

Research Article

Large Language Models in FinTech: A Systematic Review of Applications, Methodologies, Challenges, and Future Research

Abhishek Tiwari*

Computer Science (Artificial Intelligence & Machine Learning), SR Institute of Institutions, Lucknow, Uttar Pradesh, India

Abstract

Large language models (LLMs) have moved rapidly into financial technology, where they are used to analyze disclosures, answer questions over financial reports, support customers, and assist risk and compliance work. The research is spread across natural language processing, finance, and enterprise engineering, and it has not been brought together systematically. This article presents a systematic review of LLMs in FinTech, following PRISMA 2020 reporting guidance. It addresses four research questions covering applications, methodologies, challenges, and future research directions. Nine studies published between 2021 and 2024 were synthesized. They include financial benchmarks and datasets, domain-adapted and instruction-tuned financial models, an empirical study of LLMs processing corporate disclosures, surveys of foundation models and financial LLMs, and engineering studies on prompt design, guardrails, and workflow integration. The review finds that financial analysis and question answering dominate current research. Domain adaptation and benchmark development are the most mature methodologies. Numerical reasoning over tables and text, reliability, safety, and governed deployment are the main open challenges. The review proposes an integrative framework linking applications, methodologies, and controls, and sets out an agenda for future research focused on evaluation in realistic settings, customer-facing safety, and integration with risk governance.

Keywords: large language models, FinTech, systematic review, financial NLP, benchmarks, domain adaptation, instruction tuning, prompt engineering, guardrails, risk management, PRISMA

Introduction

Financial services depend on text. Annual reports, earnings calls, regulatory filings, analyst reports, customer messages, and internal policies contain much of the information on which financial decisions are based. Earlier natural language processing (NLP) methods handled narrow tasks such as sentiment classification. LLMs, trained on large volumes of text and able to follow instructions, can summarize, answer questions, extract information, and reason across many tasks (Bommasani et al., 2021).

Interest in LLMs for finance has grown quickly. Researchers have built financial benchmarks (Zhu et al., 2021; Shah et al., 2022), trained and instruction-tuned financial models (Shah et al., 2022; Xie et al., 2023), and examined how LLMs process corporate disclosures (Kim et al., 2023). Surveys have begun to organize the field (Li et al., 2023). At the same time, enterprise research has addressed how to make LLM outputs reliable (Sannidhanam, 2024a), how to constrain them with guardrails (Sannidhanam, 2024b), and how to govern the decisions they inform (Basireddy, 2022).

These strands have developed in different communities, with different methods and assumptions. Practitioners and researchers lack a consolidated view of what LLMs are used for in FinTech, how they are built and evaluated, what problems remain, and where research should go next. This systematic review addresses that need through four research questions:

- **RQ1:** What are the main applications of LLMs in FinTech?
- **RQ2:** What methodologies are used to develop, adapt, evaluate, and deploy LLMs in FinTech?
- **RQ3:** What challenges limit their effective and safe use?
- **RQ4:** What directions should future research take?

Review Methodology

Review design

The review follows the PRISMA 2020 statement for reporting systematic reviews (Page et al., 2021). Figure 1 summarizes the review process.

