. To achieve more accurate simulation for these types of instructions, we record their operands in traces.

4.2.4. MMU: Dynamic Page Table Construction

User-space programs use virtual addresses, which must be translated to physical addresses by the memory management unit (MMU) before accessing the cache or main memory. In the MMU, the virtual address first consults the L1 translation lookaside buffer (TLB). If L1 TLB hits, the physical address

is obtained directly. In case of L1 TLB miss, the virtual address will be sent to a larger L2 TLB or hardware page table walker to traverse the memory-resident page tables to find the physical address corresponding to the virtual address, which involves multiple memory accesses, especially in hypervisor environments. Page table caches are used to speed up page table walks. In summary, the hit rates of TLB and page table cache, as well as page table walker's memory latency, are crucial for MMU-sensitive programs.

To simulate the behavior of MMU and minimize the modifications on RTL modules, we need to provide a self-consistent page table for the MMU. However, traces typically contain only the physical and virtual addresses, but not the page table [49]. Therefore, we employ a dynamic page table generation method, as illustrated in Algorithm 1. By iterating over each instruction in the traces and traversing the page tables based on the virtual address, we allocate new page frames and initialize the invalid corresponding page table entries, until reaching the leaf page. The leaf entry is then initialized with the corresponding physical address. After dynamically generating the page table, when a TLB miss occurs, the memory-resident page tables are traversed.

4.3. Trace Compatibility with TraceBridge

We introduce a trace transformation methodology, TraceBridge, to bridge the incompatibilities in trace formats and instruction sets. To support trace-driven simulation, the trace must contain at least three categories of information: (1) primary instruction type, including branch types, computation, and memory operations; (2) execution guidance, including PC, branch target and conditional result, and memory address; (3) register dependencies to model instruction-level parallelism. Such information is typically included in the trace format of dynamic instrumentation tools [49] and publicly available traces [10, 35, 36], where fine-grained semantic information such as instruction opcodes are sometimes missing.

TraceBridge retains the key information from the trace, transforming its format to be compatible with the target model by refining the execution semantics. However, trace-driven RTL models pose additional low-level challenges due to their rich details: (1) *instruction correspondence and register semantics*; (2) *difference in instruction encoding size and program counter (PC) alignment constraints*; (3) *variations in branch offset ranges.*

The primary principle of TraceBridge is to maintain performance semantics consistency. This ensures that the performance characteristics of the original program are reflected in the target architecture. For confidentiality, public traces omit instruction encodings [10] or provide instruction categories [36]. To address this, we observe that an instruction can encompass multiple performance semantics, which fall into four types: ⟨Load, Computation, Store, Branch⟩. To maintain performance semantics consistency, we map each individual performance semantic to its corresponding instruction(s) in the target ISA. A single x86 instruction, which may encompass multiple micro-operations, is translated into an equivalent sequence of RISC-V instructions. For instance, the x86 RET instruction is mapped to two RISC-V instructions (LOAD and JR), and x86 memory accesses exceeding the width of a single RISC-V instruction are decomposed into multiple instructions to preserve the access range. The necessary mapping results in instruction inflation, which is analyzed in § 7.1. In the case of missing opcodes, compute instructions are mapped to representative types such as [F]ADD, [F]MUL, and CONVERT due to limited information in the traces. For ISAs with flag mechanism, such as x86, spare registers can be employed to establish inter-instruction dependencies. Furthermore, special handling for architecturally significant registers, like the return address register, guarantees the correct correspondence between x86 call/return operations and their RISC-V counterparts.

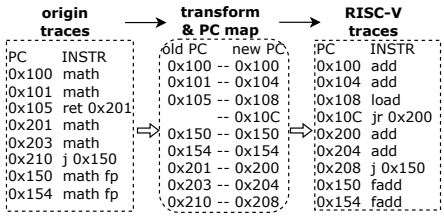


Figure 6. Example of trace transformation, consisting of PC conversion and instruction encoding mapping

To resolve differences in instruction size, PC alignment, and instruction inflation, we reorganize PCs in the traces to conform to RISC-V requirements. As illustrated in Fig. 6, we collect all instruction PCs and sequentially reassign new addresses based on RISC-V encoding size. When a PC gap is detected (e.g., from 0x105 to 0x150), the current PC is updated accordingly. A mapping from original PCs to RISC-V PCs is then constructed, and branch target addresses are updated using this mapping.

While the x86 ISA supports larger offset ranges for direct branch instructions than RISC-V, we observe that branch target computation mainly occurs in two modules: the pre-decoding unit at the fetch stage and the branch execution unit. By overriding the computation result with the target recorded in the traces, we effectively support larger branch offset ranges in the trace-driven RISC-V model.

Overall. TraceBridge provides a methodology to evaluate the microarchitectural behavior of mature, real-world software ecosystems (e.g., Google workloads) on an emerging hardware ecosystem (e.g., RISC-V). Admittedly, TraceBridge is unable to eliminate all performance discrepancies caused by inherent cross-ISA differences and missing execution information in traces, such as instruction semantics and application binary interfaces(ABIs). Furthermore, while the high-level methodology is consistent, the specific rules should adapt for source and target ISAs. For example, x86 and RISC-V differ in the number of general-purpose registers. Consequently, when translating to x86, some registers may map to memory (i.e., register spilling). It is also constrained by information missing from the trace, forcing a simplified instruction remapping, which inevitably introduces performance errors. However, according to our evaluation of missing RISC-V opcodes (§ 7.1), the accuracy is above 99% (0.95% error for SPECint2017) for early-stage performance exploration.

5. Put It All Together

TraceRTL improves the performance evaluation workflow by optimizing stages such as workload preparation, prototyping, and performance simulation. As illustrated in Fig. 7, a typical iterative workflow based on TraceRTL is employed to perform agile performance evaluation. The workflow involves the following steps: ① **Trace Preparation**: Program traces for the benchmarks or target applications are prepared

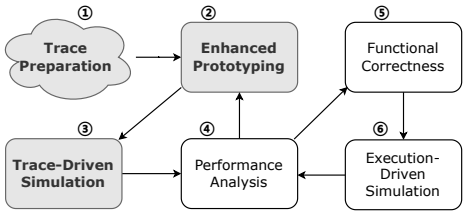


Figure 7. Agile performance evaluation workflow with TraceRTL. The workflow comprises two loops: a trace-driven loop ②→③→④→② and an execution-driven loop ④→⑤→⑥→④.

for subsequent performance evaluation. Each trace represents a program segment. Traces can be generated using a variety of tools, including dynamic instrumentation tools like Pin [46] and DynamoRIO [47], instruction-level simulators like QEMU [50] , and publicly available traces such as Google workload traces [36] and Qualcomm workload traces [35]. TraceRTL can be combined with additional techniques such as SimPoint [40] to further shorten simulation time, while also avoiding the overhead and complexity of booting. ② **Prototyping**: New microarchitectural features can be prototyped on a RTL model without full implementation, as shown in § 7.2. ③ **Trace-Driven Simulation**: The trace-formatted program segments are replayed in trace-driven simulation, yielding performance results of the CPU model. ④ **Performance Analysis**: The performance results and program behaviors are analyzed to identify performance bottlenecks. These insights inform subsequent iterations and guide prototype refinement. ⑤ **Functional Correctness**: When the design meets expected performance targets, the efforts invested in prototype development can be seamlessly carried over. TraceRTL supports compile-time mode switching between execution-driven and trace-driven simulation, enabling smooth transition to functional validation. ⑥ **Execution-Driven Simulation**: Further performance analysis and iteration are conducted through execution-driven.

TraceRTL facilitates an agile and accurate RTL-level microarchitecture design exploration process. Rather than replacing existing architectural simulators, TraceRTL serves as a complementary and reinforcing component that enhances RTL performance exploration and bridges the gap between high-level models and real RTL behavior. It targets a distinct sweet spot in the accuracy-productivity trade-off, preserving the ground-truth RTL model and accepting manageable maintenance overhead to achieve substantially higher accuracy, with comparable or potentially lower (Palladium/FPGA) simulation cost. By enabling direct performance evaluation on real RTL implementations, TraceRTL empowers architects to broaden application coverage and identify microarchitectural bottlenecks that high-level simulators may overlook.

6. Evaluation

We conduct evaluations to address two key questions:

1. Can we mitigate trace-driven simulation's performance inaccuracies (§ 6.2)?
2. Does TraceRTL achieve high performance accuracy (§6.3)?

To address these questions, we compare the performance of the original RTL model, TraceRTL, and the state-of-the-art simulator gem5 [4].

Table 2. Target system configuration.

Component	Description
Branch Predictor	uBTB, BTB, TAGE-SC, ITTAGE, RAS
Fetch/Decode/Rename Width	8/6/6
RoB/LoadQueue/StoreQueue	160/72/64
Integer/Float Register File	224/192
ALU/FMA/FDivSqrt unit	4/4/2/
Load/Store unit	3/2
L1 ICache	64KB, 4-way, 256-set
L1 DCache	64KB, 8-way, 128-set
L2 Cache	1MB, 8-way, 512-set, 4-bank
L3 Cache	16MB, 16-way, 4096-set, 4-bank
L1 ITLB/DTLB	48-entry, fully-associative
L2 TLB	2048-entry, 8-way, 32-set
DRAM	DRAMsim3, 8GB, DDR4-3200

6.1. Experimental Setup

Target System. We evaluate TraceRTL by altering an open-source high-performance RISC-V processor, XiangShan [26, 38], into a trace-driven model. TraceRTL introduces low implementation overhead while preserving RTL fidelity. It reuses the original RTL and drives existing modules by intercepting inputs and outputs. The modifications consist of three primary components. First, the simulation environment, implemented primarily in C++, manages trace file loading, instruction stream validation, and page table generation. Second, the TraceRTL module, written in Chisel, retrieves traces via the DPI and supplies instructions to the processor. Third, interface connections and execution guidance are applied to existing processor modules. The first two components are microarchitecture-agnostic, whereas the third requires tighter coupling with specific microarchitectural details. Specifically, the microarchitecture-specific modifications account for fewer than 450 LOC (lines of code). Nevertheless, the modification methodology remains portable across diverse processor designs.

XiangShan, implemented in Chisel [13], is a tape-out ready superscalar out-of-order processor. Its latest third-generation, Kunminghu, achieves a clock frequency of 3GHz and SPECint2006 score exceeding 15/GHz, demonstrating its capability as a platform for exploring high performance microarchitecture designs. We use the default configuration of XiangShan, as shown in Table 2. We take the original XiangShan's performance as the ground truth.

gem5 is widely used for CPU microarchitecture exploration and is often referenced as the ground truth in some simulator works [8, 51, 52] for its rich details. We use the XS-gem5 [53] as the baseline, which has been carefully calibrated to XiangShan through over 1,200 git commits and more than 60,000 lines of source code additions since July 2022, including XiangShan-specific adjustments.

Simulation Speed. XS-gem5 achieves a simulation speed of around 35kHz. As TraceRTL is directly derived from the original RTL model, it inherently shares a comparablesimulation speed and benefits from hardware-accelerated emulation tools. The simulation speeds are both around 6.5kHz using Verilator [14] and around 1.4MHz on Cadence Palladium, which is 40× faster than XS-gem5.

Workloads. We use SPEC CPU2006 [54] and SPEC CPU2017 [55] benchmark suites. We compare the benchmark

scores between XiangShan, TraceRTL and XS-gem5. The complete execution of SPEC CPU benchmarks takes a very long time in software simulation. A set of representative program segments are generated by sampling the SPEC CPU benchmarks using SimPoint [40]. Each segment consists of 20M instructions for warm-up and 20M instructions for performance sampling. To limit simulation time, more than 30% weight of the program segments are included for each application. NEMU [56], an instruction-level simulator, is employed to execute these segments and generate trace files to feed into TraceRTL. Both XiangShan and XS-gem5 are functionally verified against NEMU, guaranteeing they share the same execution flow.

6.2. Trace-Driven Performance Discrepancies

For the first time, we can evaluate the performance impact of the trace-driven simulation on an accurate high-performance RTL processor and the effectiveness of measures to mitigate its performance errors. We quantify the performance errors arising from wrong-path simulation, memory management unit behaviors, operand and opcode absence.

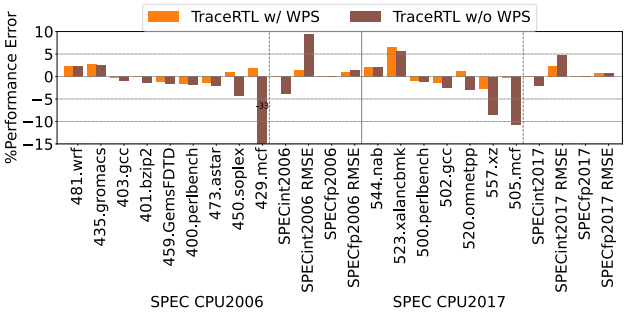


Figure 8. Performance errors of TraceRTL w/ and w/o wrong-path simulation on SPEC CPU2006 and SPEC CPU2017.

6.2.1. Wrong-path Simulation.

We adopt the mechanism detailed in § 4.2.1 to model wrong-path effects. For comparison, we also consider the basic approach where the instruction fetch halts upon encountering a mis-prediction, detailed in § 4.1.1. Fig. 8 illustrates SPEC CPU2006's and SPEC CPU2017's performance differences with and without simulating wrong-path instructions' effect, containing the sub-benchmarks whose "w/o WPS" errors are more than 1%, benchmarks' overall performance errors and RMSE (root mean squared error) metric. Although the overall performance impact of neglecting wrong paths is relatively small (-3.91% and -0.18% for SPECint2006 and SPECfp2006, -2.17% and 0.14% for SPECint2017 and SPECfp2017), certain benchmarks, such as 429.mcf and 450.soplex on SPEC CPU2006 and 505.mcf and 557.xz on SPEC CPU2017, exhibited substantial performance degradation. Our results demonstrate that simulating the impact of wrong-path instructions effectively mitigates these programs' performance discrepancies, reducing the overall performance error to 0.14% for SPECint2006 and 0.13% for SPECint2017. The RMSE of SPECint2006 and SPECint2017 falls from 9.56% and 4.87% to 1.38% and 2.38%.

6.2.2. Instruction Opcode Provisioning

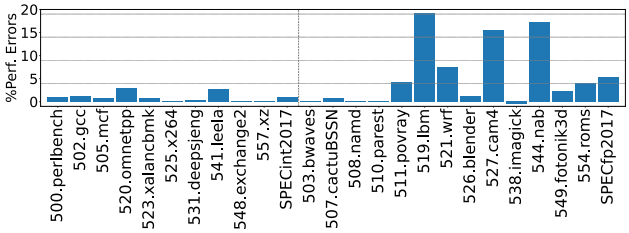
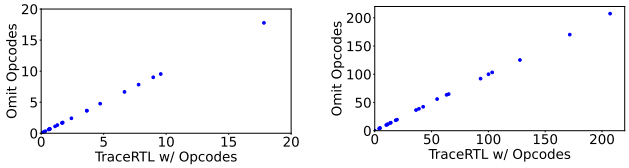


Figure 9. Performance errors of TraceRTL on SPEC CPU2017 when omitting computation instruction opcodes.



(a) Branch predictor MPKI. (b) Data cache MPKI.

Figure 10. BPU and data cache MPKI comparison between TraceRTL w/ and w/o computation instruction opcodes on SPEC CPU2017 benchmarks. Each point represents one sub-benchmark.

Coarse-grained opcode abstraction is common in trace-driven simulators without detailed execution unit modeling, or in applications that directly provide traces without instruction encoding. To quantify the performance deviations, we implemented a controlled mapping scheme within the TraceRTL framework. Specifically, the diverse array of complex computational opcodes are collapsed into a simplified set of generic operations: integer addition/multiplication (ADD/MUL) and floating-point addition/multiplication (FADD/FMUL).

The results across the SPEC CPU2017 demonstrate that the impact of opcode abstraction varies significantly between workload types. As shown in Fig. 9, SPECint2017 exhibits high resilience to coarse-grained semantic mapping, maintaining a negligible average error of 0.95%. In contrast, SPECfp2017 shows a much higher sensitivity, with the average error rising to 5.30% and peaking at 19.29% in 519.lbm. The results suggest that while coarse-grained opcode traces are sufficient for evaluating general-purpose integer architectures, they may introduce unacceptable fidelity loss for floating-point heavy workloads.

Despite the divergence, the coarse-grained abstraction effectively preserves the control-flow and memory-access characteristics of the workloads. As illustrated in Fig. 10, the MPKI metrics of branch predictor and data cache remain highly consistent between the abstracted traces and normal TraceRTL. In summary, coarse-grained opcode abstraction has limited impact on integer compute-intensive applications, frontend modules (branch prediction and instruction fetch), and memory-access related research. It is well-suited for studies where the target workloads or modules have a weak correlation with floating-point operations.

6.2.3. Uncertain-latency Operations

To model the execution latency of uncertain-latency operations, represented by floating-point division and square root (FDivSqrt), we adopt the approach that supplies operands, detailed in § 4.2.3. For comparison, we also evaluated a baseline configuration where the FDivSqrt is replaced with a fixed-latency dummy unit, with latencies varying based on the operation type and data width. As shown in Fig. 11, which

contains sub-benchmarks whose "fixed-latency" errors exceed 0.5%, the fixed-latency model resulted in overall performance errors of -1.65% and -1.22% on SPECfp2006 and SPECfp2017, respectively, with significant deviations for sub-benchmarks such as gromacs and zeusmp in SPECfp2006, and 521.wrf, 527.cam4, and 544.nab in SPECfp2017. By providing operands, we are able to improve the accuracy of performance for these applications.

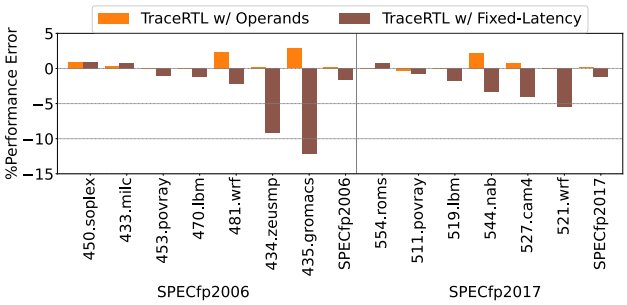


Figure 11. Performance error of simulating FDivSqrt with operand-dependent vs. fixed latency on SPECfp2006 and SPECfp2017.

6.2.4. Memory Management Unit

To evaluate the performance impact of the MMU, we employ the dynamic page table (Dynamic PT) approach detailed in § 4.2.4. For comparison, we also simulate an ideal L1 TLB which always hits and a page table walker with fixed-latency of 15 cycles. As shown in Fig. 12, which contains sub-benchmarks whose "Ideal L1TLB" errors are more than 3%, the ideal MMU introduces average performance discrepancies of 6.19% and 2.35% on SPEC CPU2017 int and fp, with 11 out of 23 benchmarks experiencing performance discrepancies exceeding 3%. When simulating a page table walker with fixed memory latency, 4 out of the 23 benchmarks have errors greater than 3%. In contrast, when simulating the actual MMU behavior, the overall performance overhead decreases to 0.13% and 0.14% for SPEC 2017 int and fp, and only 1 out of 23 benchmarks exhibits a performance error greater than 3%.

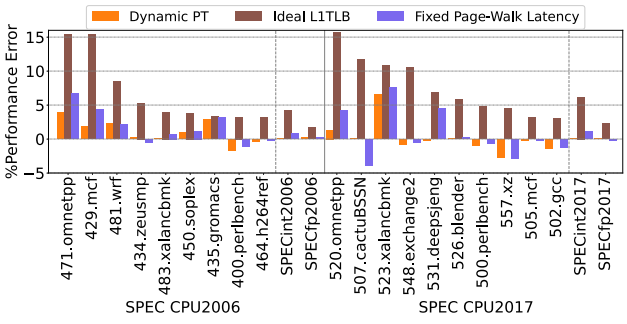


Figure 12. Performance error of simulating the MMU using different strategies on SPEC CPU2006 and SPEC CPU2017.

6.3. Overall Performance Accuracy

We evaluate the performance accuracy of TraceRTL and XS-gem5 on SPEC CPU2006 and SPEC CPU2017, with original XiangShan as the ground truth, as shown in Fig. 13.

Overall. TraceRTL achieves significantly high accuracy in overall performance. For RMSE metric, TraceRTL achieves 1.45% and 1.00% on SPECint2006 and SPECfp2006, compared to 9.85% and 19.44% for XS-gem5. Similarly, the RMSE of SPECint2017 and SPECfp2017 of TraceRTL are 2.38% and 0.67%, compared to 8.02% and 22.53% of XS-gem5.

Sub-benchmarks. TraceRTL exhibits high accuracy at both the overall and sub-benchmark levels. For XS-gem5, on SPEC CPU2006, 11 out of 29 sub-benchmarks have errors greater than 10%, and 14 out of 29 have errors greater than 3%. Similarly, on SPEC CPU2017, 7 out of 23 sub-benchmarks have errors greater than 10%, and 13 out of 23 have errors greater than 3%. These discrepancies can be attributed to the diversity of program characteristics, which makes it challenging to perfectly calibrate. In contrast, by inheriting rich details, TraceRTL effortlessly achieves high accuracy. TraceRTL achieves performance accuracy such that only 1 out of 29 on SPEC CPU2006 and 1 out of 23 on SPEC CPU2017 has an error greater than 3%.

7. Case Studies

In this section, we present case studies to demonstrate how TraceRTL facilitates agile performance evaluation:

1. **Trace Compatibility**: Using TraceBridge, we evaluate x86-based Google workload traces on the RISC-V XiangShan CPU (§ 7.1).
2. **Prototyping**: We use TraceRTL to quickly evaluate the performance impact of adopting a two-stage address translation MMU (§ 7.2) and a new floating-point unit(§ 7.3).
3. **Performance Sensitivity Accuracy**: We compare the accuracy of performance impact between TraceRTL and XS-gem5 at frontend, backend and memory (§ 7.4).

7.1. Trace Compatibility: Google Workload Traces

We evaluate datacenter workloads, the x86-based Google workload traces [36] from warehouse-scale computer workloads on the RISC-V high performance processor XiangShan to show the feasibility of TraceBridge described in § 4.3. Google workload traces consist of multiple trace groups, each containing many trace files. For each group, we select the longest trace and apply the SimPoint [40] for sampling. Applying SimPoint directly to the transformed traces can avoid errors caused by instruction inflation.

While TraceBridge maintains semantic consistency, it introduces the overhead of instruction inflation. We analyze this inflation across both static and dynamic dimensions, considering instruction count and size, as shown in Fig. 14. The inflation ratios for static and dynamic instruction counts remain stable within a narrow range, from 1.09 for arizona to 1.19 for yankee. The dynamic instruction size, an indicator of instruction cache pressure, exhibits an inflation ratio ranging from 0.95 for arizona to 1.20 for bravo.a, with 9 out of 12 applications staying within a 10% inflation margin.

We provide a Top-down [57] breakdown analysis of performance bottlenecks for both Google workload traces and SPECint2017, sorted by IPC, as shown in Fig. 15. While only 3 out of 10 SPECint2017 sub-benchmarks exhibit memory-bound

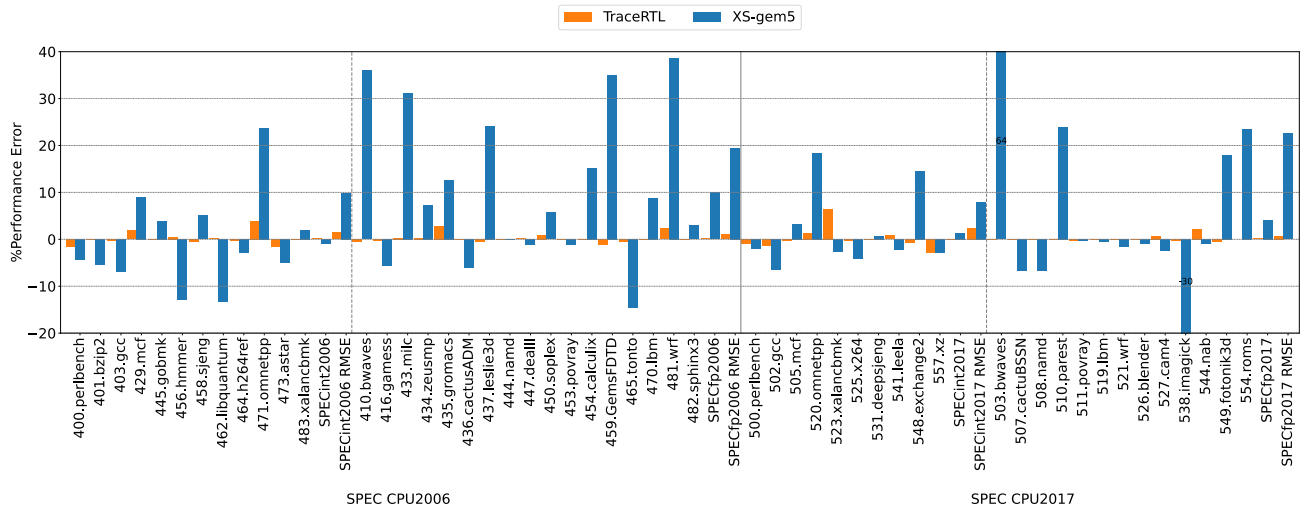


Figure 13. Performance error of TraceRTL and XS-gem5 on SPEC CPU2006 and SPEC CPU2017, using the execution-driven XiangShan as the baseline.

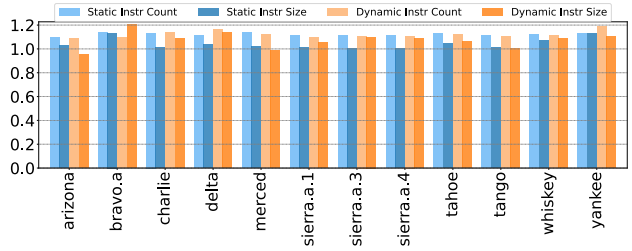


Figure 14. Instruction inflation rate of TraceBridge on Google workload traces.

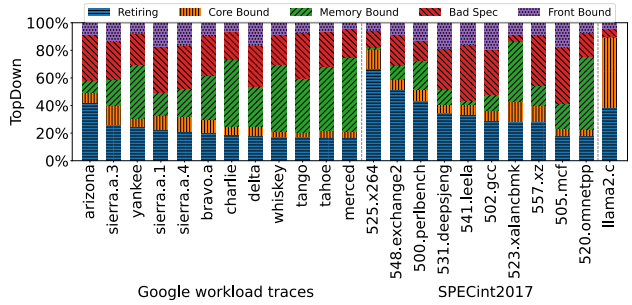


Figure 15. Top-down breakdown comparison between Google workload traces, SPECint2017 , and llama2.c.

over 20%, 8 out of 12 Google workload traces demonstrate this characteristic, with 6 reaching approximately 40%. These results highlight memory access as the primary performance bottleneck, underscoring the importance of memory optimization for warehouse-scale computing systems. TraceBridge introduces dynamic instruction size and count expansion, which primarily affects front-end and core-bound performance categories. However, since these two factors account for relatively small proportions in Google workload traces, TraceBridge has limited impact through expansion effects. Although coarse-grained instruction encoding may potentially affect floating-point workloads, the analysis in § 6.2.2 shows that it preserves accurate instruction streams and cache behavior. This indicates that the impact on memory-bound and bad-speculation categories is also minimal.

TraceRTL also streamlines porting workloads by leveraging the well-developed QEMU. It takes less than 30 minutes to compile llama2.c [58] and generate program traces by QEMU. As shown in Fig. 15, these traces are simulated on TraceRTL, and, unlike Google workload traces and SPECint2017, exhibit distinct core-bound behaviors.

7.2. Prototyping #1: Memory Management Unit

TraceRTL enables efficient prototyping and performance evaluation of complex microarchitectural modules. As a case study, we examine two-stage address translation, a key mechanism for supporting virtual machines through memory virtualization defined in the RISC-V Hypervisor extension [59].

Evaluating this module is non-trivial due to its reliance on privileged operations, complex control and status registers (CSRs), and software-managed page tables. Additionally, its performance impact is significant: address translation may trigger multiple memory accesses to page table. For instance, the RISC-V Sv39 scheme requires 3 memory accesses, while the virtualized, two-stage Sv39-Sv39x4 scheme requires up to 15 memory accesses that increase the translation latency.

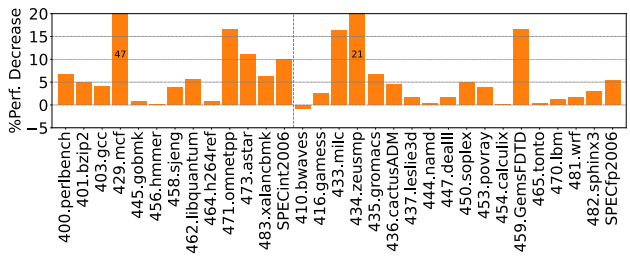


Figure 16. Performance decrease estimation when adopting two-stage address translation on SPEC CPU2006.

TraceRTL allows performance evaluation of such designs before full functional implementation is complete. By (1) directly

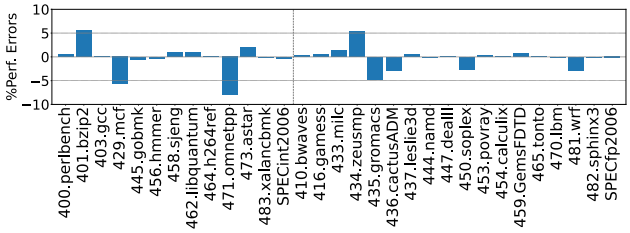


Figure 17. Performance error of TraceRTL on SPEC CPU2006 under KVM virtualization.

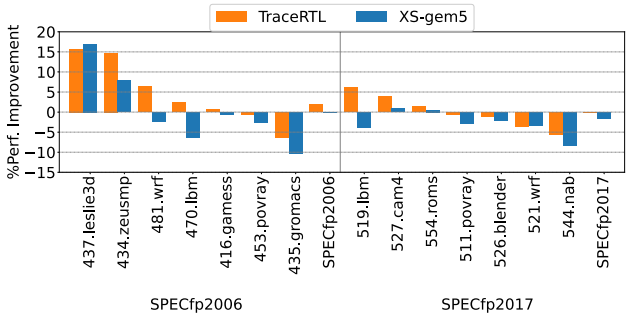


Figure 18. Performance improvement when adopting pipelined FDivSqrt on SPECfp2006 and SPECfp2017.

providing the page table following the two-stage translation scheme and (2) adding a standalone host page table walker which performs guest-physical-address to host-physical-address translation, we enable the MMU to perform the two-stage Sv39-Sv39x4 scheme, thereby obtaining the performance results of two-stage address translation. Fig. 16 illustrates the performance changes of the TraceRTL under normal address translation and two-stage address translation modes on SPEC CPU2006. The two-stage address translation results in a performance degradation of 9.99% for SPECint2006 and 5.27% for SPECfp2017. Among the 29 sub-benchmarks, 10 have a degradation exceeding 5%. In summary, TraceRTL simplifies the requirements for functional correctness and software modifications, providing a robust development platform for exploration around MMU.

To evaluate the accuracy, we compare TraceRTL-based Hypervisor against fully-functional XiangShan Hypervisor on SPEC CPU2006 under KVM virtualization. As shown in Fig. 17, TraceRTL achieves high accuracy, with performance errors below 1% for 19 of 23 sub-benchmarks. The overall error is 0.32% for SPECint2006 and 0.23% for SPECfp2006.

7.3. Prototyping #2: FDivSqrt Unit

TraceRTL enables the implementation of dummy execution units with configurable latency behavior without complex behavioral modeling. For instance, implementing a functional FDivSqrt unit in RTL entails substantial effort, as the implementation in XiangShan exceeds 2,400 LOC and requires extensive verification. To evaluate the pipelined design [60] without incurring such overhead, alternative modeling approaches are necessary. XS-gem5 adopts cycle-accurate modeling for execution units and requires more than 40 LOC of modifications. In contrast, TraceRTL enables dummy implementation with configurable latency in fewer than 10 LOC of modifications, eliminating verification overhead. Figure 18 shows the performance impact of replacing two blocking FDivSqrt units with a pipelined version across benchmarks where TraceRTL changes exceed 0.5%. Notable discrepancies between XiangShan and XS-gem5 appear in SPEC CPU2006 wrf and SPEC CPU2017 lbm. Given the differences in performance accuracy, TraceRTL results are considered more reliable.

7.4. Performance Sensitivity Accuracy

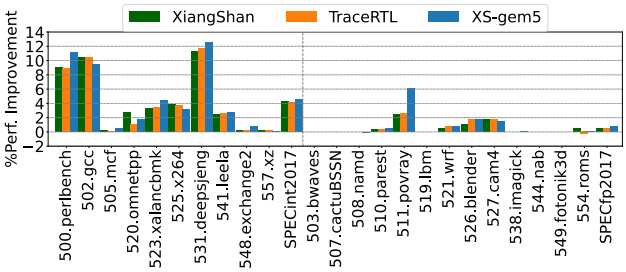
In addition to the performance accuracy of the processor model, the performance sensitivity to microarchitectural modifications is also important. To evaluate the performance sensitivity accuracy to microarchitectural modifications, we adjust key configurations in the frontend, backend, and memory subsystem. For the frontend, we compare the performance impact of different branch target buffer (BTB) sizes—specifically, from 1024 to 2048 (default) entries. For the backend, we vary the number of floating-point units FMA from 2 to 4 (default). For the memory subsystem, we evaluate performance with the best-offset prefetcher in the L2 cache both disabled and enabled (default).

As shown in Fig. 19, we compare the performance variations of XiangShan, TraceRTL and XS-gem5 on SPEC CPU2017 benchmarks under microarchitectural modifications mentioned above. The performance trends observed on TraceRTL closely match those of XiangShan better than those of XS-gem5. For instance, when enlarging BTB size, sub-benchmarks such as 500.perlbench, 502.gcc, and 511.povray exhibit similar trends between TraceRTL and XiangShan. When increasing the number of the FMA, sub-benchmarks like 507.cactuBSSN, 508.namd, and 519.lbm show consistent behavior. When adopting the best-offset prefetcher, sub-benchmarks including 500.perlbench and 507.cactuBSSN also demonstrate analogous performance improvements.

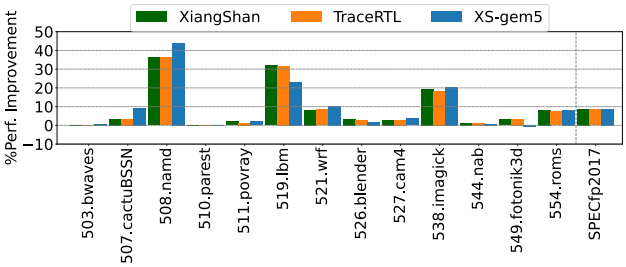
We analyze the notable performance errors of XS-gem5 and find that its prefetching subsystem is considerably more complex and finely tuned, yet lacks clear calibration against the RTL design. This mismatch diminishes the observable performance gains from new prefetchers such as best-offset. The observation highlights the fundamental calibration challenge and motivates the design of TraceRTL: while a model may overfit to the baseline configuration to reproduce similar overall performance, its performance trends for specific microarchitectural features may diverge significantly.

8. Related Work

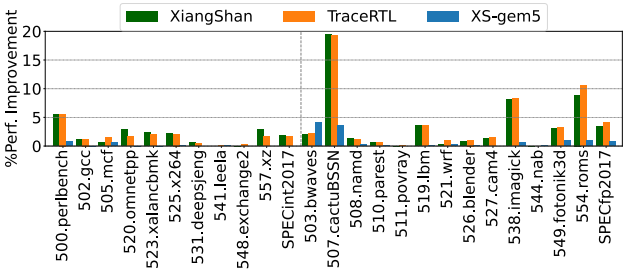
Trace-Driven Model Transformation. Prior work has explored employing trace-based methods to directly control RTL modules' behavior for functional verification, coverage analysis, and performance validation [61, 62]. These works use traces to drive separate RTL modules and the main challenge lies in the generation of traces. Some works collect the traces generated by CPU RTL models for coverage analysis [63]. In contrast, TraceRTL, centered on the whole CPU RTL model, addresses the challenges of design space exploration at the RTL level. Given that achieving high performance accuracy is both a fundamental requirement and a persistent challenge, TraceRTL provides a solution that not only supports prototyping but also enables the execution of workloads in trace form. Trace-driven methodology can be used to improve existing software simulators, such as the trace-driven gem5 mentioned at [64]. In contrast, TraceRTL enhances the RTL simulation to avoid extra model



(a) From 1024 to 2048 BTB entries on SPEC CPU2017.



(b) Increasing FMA number from 2 to 4 on SPECfp2017.



(c) Adopting L2cache best-offset prefetcher on SPEC CPU2017.

Figure 19. Performance improvements of microarchitectural modifications on XiangShan, TraceRTL and XS-gem5.

layers. Accel-Sim [65] adds a new frontend for GPGPU-Sim [66] to support trace-driven simulation. Unlike Accel-Sim's high-level GPU modeling, TraceRTL targets low-level RTL CPU models and addresses challenges of model calibration.

Trace-Driven Performance Inaccuracy. Previous works have investigated performance inaccuracy in trace-driven simulation, primarily focusing on the wrong-path simulation in single-core [67–69], multi-core [70, 71] and synchronization in multi-core simulation [72, 73]. Our methodology mainly focuses on prefetching influence of wrong paths by taking the instructions at the correct path as wrong-path instructions, to suit the RTL model and achieve high accuracy. Moreover, existing trace-driven simulators have a high level of abstraction , which may introduce performance errors thus masking some influencing factors. TraceRTL provides a platform for studying trace-driven simulation.

Error-Prone RTL Model. New RTL languages such as Bluespec SystemVerilog [12], Chisel [13], and SpinalHDL [39] provide high expressiveness and abstraction to reduce design errors. Assassyn [74] introduces a high-level abstraction for asynchronous event handling of pipelined architectures and can generate a calibrated C++ simulator. TraceRTL presents an orthogonal approach to utilizing a trace-driven methodology to decouple the functional and performance models of existing CPU models and expand the scope of workloads.

Trace Format Transformation. Prior work has explored trace format transformation, e.g., converting Arm traces into ChampSim-compatible format [7, 75]. However, ChampSim's high-level abstraction bypasses many low-level challenges, such as differences in instruction semantics, encoding size, PC alignment, and branch offset range, which become critical when executing traces on RTL models.

9. Conclusion

We propose TraceRTL, a methodology to bring trace-driven simulation to the CPU RTL model to facilitate agile performance evaluation. We evaluate TraceRTL by integrating it into XiangShan, achieving high accuracy of 99.87% and 99.86% on SPECint2017 and SPECfp2017. We propose a trace transformation strategy, TraceBridge, and evaluate x86 Google workload traces on the RISC-V XiangShan. TraceRTL mitigates the benchmarking gap between software simulators and RTL design, supports both an RTL-based performance exploration workflow and seamless integration with simulator-driven flows, serving as a bridge from early-stage exploration to last-mile RTL evaluation.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62090022, 62090023, 62172388) and the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA0320000, XDA0320300).

Declaration of competing interests

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.

Data Availability Statements

The data supporting the findings of this study are openly available in XiangShan at `https://github.com/OpenXiangShan/XiangShan/tree/dev-tracertl`.

Credit authorship contribution statement

Zifei Zhang: Conceptualization; Project administration; Methodology; Validation; Investigation; Data curation; Formal Analysis; Writing – original draft. Yinan Xu: Methodology; Writing – review & editing; Kaichen Gong: Software; Validation; Investigation. Sa Wang: Writing; Visualization. Dan Tang: Supervision; Funding acquisition; Resources. Yungang Bao: Supervision; Funding acquisition; Resources; Writing – review & editing.

References

1. Doug Burger and Todd M. Austin. The simplescalar tool set, version 2.0. *SIGARCH Comput. Archit. News*, 25(3):13–25, June 1997. doi:10.1145/268806.268810.
2. Milo M. K. Martin, Daniel J. Sorin, Bradford M. Beckmann, Michael R. Marty, Min Xu, Alaa R. Alameldeen, Kevin E. Moore, Mark D. Hill, and David A. Wood. Multifacet's general execution-driven multiprocessor simulator (gems) toolset. *SIGARCH Comput. Archit. News*, 33(4):92–99, November 2005. doi:10.1145/1105734.1105747.
3. N.L. Binkert, R.G. Dreslinski, L.R. Hsu, K.T. Lim, A.G. Saidi, and S.K. Reinhardt. The m5 simulator: Modeling networked systems. *IEEE Micro*, 26(4):52–60, 2006. doi:10.1109/MM.2006.82.
4. Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. The gem5 simulator. *SIGARCH Comput. Archit. News*, 39(2):1–7, August 2011. doi:10.1145/2024716.2024718.
5. Trevor E. Carlson, Wim Heirman, and Lieven Eeckhout. Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, New York, NY, USA, 2011. Association for Computing Machinery. doi:10.1145/2063384.2063454.
6. Daniel Sanchez and Christos Kozyrakis. Zsim: fast and accurate microarchitectural simulation of thousand-core systems. In *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ISCA '13, page 475–486, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2485922.2485963.
7. Nathan Gober, Gino Chacon, Lei Wang, Paul V. Gratz, Daniel A. Jimenez, Elvira Teran, Seth Pugsley, and Jinchun Kim. The championship simulator: Architectural simulation for education and competition, 2022. URL: https://arxiv.org/abs/2210.14324, arXiv:2210.14324.
8. Hossein Golestani, Rathijit Sen, Vinson Young, and Gagan Gupta. Calipers: a criticality-aware framework for modeling processor performance. In *Proceedings of the 36th ACM International Conference on Supercomputing*, ICS '22, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3524059.3532390.
9. Tony Nowatzki, Jaikrishnan Menon, Chen-Han Ho, and Karthikeyan Sankaralingam. Architectural simulators considered harmful. *IEEE Micro*, 35(6):4–12, 2015. doi:10.1109/MM.2015.74.
10. Cbp2025 simulator framework. https://ericrotenberg.wordpress.ncsu.edu/cbp2025-simulator-framework/, 2025.
11. Sizhuo Zhang, Andrew Wright, Thomas Bourgeat, and Arvind. Composable building blocks to open up processor design. In *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-51, page 68–81. IEEE Press, 2018. doi:10.1109/MICRO.2018.00015.
12. Thomas Bourgeat, Clément Pit-Claudel, Adam Chlipala, and Arvind. The essence of bluespec: a core language for rule-based hardware design. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2020, page 243–257, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3385412.3385965.
13. Jonathan Bachrach, Huy Vo, Brian Richards, Yunsup Lee, Andrew Waterman, Rimas Avižienis, John Wawrzynek, and Krste Asanović. Chisel: constructing hardware in a scala embedded language. In *Proceedings of the 49th Annual Design Automation Conference*, pages 1216–1225, 2012. doi:10.1145/2228360.2228584.
14. Verilator. Verilator user's guide. https://www.veripool.org/guide/latest/, 2026.
15. Haoyuan Wang and Scott Beamer. Repcut: Superlinear parallel rtl simulation with replication-aided partitioning. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS 2023, page 572–585, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3582016.3582034.
16. Kexing Zhou, Yun Liang, Yibo Lin, Runsheng Wang, and Ru Huang. Khronos: Fusing memory access for improved hardware rtl simulation. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '23, page 180–193, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3613424.3614301.
17. Haoyuan Wang, Thomas Nijssen, and Scott Beamer. Don't repeat yourself! coarse-grained circuit deduplication to accelerate rtl simulation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4*, ASPLOS '24, page 79–93, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3622781.3674184.
18. Mahyar Emami, Thomas Bourgeat, and James R. Larus. Parendi: Thousand-way parallel rtl simulation. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '25, page 783–797, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3676641.3716010.
19. Sagar Karandikar, Howard Mao, Donggyu Kim, David Biancolin, Alon Amid, Dayeol Lee, Nathan Pemberton, Emmanuel Amaro, Colin Schmidt, Aditya Chopra, Qijing Huang, Kyle Kovacs, Borivoje Nikolic, Randy Katz, Jonathan Bachrach, and Krste Asanović. Firesim: Fpga-accelerated cycle-exact scale-out system simulation in the public cloud. In *Proceedings of the 45th Annual International Symposium on Computer Architecture*, ISCA '18, page 29–42. IEEE Press, 2018. doi:10.1109/ISCA.2018.00014.
20. Sagar Karandikar, Albert Ou, Alon Amid, Howard Mao, Randy Katz, Borivoje Nikolić, and Krste Asanović. Fireperf: Fpga-accelerated full-system hardware/software performance profiling and co-design. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '20, page 715–731, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3373376.3378455.
21. Mahyar Emami, Sahand Kashani, Keisuke Kamahori, Mohammad Sepehr Pourghannad, Ritik Raj, and James R Larus. Manticore: Hardware-accelerated rtl simulation with static bulk-synchronous parallelism. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4*, pages 219–237, 2023. doi:10.1145/

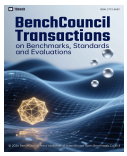
3623278.3624750.

22. Fares Elsabbagh, Shabnam Sheikhha, Victor A Ying, Quan M Nguyen, Joel S Emer, and Daniel Sanchez. Accelerating rtl simulation with hardware-software co-design. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 153–166, 2023. doi:10.1145/3613424.3614257.
23. Christopher Celio, David A Patterson, and Krste Asanovic. The berkeley out-of-order machine (boom): An industry-competitive, synthesizable, parameterized risc-v processor. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2015-167*, 2015. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-167.html.
24. Jerry Zhao, Ben Korpan, Abraham Gonzalez, and Krste Asanovic. Sonicboom: The 3rd generation berkeley out-of-order machine. In *Fourth Workshop on Computer Architecture Research with RISC-V*, volume 5, pages 1–7, 2020. URL: https://people.eecs.berkeley.edu/~krste/papers/SonicBOOM-CARRV2020.pdf.
25. Christopher Celio, Pi-Feng Chiu, Borivoje Nikolic, David A Patterson, and Krste Asanovic. BOOMv2: an open-source out-of-order RISC-V core. In *First Workshop on Computer Architecture Research with RISC-V (CARRV)*, 2017. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-157.pdf.
26. Kaifan Wang, Jian Chen, Yinan Xu, Zihao Yu, Zifei Zhang, Guokai Chen, Xuan Hu, Linjuan Zhang, Xi Chen, Wei He, Dan Tang, Ninghui Sun, and Yungang Bao. XiangShan: An Open-Source Project for High-Performance RISC-V Processors Meeting Industrial-Grade Standards . In *2024 IEEE Hot Chips 36 Symposium (HCS)*, pages 1–25, Los Alamitos, CA, USA, August 2024. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/HCS61935.2024.10665293, doi:10.1109/HCS61935.2024.10665293.
27. Chen Chen, Xiaoyan Xiang, Chang Liu, Yunhai Shang, Ren Guo, Dongqi Liu, Yimin Lu, Ziyi Hao, Jiahui Luo, Zhijian Chen, Chunqiang Li, Yu Pu, Jianyi Meng, Xiaolang Yan, Yuan Xie, and Xiaoning Qi. Xuantie-910: A commercial multi-core 12-stage pipeline out-of-order 64-bit high performance risc-v processor with vector extension: Industrial product. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 52–64. IEEE, 2020. doi:10.1109/ISCA45697.2020.00016.
28. Chen Bai, Qi Sun, Jianwang Zhai, Yuzhe Ma, Bei Yu, and Martin DF Wong. Boom-explorer: Risc-v boom microarchitecture design space exploration framework. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9. IEEE, 2021. doi:10.1109/ICCAD51958.2021.9643455.
29. Siddharth Gupta, Yuanlong Li, Qingxuan Kang, Abhishek Bhattacharjee, Babak Falsafi, Yunho Oh, and Mathias Payer. Imprecise store exceptions. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, ISCA '23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3579371.3589087.
30. Moein Ghaniyoun, Kristin Barber, Yuan Xiao, Yinqian Zhang, and Radu Teodorescu. Teesec: Pre-silicon vulnerability discovery for trusted execution environments. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, ISCA '23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3579371.3589070.
31. Luming Wang, Xu Zhang, Songyue Wang, Zhuolun Jiang, Tianyue Lu, Mingyu Chen, Siwei Luo, and Keji Huang. Asynchronous memory access unit: Exploiting massive parallelism for far memory access. *ACM Trans. Archit. Code Optim.*, 21(3), September 2024. doi:10.1145/3663479.
32. Duo Wang, Mingyu Yan, Yihan Teng, Dengke Han, Haoran Dang, Xiaochun Ye, and Dongrui Fan. A transfer learning framework for high-accurate cross-workload design space exploration of cpu. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–9, 2023. doi:10.1109/ICCAD57390.2023.10323840.
33. Hideki Ando. Performance improvement by prioritizing the issue of the instructions in unconfident branch slices. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 82–94, 2018. doi:10.1109/MICRO.2018.00016.
34. Yinan Xu, Zihao Yu, Dan Tang, Guokai Chen, Lu Chen, Lingrui Gou, Yue Jin, Qianruo Li, Xin Li, Zuojun Li, Jiawei Lin, Tong Liu, Zhigang Liu, Jiazhan Tan, Huaqiang Wang, Huizhe Wang, Kaifan Wang, Chuanqi Zhang, Fawang Zhang, Linjuan Zhang, Zifei Zhang, Yangyang Zhao, Yaoyang Zhou, Yike Zhou, Jiangrui Zou, Ye Cai, Dandan Huan, Zusong Li, Jiye Zhao, Zihao Chen, Wei He, Qiyuan Quan, Xingwu Liu, Sa Wang, Kan Shi, Ninghui Sun, and Yungang Bao. Towards developing high performance risc-v processors using agile methodology. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1178–1199, 2022. doi:10.1109/MICRO56248.2022.00080.
35. Championship value prediction. https://microarch.org/cvp1/. Accessed: 2025-02-20.
36. Google workload traces version 2. https://console.cloud.google.com/storage/browser/external-traces-v2. Accessed: 2025-02-20.
37. Wei Su, Abhishek Dhanotia, Carlos Torres, Jayneel Gandhi, Neha Gholkar, Shobhit Kanaujia, Maxim Naumov, Kalyan Subramanian, Valentin Andrei, Yifan Yuan, and Chunqiang Tang. Dcperf: An open-source, battle-tested performance benchmark suite for datacenter workloads. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, ISCA '25, page 1717–1730, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3695053.3731411.
38. OpenXiangShan. XiangShan. https://github.com/OpenXiangShan/XiangShan, 2020.
39. SpinalHDL. Scala based hdl. https://github.com/SpinalHDL/SpinalHDL, 2024.
40. Timothy Sherwood, Erez Perelman, Greg Hamerly, and Brad Calder. Automatically characterizing large scale program behavior. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS X, page 45–57, New York, NY, USA, 2002. Association for Computing Machinery. doi:10.1145/605397.605403.
41. Alen Sabu, Harish Patil, Wim Heirman, and Trevor E Carlson. Looppoint: Checkpoint-driven sampled simulation for multi-threaded applications. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 604–618. IEEE, 2022. doi:10.1109/HPCA53966.2022.00051.
42. Trevor E Carlson, Wim Heirman, Kenzo Van Craeynest, and Lieven Eeckhout. Barrierpoint: Sampled simulation

of multi-threaded applications. In *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 2–12. IEEE, 2014. doi:10.1109/ISPASS.2014.6844456.