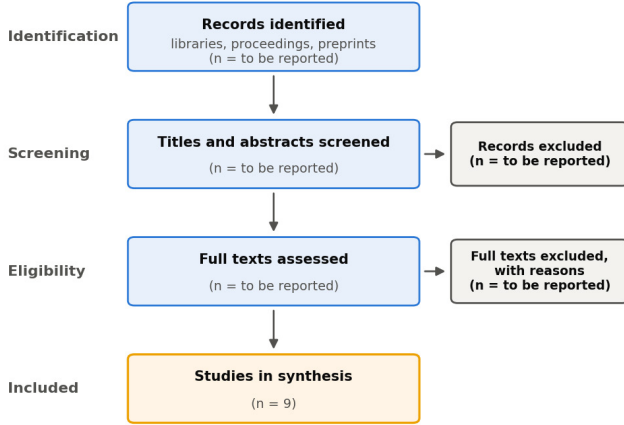


Figure 1: Review process following PRISMA 2020 stages

Search strategy

The search targeted literature on LLMs and pre-trained language models applied to finance and FinTech, together with enterprise studies on the reliability, safety, and governance of LLM applications relevant to financial services. Search terms combined model terms (“large language model”, “LLM”, “pre-trained language model”, “foundation model”, “generative AI”) with domain terms (“finance”, “financial”, “FinTech”, “banking”, “financial question answering”, “financial NLP”) and deployment terms (“prompt engineering”, “guardrails”, “risk management”, “workflow”). Sources included digital libraries of computing and finance research, conference proceedings in NLP and AI, and preprint servers, reflecting the fast pace of the field.

Eligibility criteria

Studies were included if they (a) were published between 2016 and 2024, (b) proposed, evaluated, or surveyed LLMs or pre-trained language models for financial tasks, or addressed the reliability, safety, or governance of LLM-based enterprise applications relevant to finance, and (c) reported a method, dataset, benchmark, architecture, or systematic synthesis. Opinion pieces without method, studies of non-language financial models, and duplicate versions of the same work were excluded.

Selection, extraction, and synthesis

Titles and abstracts were screened against the criteria, followed by full-text assessment. For each included study, the review extracted the year, study type, application area, methodology, main findings, and stated limitations. Findings were synthesized thematically around the four research questions.

Note on reporting: In accordance with PRISMA 2020, the number of records identified, screened, excluded (with reasons), and included at each stage should be reported

in Figure 1 and accompanied by the full search strings and dates for each source. These counts are left for completion because they depend on the exact searches run. This article reports the synthesis of the nine studies that form the final set.

3. Characteristics of Included Studies

Table 1 summarizes the nine included studies.

Figure 2 shows the included studies by type, and Figure 3 by application area.

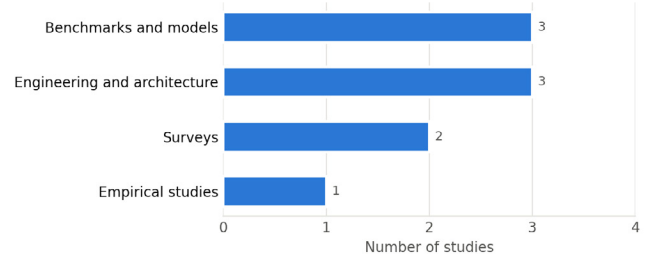


Figure 2: Included studies by study type (n = 9)

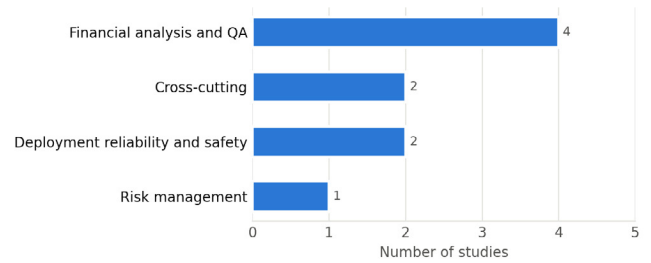


Figure 3: Included studies by application area (n = 9)

Three patterns stand out. First, benchmarks and models and engineering and architecture studies form the two largest groups. They reflect two communities, one focused on measurable progress on financial tasks and the other on making LLM applications reliable and governable. Second, financial analysis and question answering is the dominant application area. Customer-facing applications appear mainly through general deployment studies rather than finance-specific evaluations. Third, the studies are **recent**, with most published from 2021 onward and a concentration in 2023, following the wider growth of LLMs. The evidence base is therefore young and changing quickly.