43. Krste Asanovic, Rimas Avizienis, Jonathan Bachrach, Scott Beamer, David Biancolin, Christopher Celio, Henry Cook, Daniel Dabbelt, John Hauser, Adam Izraelevitz, Sagar Karandikar, Ben Keller, Donggyu Kim, and John Koenig. The rocket chip generator. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2016-17*, 4:6–2, 2016. URL: https://aspire.eecs.berkeley.edu/wp/wp-content/uploads/2016/04/Tech-Report-The-Rocket-Chip-Generator-Beamer.pdf.
44. Bruno Sá, Luca Valente, José Martins, Davide Rossi, Luca Benini, and Sandro Pinto. CVA6 RISC-V virtualization: Architecture, microarchitecture, and design space exploration. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2023. doi:10.1109/TVLSI.2023.3302837.
45. RISC-V community. Olympia. https://github.com/riscv-software-src/riscv-perf-model, 2026.
46. Chi-Keung Luk, Robert Cohn, Robert Muth, Harish Patil, Artur Klauser, Geoff Lowney, Steven Wallace, Vijay Janapa Reddi, and Kim Hazelwood. Pin: building customized program analysis tools with dynamic instrumentation. In *Proceedings of the 2005 ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '05, page 190–200, New York, NY, USA, 2005. Association for Computing Machinery. doi:10.1145/1065010.1065034.
47. Derek Bruening, Evelyn Duesterwald, and Saman Amarasinghe. Design and implementation of a dynamic optimization framework for windows. In *4th ACM workshop on feedback-directed and dynamic optimization (FDDO-4)*, page 20, 2001.
48. Nicholas Nethercote and Julian Seward. Valgrind: a framework for heavyweight dynamic binary instrumentation. In *Proceedings of the 28th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '07, page 89–100, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1250734.1250746.
49. DynamoRIO. Dynamorio trace format. https://dynamorio.org/sec_drcachesim_format.html. Accessed: 2026-02-07.
50. Fabrice Bellard. QEMU, a fast and portable dynamic translator. In *USENIX annual technical conference, FREENIX Track*, volume 41, pages 10–5555. California, USA, 2005. URL: https://www.usenix.org/legacy/event/usenix05/tech/freenix/full_papers/bellard/bellard.pdf.
51. Santosh Pandey, Amir Yazdanbakhsh, and Hang Liu. Tao: Re-thinking dl-based microarchitecture simulation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(2):1–25, 2024. doi:10.1145/3656012.
52. Muhammad E. S. Elrabaa, Ayman Hroub, Muhamed F. Mudawar, Amran Al-Aghbari, Mohammed Al-Asli, and Ahmad Khayyat. A very fast trace-driven simulation platform for chip-multiprocessors architectural explorations. *IEEE Transactions on Parallel and Distributed Systems*, 28(11):3033–3045, 2017. doi:10.1109/TPDS.2017.2713782.
53. OpenXiangShan. XS-gem5. https://github.com/OpenXiangShan/GEM5, 2020.
54. John L Henning. Spec cpu2006 benchmark descriptions. *ACM SIGARCH Computer Architecture News*, 34(4):1–17, 2006. doi:10.1145/1186736.1186737.
55. James Bucek, Klaus-Dieter Lange, and Jóakim v. Kistowski. SPEC CPU2017: Next-generation compute benchmark. In *Companion of the 2018 ACM/SPEC International Conference on Performance Engineering*, pages 41–42, 2018. doi:10.1145/3185768.3185771.
56. OpenXiangShan. NEMU. https://github.com/OpenXiangShan/NEMU, 2019.
57. Ahmad Yasin. A top-down method for performance analysis and counters architecture. In *2014 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 35–44, 2014. doi:10.1109/ISPASS.2014.6844459.
58. Andrej Karpathy. llama2.c: Inference Llama 2 in one file of pure C. https://github.com/karpathy/llama2.c. Accessed: 2026-02-07.
59. RISC-V. RISC-V Instruction Set Manual. https://github.com/riscv/riscv-isa-manual. Accessed: 2026-02-07.
60. Javier D. Bruguera. Low-latency and high-bandwidth pipelined radix-64 division and square root unit. In *2022 IEEE 29th Symposium on Computer Arithmetic (ARITH)*, pages 10–17, 2022. doi:10.1109/ARITH54963.2022.00012.
61. Vivekananda M Vedula, Jacob A Abraham, Jayanta Bhadra, and Raghuram Tupuri. A hierarchical test generation approach using program slicing techniques on hardware description languages. *Journal of Electronic Testing*, 19:149–160, 2003. doi:10.1023/A:1022885523034.
62. Lingyi Liu and Shobha Vasudevan. Efficient validation input generation in rtl by hybridized source code analysis. In *2011 Design, Automation & Test in Europe*, pages 1–6, 2011. doi:10.1109/DATE.2011.5763253.
63. Biruk Mammo, Jim Larimer, Matthew Morgan, Dave Fan, Eric Hennenhoefer, and Valeria Bertacco. Architectural trace-based functional coverage for multiprocessor verification. In *2012 13th International Workshop on Microprocessor Test and Verification (MTV)*, pages 1–5, 2012. doi:10.1109/MTV.2012.12.
64. Sotiris Apostolakis, Chris Kennelly, Xinliang David Li, and Parthasarathy Ranganathan. Necro-reaper: Pruning away dead memory traffic in warehouse-scale computers. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '25, page 689–703, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3676641.3716007.
65. Mahmoud Khairy, Zhesheng Shen, Tor M. Aamodt, and Timothy G. Rogers. Accel-sim: An extensible simulation framework for validated gpu modeling. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 473–486, 2020. doi:10.1109/ISCA45697.2020.00047.
66. Ali Bakhoda, George L. Yuan, Wilson W. L. Fung, Henry Wong, and Tor M. Aamodt. Analyzing cuda workloads using a detailed gpu simulator. In *2009 IEEE International Symposium on Performance Analysis of Systems and Software*, pages 163–174, 2009. doi:10.1109/ISPASS.2009.4919648.
67. Onur Mutlu, Hyesoon Kim, David N Armstrong, and Yale N Patt. An analysis of the performance impact of wrong-path memory references on out-of-order and runahead execution processors. *IEEE Transactions on Computers*, 54(12):1556–1571, 2005. doi:10.1109/TC.2005.190.

68. Stijn Eyerman, Sam Van den Steen, Wim Heirman, and Ibrahim Hur. Simulating wrong-path instructions in decoupled functional-first simulation. In *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 124–133. IEEE, 2023. doi:10.1109/ISPASS57527.2023.00021.
69. Bhargav Reddy Godala, Sankara Prasad Ramesh, Krishnam Tibrewala, Chrysanthos Pepi, Gino Chacon, Svilen Kanev, Gilles A Pokam, Daniel A Jiménez, Paul V Gratz, and David I August. Correct wrong path. *arXiv preprint arXiv:2408.05912*, 2024. URL: https://doi.org/10.48550/arXiv.2408.05912.
70. Resit Sendag, Ayse Yilmazer, Joshua J. Yi, and Augustus K. Uht. The impact of wrong-path memory references in cache-coherent multiprocessor systems. *Journal of Parallel and Distributed Computing*, 67(12):1256–1269, 2007. Best Paper Awards: 20th International Parallel and Distributed Processing Symposium (IPDPS 2006). URL: https://www.sciencedirect.com/science/article/pii/S0743731507000457, doi:10.1016/j.jpdc.2007.03.005.
71. R. Sendag, A. Yilmazer, J.J. Yi, and A.K. Uht. Quantifying and reducing the effects of wrong-path memory references in cache-coherent multiprocessor systems. In *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, pages 10 pp.–, 2006. doi:10.1109/IPDPS.2006.1639260.
72. Stephen R Goldschmidt and John L Hennessy. The accuracy of trace-driven simulations of multiprocessors. *ACM SIGMETRICS Performance Evaluation Review*, 21(1):146–157, 1993. doi:10.1145/166962.167001.
73. Karthik Sangaiah, Michael Lui, Radhika Jagtap, Stephan Diestelhorst, Siddharth Nilakantan, Ankit More, Baris Taskin, and Mark Hempstead. Synchrotrace: Synchronization-aware architecture-agnostic traces for lightweight multicore simulation of cmp and hpc workloads. *ACM Trans. Archit. Code Optim.*, 15(1), March 2018. doi:10.1145/3158642.
74. Jian Weng, Boyang Han, Derui Gao, Ruijie Gao, Wanning Zhang, An Zhong, Ceyu Xu, Jihao Xin, Yangzhixin Luo, Lisa Wu Wills, and Marco Canini. Assassyn: A unified abstraction for architectural simulation and implementation. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, ISCA '25, page 1464–1479, New York, NY, USA, 2025. Association for Computing Machinery. doi:10.1145/3695053.3731004.
75. Josué Feliu, Arthur Perais, Daniel A. Jiménez, and Alberto Ros. Rebasing microarchitectural research with industry traces. In *2023 IEEE International Symposium on Workload Characterization (IISWC)*, pages 100–114, 2023. doi:10.1109/IISWC59245.2023.00027.



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/xqpw3878
Review Article

REVIEW ARTICLE

Mapping the Intellectual Landscape of Blockchain in the Banking Industry: A Hybrid Bibliometric and Systematic Review (2015–2025)

Sadeq Abdullah Aladeeb[1,2,*] and Fatima Zohra Sossi Alaoui[1]

[1]Laboratory of Economics, Finance, Management and Innovation, Faculty of Economics and Management, Ibn Tofail University, Kenitra, Morocco and [2]Department of Accounting and Auditing, Faculty of Commerce and Economics, Sana'a University, Sana'a, Yemen
*Corresponding author. Email: sadeqabdullahhasan.al-adeeb@uit.ac.ma
Received on 8 August 2025; Accepted on 13 March 2026

Abstract

The advent of blockchain technology has introduced new alternatives to traditional banking systems, providing a decentralized, secure, and transparent framework. However, its adoption is still complex and uneven for many reasons. This study provides a comprehensive mapping of the intellectual trajectory, thematic structure, and development of blockchain technology research in the banking sector. Using a hybrid literature review methodology that combines bibliometric analysis and systematic content review, the study analyzes 389 peer-reviewed publications retrieved from Scopus (2015–May 2025). VOSviewer was employed to conduct performance analysis and science mapping, including co-authorship, co-citation, keyword co-occurrence, and bibliographic coupling analyses. In parallel, qualitative thematic analysis identified six clusters: (1) blockchain in banking and financial intermediation to enhance operational efficiency, (2) decentralized finance and cryptocurrencies, (3) integration of blockchain with other digital innovations, (4) trust-related dimensions, (5) institutional and regulatory aspects, and (6) strategies for modernizing banking business models. The findings reveal a steady rise in research output, regional disparities in collaboration, and thematic evolution from early conceptualization to recent signs of diversification of applied research. By integrating quantitative and qualitative insights, this study highlights key research gaps, offers directions for future work, and provides guidance for academics, practitioners, and policymakers on the transformative potential and challenges of blockchain in banking.

Key words: Blockchain Technology, Banking Sector, Bibliometric Analysis, Systematic Content Review, Financial Technology, Decentralized Finance

1. Introduction

In recent years, the banking sector has undergone a significant transformation, driven by the rapid advancement of emerging technologies, particularly blockchain. The widespread adoption of smartphones and high-speed data transmission has not only disrupted social interactions but also traditional business operations. However, legacy banking systems have faced challenges in adapting to these technological advancements due to factors such as structural rigidity, high operational costs, and inadequate transaction processing speeds. For instance, cross-border remittances frequently necessitate several days to complete, an inefficiency that starkly contrasts with the near-instantaneous nature of digital communication [1].

In this context, blockchain technology emerged in 2008 as the technology underpinning Bitcoin, a peer-to-peer digital currency eliminating the intermediary [2]. Its decentralized nature allows for secure, anonymous, and cost-effective transactions. This has led to the conclusion that it possesses considerable potential as an off-balance sheet replacement for conventional banking systems [3].

The theory behind blockchain, however, goes back to the early 1990s when Stuart Haber and W. Scott Stornetta developed a cryptographically secure method of time-stamping digital documents [4]. Their subsequent introduction of Merkle trees enabled data to be gathered into chained blocks, significantly enhancing security as well as efficiency [5]. The modern blockchain, conceptualized by Satoshi Nakamoto, is a form of Distributed Ledger Technology (DLT) that utilizes consensus algorithms on distributed nodes to record transactions [6, 7]. The design of the blockchain, which consists of a series of blocks that are cryptographically linked, ensures immutability and tamper-proofing. Consequently, it establishes a highly reliable digital record-keeping system [8]. The application of blockchain technology has expanded beyond its initial implementation in the domain of cryptocurrency. It has been adopted in various

In the banking sector, blockchain is increasingly seen as an innovative way to transform the trustworthiness and reliability of data management [15]. As digital technology continues to penetrate daily life and concern about data security grows, blockchain's significance will continue to rise. It may become as integral to daily life as the internet [6]. Furthermore, the emergence of newer technologies, such as blockchain, will transform the banking sector in the near future [16]. For example, banks are expected to save $10 billion in cross-border payment fees by 2030 by adopting blockchain technology [17]. According to World Economic Forum projections, blockchain technology will reach a significant milestone by 2027, becoming integrated into various sectors of the global economy. A considerable augmentation in the financial sector, including the banking industry, is projected to increase GDP by 10% [18].

Among the most prominent manifestations of this transformation is the rise of decentralized finance (DeFi), which uses blockchain technology to facilitate peer-to-peer financial services without the need for intermediaries such as conventional banks. This setup transcends geographical locations and provides basic financial services, such as savings, loans, and investment products to poor communities in emerging economies [19, 20].

As a revolutionary innovation, blockchain technology offers numerous benefits: enhanced security, privacy, operational transparency, and increased efficiency. This is all a result of its decentralized nature and the use of cryptographic algorithms, which significantly reduce the risk of cyberattacks and fraud while ensuring traceability and data integrity [21–24]. Consequently, banks are increasingly exploring blockchain technology for applications such as cross-border payments, streamlined Know Your Customer (KYC) processes, enhanced anti-money laundering (AML) measures, and automated contract enforcement through smart contracts. These innovations collectively contribute to lowering operational costs and improving overall efficiency [25, 26].

Despite the promise of blockchain technology, its adoption by banks faces limiting factors. These factors include regulatory uncertainty, technical complexity, and resistance to change at the organizational level. A meticulous examination of the opportunities and limitations presented by this technology is imperative, accompanied by a thorough assessment of awareness, readiness, and acceptance levels among banks and customers [27–29].

The timing, evolution trajectory, and possible impact of blockchain technology on banking have garnered considerable interest among academics and practitioners. In recent years, academic interest in the topic has increased markedly, resulting in a large and diverse body of literature. No study, to the best of my knowledge, has ever carried out a detailed systematic mapping of the intellectual structure, theme development, and future research trends of blockchain literature in the banking sector using a systematic integration of bibliometric analysis and systematic content review methods. Consequently, there is a need to identify and assess the current state of the art and prevailing research trends in this domain.

To fill this gap, this study utilizes a hybrid methodology of literature review, combining the bibliometric analysis and systematic content review to answer the following research questions:

RQ1: *What are the prevailing research trends and patterns of scholarly collaboration in the domain of blockchain technology in the banking sector between 2015 and 2025?*

RQ2: *What are the core thematic clusters and intellectual structures underpinning blockchain research in the banking sector?*

RQ3: *What key research gaps and future directions can be identified to advance the understanding and application of blockchain technology in the banking sector?*

Amidst the accelerating digitalization of banking and financial systems, blockchain technology is revolutionizing how banking services are produced and disseminated. The primary aim of this research is to synthesize the current academic literature on the impact of blockchain on banking by identifying the key concepts, emerging research trends, and prevailing themes. To this end, the study adopts a mixed-method research approach combining quantitative bibliometric analysis with a qualitative systematic content analysis to map the intellectual landscape of blockchain studies in banking. The combination strengthens the validity and credibility of the findings, offering an overarching perspective on how blockchain is reshaping the industry. Besides mapping the literature, the study provides critical reflections on academic and institutional responses to the emergence of blockchain and indicates avenues for further research in a bid to advance its revolutionary potential in the banking sector.

By doing so, this study contributes to a deeper understanding of the significant development of the field and guides future academic and practical engagement with blockchain innovation in the banking sector. The novelty of the study lies in its explicit framework of triangulation and cross-validation that combines bibliometric science mapping and qualitative thematic analysis, providing valuable and actionable insights for academics, practitioners, and policymakers. In addition, the study contributes to the development of transparent, safe, and effective banking and financial systems by identifying the advantages and obstacles related to the adoption of blockchain technology systematically and the proposal of an organized agenda for future research in this field.

The present article is structured as follows. Subsequent to this introduction, Section 2 delineates the hybrid review methodology, meticulously expounding the bibliometric and systematic content analysis approaches. In Section 3, the results of the performance analysis and science mapping of 389 publications on blockchain in banking are presented, and the six main thematic clusters identified are discussed. Section 4 identifies the managerial and practical implications of the findings. Finally, Section 5 offers the main conclusions, which include a summary of the key findings, a proposal of directions for future research, and an acknowledgement of the study's limitations.

2. Research Methodology

To achieve the purposes of this study, we use a hybrid review methodology that integrates bibliometric analysis and systematic content analysis. The mixed-method approach combines quantitative analysis with a substantial emphasis on qualitative analysis.

A hybrid review approach, as described by Paul and Criado [30], is a method that facilitates a comprehensive examination of the literature by combining quantitative and qualitative approaches, with the aim of organizing, analyzing, and interpreting data in a meaningful way. The objective is to provide a comprehensive summary of the scholarly literature on the adoption of blockchain technology (BCT) in the banking industry and to offer an integrative review of the main topics, major findings, and research agendas for the future in this domain.

Bibliometric analysis, which relies on the statistical evaluation of academic production [31], is complemented in this study by content analysis, a qualitative technique used for the systematic analysis of textual information and disclosure structure of existing knowledge within a given discipline [32]. The methodology stages and analytical tools adopted to fulfill the objectives of the study are outlined in Figure 1.

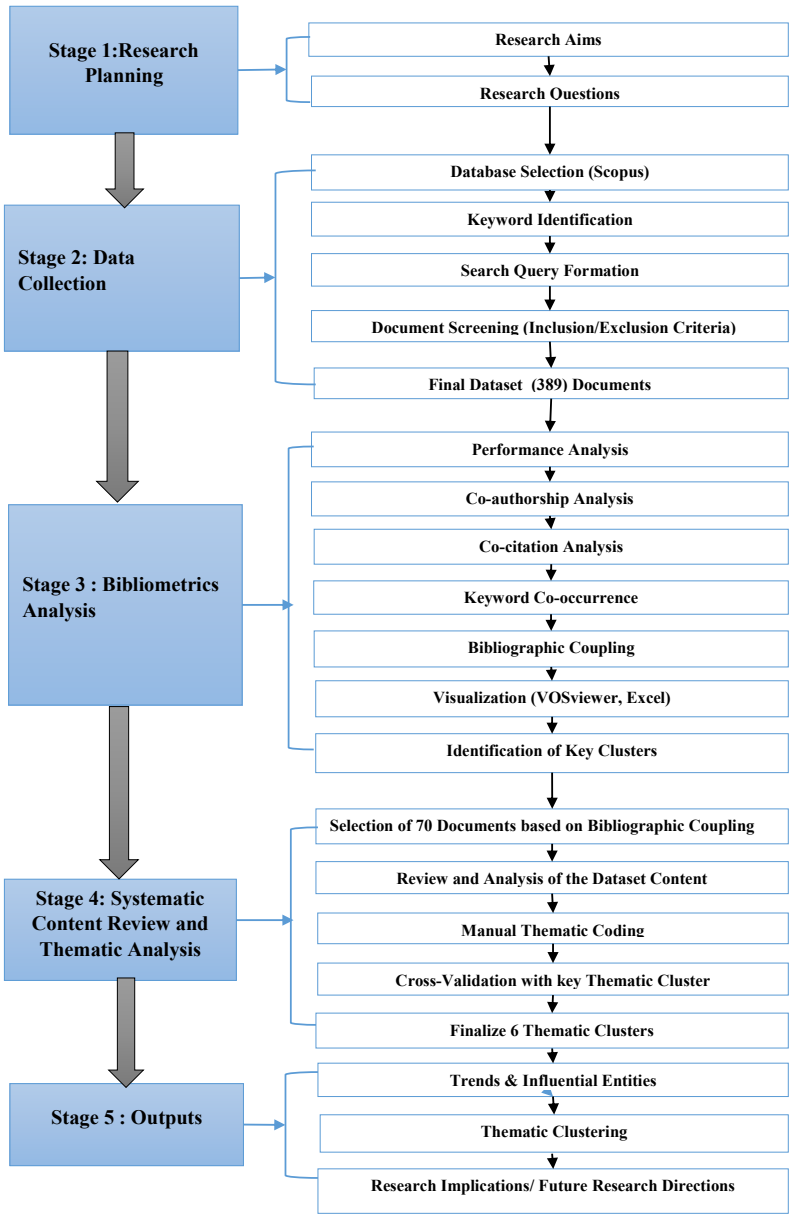


Figure 1. Research Design of the Hybrid Bibliometric–Systematic Literature Review (Developed by the Authors)

2.1. Data Collection

2.1.1. Database Selection

In this stage, data were collected from the Scopus database, a widely recognized and reputable source for bibliometric research [18, 33]. Although there are other databases, such as Web of Science, IEEE Xplore, and Google Scholar, Scopus was selected because it offers the largest curated abstract and citation database of peer-reviewed social science and business publications, indexing over 27,000 active titles from more than 7,000 international publishers, with particularly strong coverage in Finance, Management, Economics, and Information Systems disciplines[34, 35]. Prior bibliometric methodology research indicates that Scopus retrieves broader journal coverage and comparable citation structures to Web of Science for management and interdisciplinary technology studies, while offering superior metadata consistency for science mapping analyses. Its extensive coverage also makes it convenient for research in corporate finance, such as the adoption of blockchain in the banking sector. A defined inclusion criterion was applied for the selection of relevant keywords and the extraction of the dataset for bibliometric analysis and systematic literature review.

2.1.2. Keyword Identification

To identify the appropriate keywords for retrieving the dataset of our research, we have carried out a comprehensive review of the previous literature on blockchain in the banking sector. The focus was on determining the most frequent words used in the current literature [18, 36, 37]. For this purpose, Google Scholar was used in the search using the keyword phrase "Blockchain Technology in the Banking Sector," and related studies were referenced to determine common keywords. Besides, previous bibliometric and systematic literature reviews were also examined to confirm that the selected keywords were inclusive and specific.

Based on this literature review, we identified several frequently used search terms, such as "Blockchain in Bank," "Blockchain Technology in Bank," "Blockchain in Finance," and "Blockchain Technology in Finance." Additionally, Boolean search strings such as (blockchain AND banking) and (block-chain AND adoption AND banking) were identified. Furthermore, consultation with two academic experts in finance and block-chain confirmed that the keywords "Blockchain AND Banking" are frequently used to describe studies where both blockchain and banking are major foci, rather than merely contextually related.

Although this exploratory phase did identify a number of related terms, we purposely restricted the scope of the final retrieval query to just "Blockchain AND Banking" to make sure that both the blockchain and banking domains are the primary focus of our analysis and that the studies retrieved are focused products of those two areas of study. Moreover, the selection of these keywords is congruent with our research objectives, particularly in developing an intellectual structure and determining the main contributions towards the understanding of the impact of blockchain technology on banking.

2.1.3. Search Criteria and Data Extraction

The data collection process during the study was conducted systematically, following standard bibliometric study practices [38] and PRISMA guidelines for transparent reporting [39]. On 12 May 2025, a search was made using the Scopus database with the search term "Blockchain AND Banking" and yielded 1,641 documents published between 2015 and 12 May 2025. Although blockchain technology emerged in 2008, until 2015, academic interest in adopting blockchain technology in the banking sector was significantly nonexistent. Therefore, the selected time frame (2015–2025) indicates the evolution of scientific production in the field.

Because of the novelty and rapid progress of the research field, formal inclusion criteria were applied to ensure the analytical relevance and dataset quality. Only peer-reviewed articles, conference papers, and review articles were chosen, restricting analysis to the most relevant subject areas: Business, Management, and Accounting; Economics, Econometrics, and Finance; and Social Sciences. Publications that focused primarily on technical or computational aspects without a substantial connection to banking, economic, or financial applications were excluded to maintain thematic consistency with the study's objectives. Additionally, only English-language documents were included to ensure conceptual consistency and facilitate systematic review. The final search string was as follows:

TITLE-ABS-KEY (Blockchain AND Banking) AND PUBYEAR > 2015 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , "BUSI") OR LIMIT-TO (SUBJAREA , "ECON") OR LIMIT-TO (SUBJAREA , "SOCI")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "re"))

This search strategy prioritizes thematic specificity over broad recall, which is a typical approach used in bibliometric mapping studies that strive for clearer concepts and greater analytical coherence. In addition, exploratory pilot studies utilizing broader terminology (fintech, financial services, distributed ledgers, etc.) resulted in the retrieval of an excessive number of records (more than double) that either only slightly or not substantially referenced Blockchain and/or Banking. As a result, continued usage of this focused search strategy was maintained to ensure precision and maintain the analytical quality of the results of bibliometric and thematic analyses of this research.

Following the removal of duplicates and non-relevant documents using filters, the final dataset of 389 documents was attained. The records were saved in CSV (Comma-Separated Values) format for subsequent bibliometric analysis. To ensure replicability and transparency, the dataset has been publicly released in a special repository and is in line with the banking and finance literature's standard practices.

2.2. Data Refinement and Analysis

The second stage of the systematic protocol involved refining retrieved data after the previous step to generate a dataset for bibliometric mapping and thematic synthesis of the literature. It is important to note that this stage did not change the content or makeup of the dataset retrieved from earlier stages; rather, it enhanced the reliability and interpretability of analyses of keyword-based data, specifically co-occurrence networks and clustering themes from keywords.

Data preparation involved the systematic elimination of false positives, the cleaning of metadata fields, and the normalization of author-provided keywords. Keyword optimization was accomplished by merging singular and plural forms, standardizing spelling differences, and consolidating synonymous terms into cohesive conceptual labels. For instance, terms like "cryptocurrency" and "cryptocurrencies," "smart contracts" and "smart contract," as well as "bank" and "banks," were standardized into singular keyword inputs. Additionally, terminology standardization was employed to harmonize overlapping definitions typically found in the diverse blockchain literature. This process included incorporating equivalent phrases such as "distributed ledger," "distributed ledger technology," "fintech," "decentralized finance," and "DeFi," along with "banking sector"

and "banking industry." Terms that were irrelevant or contextually unsuitable (such as "bibliometric analysis and COVID-19") were eliminated to uphold thematic consistency. After optimization and normalization, two complementary analytical methods were executed: - Descriptive bibliometric analysis, which evaluated publication trends, citation patterns, source productivity, and networks of key contributors across various fields; - Systematic content analysis, which pinpointed key research themes, core theme groups, and predominant scholarly discussions arising from the literature.

As a result, this refinement process ensured the statistical reliability of bibliometric network structures and the conceptual clarity of thematic interpretations while not impacting article inclusion or research coverage.

2.2.1. Bibliometric Analysis

Following the collection and preparation of the research dataset, a scientometric analysis was conducted for the purpose of examining the structure and dynamics of the research field. In this study, VOSviewer [40], a specialized computer software program for constructing and visualizing large-scale bibliometric networks [41], was employed. VOSviewer software was selected due to its proven capabilities in the management of large networks, as well as its inbuilt text-mining functionality, which enables the extraction and analysis of valuable terms and concepts from the literature [42]. Additionally, Microsoft Excel was used for statistical analysis and data visualization, including examining publication trends by year, conducting citation analysis, and determining keyword frequency.

Bibliometric analysis provides a comprehensive approach for tracing the development of a research theme with established and reproducible methods. These methods are largely recognized as objective, accurate, and reproducible [43]. Two main bibliometric approaches were applied in this study: performance analysis and science mapping.

Performance analysis is a fundamental component of bibliometric analysis, focused on the quantitative assessment of scientific productivity and impact. It provides a data-driven perspective of scholarly output and the growth of a scientific discipline over time. This technique encompasses the analysis of annual publication trends, identification of highly cited publications, evaluation of leading scientific journals, and assessment of the research contributions by institutions, countries, and individual authors [44]. By examining these indicators, performance analysis provides a comprehensive understanding of the intellectual evolution of the field and uncovers its most influential contributors.

Science mapping, on the other hand, provides a graphic and structural representation of the intellectual architecture of the field [45]. This technique involves advanced bibliometric techniques such as co-authorship analysis, citation and co-citation analysis, keyword co-occurrence, and bibliographic coupling. These analyses help to uncover primary research areas, common keywords, and the thematic clusters that form the landscape of the field [46]. In particular, bibliographic coupling was used to study thematic clusters and current fronts of research to identify emeging topics, research gaps, and directions for future research. Certain previous bibliometric research has utilized similar approaches to examine blockchain research within the banking sector [18], [36], [37], confirming the relevance and propriety of the methodology used herein.

2.2.2. Systematic Content Review

To comprehensively explore the emerging themes of block-chain technology in banking, this study adopted a two-phase methodological design. Specifically, it combined bibliometric analysis with systematic content review. The mixed-method approach was used to address the research questions RQ2 and RQ3. By integrating quantitative and qualitative techniques, the study aimed to synthesize dominant research themes, assess how blockchain would impact banking operations, and ascertain dominant scholarly trends. Systematic content analysis not only contributed to complementing bibliometric findings but also to enhancing the interpretative depth of the results.

In the first phase, bibliometric techniques were applied using VOSviewer in order to visualize the intellectual structure of the field. Following the procedure outlined by [38], two science mapping techniques were used. First, an analysis of keyword co-occurrence was performed to identify words that frequently co-occur together in the metadata of articles' titles, abstracts, and keywords [47]. Consequently, the main research areas, key themes, and emerging research topics were identified [45]. Second, bibliographic coupling was employed to cluster articles that share common cited references, thereby revealing thematically related research streams [48]. A minimum citation threshold of 30 citations per document was applied to exclude publications with limited scholarly impact. This resulted in the selection of 69 highly cited papers. Additionally, the 10 highly cited papers were manually added to ensure conceptual comprehensiveness. After duplicate removal and further manual filtering, the final dataset of 70 peer-reviewed papers was established as the foundation for the subsequent qualitative review.

In the second phase, a qualitative content analysis was performed using Braun and Clarke's framework [49] as follows. First, each article in the final dataset was examined and coded to extract relevant information regarding research objectives, methodological approaches, core themes, principal findings, and identified research gaps. Second, the coded content was grouped into preliminary thematic categories based on their conceptual similarities. Thereafter, these thematic categories were manually refined to ensure conceptual relevance and logical coherence within the categorization. For instance, thematic clusters that have similar central themes (e.g., blockchain and cryptocurrency, blockchain and DeFi) were consolidated into a common thematic cluster. Finally, the outcomes derived from the qualitative analysis were cross-validated against those generated through keyword co-occurrence analysis to enhance the results of the study.

Methodological Novelty and Contribution

The novelty of the methodological approach of this study resides in its explicit triangulation and cross-validation framework that combines bibliometric maps of science with qualitative thematic analysis. Previous studies in this field either used descriptive bibliometric mapping or a qualitative synthesis, but these studies were typically based on small samples and treated these methods separately. In contrast, this research is designed in three stages: (i) to identify macro-level thematic structure through quantitative bibliometric mapping; (ii) to use systematic qualitative content analysis to capture in-depth conceptual patterns and research gaps; and (iii) to cross-validate the results of quantitative and qualitative analysis to determine both the statistical relationship and conceptual alignment of those analyses.

The triangulation approach provides greater methodological strength because it provides a more extensive and functionally reliable representation of the research landscape, helping to better establish a framework for developing theories, as well as planning future investigations into the impact of blockchain on banking studies.

Methodological Challenges and Mitigation

The rapid growth of the literature on blockchain applications in banking presents several methodological challenges. These issues arise from four dimensions of interrelated challenges: disciplinary fragmentation, terminological inconsistency, publication volume/size/overview, and methodological heterogeneity. Blockchain research in banking spans broad disciplinary areas, including finance, computer science, information systems, law, and regulatory research, making it difficult to align themes and integrate theories. Additionally, many overlapping terms exist, including fintech, digital banking, cryptocurrencies, decentralized finance (DeFi), central bank digital currency (CBDC), etc. These multiple terms significantly increase the potential for conceptual confusion and misclassification.

In addition to these difficulties caused by the rapid growth in the number of publications, many difficult processes of literature screening and synthesis occur when hundreds of literature articles are reviewed while attempting to keep the reviews analytically sound. Moreover, the research literature reviewed exhibited considerable methodological differences, ranging from technical system architectures and research analysis to policy-oriented and conceptual frameworks, complicating the synthesis of cross-study data.

To alleviate these issues, the present research develops a triangulated methodological approach that employs bibliometric mapping of literature using computer analysis tools as well as systematic qualitative analysis and manual validation [50–52]. Triangulating the methodology allows for increased coverage of the literature reviewed while providing for increased assurance of analytical integrity and credibility in addition to conceptual consistency.

In summary, this study's integrated methodology enhances the validity and reliability of its findings by combining quantitative mapping, qualitative thematic interpretation, and cross-validation. This integrative approach strengthens the robustness of the findings and enables a holistic understanding of blockchain's role in banking. It also provided a solid foundation for identifying future research directions in this evolving field.

3. Results and Discussion

3.1. General information and performance analysis

The bibliometric analysis revealed 389 documents, published in 269 sources between 2015 and May 2025, authored or co-authored by 1,077 scholars. The principal purpose of collecting this bibliographic dataset is to provide an overall picture of the scientific literature that addresses the application of blockchain in banking during this period. This overview not only identifies key publication patterns but also brings an understanding of the evolution of the field. Such mapping is essential for understanding the development of the topic, as it helps to identify publication patterns, collaborative networks, and the most active research domains. Moreover, it underscores the academic relevance of the dataset and provides a foundation for further analysis.

Table 1 presents the descriptive statistics summarizing the dataset. In addition, the results illustrate key aspects of research productivity and collaboration, such as annual publication trends (Table 2; Figure 2), top productive scientific journals publishing in the field (Table 3), top contributing authors (Table 4), and most active institutions (Table 5),leading countries in publication output (Table 6), and the highly cited documents (Table 7). These analyses collectively provide a detailed account of the scientific landscape and support the evaluation of scholarly performance in the field.

Table 1. Main Information of the Dataset

Description	Results
Retrieval Date	12 May 2025
Time-Span	2015–May 2025
Total Publications	389.00
Subject Area:	
Business, Management, and Accounting	
Economics, Econometrics, and Finance	
Social Sciences	
Document Type:	
Article	274.00
Conference Paper	84.00
Review	31.00
Number of Cited Publications	313.00
Number of Non-Cited Publications	76.00
Total Citations	9354.00
Average Citations per Publication	24.05
Average Citations per Cited Publication	29.89
Average Years from Publication	3.10
Average Citations per Year per Document	4.63
Sources (Journals, Books, etc.)	269.00
Affiliations	786.00
Countries	88.00
References	18437.00
Keywords Plus (ID)	1818.00
Author's Keywords (DE)	1131.00
Authors	1077.00
Publications per Author	0.36
Authors per Publication	2.77

3.1.1. Publication Trends Over Time

Table 2 and Figure 2 illustrate the publication trends over a year from 2016 to May 2025. The analysis comprises metrics such as total publications (TP), cumulative publications (CTP), total citations (TC), and average citations per publication (TC/CTP and TC/TCP). The data reveal three distinct phases in the evolution of the research field: (1) Early emergence and foundational impact (2016–2018), (2) Expansion and thematic diversification (2019–2021), and (3) Peak production with initial signs of saturation (2022–2025).

The initial phase (2016–2018) reflects the inception of academic activity, with four articles published in 2016 being cited 1,249 times (312.25 per article), indicating foundational significance. The number of publications increased from eight in 2017 to 20 in 2018, reflecting growing interest in the potential of blockchain technology in the banking sector.

In the second phase, between 2019 and 2021, production increased sharply, from 30 in 2019 to 47 papers in 2020, though decreasing slightly to 41 papers in 2021. Despite the growth being notable, average citations declined (TC/CTP fell from 8.85 in 2019 to 5.99 in 2021), most likely due to higher participation and decline of the novelty. This is the stage that points towards the diversification of research themes and the decline in the productivity of 2021, possibly impacted by global disruptions such as the COVID-19 pandemic.

The third phase (2022–2025) represents the most productive period in terms of publication volume, with annual outputs increasing from 54 in 2022 to a peak of 78 in 2024. Publications during this phase constitute over half of the total output, highlighting the area's rapid expansion and highest level of publication activity. Although the TC/CTP ratio fell from 6.53 in 2022 to 1.17 in 2024, this decline

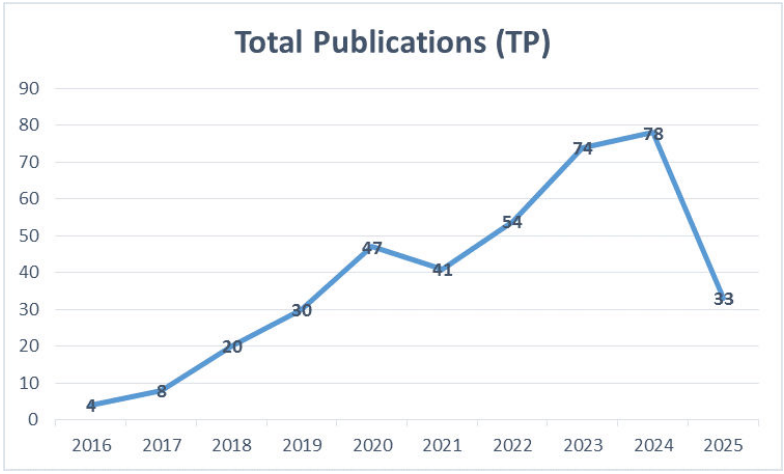


Figure 2. Total Publications (TP) over time (2016–2025). *Note: The data for 2025 (33 publications) is incomplete, reflecting the data cutoff date of May 12, 2025, and therefore does not accurately represent a decline in annual output.*

Table 2. Publication Trends Over Time

Year	TP	PTP	CTP	TCP	TC	TC/CTP	TC/TCP
2016	4.00	1.00%	4.00	4.00	1249.00	312.25	312.25
2017	8.00	2.00%	12.00	8.00	850.00	70.83	106.25
2018	20.00	5.00%	32.00	19.00	1033.00	32.28	54.37
2019	30.00	12.00%	62.00	28.00	549.00	8.85	19.61
2020	47.00	12.00%	109.00	44.00	2321.00	21.29	52.75
2021	41.00	11.00%	150.00	38.00	899.00	5.99	23.66
2022	54.00	14.00%	204.00	52.00	1333.00	6.53	25.63
2023	74.00	19.00%	278.00	56.00	656.00	2.36	11.71
2024	78.00	20.00%	356.00	53.00	418.00	1.17	7.89
2025	33.00	8.00%	389.00	11.00	46.00	0.12	4.18

TP = Total Publications, PTP = Percentage of Total Publications, CTP = Cumulative Total Publications, TCP = Total Cited Publications, TC = Total Citations.

is largely attributable to the recency effect, as newer articles have not yet accumulated significant citations.

The data for 2025 is partial and represents an artifact of the data cutoff. As of May 12, 2025, only 33 publications were indexed at this time. So, the apparent decline in output for 2025 constitutes a methodological artifact rather than a substantive downturn. While this figure is expected to increase significantly by the end of the year, annual publication counts, rather than citation-based indicators, reflect a consistent rise in research activity. Moreover, the increasing diversification of research themes, particularly applied studies integrating blockchain with AI, IoT, and FinTech, is likely to influence future citation patterns.

Overall, while publication volumes have risen exponentially, falling citation metrics indicate the need for yet more innovative and theory-driven studies. Future studies need to undertake interdisciplinary, problem-based approaches to advance the practical uptake of blockchain in banking contexts.

3.1.2. Leading Scientific Journals Publishing Blockchain and Banking Research

The most impactful journals that publish research on block-chain technology within the banking sector are detailed in Table 3, which presents both productivity measures (Total number of publications) and impact indicators (Total citations, Average citations per article, Average publication year, and normalized citation metrics). These combined measures enable the evaluation of not only the quantity of output but also the intellectual impact of each journal within the rapidly changing research environment.

An important observation in Table 3 is that, while Technological Forecasting and Social Change and Sustainability (Switzerland) are at the forefront journals in terms of volume, each journal's scholarly impact varies significantly. Technological Forecasting and Social Change exhibits a notably superior citation profile (648 total citations; 92.57 citations per article), which underscores the journal's strong focus on technology adoption, innovation dissemination, and socio-economic changes, subjects that closely relate to block-chain research in the financial sector. Its wide interdisciplinary readership and emphasis on theory-driven forecasting likely enhance its visibility and citation across various fields. In comparison, although Sustainability frequently covers block-chain topics, its more practical and policy-oriented focus, often aimed at specific sustainability audiences, leads to lower average citation rates (18.86 per article), indicating a more localized rather than broad academic influence.

In contrast, Financial Innovation, despite having published only five articles, boasts the highest overall citations (922) and greatest average impact per article (184.40). This remarkable achievement illustrates that thematic relevance of a journal, rather than just the volume of publications, drives academic influence. The journal's concentrated focus on financial technologies, digital currencies, and banking change positions it as a primary outlet for significant theoretical and empirical contributions, making its articles particularly prominent and often cited across finance, economics, and policy research communities.

Table 3. Leading Scientific Journals Publishing Blockchain in Banking Research

Rank	Source	Documents	Citations	Avg. Citations	Avg. Year	Avg. Norm. Citations
1	Technological Forecasting and Social Change	7.00	648.00	92.57	2022.29	4.04
2	Sustainability (Switzerland)	7.00	132.00	18.86	2021.57	0.98
3	International Journal of Scientific and Technology Research	6.00	57.00	9.50	2019.67	0.21
4	Financial Innovation	5.00	922.00	184.40	2020.40	1.99
5	Technology Analysis and Strategic Management	4.00	101.00	25.25	2022.75	3.55
6	Journal of Risk and Financial Management	4.00	72.00	18.00	2022.75	1.67
7	IEEE Transactions on Engineering Management	4.00	204.00	51.00	2022.00	6.44
8	Frontiers in Blockchain	4.00	93.00	23.25	2021.00	2.01
9	New Economic Windows	3.00	543.00	181.00	2016.00	0.58
10	Journal of Money Laundering Control	3.00	116.00	38.67	2020.00	2.00
11	Journal of Financial Stability	3.00	83.00	27.67	2020.33	0.99
12	Fintech	3.00	85.00	28.33	2023.00	1.15

A similar trend is evident in New Economic Windows, which attained 543 citations with just three publications. Its early exploration of blockchain topics (with an average publication year of 2016) enabled its articles to gather citations over an extended period, demonstrating the benefits of early involvement in emerging research areas. These foundational studies often serve as essential reference points for subsequent scholarship.

On the other hand, journals like the International Journal of Scientific and Technology Research, while comparatively productive (six publications), exhibit limited citation impact (averaging 9.5 citations per article). This variance likely stems from the journal's broader technical audience and its less focused engagement with financial or banking communities, leading to reduced citation engagement within social science and finance-oriented research networks.

Normalized citation metrics further enhance impact evaluation by considering publication age. Journals such as IEEE Transactions on Engineering Management (6.44) and Technology Analysis and Strategic Management (3.55) show strong relative citation performance given their more recent publication schedules. Their heightened normalized impact emphasizes the increasing importance of management- and gover-nance-related perspectives in blockchain research, particularly at the crossroads of engineering innovation, organizational strategy, and transformation in the financial sector.

Overall, these trends suggest that scholarly influence in the realm of blockchain-banking research is more influenced by journal thematic alignment, multidisciplinary engagement, early positioning in specific topics, and theoretical focus rather than merely by publication frequency. Journals that contextualize blockchain within wider discussions on financial governance, innovation management, regulatory adjustment, and socio-economic change achieve greater citation visibility than journals that are technically oriented or narrowly focused on sustainability. This uneven distribution of influence indicates that the intellectual essence of the field is anchored in publications that connect financial theory, policy analysis, and studies of innovation rather than solely in technically driven or sustainability-centric journals.

3.1.3. The 10 Most Influential Authors

Table 4 shows the most prolific authors who have made the largest academic contributions to blockchain research in the banking sector. This evaluation considers their productivity, citation impact, normalized influence, and the strength of their collaborative networks. These metrics not only identify the most visible researchers but also show how intellectual leadership and collaboration patterns shape the field.

It can be seen that Devi, N. Chitra and Kumari, Anitha are the most prolific with three papers each and the same citation count of 105 and 35 average citations per paper. While both have the same normalized citation score (2.14), only Devi has a sizable total link strength (19), suggesting more robust collaboration networks. This suggests that Devi's influence goes beyond citation metrics to include a bridging function between various research teams, encouraging cross-pollination of ideas related to adoption, operational efficiency, and governance in blockchain.

In contrast, although Mbaidin, Hisham O. has the same number of publications of Devi and Kumari, he has lower citations and average citation per document wit 43 and 14.33 respectively. Moreover, the author is strongly linked (link strength: 23), suggesting broad collaborative activity in the field. This pattern highlights authors whose main contributions are in interdisciplinary collaboration and empirical research across multiple countries. This fosters methodological diversity but may not yet result in highly cited conceptual breakthroughs.

A different type of intellectual leadership is seen in authors like Ramzi El-Haddadeh, Nitham Hindi, Vishanth Weerakkody, and especially Uthayasankar Sivarajah. They achieve notable citation efficiency despite fewer publications. Each of them had two high-impact papers with over than 100 citations, an average of 55 citations per article, and 2.54 normalized scores, indicating influence and visibility. However, Uthayasankar Sivarajah has the highest citation average (109.5) and a 5.03 normalized citation score, showing exceptional scholarly impact with fewer papers. He particularly focuses on governance, data management, and digital transformation strategies within financial institutions. These authors help consolidate theory by presenting models that link blockchain adoption with organizational readiness and regulatory issues.

Emerging researchers like Gan, Qingqiu, and Lau, Raymond Yiu Keung, show strong normalized citation rates of 4.78 despite their recent publication activity, with an average publication year of 2024.5. Their rapid accumulation of citations highlights a growing second wave of leadership focused on algorithmic finance, data analytics, and the convergence of emerging fintech. This trend indicates a shift in the field from foundational theoretical work to application-oriented and interdisciplinary growth.

In summary, the author network structure illustrates a layered knowledge ecosystem that balances established theorists, network connectors, and rapidly advancing innovators. Leadership in this field is defined not just by the number of publications but also by the ability to present impactful conceptual frameworks, provide scalable empirical evidence, and foster new research initiatives through collaborative networks. This evolving profile of authorship shows the maturation of blockchain and banking research into a more unified yet methodologically diverse academic domain.

3.1.4. The Top 10 Most Productive Institutions

Table 5 presents the leading institutions that have contributed most to blockchain research in banking in terms of productivity, impact,

Table 4. The Most Influential Authors

Rank	Author	TP	TC	APY	ACPP	ANC	TLS
1	Devi, N. Chitra	3.00	105.00	2022.33	35.00	2.14	19.00
2	Kumari, Anitha	3.00	105.00	2022.33	35.00	2.14	0.00
3	Mbaidin, Hisham O.	3.00	43.00	2023.67	14.33	1.91	23.00
4	Choo, Kim-Kwang Raymond	2.00	63.00	2022.00	31.50	2.14	3.00
5	El-Haddadeh, Ramzi	2.00	110.00	2022.00	55.00	2.54	27.00
6	Gan, Qingqiu	2.00	37.00	2024.50	18.50	4.78	8.00
7	Hindi, Nitham	2.00	110.00	2022.00	55.00	2.54	10.00
8	Lau, Raymond Yiu Keung	2.00	37.00	2024.50	18.50	4.78	3.00
9	Sivarajah, Uthayasankar	2.00	219.00	2022.00	109.50	5.03	13.00
10	Weerakkody, Vishanth	2.00	110.00	2022.00	55.00	2.54	10.00

TP = Total Publications; TC = Total Citations; APY = Average Publication Year; ACPP = Average Citations Per Publication; ANC = Average Normalized Citations; TLS = Total Link Strength.

and other important indicators such as, citations, average publication year, average citations per document, and average normalized citations.