Applications (RQ1)

Financial analysis and question answering

The most studied application is analysis of financial documents. Zhu et al. (2021) introduced TAT-QA, a benchmark of questions over financial reports that combine tables and text and often require numerical reasoning such as addition, comparison, and ratio calculation. Shah et al. (2022) assembled FLUE, a

Table 1: Characteristics of included studies

<i>Study</i>	<i>Year</i>	<i>Type</i>	<i>Application area</i>	<i>Main contribution</i>
Zhu et al.	2021	Benchmark	Financial analysis and question answering	TAT-QA benchmark for question answering over tables and text in financial reports
Bommasani et al.	2021	Survey	Cross-cutting	Opportunities and risks of foundation models
Shah et al.	2022	Benchmark and model	Financial analysis and question answering	FLUE benchmark suite and FLANG finance-specific language model
Basireddy	2022	Architecture	Risk management	Configurable enterprise workflow architecture for digitized risk management
Xie et al.	2023	Benchmark and model	Financial analysis and question answering	PIXIU: instruction data, instruction-tuned financial LLM, and evaluation benchmark
Li et al.	2023	Survey	Cross-cutting	Survey of LLMs in finance, approaches, and adoption guidance
Kim et al.	2023	Empirical study	Financial analysis and question answering	LLM summaries of corporate disclosures and their informativeness
Sannidhanam	2024a	Engineering patterns	Deployment reliability	Prompt engineering patterns for reliable LLM outputs in production
Sannidhanam	2024b	Engineering mechanisms	Deployment safety	Guardrails and safety mechanisms for LLM-powered enterprise applications

benchmark covering financial tasks including sentiment analysis, headline classification, named entity recognition, structure boundary detection, and question answering. Xie et al. (2023) presented PIXIU, which includes a financial instruction dataset, an instruction-tuned financial LLM, and an evaluation benchmark spanning financial NLP and prediction tasks. Kim et al. (2023) examined how an LLM processes corporate disclosures. They found that LLM-generated summaries were much shorter than the original text and that the sentiment of the summaries was more informative about market reactions than the sentiment of the full documents. This suggests that LLMs can help investors cut through lengthy, uninformative disclosures.

Customer service and interaction

The finance-specific studies in the review focus mainly on analysis, but the surveys identify customer interaction as a major application area (Li et al., 2023). This includes chat assistants, product explanations, and support. The enterprise studies address the reliability and safety requirements that such customer-facing uses impose, including structured prompting to produce consistent and grounded responses (Sannidhanam, 2024a) and guardrails to block unsafe outputs and escalate to humans (Sannidhanam, 2024b).

Risk management and compliance

Risk and compliance applications include monitoring communications and complaints, summarizing incidents, and mapping regulations to controls. The included finance NLP studies address this area only indirectly. Basireddy (2022) contributes the architectural foundation. Configurable workflows route identified risks through

assessment, approval, escalation, and remediation with audit trails, which is the structure into which LLM outputs must fit if they are to inform risk decisions.

Methodologies (RQ2)

The review identified five methodological approaches.

Domain adaptation. Pre-training or continued training on financial text improves performance on financial tasks. Shah et al. (2022) showed that FLANG, trained with finance-specific objectives on financial data, outperformed general models on the FLUE benchmark.

Instruction tuning. Xie et al. (2023) fine-tuned a general open LLM on financial instruction data, producing a model that follows financial task instructions across multiple tasks. Instruction tuning adapts a general model to financial tasks at far lower cost than training from scratch.

Benchmark-based evaluation. Benchmarks such as TAT-QA, FLUE, and the PIXIU evaluation suite provide shared tasks and metrics (Zhu et al., 2021; Shah et al., 2022; Xie et al., 2023). Li et al. (2023) organize approaches into zero-shot and few-shot use of general models, fine-tuning, and training domain-specific models from scratch. They suggest choosing among these according to task requirements, data availability, and cost.

Prompt engineering. Kim et al. (2023) used prompting to obtain summaries from a general model without fine-tuning. Sannidhanam (2024a) sets out prompt patterns for reliable production outputs, including explicit instructions, structured output formats, grounding in supplied context, constraints, and output validation.

Guardrails and workflow integration. Sannidhanam (2024b) describes safety mechanisms around LLM

applications: input filtering, output validation, action restrictions, protection of sensitive data, and human escalation. Basireddy (2022) describes configurable workflows that bring AI-informed decisions under governance.