Foremost among them is the Department of Management Studies at the Indian Institute of Technology Delhi, with 3 papers that garnered 110 citations, achieving an average of 36.67 citations per paper and an average normalized citation score of 1.32. This reflects a high academic impact and research quality in the feild. Conversely, the Adnan Kassar School of Business at the Lebanese American University, despite being equally prolific with 3 papers, has a lower average citation (3.67) and normalized citation score (0.49), revealing a less widespread scholarly impact.

In addition, certain institutions such as Al Qasimia University, Mutah University, Abu Dhabi University, and independent institutions such as the Financial and Taxation Consultant, Jordan, both of which have 2 papers of low citation frequency (average number of citations per paper of 6) but relatively high normalized citation scores (1.12), showing greater engagement and increasing popularity over the last few years (average year of publication: 2024), were also taken into account.

Most prominently, Spiru Haret University of Romania, with only 2 publications, received 56 citations and the highest normalized citation score (3.23), indicating the influence of its work in the discipline. Similarly, Symbiosis Institute of Digital and Telecom Management achieved a moderate impact with 21 citations from 2 publications.

In general, the results show geographically widespread and institutionally varied research efforts. Productivity is spread across institutions, but citation impact is concentrated in a few, indicating the distinction between quantity and quality of scholarly production.

3.1.5. The Most Productive and Influential Countries

Table 6 illustrates the significant geographical variation in research contributions, citation impact, and other major indicators, such as average publication year, average citations, average normalized citations, and total link strength of blockchain research in the banking sector.

As shown in Table 6, India is the most prolific and productive country with 94 documents, but it is lower ranked in citation impact (average citations per paper with 17.33) and normalized citation score (1.28). This indicates that while it leads in quantity, the overall impact remains moderate.

In contrast, the United States, with 51 papers, has the highest total citations (2,847) and a high average citation score (55.82), thus indicating a high academic impact. Likewise, the United Kingdom, with a lower productivity of 25 publications, achieves the top average citations (64.44) and normalized citation score (2.42), reflecting high-quality and highly recognized research output.

China also demonstrates a balanced profile with 24 papers and an average citation of 47.79, showing a good compromise between productivity and impact. The United Arab Emirates shows emerging activity with 21 papers and a good normalized score (1.41), yet still a moderate average citation per document (12.19).

Other countries, such as Germany, Italy, and Malaysia are moderately impactful and productive. Jordan and Switzerland, in contrast, while producing smaller volumes of output (12 and 10 papers, respectively), stand at competitive normalized citation averages (1.15 and 0.82, respectively), indicating quite high-impact research. Surprisingly, Spain and the Russian Federation have lower normalized and average citation indicators, reflecting limited impact despite modest research production.

Overall, India produces the most research in quantity, but other countries like the UK, the US, and China have a greater scientific impact. These patterns show that there is a global contribution, but the quality and visibility of research in the field of blockchain in banking are uneven.

3.1.6. The Top 10 Most Cited Documents

As we stated above, the dataset is retrieved from the Scopus database, and as we know, the topic has been investigated in various contexts by authors from Business, Management and Accounting, Economics, Econometrics and Finance, and Social Sciences. The analysis of the top 10 most cited documents in blockchain and banking research identifies the seminal works that have influenced academic investigation and applied applications in this multidisciplinary research area. These documents span various areas of study, ranging from financial innovation to accounting, regulatory studies, and information systems. Citation counts indicate academic and intellectual interest, while more complex metrics, such as average citations per year and normalized citation score, provide a better indication of the significant documents and their comparative influence over time and across research fields [53].

In view of this, Table 7 presents the ten most highly cited documents in our research field, according to the Scopus database. It is noted that, nine of the ten most highly cited papers received more than 200 citations, even though most of them were published less than four years ago.

Leading the list is Guo and Liang [26] pioneering document entitled "Blockchain application and outlook in the banking industry," published in the Financial Innovation journal, with a total of 706 citations as the most cited document in finance. Its average annual citation rate of 78.44 indicates a consistently high impact since its publication, although its normalized citation score of 2.26 suggests that, despite its high number of citations, its performance compared to other publications in its field is more moderate. Nonetheless,

Table 5. The Most Influential Institutions

Rank	Institution	TP	TC	APY	ACPP	ANC
1	Adnan Kassar School of Business, Lebanese American University, Beirut, Lebanon	3.00	11.00	2023.67	3.67	0.49
2	Dept. of Management Studies, Indian Institute of Technology Delhi, New Delhi, India	3.00	110.00	2023.00	36.67	1.32
3	Al Qasimia University, United Arab Emirates	2.00	12.00	2024.00	6.00	1.12
4	Business Intelligence and Data Analytics Dept., Business School, Mutah University, Jordan	2.00	12.00	2024.00	6.00	1.12
5	Dept. of Economics, College of Economics and Management, Al Qasimia University, Sharjah, UAE	2.00	12.00	2024.00	6.00	1.12
6	Faculty of Economics, Kharazmi University, Tehran, Iran	2.00	13.00	2022.50	6.50	0.37
7	Faculty of IT, Abu Dhabi University, UAE	2.00	12.00	2024.00	6.00	1.12
8	Financial and Taxation Consultant, Jordan	2.00	12.00	2024.00	6.00	1.12
9	Spiru Haret University, Romania	2.00	56.00	2023.50	28.00	3.23
10	Symbiosis Institute of Digital and Telecom Mgmt., Symbiosis Intl. (Deemed Univ.), Pune, India	2.00	21.00	2022.00	10.50	0.43

TP = Total Publications; TC = Total Citations; APY = Average Publication Year; ACPP = Average Citations Per Publication; ANC = Average Normalized Citations.

Table 6. The Most Productive Countries

Rank	Country	TP	TC	APY	ACPP	ANC	TLS
1	India	94.00	1629.00	2022.60	17.33	1.28	48.00
2	United States	51.00	2847.00	2021.35	55.82	1.62	31.00
3	United Kingdom	25.00	1611.00	2022.08	64.44	2.42	43.00
4	China	24.00	1147.00	2022.50	47.79	1.30	20.00
5	United Arab Emirates	21.00	256.00	2022.71	12.19	1.41	12.00
6	Italy	20.00	327.00	2021.70	16.35	0.90	16.00
7	Russian Federation	19.00	169.00	2019.58	8.89	0.29	0.00
8	Germany	18.00	740.00	2021.44	41.11	1.31	8.00
9	Malaysia	14.00	186.00	2022.36	13.29	0.97	19.00
10	Jordan	12.00	103.00	2023.75	8.58	1.15	18.00
11	Spain	11.00	145.00	2021.55	13.18	0.81	0.00
12	Indonesia	10.00	137.00	2022.30	13.70	0.33	2.00
13	Switzerland	10.00	280.00	2021.80	28.00	0.82	5.00

TP = Total Publications; TC = Total Citations; APY = Average Publication Year; ACPP = Average Citations Per Publication; ANC = Average Normalized Citations; TLS = Total Link Strength.

the work is still influential owing to its pioneering and general introduction of the revolutionary nature of blockchain for banking, specifically as it pertains to operational efficiency and transparency and transactional security.

On the contrary, Thakor's [54] article entitled "Fintech and banking: What do we know? ranks second in terms of total citations (601), but outperforms all other documents in terms of average annual citations (120.20) and the number of normalized citations (12.17). This suggests that the study has quickly become a leading reference in its field, although it has just 4 years since its publication. This suggests that the study is already a classic reference in the area. The Journal of Financial Intermediation presents a solid theoretical model on how fintech, including blockchain technology, is transforming long-established paradigms in banking. Its very high normalized citation score also indicates high influence and cross-disciplinary adoption, especially in finance, economics, and regulation studies in banking.

Its third most cited paper, authored by Dai and Vasarhelyi [55], entitled "Toward blockchain-based accounting and assurance," published in the Journal of Information Systems, has been cited 532 times. It has a high average of 66.5 yearly citations and a normalized score of 5.01, attesting to its contributory quality as a connecting publication between accounting theory and blockchain technology. It offers research that informs discussion about the use of blockchain to enable auditability and trust in financial reports, and is thus a reference work on the research of financial assurance with blockchain-based.

An equally significant contribution is made by Peters and Panayi [56], entitled "Understanding modern banking ledgers using blockchain technologies," cited 452 times. Its 50.22 times per year citation rate indicates ongoing interest by researchers, while its normalized citation of 1.45 indicates moderate impact in its broader research field. The significance of this work lies in its specific contribution to addressing distributed ledger technology and smart contracts, and offering insight into blockchain's technology foundation from a banking industry perspective.

Additionally, the International Journal of Information Management published research by Schuetz and Venkatesh [57] on using blockchain to drive financial inclusion in India. The article was cited 297 times with an average annual citation rate of 59.4 and a normalized citation rate of 6.01. This article is clearly very interdisciplinary in applicability. Its focus on social and developmental implications of blockchain makes it more relevant in policy development and financial inclusion policies, particularly in emerging economies.

With regard to infrastructure and security, although not banking-focused, Minoli and Occhiogrosso's [58] article entitled "Blockchain Mechanisms for IoT Security," has garnered 285 citations, an average annual citation of 40.71, and a normalized

Table 7. The Top 10 Most Cited Documents

Rank	Authors	Year	Title	Source	Document Type	TC	ACPY	NC
1	Ye Guo & Chen Liang	2016	Blockchain application and outlook in the banking industry	Financial Innovation, 2(1)	Original research article	706.00	78.44	2.26
2	Anjan V. Thakor	2020	Fintech and banking: What do we know?	Journal of Financial Intermediation, 41	Review article	601.00	120.20	12.17
3	Dai J.; Vasarhelyi M.A.	2017	Toward blockchain-based accounting and assurance	Journal of Information Systems, 31(3)	Conceptual research article	532.00	66.50	5.01
4	Gareth W. Peters & Efstathios Panayi	2016	Understanding Modern Banking Ledgers Through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money	New Economic Windows (NEW), pp. 239–278	Book chapter	452.00	50.22	1.45
5	Schuetz S.; Venkatesh V.	2020	Blockchain, adoption, and financial inclusion in India: Research opportunities	International Journal of Information Management, 52	Original research article	297.00	59.40	6.01
6	Daniel Minoli & Benedict Occhiogrosso	2018	Blockchain mechanisms for IoT security	Internet of Things (Netherlands), 1–2, 1–13	Original research article	285.00	40.71	5.52
7	Zetzsche D.A.; Arner D.W.; Buckley R.P.	2020	Decentralized Finance	Journal of Financial Regulation, 6(2), 172–203	Conceptual/policy article	264.00	52.80	5.35
8	Poonam Garg et al.	2021	Measuring the perceived benefits of implementing blockchain technology in the banking sector	Technological Forecasting and Social Change, 163	Empirical research article	218.00	54.50	9.94
9	Saurabh Ahluwalia et al.	2020	Blockchain technology and startup financing: A transaction cost economics perspective	Technological Forecasting and Social Change, 151	Empirical research article	212.00	42.40	4.29
10	Mohd Javaid et al.	2022	A review of Blockchain Technology applications for financial services	BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2(3)	Review article	207.00	69.00	8.39

TC = Total Citations; ACPY = Average Citations per Year; NC = Normalized Citations.

score of 5.52. Its interdisciplinary contribution comes in the form of providing data transmission protocols that are secure, something that would be essential to highly technologically advanced banking systems that are based on Internet-of-Things (IoT) incorporation.

Furthermore, regulatory aspects of blockchain are analyzed in the most highly-cited paper by Zetzsche, Arner, and Buckley [59], entitled "Decentralized Finance," which has been cited 264 times. The article has a yearly average of 52.8 citations and a normalized citation of 5.35, and it illustrates increasing academic interest in legal and compliance matters of decentralized financial systems. Published in the Journal of Financial Regulation, it offers a critical framework for the examination of blockchain's legal and systemic issues and thus is extremely useful to researchers as well as policymakers.

Empirical understanding of blockchain adoption is presented in their article "Measuring the perceived benefits of implementing blockchain in the banking sector," which has been cited 218 times, by Garg et al. [60]. Interestingly, it has a high average citation rate of 54.5 per year and a significant normalized citation score of 9.94, which indicates high use and strong cross-field influence. Using structural equation modeling, the authors assign a numeric value to the benefits of blockchain, such as trust, transparency, and efficiency, and make this study highly applicable to banking professionals.

Parallel to this is the work of Ahluwalia, Mahto, and Guerrero [61] enhances the knowledge of blockchain technology within the entrepreneurial finance context through their empirical article titled "Blockchain and Startup Finance." The paper has been cited 212 times at an average rate of 42.4 citations per annum, besides a normalized citation count of 4.29. The article extends the use of blockchain from traditional banking institutions to its impact on startup and venture capital environments through the adoption of transaction cost economics as a conceptual building block. Rounding out the list is the most recent contribution by Javaid et al. [62], titled "A Review of Blockchain Applications in Financial Services," which accumulated 207 citations within a brief period. With an annual average of 69.0 citations and a normalized citation score of 8.39, the article's direct impact and growing importance are evident. The article summarizes the various applications of blockchain technology in financial services, reflecting the growing demand from academics and industry experts for comprehensive reviews amid the rapid development of the Fintech sector.

Taken together, the citation patterns observed suggest that the influence within the blockchain-banking literature is more linked to the capacity to relate technological advancements to broader institutional, accounting, regulatory, and socio-economic issues than to purely technological innovation. Works that receive a high number of citations bring together conceptual theorization (like fintech transformation), incorporate insights from multiple disciplines (such as accounting, law, and information systems), and present empirical evidence that tackles real-world banking issues, including trust, financial inclusion, compliance, and efficiency. This highlights that the most impactful articles in academia frame blockchain not merely as a technical tool, but as a driver for significant changes in banking ecosystems. As a result, the structure of citations indicates a mature field that is progressively focusing on governance frameworks, adoption processes, regulatory legitimacy, and organizational transformation instead of isolated demonstrations of technology.

3.2. Science Mapping

Science mapping examines the relationships among contributors in a research field. Particularly, it focuses on patterns of intellectual interaction and structural connections between key scholarly constituents, such as how sources, countries, institutions, authors, references, keywords, and publications relate to each other [46, 63, 64].

The present study uses a range of science mapping techniques, including co-authorship analysis, co-citation analysis, co-occurrence analysis, and bibliographic coupling analysis. These methods facilitate gaining in-depth knowledge about the evolution of the field, the collaborative patterns that characterize it, and the thematic structure that underpins it [46]. When paired with network visualization software such as VOSviewer, these methods illustrate the bibliometric and intellectual structure of the research landscape [41, 45], as outlined below.

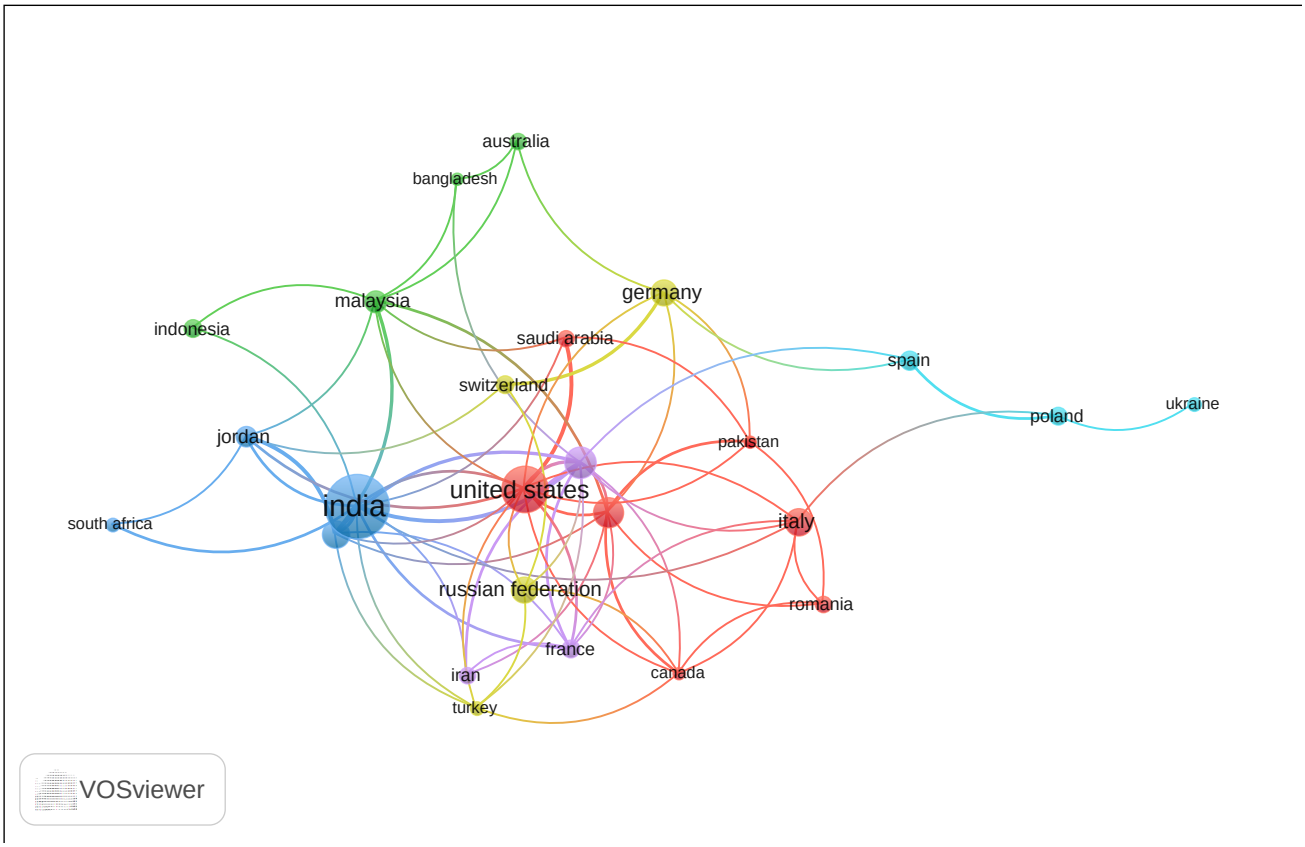


Figure 3. International Co-authorship Network of Countries in Blockchain and Banking Research. *Node size represents publication volume, link thickness indicates collaboration intensity, and colors denote distinct collaboration clusters.*

3.2.1. Co-authorship of Countries

Co-authorship analysis is a bibliometric technique that is employed to study patterns of collaboration among authors, institutions, and countries based on joint publications [65, 66]. At the national level, it reveals international research collaboration, mapping the global dispersion of scientific production and the transnational network structure [67, 68]. Particularly, the analysis reveals leading countries, maps geographical patterns of collaboration, and illustrates the effect of international networks on knowledge production [69, 70].

To explore global collaboration in blockchain research in the banking sector, we conducted a co-authorship analysis at the country level using VOSviewer. We included countries that had at least five documents and 30 citations. This led to 25 out of 88 countries meeting the criteria, with 72 links and a total link strength (TLS) of 105. As shown in Figure 3, the visualization displays six color-coded clusters, where nodes represent countries and links indicate the strength and frequency of co-authorships. Node size reflects publication volume, while link thickness shows collaboration intensity, and TLS quantifies a country's total collaborative strength.

The blue cluster, led by India, comprises the United Arab Emirates, Jordan, and South Africa, indicating close cooperation between South Asia and the Middle East. The central position and large node size of India highlight its high research productivity and its role as a regional leader in blockchain innovation. The participation of the United Arab Emirates and South Africa signifies an escalating level of interest in the financial applications of blockchain among digitally transforming economies.

The red cluster comprises the United States, China, Italy, Romania, Saudi Arabia, Pakistan, and Canada, forming a wide intercontinental network. The U.S. stands out for its high research volume and multiple collaborative ties. This cluster spans North America, Europe, the Middle East, and South Asia, indicating rich interdisciplinary exchanges. China and Italy are major contributors to the technological and regulatory aspects of blockchain, while Saudi Arabia and Pakistan can point to stronger academic connections with the West, possibly underpinned by digitization reforms and plans like the Vision 2030 of Saudi Arabia.

The yellow cluster includes the Russian Federation, Germany, Switzerland, and Turkey. Though geographically spread across Europe and Eurasia, these countries show strategic interest in digital finance and decentralization. Germany and Switzerland lead in fintech, while Russia and Turkey focus on modernizing financial systems, suggesting collaboration based on national strategies for digital transformation.

Moreover, the purple cluster consists of the United Kingdom, France, and Iran. The UK is the middle connection between the Middle East and Western Europe, showing high intra-European cooperation along with historical scholarly ties to the region. France and the UK are high-output researchers, while Iran shows up as a leading Middle Eastern producer of blockchain research. However, the light blue cluster includes Poland, Spain, and Ukraine. The nations, while not central, are reflective of increasing Eastern and Southern European engagement in blockchain research. Their inclusion is reflective of increased cross-border collaboration as well as a willingness to adopt blockchain towards economic modernization.

Overall, the findings of this analysis reveal a dispersed worldwide and interconnected research landscape. Developed and emerging economies are actively engaging with blockchain research in banking.

3.2.2. Co-citation of Authors

Co-citation analysis is a bibliometric technique that is applied to examine the intellectual landscape of a research area through analyzing how frequently two documents, authors, or sources are cited together in subsequent works [71]. A specific type of this analysis, Author Co-citation Analysis (ACA), examines how frequently two authors appear cited in tandem, therefore reflecting the conceptual structure underlying scholarly communication and conceptual evolution in an area [72, 73]. An increased frequency of co-citation between two authors implies a tight thematic correspondence or common influence on the shaping of specific streams of research [74].

In the current study, to better understand intellectual foundations and underlying blockchain research in the banking context, an author co-citation analysis was conducted using VOSviewer software. We applied a minimum threshold of 25 citations per author, resulting in the identification of 102 prominent authors out of a total of 25,779 who met the predefined criteria.

As shown in the network map in Figure 4, the authors were distributed to four distinct clusters, each represented by a different color. This network included 4,921 co-citation links and a total link strength of 56,680. The authors are shown as nodes within the clusters, while the edges illustrate how they have been co-cited. The sizes of the nodes indicate the extent of their co-citation. As a result, authors who are frequently co-cited appear as larger nodes. This pattern reveals a strong trend in scholarly relationships and co-citations, along with the overall growth in research for this field.

The red color is the first cluster in the network map. It is the largest and most central cluster and consists of authors like Chen Y., Chen S., Wang Y., Wang H., Liu J., Zhang Y., and Xu X. These authors have made major contributions in applying blockchain technology, digital technology, and information systems to banking and finance. They are most frequently cited in academic literature, i.e., they are the foundation of theoretical and empirical research on blockchain technology in the field. This cluster is also highly linked to other clusters, which indicates the intellectual power of the cluster over other fields.

In contrast, the second cluster, as can be shown by the blue color, includes prominent authors Kumar S., Khan S., Arner D.W., Zetzsche D.A., Thakor A.V., Kauffman R.J., and Hassan M.K. These authors are mainly involved with financial regulation, law, and policy matters related to blockchain technology. Their co-citation network indicates that they concentrate on the risk, governance, and legal concerns of blockchain implementation in banks. The uniqueness of the cluster indicates the interdisciplinary connection of information systems, law, and finance.

The third cluster, shown in green color, consists of authors such as Nakamoto S., Tapscott D., De Filippi P., Eyal I., Zhang Z., Hassani H., Janssen M., Potts J., and El-haddadeh R. They provide an all-round perspective of the revolutionary role of blockchain technology in banks. They examine cryptocurrencies, decentralization, governance, and innovation. Additionally, their co-citation suggests blockchain research covers a wide range of themes, from technical to legal, economic, and regulatory domains.

Finally, the fourth yellow color cluster comprises the following authors: Dwivedi Y.K., Kshetri N., Gupta S., Gunasekaran A., and Venkatesh V. This cluster also appears to be talking about information systems, models of technology adoption, and regulatory effects of blockchain technology. The cluster suggests that there is a widening of the research landscape on the implementation of blockchain technology in bank operations and business designs, with a concentration on technology adoption and strategic management.

In summary, these findings will be valuable to other researchers, IT professionals, financial service firms, practitioners, and banking professionals looking to consult with the right experts in related services.

3.2.3. Keyword Co-occurrence Analysis

Keyword co-occurrence analysis is a widely used bibliometric method that is employed to map and identify the intellectual structure and thematic evolution of a research field. It measures the frequency with which co-occurring pairs of keywords appear in the same papers, based on the assumption that higher co-occurrence indicates a stronger conceptual relationship between the terms [47]. This technique enables researchers to identify the primary research themes, evaluate the conceptual associations, and detect emerging topics in the literature [45, 75].

In the present study, we conducted a keyword co-occurrence analysis using VOSviewer software to gain a more profound understanding of the thematic context of blockchain technology in the banking sector. This approach has been demonstrated to be effective in identifying the leading research clusters and their connections. This is based on the frequency of using keywords and how they co-occur across publications' titles, abstracts, and keywords.

For this study, a minimum of five occurrences for an author-keyword was applied as an inclusion criterion. This was used to ensure an analytical focus on the most relevant and frequently occurring terms. Of the 1,131 keywords examined, 52 satisfied this initial criterion. In the second stage of our research protocol, we manually refined the dataset of the selected keywords by merging singular and plural terms, such as "cryptocurrency" and "cryptocurrencies," "smart contract" and "smart contracts," and "bank" and "banks." We also consolidated and standardized synonyms, including "distributed ledger" and "distributed ledger technology," "fintech" and "financial technology," "decentralized finance" and "DeFi," and "banking industry" and "banking sector." Furthermore, we eliminated keywords that were not related to our topic, such as "bibliometric analysis and COVID-19." Following the data refinement, the 42 keywords were included in the final analysis. Table 8 presents the most frequently occurring keywords and the data needed to ascertain areas related to blockchain research in banking. The 42 keywords yielded 289 links, with a total TLS of 799, and were organized into six distinct thematic clusters.

Table 8. Top Keywords by Occurrence

Rank	Keyword	Occurrences	TLS
1	blockchain	206.00	373.00
2	fintech	68.00	171.00
3	banking	49.00	125.00
4	blockchain technology	49.00	49.00
5	cryptocurrency	40.00	104.00
6	bitcoin	33.00	92.00
7	artificial intelligence	19.00	54.00
8	financial inclusion	16.00	39.00
9	smart contracts	16.00	30.00
10	finance	14.00	39.00
11	digital banking	13.00	27.00
12	financial services	13.00	36.00
13	innovation	13.00	40.00
14	security	13.00	31.00

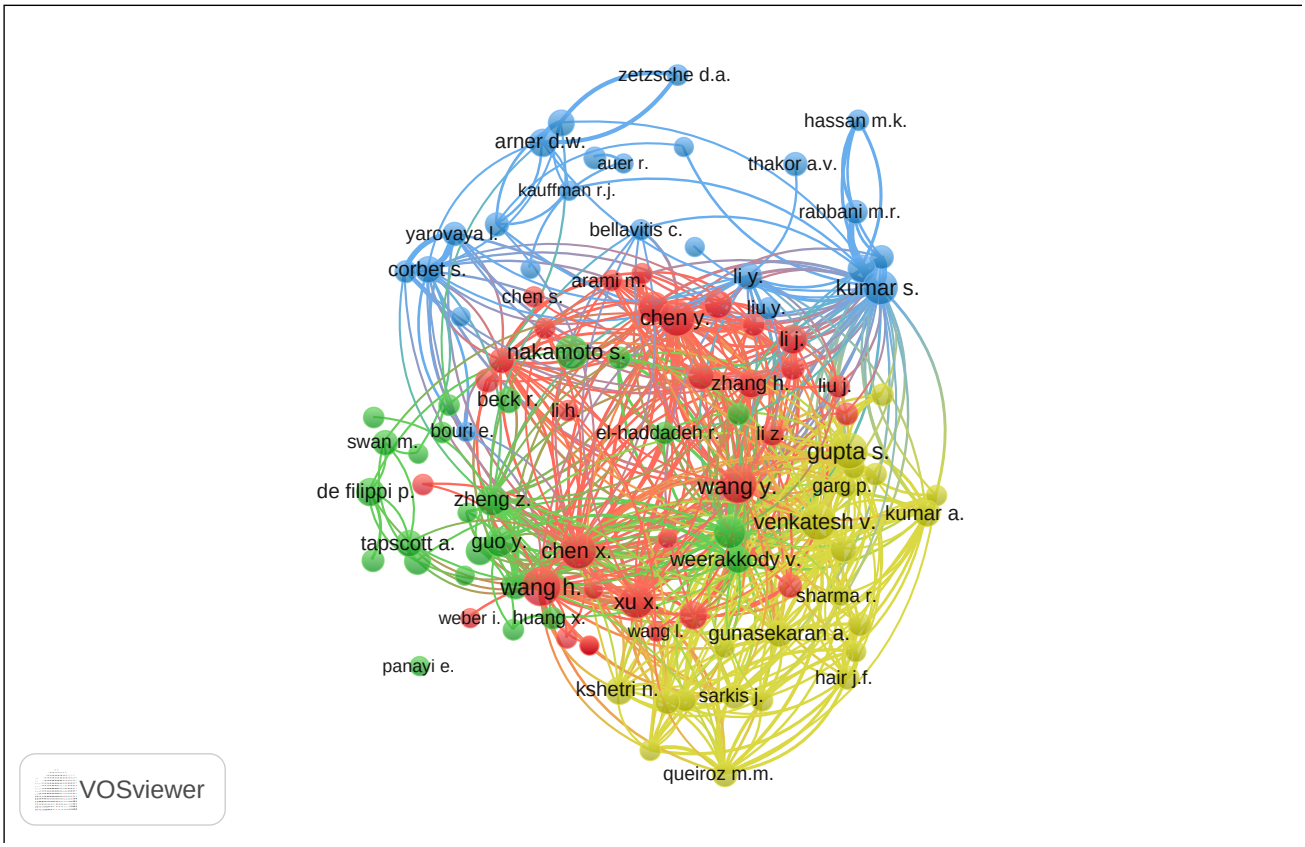


Figure 4. Author Co-citation Network in Blockchain and Banking Research. *Node size corresponds to citation influence, while links indicate co-citation strength. Colors denote major intellectual clusters.*

The network visualization produced (Figure 5) presents these clusters with each node representing a keyword, the node size representing frequency of occurrence, and lines (edges) representing co-occurrence relationships. The thickness of the lines is indicative of the strength of the relationship between terms, with thicker lines denoting a stronger relationship. The closeness of the lines to each other is also a helpful way to determine how related they are.

As shown in the network map, the keyword ”blockchain” is the most central node in terms of frequency of occurrence and interconnectivity. This is indicative of its central position in scientific discourse. Secondary keywords such as ”banking,” ”fintech,” ”cryptocurrency,” and ”Bitcoin,” which are also highly frequent and highly interconnected, emphasize blockchain’s central position in discourse regarding digital change in the finance and banking sector. The visualization (Figure 5) breaks down six distinct thematic clusters based on the following:

The initial cluster (blue) focuses on cryptocurrencies and decentralization, as evidenced by the terms ”blockchain,” ”Bitcoin,” ”cryptocurrencies,” ”decentralization,” ”Ethereum,” ”money,” and ”regulation.” The strong interconnection between these keywords and ”blockchain” indicates the inherent relationship of blockchain technology with digital currencies, particularly Bitcoin and Ethereum, which have always been of academic interest and a research topic in this field. Furthermore, the cluster groups critical words that define the world of cryptocurrency, since Bitcoin, Ethereum, and cryptocurrencies in general have a very close link with terms such as ”decentralization” and ”money.” It is clear that the literature in this cluster provides a comprehensive overview of the history and evolution of blockchain technology as applied to decentralized digital currencies. In addition to this, it provides a detailed discourse on the regulation of crypto assets, which is an inevitable consequence of the disruptive effect that these assets have on traditional financial institutions. This cluster reflects a wide range of studies on how blockchain technology can reshape the structure of money and payment systems, indicating sustained academic interest in decentralized money innovations.

Conversely, the second cluster (red) focuses on banking innovation and technology adoption. This cluster includes both emerging technology keywords, such as machine learning, artificial intelligence, big data, and the Internet of Things, as well as banking applications, including technology adoption, cybersecurity, sustainability, and digital banking. Together, these keywords encapsulate the technological infrastructure necessary to integrate blockchain technology into banking. Furthermore, this suggests that researchers are progressively interested in examining the combination of blockchain with other emerging technologies to re-engineer banking operations and service delivery. The emphasis on cybersecurity and sustainability indicates great concerns about the security and sustainability of innovation within financial institutions.

Similarly, the third cluster (in green) includes the keywords ”banking,” ”fintech,” ”finance,” ”financial services,” ”financial inclusion,” ”crowdfunding,” and ”peer-to-peer lending.” This indicates an awareness of blockchain technology’s macro-level ramifications on the augmentation of access to and efficiency of financial systems. The prevalence of the term ”fintech” in this cluster captures the essence of the transformation in financial intermediation, highlighting the pivotal role of blockchain technology in reengineering financial services. Additionally, the intersection of ”fintech” and

"financial inclusion" suggests a promising research area exploring blockchain's potential to address gaps in the banking sector.

Another notable cluster, marked in purple, focuses on trust-related issues and includes terms such as "trust," "transparency," "security," "privacy," "smart contracts," and "banking." The prevalence of these keywords indicates a persistent academic interest in the technological and ethical dimensions of blockchain technology. Specifically, the focus is on the potential impact of blockchain technology on trust, privacy, and security in banking and financial institutions. This thematic emphasis highlights blockchain technology's central role in addressing data integrity and user trust challenges, both of which are key to maximizing its value in banking applications.

The fifth cluster is represented by light blue and comprises keywords such as "digitization," "innovation," "digital transformation," "banking services," and "Islamic banking." These terms pertain to digital transformation and innovation in the banking sector. This thematic cluster indicates research trends that investigate the impact of blockchain technology on contemporary banking models with the aim of diversification and modernization.

The yellow cluster is particularly significant because it includes the keywords "distributed ledger technology," "decentralized finance," "financial regulation," "central bank digital currency," "cryptocurrencies," and "RegTech." These terms are poised to dominate future discourse concerning regulation and decentralized finance (DeFi). "Regtech" signifies the integration of regulatory control and compliance in blockchain-based banking. The cluster also highlights the pivotal role of policy and governance mechanisms in the adoption of blockchain technology in financial markets.

The network visualization of keyword co-occurrence in (Figure 5) led to the identification of six major clusters, confirming the thematic structure of the field. These clusters show the current research frontiers and common terms used by scholars. For the final synthesis and interpretation of these thematic clusters, please see Section 3.3.

3.2.4. Bibliographic Coupling of Documents

Bibliographic coupling is a bibliometric technique that measures the similarity between two documents based on their shared references. The extent of the overlap between references is indicative of the strength of the implied connection among the documents. This is because it is assumed that they are discussing the same topics or drawing on identical intellectual structures [48]. This technique is particularly useful for identifying stable research streams and the underlying intellectual structure of a research field.

The present study used VOSviewer to perform bibliographic coupling analysis and to visualize the intellectual structure of blockchain literature in the banking sector. Two documents are considered to be bibliographically coupled if they cite one or more of the common references. To enhance interpretability and focus on influential contributions, a minimum of 30 citations per document and a minimum cluster size of 10 documents were applied to be analytically significant. The application of this criterion resulted in the selection of 65 articles, which were subsequently organized into four clusters, each distinguished by a distinct color as shown in Figure 6.

In the resulting network visualization, each node represents an individual academic paper that has been used in the analysis. The size of a node is directly proportional to the number of citations it has received. The presence of larger nodes is indicative of a greater level of scientific influence. Lines linking nodes indicate bibliographic coupling relationships, while the thickness of the lines signifies the number of common citations between the two documents. The thickness of the line is indicative of the strength of the connection, with thicker lines denoting closer intellectual or thematic relationships. Moreover, the visualization map supports two important quantitative indicators. It produced 603 bibliographic links among the 65 documents that have demonstrated exceptional scholarly impact, as evidenced by their substantial citation counts. Additionally, the total link strength (TLS), calculated as the sum of all individual link strengths, is 1,226, reflecting high levels of connectivity and a comprehensive set of blockchain banking studies. The network map in this case provides valuable insight into thematic connectivity among highly cited articles. The clustering reflects how closely related the topics are and how references are linked between publications, which in turn highlights the main themes across the field.

Figure 6 demonstrates that Thakor's (2020) work exhibits considerable scholarly influence, characterized by its substantial node and cross-cluster edges, thereby establishing a significant connection between the domains of mainstream banking and blockchain literature. Dai (2017), Minoli (2018), and Schuetz (2020) have also been revealed to be central and highly connected nodes, forming a dense core within the red cluster. The significant overlap between these fields could potentially indicate an underlying contribution, particularly to blockchain technology and financial applications. In contrast, Auer (2022), Rehman (2023), and Kumar (2018) have focused their attention on peripheral areas, suggesting the existence of niches or novel research avenues that are less directly connected to the central literature. The peripheral nodes in this case reflect the growing bifurcation of topics such as DeFi and cryptocurrency regulation. As shown in Figure 6 6, the map visualization demonstrates the following clusters:

Cluster 1 (Red) is dominated by influential documents, including Dai (2017), Peters (2016), Schuetz (2020), Alhuwalia (2020), Hooper (2020), Shoaib (2020), and Cuccuru (2017). The cluster forms the theoretical basis of the field and focuses on blockchain technology infrastructure, settlement processes, transparency, auditability, and value creation within financial systems. This cluster constitutes a pivotal theoretical construct, establishing intricate internal relationships and exhibiting notable coupling strength.

Cluster 2 (Green), to which Thakor (2020), Minoli (2028), Chen (2017), Bayram (2022), Naimi-Sadigh (2022), Sangwan (2020), and Kimani (2020) belong, is characterized by its high level of interconnectedness and its tendency to explore blockchain convergence with FinTech innovation and financial inclusiveness for transforming banking services. This tendency is underpinned by a focus on empirical rationales and case-study findings.

Cluster 3 (Blue) consists of the following documents: Javaid (2022), Khalil (2022), Menon (2024), Elbashbishy (2022), Choo (2020), Rehma (2023), and Schlatt (2022). This cluster emphasizes digital transformation, service innovation, and customer-focused approaches to blockchain banking. The significant number of connections within this cluster indicates the presence of an emergent yet cohesive scholarly conversation.

Cluster 4 (yellow) is led by Garg (2021), Osmani (2021), Kumar (2018), Le Nguyen (2018), and Auer (2022). This cluster presents a network that also focuses on blockchain adoption models, consumer trust, and theories of innovation diffusion. This cluster is grounded in extant literature on the behavior and diffusion of innovation, thereby establishing a relationship at both technical and organizational levels.

The high interconnectivities among clusters emphasize the interdisciplinary nature of blockchain research in banking, due to the convergence of technology, economics, regulation, and behavioral perspectives. The prevalence of strong coupling relationships and numerous thematic avenues also suggests that, despite the fact that the field is still in its infancy, it has attained some level of maturity with well-defined but interrelated subfields.

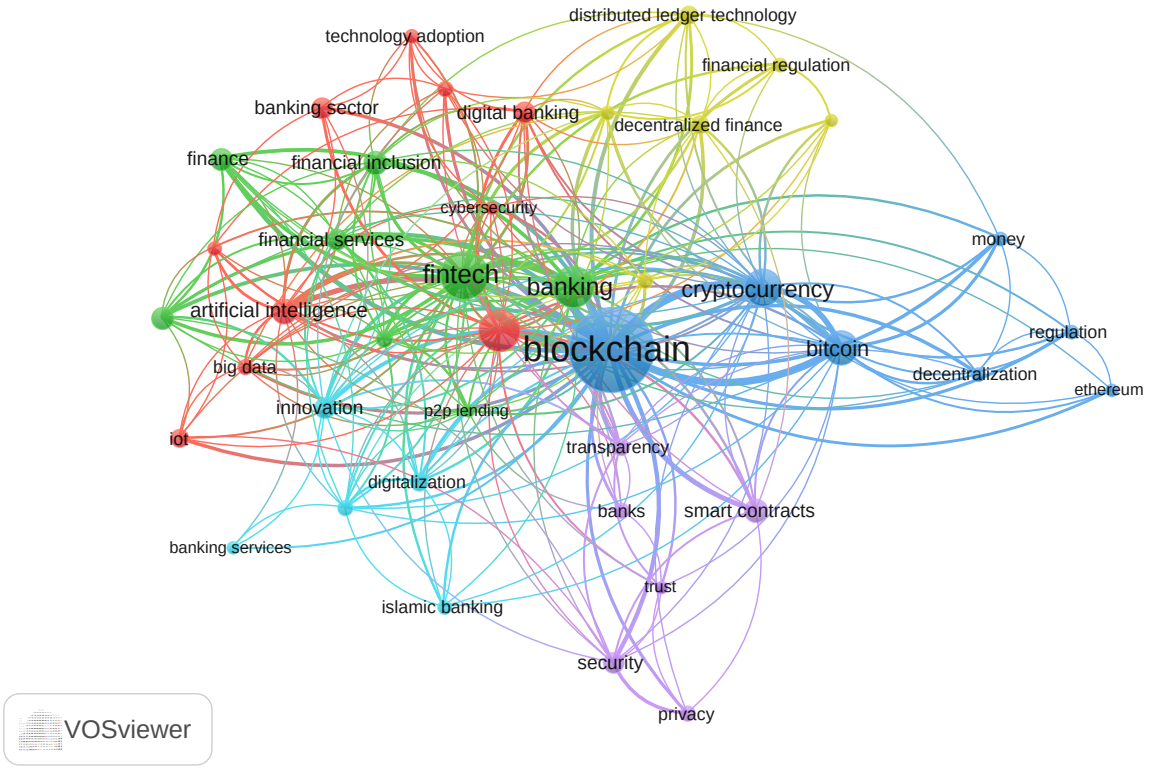


Figure 5. Keyword Co-occurrence Network of Blockchain and Banking Research. *Node size reflects keyword frequency, link strength indicates co-occurrence intensity, and clusters represent dominant thematic areas.*

3.3. Content Analysis and Thematic Clustering

In order to address Research Question 2 (RQ2), this section presents a qualitative thematic analysis of literature on blockchain technology in banking, which was systematically performed with the aim of identifying the main themes and providing a comprehensive understanding of the research landscape. Instead of repeating the bibliometric analyses provided in Section 3.2, this section builds on those quantitative findings to provide contextual interpretation, conceptual validation, and thematic coherence.

As outlined in the research methodology section, the initial identification of the major themes was derived using two methods: keyword co-occurrence and bibliographic coupling (as described in Sections 3.2.3 and 3.2.4). Keyword co-occurrence and bibliographic coupling highlight the closest relationships in terms of their relationship, or co-occurrence with each other, based on the frequency with which they were cited by authors and published in peer-reviewed journals [48, 76].

The previously described bibliometric research methods are helpful for creating a high-level map of the research domain. However, they do not provide a detailed explanation of the substantive content within the identified thematic clusters. To address this gap in the literature, we conducted a qualitative content analysis to synthesize, triangulate, and validate the mapped bibliometric thematic clusters.

To achieve this qualitative synthesis, we employed Braun and Clarke's thematic analysis framework [49]. First, we identified a dataset of 70 articles selected in previous sections of this paper. We then reviewed the articles manually to determine if the thematic clusters contained similar semantic meanings and if the cluster contents were conceptually consistent. Finally, we examined whether the thematic clusters contained relevant theories.

As illustrated in Table 9, this methodological approach yielded six robust thematic clusters that collectively define the intellectual structure of blockchain research in the banking sector from 2015 to May 2025. These thematic clusters represent the analytical framework through which to understand how blockchain influences financial intermediation, business processes, compliance with regulation, innovation strategy, trust creation, and integration with next-generation technology.

The following discussion focuses on the conceptual significance of these themes, providing a concentrated and analytical summary of the key intellectual trends in the field. This fulfills the mandate of the systematic content review phase.

3.3.1. Cluster 1: Blockchain Applications for Transforming Banking Operations and Financial Intermediation.

This cluster represents the most foundational and established body of literature on blockchains in banking, focusing on their ability to improve efficiency, smooth transaction frictions, or transform core banking systems. More conceptually, the literature in this stream is concerned with the idea that the value of blockchains is not found in isolated pilot projects but rather in their integration into back-office functions, interbank settlement mechanisms, and audit and compliance processes.

Foundational studies, such as [55, 56], establish the theoretical frameworks that describe how blockchain technology helps automate

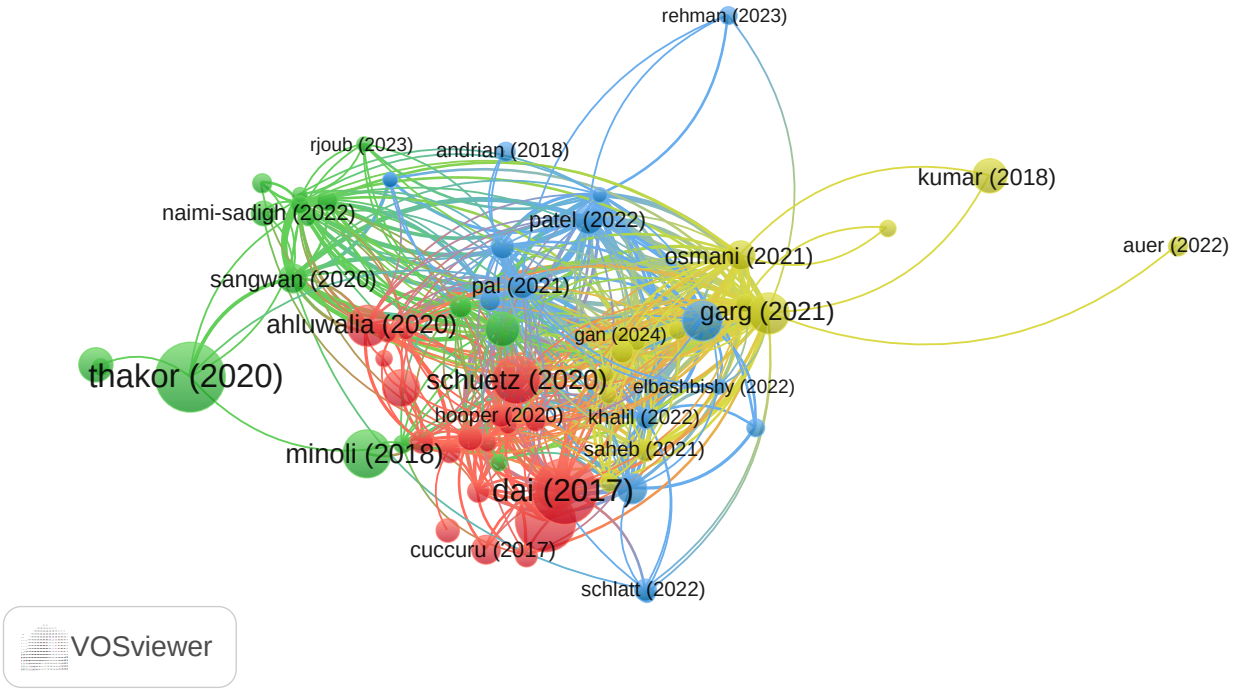


Figure 6. Document–Bibliographic Coupling in Blockchain and Banking Research. *Nodes represent documents, node size reflects citation influence, links indicate bibliographic coupling strength, and colors represent major intellectual and thematic clusters.*

Table 9. Identified Thematic Clusters in Blockchain Banking Research.

No.	Cluster Thematic	Main Themes	Sample References
1	Blockchain Applications for Transforming Banking Operations and Financial Intermediation	Real-time accounting, automation, reconciliation, operational efficiency, startup finance, and cost reduction.	[26], [55, 56],[61, 62],[77–80]
2	Decentralized Finance (DeFi) and Cryptocurrencies Enabled by Blockchain	DeFi, ICOs, remittances, financial decentralization, ethics of crypto, speculative behavior.	[81–87]
3	Blockchain as an Enabler of Digital and Financial Technology Convergence	Integration with IoT, AI, ML, FinTech, KYC, smart contracts, and digital ID; enhancing automation and inclusion.	[46], [54], [58], [88–92]
4	Trust-Related Dimensions in Blockchain-Based Banking	Trust, transparency, data privacy, organizational confidence, strategic alignment, adoption barriers.	[1], [60], [91], [93–96],
5	Regulatory, Legal, and Institutional Frameworks for Blockchain Governance	Smart contracts and law, compliance, anti-money laundering (AML), CBDCs, DeFi regulation, policy adaptation.	[59], [97–101]
6	Strategic Modernization of Banking Business Model Enabled by Blockchain	Disruption, competitive strategy, sandboxes, sustainable development.	[16], [37], [102]

banking ledgers, enhances settlement efficiency and reconciliation accuracy, and improves auditability. This technology also enables continuous quality assurance and real-time accounting systems. Building on this conceptual foundation, empirical evidence, notably from the Sponta Banca initiative, shows that blockchain frameworks significantly reduce settlement timeframes, enhance data traceability, and increase the reliability of interbank data exchange [77].

A large body of literature on this cluster, such as works by [26, 62, 80, 103, 104], consistently emphasizes the advantages of blockchain technology. Compared to traditional systems, blockchain technology enhances operational efficiency in terms of cost savings, risk mitigation, transaction security, transparency, and privacy. Also, this technology helps minimize information asymmetry and startup capital costs [61]. These benefits extend beyond payments to credit information systems, international settlements, and broader financial data networks. This reinforces the idea that blockchain technology is fundamental rather than limited in application.