Comparing adaptation strategies. The included studies point to a trade-off between specialization and cost. Training or continuing to train a domain-specific model, as with FLANG (Shah et al., 2022), yields strong performance on financial language but requires substantial financial data and computing resources, and the resulting model must be maintained as general models advance. Instruction tuning an open general model, as with PIXIU (Xie et al., 2023), is cheaper and allows institutions to benefit from improvements in the underlying model, but depends on the quality and coverage of the instruction data. Prompting a general model without further training, as in Kim et al. (2023), is the fastest and least costly option and can be surprisingly effective for summarization and sentiment tasks. Its reliability, however, depends heavily on prompt design and output checks (Sannidhanam, 2024a). Li et al. (2023) suggest that institutions begin with prompting and move to fine-tuning or domain-specific models only when task performance, data sensitivity, or cost at scale justify it. This view is consistent with the evidence reviewed here. Figure 4 integrates the applications, methodologies, and controls identified in the review into a single framework.

Challenges (RQ3)

Table 2 summarizes the main challenges identified in the included studies, with the evidence for each and the mitigations proposed.

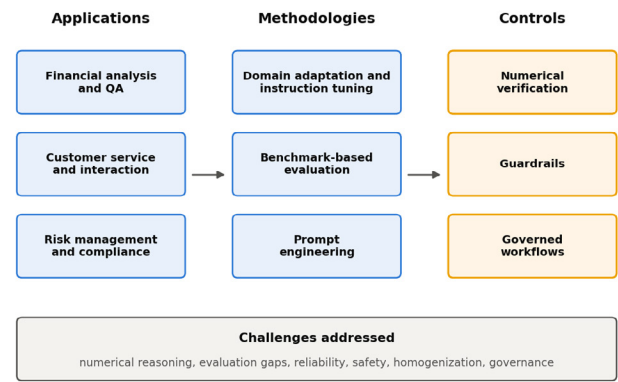


Figure 4: Integrative framework linking applications, methodologies, and controls for LLMs in FinTech

Two challenges stand out. First, numerical reasoning is a persistent weakness that matters greatly in finance, where small calculation errors can have large consequences. Second, there is a gap between capability research and deployment research. The finance-specific studies measure what models can do on benchmarks, while the enterprise studies address reliability, safety, and governance. Few studies connect the two in real financial deployments.

Future Research Directions (RQ4)

Table 3 sets out a research agenda derived from the gaps identified.

Among these directions, three deserve priority. First, realistic evaluation is foundational. Benchmark results on curated question sets may overstate performance

Table 2: Challenges of LLMs in FinTech, evidence, and proposed mitigations

Challenge	Evidence from included studies	Proposed mitigations
Numerical reasoning over tables and text	Questions requiring calculation over hybrid financial data remain difficult for models (Zhu et al., 2021)	Specialized models, external calculation, verification against source
Domain mismatch	General models underperform domain-adapted models on financial tasks (Shah et al., 2022)	Domain pre-training, instruction tuning (Shah et al., 2022; Xie et al., 2023)
Evaluation gaps	Benchmarks focus on analysis tasks; customer and risk uses are rarely evaluated in finance-specific settings (Li et al., 2023)	Broader benchmarks, evaluation in realistic deployments
Reliability of outputs	Inconsistent or ungrounded outputs in production (Sannidhanam, 2024a)	Structured prompt patterns, grounding, output validation
Safety and misuse	Harmful, misleading, or sensitive outputs; broad risks of foundation models (Bommasani et al., 2021; Sannidhanam, 2024b)	Guardrails, access controls, escalation
Homogenization and systemic risk	Many applications built on few foundation models share their flaws (Bommasani et al., 2021)	Model diversity, independent validation
Governance of decisions	LLM outputs influencing consequential decisions without accountability	Configurable workflows with approval and audit (Basireddy, 2022)
Rapid obsolescence	Fast model turnover makes findings date quickly (Li et al., 2023)	Continuous re-evaluation, reporting of model versions