Furthermore, this cluster emphasizes the strategic and organizational factors that facilitate successful blockchain implementation.

Research using technology adoption models [79] and innovation capability frameworks [105] identifies critical factors that mediate the operational effectiveness of blockchain technology, including trust, management commitment, and resource readiness. Furthermore, studies focusing on emerging markets [78, 106] indicate that banks' ability to achieve efficiency improvements is significantly affected by institutional maturity and technological infrastructure.

Overall, these findings underscore the importance of blockchain technology as a key tool capable of reducing operational costs, automating complex verification tasks, and promoting resilient financial systems with low response times. However, the studies also point to ongoing challenges, particularly with regard to scalability and institutional readiness, which continue to affect the speed and scope of practical implementation.

3.3.2. Cluster 2: Decentralized Finance (DeFi) and Cryptocurrencies Enabled by Blockchain

This thematic cluster focuses on an increasingly significant body of research that examines blockchain technology as the core infrastructure for decentralized finance (DeFi) and cryptocurrency-driven financial systems. Theoretically, research presents blockchain as a tool that eliminates intermediaries in conventional financial operations by allowing direct peer-to-peer value exchange, automating processes thr-ough smart contracts, and fostering transparent financial frameworks that function independently of central authorities. The core idea that emerges from this stream is that decentralized finance not only improves current banking processes but also radically challenges traditional models of financial intermediation.

Groundbreaking research indicates that decentralized finance offers an alternative financial structure capable of replicating essential banking activities or services, such as lending, borrowing, and asset trading, through decentralized protocols that are governed by code rather than traditional institutions [81]. This viewpoint is further reinforced by theoretical contributions that depict blockchain as a "trust protocol," highlighting its function in enabling transparency, immutability, and the automated execution of financial transactions [86]. Collectively, this body of literature lays the theoretical groundwork for comprehending how decentralized systems challenge conventional banking frameworks.

Furthermore, empirical and analytical studies within this thematic cluster reveal a more nuanced and diverse landscape. While blockchain-based money transfer systems and tokenized financial instruments show the potential to reduce costs and increase efficiency, evidence suggests that adoption of cryptocurrencies is often driven by speculative behavior rather than dissatisfaction with traditional banking services [85, 87]. Furthermore, research highlights persistent concerns about market volatility, governance ambiguity, and regulatory uncertainty, which continue to shape the risk profile of decentralized finance (DeFi) systems [82–84].

In addition to technical and economic factors, this research highlights the ethical, behavioral, and institutional consequences of DeFi. Studies focusing on accountability, financial inclusion, and ethical responsibilities caution that the decentralization of financial authority introduces new challenges associated with consumer protection, systemic risk, and regulatory supervision [82, 83, 107]. These findings imply that the transformative capacity of DeFi is closely connected to governance and public policy factors.

In summary, these studies affirm that decentralized finance (DeFi) and cryptocurrencies signify a groundbreaking extension of blockchain technology with the potential to transform financial intermediation. However, the research notes that the long-term viability of DeFi and its integration into mainstream banking systems depends on establishing regulatory frameworks, governance structures, and empirical evaluations of systemic risks.

3.3.3. Cluster 3: Blockchain as an Enabler of Digital and Financial Technology Convergence

This thematic cluster includes studies that look at blockchain as a fundamental infrastructure that supports and enhances the functionality of other emerging digital and financial technologies, such as artificial intelligence (AI), machine learning (ML), the Internet of Things (IoT), FinTech platforms, smart contract applications, and digital identity systems. Conceptually, the literature in this stream portrays blockchain not as an isolated solution but as a coordination and trust layer that improves interoperability, automation, and data integrity across complex digital ecosystems.

Key contributions within this cluster highlight the potential of blockchain to reshape financial value chains by facilitating decentralized data exchange, automated decision-making, and secure identity management [54]. In addition to financial services, this stream also highlights the role of blockchain technology in securing Internet of Things (IoT) systems by preventing data manipulation and enabling decentralized control, particularly in environments that require high levels of reliability and trust [58]. These studies portray blockchain as a complementary technology that enhances the reliability and transparency of data-driven financial services while paving the way for innovative digital intermediation. In this context, blockchain technology enables the secure integration of diverse technologies that typically operate in isolated locations.

Empirical studies further indicate that the convergence of blockchain with AI, big data analytics, cloud computing, and mobile banking technologies can lead to significant performance enhancements in the delivery of financial services, especially in lending, risk assessment, and customer onboarding processes [89, 90, 92, 107]. Evidence from banking applications points out that such technological convergence improves predictive accuracy, operational scalability, and financial inclusion, particularly for small and medium-sized enterprises and underrepresented populations.

A notable subtheme within this cluster focuses on digital identity management and automated compliance processes. Research on blockchain-based self-sovereign identity and smart contract–enabled Know Your Customer (KYC) processes shows significant advances in privacy protection, cost-effectiveness, and regulatory compliance [91]. Similarly, research on block-chain-enabled access control mechanisms highlights its potential to improve data management and security across interconnected digital platforms [88]. These applications demonstrate blockchain's potential to address long-standing inefficiencies in identity verification and data governance within financial institutions.

In summary, this cluster emphasizes the importance of blockchain as a key driver of technological convergence in digital finance. However, the literature also indicates ongoing challenges related to the system's compatibility, organizational coordination, and institutional compatibility. However, the literature also emphasizes ongoing challenges concerning system interoperability, regulatory harmonization, and compatibility. These limitations imply that the advantages of blockchain integration hinge on supportive institutional frameworks and the maturity of related technologies.

3.3.4. Cluster 4: Trust-Related Dimensions in Blockchain-Based Banking

This cluster synthesizes literature examining the impact of trust on the use of blockchain technology in banking. Cryptographic

verification and decentralized consensus have often led to characterizing blockchain as a "trustless" technology; however, existing studies continually highlight the importance of trust between organizations, user confidence, and the legitimacy of institutions when implementing blockchain within the financial sector. From a conceptual framework, research within this stream illustrates that whilst blockchain does not remove trust, it re-establishes it, moving it from centralized intermediaries to the technology itself, to governance structures, and to the institutions.

Empirical research indicates that, for both users and banks, perceived usefulness, transparency, and security are strong motivators for the adoption of blockchain technology, while technical capability is not as significant [1, 60, 93]. These results indicate that both types of trust (in technology and in the institution) interact with each other rather than exist separately.

A second theme in this cluster discusses blockchain's effects on increasing transparency and providing customers with greater data integrity and privacy. The research has indicated that the implementation of blockchain-based architectures can decrease the level of information asymmetry between lenders and borrowers, provide an increased level of auditing capabilities, and create greater levels of confidence in financial transactions through various regulatory processes, including lending [95, 96]. However, the literature points out that certain organizational barriers to establishing trust exist within financial industries, such as resistance to changing current ways of distributing credit and the lack of standardization, and the uncertainty regarding accountability, i.e., which party or parties are ultimately responsible in any given transaction [94].

Another prominent sub-theme is connected to digital identity and the privacy-preserving elements of the associated trust mechanism. Research examining digital identity management through blockchain and the application of KYC frameworks has illustrated that decentralized identity models will augment user control over their personal data and enable compliance with regulatory KYC requirements, while also assisting banks and regulators in forming a greater degree of institutional trust [91]. Furthermore, current research has demonstrated that the manner in which a digital identity is constructed has a direct impact on the level of trust between banks, regulatory authorities, and their customers.

In summary, the literature supporting this stream clearly establishes trust as a multi-dimensional construct that acts as an intermediary factor in the adoption of blockchain technology in the financial industry. While blockchain technologies provide a structure to increase transparency and security, the literature confirms that accepting blockchain technology into an organization must reach the appropriate balance between the institution's expectations regarding the reliability of the technology, the organizational readiness to use the technology, the availability of clear laws and regulations related to the use of the technology, and the overall level of acceptance by society at large.

3.3.5. Cluster 5: Regulatory, Legal, and Institutional Frameworks for Blockchain Governance

This thematic cluster synthesizes research on the impact of regulations, laws, and institutional frameworks on the use and adoption of blockchain technology in financial institutions. In theory, and according to the literature in this stream, blockchain technology contributes to greater transparency, process automation, and increased efficiency. However, institutions are unable to fully leverage this potential due to the uncertainty surrounding the regulation of this technology and because the current limited regulatory and legal structures are unable to keep pace with the transformation brought about by blockchain technology.

This theme focuses primarily on how enforcement and governance issues related to blockchain applications and the use of smart contracts are evolving. Many researchers point to numerous areas where the mechanisms for creating automatically enforced records conflict, as well as many unresolved issues related to accountability, jurisdiction, and the enforceability of programming-based agreements [97]. These challenges illustrate the difficulty of applying standard regulatory structures to decentralized financial systems.

Other important areas in this thematic cluster are compliance and risks that may threaten the integrity of the financial system and the systemic risks of blockchain technology. The use of blockchain technology for pseudonyms in financial transactions poses a potential dilemma for financial regulators [73] [99, 74]. The ease of creating anonymous accounts gives users easier access to money laundering [98]. At the same time, technologies enabled by blockchain, such as automated reporting, automated audit trails, and early warning systems, enhance transparency and regulatory effectiveness [100].

Additionally, studies indicate that regulatory approaches are necessary to support blockchain technology innovations. Regulatory sandboxes serve as tools for managing the relationship between innovation and risk through controlled testing, contributing to opportunities for learning, public policy development, and institutional adaptation [26]. Researchers are also exploring ways to integrate regulation into decentralized finance (DeFi) applications, emphasizing the need to incorporate governance and compliance mechanisms into system design as a means of mitigating the risks associated with decentralization [59].

Finally, studies on central bank digital currencies (CBDCs) show how the introduction of blockchain technology has prompted public authorities to develop hybrid governance models. Evidence from digital currency projects shows central banks' efforts to combine technological advances with centralized oversight to achieve financial stability objectives and ensure the effective transmission of monetary policy [101].

In conclusion, this research corpus reinforces the three essential ingredients for the long-term success of blockchain technology in the banking sector: regulatory clarity, institutional flexibility, and adaptive governance. In all three areas, the literature shows that sustainable governance of blockchain technology requires a balance between providing an environment conducive to innovation, ensuring legal certainty for consumers, and maintaining systemic financial stability.

3.3.6. Cluster 6. Strategic Modernization of Banking Business Model Enabled by Blockchain

This thematic cluster focuses on the role of blockchain as a mechanism for modernizing conventional business models in the banking industry, specifically as a type of strategic transformation. While much research refers to blockchain for its greater operational efficiency, the literature in this stream points out that blockchain's influence will create long-term economic and social governance systems by creating new biases toward competition and allowing for entirely new financial service architectures.

The literature in this cluster collectively conceptualizes blockchain technology as disruptive rather than complementary to existing systems and processes. This body of literature recognizes that blockchain platforms disrupt the traditional centralized structure of the banking industry by enabling the delivery of new services to customers and allowing peer-to-peer interactions to create value without going through a bank or intermediary. Consequently, banks are under increased pressure to reevaluate their strategic positioning, organizational structure, and competitive response to their evolving roles in the digital financial service environment [16, 37].

In addition, this cluster of research has further explored how business model innovation driven by blockchain technology can support financial inclusion and global sustainable development.

Studies have identified ways in which blockchain can provide expanded access to financial services, decrease transaction costs, and improve transparency in areas such as payments, savings, credit, and insurance, particularly inunderserved areas and regions [102]. However, the literature of this cluster has also emphasized that, for these potential strategic benefits to be realized, a supporting institutional framework is necessary.

In general, this cluster validates the assertion that blockchain is a strategic enabler of banking modernization, encompassing more than incremental process improvements. That said, the findings also indicate that the effect of blockchain on banking ultimately depends on how well financial institutions use and integrate the new technology into their organizational strategies and adapt to achieve organizational compliance and advance organizational goals in a changing economy and broader social structure.

4. Research Implications

4.1. Theoretical Implications

This review enhances blockchain adoption theory by broadening primarily individual-level acceptance models (e.g., TAM, UTAUT) and organization-centered readiness viewpoints (e.g., TOE, RBV) into a multi-tiered, ecosystem-based comprehension of blockchain dissemination in tightly regulated financial contexts. The bibliometric clustering demonstrates that blockchain adoption in the banking sector is influenced not only by technological preparedness or perceived value but also by the interplay of regulatory legitimacy, institutional trust, cross-organizational interoperability, and strategic resource management throughout financial networks. This observation refines traditional technology adoption models by highlighting that disruptive financial technologies face diffusion constraints imposed by governance frameworks and regulatory compliance demands, resulting in adoption pathways that are fundamentally different from those seen in cons-umer-oriented digital technologies.

Additionally, the thematic evolution indicates a theoretical shift within the literature from initial techno-optimistic narratives to analytical perspectives that focus on institutional, risk-oriented, and governance issues. This progression marks a shift from exploratory research on technology diffusion to integrated frameworks that regard blockchain as a facilitator of organizational transformation rather than simply a discrete operational tool. Therefore, this review presents a cohesive conceptual framework that incorporates technological, organizational, regulatory, and ecosystem dynamics into a comprehensive explanatory model for blockchain-driven financial innovation.

By synthesizing bibliometric findings with qualitative thematic analysis, this study presents an established, multi-level framework that describes the process by which the banking sector adopts blockchain technology as a broader ecosystemic and governance-driven process rather than a technology-driven phenomenon.

4.2. Managerial Implications

In addition to outlining technological advantages, the current findings suggest a strategic rethinking of blockchain as a tool for organizational transformation rather than a mere digital upgrade. By synthesizing insights from bibliometric and thematic clusters, this research shows that the success of adoption is more dependent on banks' capacity to implement coordinated process reengineering, cross-unit integration, and alignment of institutional governance than on technical installation.

The thematic clusters that highlight operational efficiency, cost savings, and process automation imply that blockchain should be viewed not just as a technological asset but also as a driver of operational reorganization. Therefore, banking managers are urged to re-evaluate current workflows and identify areas where distributed ledger technologies can optimize accounting processes, enhance reconciliation accuracy, and decrease overhead expenses through smart contract automation [55, 62].

Moreover, the findings stress the growing importance of security, transparency, and trust in modern banking practices. With increasing cyber threats and regulatory compliance demands, blockchain-based systems provide solutions for ensuring data integrity, tracing audit trails, and automating contract enforcement. These features are particularly pertinent to Know Your Customer (KYC) and Anti-Money Laundering (AML) compliance frameworks, where blockchain applications can support regulatory adherence while simultaneously enhancing institutional credibility [98, 103].

Similarly, the rise of decentralized finance (DeFi) and token-based ecosystems indicates a fundamental shift in banking business models. As a result, managers must look beyond incremental enhancements to investigate new service architectures, such as peer-to-peer intermediation platforms, blockchain-enabled payment systems, and digital asset tokenization. This shift requires innovation-driven leadership cultures, investment in blockchain-related expertise, and strategic alliances with fintech developers to maintain a competitive advantage.

Lastly, the noted decrease in citation impact alongside increasing publication volumes highlights the need for more practically oriented blockchain initiatives. Banking leaders must connect blockchain adoption to clearly defined institutional objectives, quantifiable performance metrics, and stepwise implementation strategies to ensure that investments yield tangible benefits rather than remaining symbolic or experimental.

In summary, these managerial implications illustrate that the adoption of blockchain is primarily a challenge of leadership, governance, and change management, rather than solely a decision related to technological procurement.

4.3. Practical Implications

From a practical viewpoint, this review indicates that blockchain technology generates its most significant benefits when it is integrated within regulatory and transactional frameworks rather than operated as a standalone pilot initiative. The most pronounced empirical focus in the literature pertains to cross-border settlements and interbank transaction clearing, where inefficiencies are still common. Incorporating blockchain into these areas has the ability to speed up settlement times, lower operational expenses, and reduce the risks of fraud [26, 56, 60].

Concurrently, blockchain provides capabilities for automating regulatory processes and managing identities. Smart contracts and decentralized identity systems can improve compliance precision and operational transparency, yielding considerable cost savings in fulfilling KYC, AML, and financial reporting requirements [80, 98]. Therefore, regulatory bodies and financial institutions are urged to consider RegTech-driven blockchain solutions not merely as additional controls but as comprehensive compliance frameworks.

The literature also highlights the inclusive potential of blockchain, especially via DeFi-enabled microfinance platforms, crowdfunding opportunities, and mobile-focused peer lending initiatives[59, 85, 108]. Such models create avenues for underserved communities to obtain financial services without reliance on traditional intermediaries. Implementation efforts should, therefore, prioritize areas with high rates of financial exclusion, particularly

in emerging and developing economies. For technology developers and consulting agencies, the insights point to key areas for development that include secure audit platforms, green finance traceability systems, decentralized asset management frameworks, and interoperable payment solutions. Collaborative design partnerships with financial institutions are essential to ensure that technological models closely correspond with sector-specific regulatory and operational needs.

Ultimately, the effective implementation of blockchain in the banking sector necessitates not only experimental adoption but also ongoing institutional coordination that encompasses regulatory dialogue, workforce education, governance adaptation, and strategic oversight. Thus, the full potential of blockchain is realized when technical advancements are aligned with organizational preparedness and policy coherence.

5. Conclusion and Future Research

5.1. Conclusion

This research comprehensively examined the evolving intellectual landscape and thematic development of blockchain studies within the banking industry from 2015 to 2025 using a hybrid approach that combines bibliometric analysis with qualitative systematic synthesis. The analysis of 389 peer-reviewed articles highlighted distinct developmental stages—from initial conceptual exploration to thematic broadening and into the current phase of applied governance and integration studies.

An analysis of geographic contributions revealed disparities, with the majority coming from India, the United States, and the United Kingdom, while newer research centers in China, the United Arab Emirates, and various parts of Europe are progressively influencing the empirical direction of the field. At the levels of institutions and authorship, research networks show both fragmentation and cross-regional collaboration, indicating that global integration in research is inconsistent.

Six key thematic clusters delineate the structure of disciplinary knowledge: financial intermediation and operational efficiency, decentralized finance (DeFi) and cryptocurrencies, convergence of blockchain technology, infrastructures for trust and transparency, regulatory and governance frameworks, and modernization strategies in banking. Together, these aspects characterize blockchain as not just a standalone technological fix but as an integrated transformation platform that concurrently impacts organizational frameworks, regulatory systems, and financial ecosystems.

Although the volume of publications is on the rise, the literature remains empirically scattered. Studies focusing on large-scale industry adoption are limited, the interactions between blockchain and complementary technologies (such as AI and IoT) are insufficiently theorized, and long-term evaluations of financial stability and systemic risk are scarce. Governance research, especially in areas of regulatory enforcement and international coordination, is also still underexplored.

In addition to mapping thematic growth, this review offers an integrative theoretical framework based on our synthesis of the six thematic clusters. This framework improves our understanding of blockchain adoption by presenting it as an innovation process influenced by regulatory legitimacy, organizational governance, and ecosystem interoperability, rather than merely a technical event. This integrative view distinguishes the current review from previous bibliometric analyses because it clearly articulates the causal relationships connecting our validated knowledge structure to the broader agenda of organizational transformation, regulatory alignment, and strategic value creation in finance.

As a result, this study provides a cohesive theoretical groundwork for future empirical research and offers practical insights for banking professionals and policymakers as they navigate the implementation of blockchain technologies in regulatory environments undergoing transition.

5.2. Future Research Directions

To improve our understanding of this research area, more research should be conducted on the six thematic clusters discussed earlier. Since research on blockchain technology in the banking sector is in its infancy, identifying and defining possible areas for future research is crucial. These research directions are derived from existing literature and reflect the gaps, constraints, and prospects identified by previous researchers.

Existing studies have identified that blockchain has the potential to transform operational processes in the banking sector for greater effectiveness, financial inclusion, and decentralized finance (DeFi), as well as to completely modernize business models [55, 81, 86]. However, serious issues remain regarding regulatory ambiguity [59], interoperability [26], adoption of trust [93], and integration into future-proof technologies [90]. Thus, future research must bridge these gaps through empirical, interdisciplinary, and cross-regional studies.

Table 10 shows directions reflecting both conceptual and practical priorities. These directions provide a research map for charting blockchain scholarship and positioning policymakers, financial institutions, and technology providers toward the development of secure, ethical, and scalable distributed ledger technology applications.

5.3. Limitations of the Study

Despite providing an overall bibliometric and thematic analysis, this study has some limitations that should be acknowledged. First, the dataset was derived exclusively from the Scopus database. Although Scopus provides the widest coverage of peer-reviewed journals related to finance, management, and information systems research, the exclusion of other databases (such as Web of Science, IEEE Xplore, and Google Scholar) may have resulted in the omission of some relevant publications, particularly conference proceedings and technically oriented studies. Nevertheless, this review focuses primarily on the social, economic, managerial, and organizational aspects of blockchain technology in the banking sector, rather than on the development of engineering or cryptographic systems, which are typically covered in technical databases.

Furthermore, the study examined 389 peer-reviewed articles from 2015 to May 2025. Due to Scopus's dynamic nature, the database used for the study might not include the newest publications at the cutoff time of the final submission, which could slightly affect the bibliometric results. The study only used VOSviewer to map and visualize bibliometric networks. Although VOSviewer is a popular tool, other tools, such as Gephi or CiteSpace, could have been used to provide additional bibliometric measures, including network centrality, modularity, and mediation scores.

Furthermore, this research did not propose a conceptual model for how banks adopt blockchain technology. Therefore, subsequent studies can build on this research to develop a more extensive model that encapsulates the multidimensionality of blockchain applications. Despite its limitations, the research provides a preliminary examination of the intellectual structure and thematic history of blockchain research in the banking sector.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

Funding

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

Declaration of competing interests

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 article. Moreover, they assert that no conflicts of interest exist.

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

During the preparation of this manuscript, the author(s) used language editing tools/services, including DeepL and Grammarly, to improve grammatical accuracy and readability. The author(s) subsequently reviewed and edited the contents thoroughly for accuracy and integrity after utilizing these tools/services and are fully responsible for the final version of the manuscript.

Data Availability Statement

The bibliometric dataset supporting the findings of this study, including the Scopus CSV file used for VOSviewer analyses, is publicly available on Zenodo at:

https://doi.org/10.5281/zenodo.17992285

Credit authorship contribution statement

[Sadeq Abdullah Aladeeb]: Conceptualization, Software, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft & Editing. [Fatima Zohra Sossi Alaoui]: Supervision, Validation, review & editing.

Table 10. Blockchain Themes and Future Research Directions in Banking.

No.	Cluster Theme	Future Research Directions	References
1	Blockchain Applications for Transforming Banking Operations and Financial Intermediation	• Carry out comparative empirical research assessing the effects of smart contracts on transaction settlement durations and operational costs in various banks. • Create process-mapping models to quantify the reduction of reconciliation steps in interbank clearing attributable to blockchain (utilizing Business Process Model and Notation "BPMN" and time–motion analysis). • Perform cross-country econometric evaluations to determine how blockchain-based remittance solutions impact transfer expenses and delivery times in developing compared to developed nations. • Employ UTAUT2 or TOE frameworks to pinpoint the factors influencing blockchain adoption in retail versus corporate banking sectors. • Employ UTAUT2 or TOE frameworks to pinpoint the factors influencing blockchain adoption in retail versus corporate banking sectors. • Conduct case studies in low-income nations to uncover obstacles to scalability, interoperability, and institutional integration.	[56],[61, 62], [77],[79], [103],[105]
2	Decentralized Finance (DeFi) and Cryptocurrencies Enabled by Blockchain	• Model contagion and systemic risks within DeFi ecosystems through network analytics and simulation methodologies (e.g., agent-based modeling). • Conduct studies on regulatory impacts, comparing the effectiveness of various legal frameworks in mitigating fraud and protecting consumers in DeFi lending platforms. • Conduct studies on regulatory impacts, comparing the effectiveness of various legal frameworks in mitigating fraud and protecting consumers in DeFi lending platforms • Evaluate the influence of DeFi credit markets on the liquidity, profitability, and risk parameters of commercial banks. • Carry out behavioral studies to examine how cultural differences shape motivations for adopting cryptocurrencies (speculation versus utility).	[81–83],[85–87]
3	Blockchain as an Enabler of Digital and Financial Technology Convergence	• Design and evaluate blockchain–IoT prototypes for real-time Know Your Customer (KYC) / Anti-Money Laundering (AML) monitoring within banking data streams. • Assess the effectiveness of AI-enhanced smart contracts in dynamic access control through penetration testing and cybersecurity evaluations. • Create machine-learning models using blockchain transaction data to forecast credit risk or fraud patterns, and validate using actual banking datasets. • Develop and assess (Self-Sovereign Identity) SSI-based identity frameworks in partnership with banks to gauge improvements in onboarding efficiency and KYC compliance.	[58], [88],[90],[91]
4	Trust-Related Dimensions in Blockchain-Based Banking	• mixed-methods surveys and interviews to evaluate the impact of human trust and organizational culture on blockchain adoption within banks. • Establish a standardization readiness index to evaluate how system compatibility, legacy systems, and regulations impede blockchain integration. • Design blockchain-based credit scoring prototypes and assess their effectiveness in diminishing information asymmetry in SME lending. • Implement longitudinal studies to track how increased transparency through blockchain influences customer trust over time.	[60],[93], [95],[96]
5	Regulatory, Legal, and Institutional Frameworks for Blockchain Governance	• Propose and evaluate blockchain-enabled AML/CFT (Countering the Financing of Terrorism) monitoring systems and measure their detection accuracy compared to traditional systems. • Examine the efficacy of regulatory sandboxes by monitoring innovation outputs (patents, pilots, startups) preceding and following sandbox involvement. • Develop automated reporting and cryptographic proof systems for embedded supervision models in DeFi. • Analyze real-world CBDC pilot projects (e.g., e-CNY) to gauge privacy risks, transaction speeds, and impacts on monetary policy using macro-financial models.	[26],[59],[97],[98],[101]
6	Strategic Modernization of Banking Business Model Enabled by Blockchain	• Employ scenario analysis to illustrate how blockchain influences competition between neobanks and traditional banks. • Perform studies on the effects of financial inclusion by evaluating blockchain-based microfinance initiatives in rural or underserved areas. • Chart out policy, infrastructure, and institutional elements that contribute to successful blockchain-driven transformation using the PESTEL (Political, Economic, Social, Technological, Environmental, and Legal) framework and multi-country case research.	[16],[37],[102]

References

1. A. Kumari and C. Devi. The impact of fintech and blockchain technologies on banking and financial services. *Technology Innovation Management Review*, 12(1/2):22010204, 2022. doi:10.22215/timreview/1481.
2. S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008. Social Science Research Network, Rochester, NY: 3440802. doi:10.2139/ssrn.3440802.
3. T. Philip, R. G. Brown, and Y. Danny. Blockchain technology in finance. *Computer*, 50(9):14–17, 2017. doi:10.1109/mc.2017.3571047.
4. S. Sarmah. Understanding blockchain technology. *American Journal of Computer Science and Technology*, 8(2):23–29, August 2018. doi:10.5923/j.computer.20180802.02.
5. A. T. Sherman, F. Javani, H. Zhang, and E. Golaszewski. On the origins and variations of blockchain technologies. *IEEE Security & Privacy*, 17(1):72–77, 2019. doi:10.1109/MSEC.2019.2893730.
6. A. A. Monrat, O. Schelén, and K. Andersson. A survey of blockchain from the perspectives of applications, challenges, and opportunities. *IEEE Access*, 7:117134–117151, 2019. doi:10.1109/ACCESS.2019.2936094.
7. J. Li and M. Kassem. Applications of distributed ledger technology (dlt) and blockchain-enabled smart contracts in construction. *Automation in Construction*, 132:103955, 2021. doi:10.1016/j.autcon.2021.103955.
8. M. S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli, and M. H. Rehmani. Applications of blockchains in the internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 21(2):1676–1717, 2019. doi:10.1109/COMST.2018.2886932.
9. M. H. Joo, Y. Nishikawa, and K. Dandapani. Cryptocurrency, a successful application of blockchain technology. *Managerial Finance*, 46(6):715–733, 2019. doi:10.1108/MF-09-2018-0451.
10. T. K. Mackey et al. 'fit-for-purpose?' – challenges and opportunities for applications of blockchain technology in the future of healthcare. *BMC Medicine*, 17(1):68, 2019. doi:10.1186/s12916-019-1296-7.
11. C. Laroiya, D. Saxena, and C. Komalavalli. Chapter 9 - applications of blockchain technology. In S. Krishnan, V. E. Balas, E. G. Julie, Y. H. Robinson, S. Balaji, and R. Kumar, editors, *Handbook of Research on Blockchain Technology*, pages 213–243. Academic Press, 2020. doi:10.1016/B978-0-12-819816-2.00009-5.
12. A. Pal, C. K. Tiwari, and A. Behl. Blockchain technology in financial services: a comprehensive review of the literature. *Journal of Global Operations and Strategic Sourcing*, 14(1):61–80, March 2021. doi:10.1108/JGOSS-07-2020-0039.
13. Niels Hackius and Moritz Petersen. Blockchain in logistics and supply chain: Trick or treat? In Wolfgang Kersten, Thorsten Blecker, and Christian M. Ringle, editors, *Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 23*, pages 3–18, Berlin, 2017. epubli GmbH. doi:10.15480/882.1444.
14. Marcin Hernes, Artur Rot, and Dorota Jelonek, editors. *Towards Industry 4.0 — Current Challenges in Information Systems*, volume 887 of *Studies in Computational Intelligence*. Springer International Publishing, Cham, Switzerland, 1st edition, 2020. doi:10.1007/978-3-030-40417-8.
15. L. Cocco, A. Pinna, and M. Marchesi. Banking on blockchain: Costs savings thanks to the blockchain technology. *Future Internet*, 9(3):25, September 2017. doi:10.3390/fi9030025.
16. W. L. Harris and J. Wonglimpiyarat. Blockchain platform and future bank competition. *Foresight*, 21(6):625–639, July 2019. doi:10.1108/FS-12-2018-0113.
17. H. S. Ali, F. Jia, Z. Lou, and J. Xie. Effect of blockchain technology initiatives on firms' market value. *Financial Innovation*, 9(1):1–35, 2023. doi:10.1186/s40854-023-00456-8.
18. S. M. M. Rahman et al. Blockchain in the banking industry: Unravelling thematic drivers and proposing a technological framework through systematic review with bibliographic network mapping. *IET Blockchain*, 5(1):e12093, 2025. doi:10.1049/blc2.12093.
19. A. Ali. Decentralized finance (defi) and its impact on traditional banking systems: Opportunities, challenges, and future directions. Social Science Research Network, August 2024. SSRN ID: 4942313. doi:10.2139/ssrn.4942313.
20. A. Alamsyah, G. N. W. Kusuma, and D. P. Ramadhani. A review on decentralized finance ecosystems. *Future Internet*, 16(3):76, March 2024. doi:10.3390/fi16030076.
21. D. Tapscott and A. Tapscott. *Blockchain Revolution: How the Technology Behind Bitcoin Is Changing Money, Business, and the World*. Penguin Publishing Group, 2016.
22. T. Ahram, A. Sargolzaei, S. Sargolzaei, J. Daniels, and B. Amaba. Blockchain technology innovations. In *2017 IEEE Technology & Engineering Management Conference (TEMSCON)*, pages 137–141, June 2017. doi:10.1109/TEMSCON.2017.7998367.
23. V. K. Vemuri. Blockchain: a practical guide to developing business, law, and technology solutions. *Journal of Information Technology Case and Application Research*, 2018. Accessed: Jun. 18, 2025. doi:10.1080/15228053.2019.1588546.
24. R. Zhang, R. Xue, and L. Liu. Security and privacy on blockchain. *ACM Computing Surveys*, 52(3):1–34, May 2020. doi:10.1145/3316481.
25. E. O. Manu A. D. Bello A. O. Leo C. E. Ukatu A. A. Bello, D. A. Oduro and N. Okika. Enhancing know your customer (kyc) and anti-money laundering (aml) compliance using blockchain: A business analysis approach. Iconic Research And Engineering Journals, 8(9):297–305, 2025. [Online]. URL: https://www.irejournals.com/paper-details/1707440.
26. Y. Guo and C. Liang. Blockchain application and outlook in the banking industry. *Financial Innovation*, 2(1):24, December 2016. doi:10.1186/s40854-016-0034-9.
27. T.-G. Budisteanu. Blockchain and the banking sector: Benefits, challenges and perspectives. *Journal of Service Science (JSS)*, 13(03):288–300, 2025. doi:10.4236/jss.2025.133019.
28. R. Mishra, R. K. Singh, S. Kumar, S. K. Mangla, and V. Kumar. Critical success factors of blockchain technology adoption for sustainable and resilient operations in the banking industry during an uncertain business environment. *Electronic Commerce Research*, 25(1):595–629, February 2025. doi:10.1007/s10660-023-09707-3.
29. H. Taherdoost. A critical review of blockchain acceptance models—blockchain technology adoption frameworks and applications. *Computers*, 11(2):24, February 2022. doi:10.3390/computers11020024.
30. J. Paul and A. R. Criado. The art of writing literature review: What do we know and what do we need to know? *International Business Review*, 29(4):101717, Aug 2020. doi:10.1016/j.ibusrev.2020.101717.
31. J. Pritchard. Statistical-bibliography or bibliometrics? *Journal of Documentation*, 25(4):348–349, 1969.
32. L. Haggarty. What is content analysis? *Medical Teacher*, 18(2):99–101, Jan 1996. doi:10.3109/01421599609034141.

33. Y. Feng, Q. Zhu, and K.-H. Lai. Corporate social responsibility for supply chain management: A literature review and bibliometric analysis. *Journal of Cleaner Production*, 158:296–307, Aug 2017. doi:10.1016/j.jclepro.2017.05.018.
34. T. Anushree, K. Puneet, M. Matti, and D. Amandeep. Blockchain applications in management: A bibliometric analysis and literature review. *Technological Forecasting and Social Change*, 166:120649, May 2021. doi:10.1016/j.techfore.2021.120649.
35. M. M. Alshater, M. Joshipura, R. E. Khoury, and N. Nasrallah. Initial coin offerings: a hybrid empirical review. *Small Business Economics*, 61(3):891–908, Oct 2023. doi:10.1007/s11187-022-00726-2.
36. S. M. M. Rahman, K. J. Yii, E. K. Masli, and M. L. Voon. The blockchain in the banking industry: a systematic review and bibliometric analysis. *Cogent Business & Management*, 11(1):2407681, Dec 2024. doi:10.1080/23311975.2024.2407681.
37. R. Patel, M. Migliavacca, and M. E. Oriani. Blockchain in banking and finance: A bibliometric review. *Research in International Business and Finance*, 62:101718, Dec 2022. doi:10.1016/j.ribaf.2022.101718.
38. N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296, Sep 2021. doi:10.1016/j.jbusres.2021.04.070.
39. M. J. Page and et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Revista Española de Cardiología (English Edition)*, 74(9):790–799, Sep 2021. doi:10.1016/j.rec.2021.07.010.
40. N. J. van Eck and L. Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, Aug 2010. doi:10.1007/s11192-009-0146-3.
41. N. J. van Eck and L. Waltman. Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, 111(2):1053–1070, May 2017. doi:10.1007/s11192-017-2300-7.
42. A. M. Alawag, W. S. Alaloul, B. N. Saleh Al-dhawi, A. O. Baarimah, M. A. Bazel, and A. W. Mushtaha. A review and bibliometric analysis of blockchain adoption within the context of smart construction projects. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, pages 805–811, Jan 2024. doi:10.1109/ICETSIS61505.2024.10459703.
43. Öguzhan Öztürk, Rıdvan Kocaman, and Dominik K. Kanbach. How to design bibliometric research. *Review of Managerial Science*, 18:3333–3361, 2024. doi:10.1007/s11846-024-00738-0.
44. Rahul Kumar. Bibliometric analysis: comprehensive insights into tools, techniques, applications, and solutions for research excellence. *Spectrum of Engineering and Management Sciences*, 3(1):45–62, 2025. doi:10.31181/sems31202535k.
45. C. M. J., L.-H. A. G., H.-V. E., and H. F. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7):1382–1402, May 2011. doi:10.1002/asi.21525.
46. Chaomei Chen. Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2), May 2017. doi:10.1515/jdis-2017-0006.
47. M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2):191–235, Mar 1983. doi:10.1177/053901883022002003.
48. M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963. doi:10.1002/asi.5090140103.
49. V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, Jan 2006. doi:10.1191/1478088706qp063oa.
50. M. Cheng, D. Edwards, S. Darcy, and K. Redfern. A tri-method approach to a review of adventure tourism literature: Bibliometric analysis, content analysis, and a quantitative systematic literature review. *Journal of Hospitality & Tourism Research*, 42(6):997–1020, Aug 2018. doi:10.1177/1096348016640588.
51. M. D. Moon. Triangulation: A method to increase validity, reliability, and legitimation in clinical research. *Journal of Emergency Nursing*, 45(1):103–105, Jan 2019. doi:10.1016/j.jen.2018.11.004.
52. W. M. Lim, T. Rasul, S. Kumar, and M. Ala. Past, present, and future of customer engagement. *Journal of Business Research*, 140:439–458, Feb 2022. doi:10.1016/j.jbusres.2021.11.014.
53. D. W. Aksnes, L. Langfeldt, and P. Wouters. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1), 2019. doi:10.1177/2158244019829575.
54. A. V. Thakor. Fintech and banking: What do we know? *Journal of Financial Intermediation*, 41:100833, Jan 2020. doi:10.1016/j.jfi.2019.100833.
55. J. Dai and M. A. Vasarhelyi. Toward blockchain-based accounting and assurance. *Journal of Information Systems*, 31(3):5, 2017.
56. G. Peters and E. Panayi. Understanding modern banking ledgers through blockchain technologies: Future of transaction processing and smart contracts on the internet of money. Social Science Research Network, Nov 2015. SSRN ID: 2692487. doi:10.2139/ssrn.2692487.
57. S. Schuetz and V. Venkatesh. Blockchain, adoption, and financial inclusion in india: Research opportunities. *International Journal of Information Management*, 52:101936, Jun 2020. doi:10.1016/j.ijinfomgt.2019.04.009.
58. D. Minoli and B. Occhiogrosso. Blockchain mechanisms for iot security. *Internet of Things*, 1–2:1–13, Sep 2018. doi:10.1016/j.iot.2018.05.002.
59. D. A. Zetzsche, D. W. Arner, and R. P. Buckley. Decentralized finance (defi), 2020. doi:10.2139/ssrn.3539194.
60. P. Garg, B. Gupta, A. K. Chauhan, U. Sivarajah, S. Gupta, and S. Modgil. Measuring the perceived benefits of implementing blockchain technology in the banking sector. *Technological Forecasting and Social Change*, 163:120407, Feb 2021. doi:10.1016/j.techfore.2020.120407.
61. S. Ahluwalia, R. V. Mahto, and M. Guerrero. Blockchain technology and startup financing: A transaction cost economics perspective. *Technological Forecasting and Social Change*, 151:119854, Feb 2020. doi:10.1016/j.techfore.2019.119854.
62. M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Khan. A review of blockchain technology applications for financial services. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(3):100073, Jul 2022. doi:10.1016/j.tbench.2022.100073.
63. Katy Börner, Chaomei Chen, and Kevin W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255, 2003. doi:10.1002/aris.1440370106.
64. Katy Börner, T. N. Theriault, and Kevin W. Boyack. Mapping science introduction: Past, present and future. *Bulletin of the*

Association for Information Science and Technology, 41(2):12–16, 2015. doi:10.1002/bult.2015.1720410205.

65. G. Wolfgang and S. András. Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research*, pages 257–276. Springer, 2005. doi:10.1007/1-4020-2755-9_12.
66. A. Isfandyari-Moghaddam, M. K. Saberi, S. Tahmasebi-Limoni, S. Mohammadian, and F. Naderbeigi. Global scientific collaboration: A social network analysis and data mining of the co-authorship networks. *Journal of Information Science*, 49(4):1126–1141, August 2023. doi:10.1177/01655515211040655.
67. T. Luukkonen, O. Persson, and G. Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, 17(1):101–126, 1992.
68. Q. Gui, C. Liu, and D. Du. Globalization of science and international scientific collaboration: A network perspective. *Geoforum*, 105:1–12, October 2019. doi:10.1016/j.geoforum.2019.06.017.
69. J. S. Katz and B. R. Martin. What is research collaboration? *Research Policy*, 26(1):1–18, March 1997. doi:10.1016/S0048-7333(96)00917-1.
70. C. Wagner and L. Leydesdorff. Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, 1(2):185–208, 2005. doi:10.1504/IJTG.2005.007050.
71. H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973. doi:10.1002/asi.4630240406.
72. H. D. White and B. C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3):163–171, 1981. doi:10.1002/asi.4630320302.
73. D. Zhao and A. Strotmann. Intellectual structure of information science 2011–2020: an author co-citation analysis. *Journal of Documentation*, 78(3):728–744, 2021. doi:10.1108/JD-06-2021-0119.
74. K. W. McCain. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6):433–443, 1990. doi:10.1002/(SICI)1097-4571(199009)41:6<433::AID-ASI11>3.0.CO;2-Q.
75. M. Sedighi. Application of word co-occurrence analysis method in mapping of the scientific fields (case study: the field of informetrics). *Library Review*, 65(1/2):52–64, 2016. doi:10.1108/LR-07-2015-0075.
76. L. Leydesdorff and A. Nerghes. Co-word maps and topic modeling: A comparison using small and medium-sized corpora (n < 1,000). *Journal of the Association for Information Science and Technology*, 68(4):1024–1035, 2017. doi:10.1002/asi.23740.
77. N. Cucari, V. Lagasio, G. Lia, and C. Torriero. The impact of blockchain in banking processes: the interbank spunta case study. *Technology Analysis Strategic Management*, 2022. Accessed: Jun. 25, 2025. doi:10.1080/09537325.2021.1891217.
78. M. Shoaib, M. K. Lim, and C. Wang. An integrated framework to prioritize blockchain-based supply chain success factors. *Industrial Management Data Systems*, 120(11):2103–2131, 2020. doi:10.1108/IMDS-04-2020-0194.
79. R. K. Jena. Examining the factors affecting the adoption of blockchain technology in the banking sector: An extended utaut model. *International Journal of Financial Studies*, 10(4):90, 2022. doi:10.3390/ijfs10040090.
80. L. Mishra and V. Kaushik. Application of blockchain in dealing with sustainability issues and challenges of financial sector. *Journal of Sustainable Finance Investment*, 13(3):1318–1333, 2023. doi:10.1080/20430795.2021.1940805.
81. P. Schueffel. Defi: Decentralized finance - an introduction and overview. *Journal of Innovation Management*, 9:i, Dec 2021. doi:10.24840/2183-0606_009.003_0001.
82. C. Dierksmeier and P. Seele. Cryptocurrencies and business ethics. *Journal of Business Ethics*, 152(1):1–14, Sep 2018. doi:10.1007/s10551-016-3298-0.
83. M. Fauzi and N. Paiman. Bitcoin and cryptocurrency: Challenges, opportunities and future works. *Journal of Asian Finance Economics and Business*, 7:695–704, Aug 2020. doi:10.13106/jafeb.2020.vol7.no8.695.
84. J. Campino, A. Brochado, and Á. Rosa. Initial coin offerings (icos): Why do they succeed? *Financial Innovation*, 8(1):17, Jan 2022. doi:10.1186/s40854-021-00317-2.
85. T. MacDonald, D. W. E. Allen, and J. Potts. Blockchains and the boundaries of self-organized economies: Predictions for the future of banking. Social Science Research Network, Mar 2016. SSRN working paper No. 2749514. doi:10.2139/ssrn.2749514.
86. Alex Tapscott and Don Tapscott. How blockchain is changing finance. Harvard Business Review, Mar. 1 2017. [Online]. URL: https://hbr.org/2017/03/how-blockchain-is-changing-finance.
87. Raphael Auer and David Tercero-Lucas. Distrust or speculation? the socioeconomic drivers of u.s. cryptocurrency investments. *Journal of Financial Stability*, 62:101066, 2022. doi:10.1016/j.jfs.2022.101066.
88. Ravi Gupta, V. K. Shukla, S. S. Rao, Shahbaz Anwar, Pooja Sharma, and Ritu Bathla. Enhancing privacy through 'smart contract' using blockchain-based dynamic access control. In *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, pages 338–343, 2020. doi:10.1109/ICCAKM46823.2020.9051521.
89. Amin Naimi-Sadigh, Tooraj Asgari, and Majid Rabiei. Digital transformation in the value chain disruption of banking services. *Journal of the Knowledge Economy*, 13(2):1212–1242, 2022. doi:10.1007/s13132-021-00759-0.
90. Husam Rjoub, Temitope S. Adebayo, and Dervis Kirikkaleli. Blockchain technology-based fintech banking sector involvement using adaptive neuro-fuzzy-based k-nearest neighbors algorithm. *Financial Innovation*, 9(1):65, 2023. doi:10.1186/s40854-023-00469-3.
91. Vincent Schlatt, Jonas Sedlmeir, Sven Feulner, and Nils Urbach. Designing a framework for digital kyc processes built on blockchain-based self-sovereign identity. *Information & Management*, 59(7):103553, 2022. doi:10.1016/j.im.2021.103553.
92. Syed Umar Rehman et al. Fintech adoption in smes and bank credit supplies: A study on manufacturing smes. *Economies*, 11(8):213, 2023. doi:10.3390/economies11080213.
93. Qian Gan and Raymond Y. K. Lau. Trust in a 'trust-free' system: Blockchain acceptance in the banking and finance sector. *Technological Forecasting and Social Change*, 199:123050, 2024. doi:10.1016/j.techfore.2023.123050.
94. T. Saheb and F. H. Mamaghani. Exploring the barriers and organizational values of blockchain adoption in the banking industry. *The Journal of High Technology Management Research*, 32(2):100417, November 2021. doi:10.1016/j.hitech.2021.100417.

95. J. G. Umarovich and R. K. Bakhtiyorovich. Modeling the decision-making process of lenders based on blockchain technology. In *2021 International Conference on Information Science and Communications Technologies (ICISCT)*, pages 1–5, November 2021. doi:10.1109/ICISCT52966.2021.9670211.
96. D. Kimani, K. Adams, R. Attah-Boakye, S. Ullah, J. Frecknall-Hughes, and J. Kim. Blockchain, business and the fourth industrial revolution: Whence, whither, wherefore and how? *Technological Forecasting and Social Change*, 161:120254, December 2020. doi:10.1016/j.techfore.2020.120254.
97. P. Cuccuru. Beyond bitcoin: an early overview on smart contracts. *International Journal of Law and Information Technology*, 25(3):179–195, September 2017. doi:10.1093/ijlit/eax003.
98. C. Albrecht, K. M. Duffin, S. Hawkins, and V. M. M. Rocha. The use of cryptocurrencies in the money laundering process. *Journal of Money Laundering Control*, 22(2):210–216, May 2019. doi:10.1108/JMLC-12-2017-0074.
99. K.-K. R. Choo, S. Ozcan, A. Dehghantanha, and R. M. Parizi. Editorial: Blockchain ecosystem—technological and management opportunities and challenges: Part ii. *IEEE Transactions on Engineering Management*, 69(3):773–775, June 2022. doi:10.1109/TEM.2022.3147274.
100. S. Dashottar and V. Srivastava. Corporate banking—risk management, regulatory and reporting framework in india: a blockchain application-based approach. *Journal of Banking Regulation*, 22(1):39–51, March 2021. doi:10.1057/s41261-020-00127-z.
101. Jianguo Xu. Developments and implications of central bank digital currency: The case of china e-cny. *Asian Economic Policy Review*, 17(2):235–250, 2022. doi:10.1111/aepr.12396.
102. D. Mhlanga. Block chain technology for digital financial inclusion in the industry 4.0, towards sustainable development? *Frontiers in Blockchain*, 6, February 2023. doi:10.3389/fbloc.2023.1035405.
103. M. Osmani, R. El-Haddadeh, N. Hindi, M. Janssen, and V. Weerakkody. Blockchain for next generation services in banking and finance: cost, benefit, risk and opportunity analysis. *Journal of Enterprise Information Management*, 34(3):884–899, Jun 2020. doi:10.1108/JEIM-02-2020-0044.
104. R. Weerawarna, S. Miah, and X. Shao. Emerging advances of blockchain technology in finance: a content analysis. *Personal and Ubiquitous Computing*, 27:1–14, Feb 2023. doi:10.1007/s00779-023-01712-5.
105. M. Khalil, K. F. Khawaja, and M. Sarfraz. The adoption of blockchain technology in the financial sector during the era of fourth industrial revolution: a moderated mediated model. *Qualitative and Quantitative*, 56(4):2435–2452, Aug 2022. doi:10.1007/s11135-021-01229-0.
106. H. O. Mbaidin, M. A. K. Alsmairat, and R. Al-Adaileh. Blockchain adoption for sustainable development in developing countries: Challenges and opportunities in the banking sector. *International Journal of Information Management Data Insights*, 3(2):100199, Nov 2023. doi:10.1016/j.jjimei.2023.100199.
107. Luca Rella. Blockchain technologies and remittances: From financial inclusion to correspondent banking. *Frontiers in Blockchain*, 2, 2019. doi:10.3389/fbloc.2019.00014.
108. Zhuo Chen, Yanhui Li, Yujie Wu, and Jie Luo. The transition from traditional banking to mobile internet finance: an organizational innovation perspective - a comparative study of citibank and icbc. *Financial Innovation*, 3(1):12, 2017. doi:10.1186/s40854-017-0062-0.