Table 3: Future research agenda for LLMs in FinTech

<i>Direction</i>	<i>Gap addressed</i>	<i>Example research questions</i>
Realistic evaluation	Benchmarks do not reflect deployment conditions	How do benchmark scores relate to accuracy and value in live financial workflows?
Numerical reliability	Weak calculation over tables and text	How effectively do tool use and verification reduce numerical errors in financial answers?
Customer-facing safety	Few finance-specific evaluations of customer applications	What error and escalation rates are acceptable for LLM assistants in consumer finance?
Risk and compliance applications	Limited direct study	Can LLMs detect emerging conduct risks from complaints earlier than existing methods?
Governance integration	Capability and governance studied separately	How should LLM outputs be integrated into approval workflows and audit records?
Prompting and guardrail evaluation	Engineering patterns rarely evaluated on financial tasks	Which prompt patterns and guardrails measurably improve reliability in finance?
Systemic effects	Homogenization risk from shared foundation models	Does widespread use of similar models create correlated errors across institutions?
Reproducibility	Rapid model change	What reporting standards would make financial LLM studies comparable over time?

on messy, changing, real documents, and without deployment evidence it is hard to judge any other direction. Second, numerical reliability is specific to finance and consequential. Studies that compare model-only answers with answers supported by calculation tools and source verification, using benchmarks such as TAT-QA (Zhu et al., 2021), would directly inform practice. Third, **governance integration** connects the two research communities identified in this review. Studies that follow LLM outputs from generation through guardrails (Sannidhanam, 2024b) into approval workflows (Basireddy, 2022), and measure where errors are caught, would show how capability and control combine in practice.

Discussion

Summary of findings

For RQ1, financial analysis and question answering dominate. Customer service and risk management are recognized applications but less studied in finance-specific terms. For RQ2, domain adaptation, instruction tuning, and benchmark-based evaluation are the most developed methodologies, and prompt engineering, guardrails, and workflow integration address deployment. For RQ3, numerical reasoning, evaluation gaps, reliability, safety, homogenization, and governance are the main challenges. For RQ4, the most pressing need is research that connects model capability with reliable, safe, and governed deployment in real financial settings.

Implications

For researchers, the review shows that the field would benefit from evaluations that move beyond benchmarks to real workflows, and from studies that combine financial NLP with deployment engineering and governance. For practitioners, it suggests choosing among general, fine-tuned, and domain-specific models by task (Li et al., 2023). It also suggests treating numerical outputs with special caution and placing prompt patterns, guardrails, and governed workflows around any LLM that informs customer or risk decisions.

Limitations

This review has important limitations. The final set of nine studies is small, and the field is changing quickly, so newer work may alter the picture. The inclusion of enterprise engineering studies broadens the review beyond finance-specific NLP, which is intentional but may affect comparability. Several included studies are preprints. Finally, the complete PRISMA record counts and search strings must be reported for the review to be fully reproducible.

Conclusion

This systematic review synthesized research on LLMs in FinTech across applications, methodologies, challenges, and future directions. It found a field in which financial analysis and question answering are well studied, domain adaptation and benchmarks are maturing, and

numerical reasoning and deployment governance remain major challenges. The integrative framework connects applications with the methodologies used to build them and the controls needed to deploy them safely. The central conclusion is that the next stage of research should connect what LLMs can do on financial benchmarks with how they perform reliably, safely, and accountably in real financial workflows.

References

- Basireddy, S. R. (2022). Digitizing risk management through configurable enterprise workflow architecture. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 14(4), 231–239.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv.
- Kim, A. G., Muhn, M., & Nikolaev, V. V. (2023). *Bloated disclosures: Can ChatGPT help investors process information?* (arXiv:2306.10224). arXiv.
- Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF '23)* (pp. 374–382).
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- Sannidhanam, A. H. (2024a). Prompt engineering patterns for reliable LLM outputs in production environments. *Journal of Multidisciplinary Knowledge*, 4(1), 29–39. <https://jmk.datatablets.com/index.php/j/article/view/168>
- Sannidhanam, A. H. (2024b). Guardrails and safety mechanisms for LLM-powered enterprise applications. *International Journal of Engineering Science & Humanities*, 14(3), 273–285. <https://www.ijesh.com/j/article/view/933>
- Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., & Yang, D. (2022). When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2322–2335).
- Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). *PIXIU: A large language model, instruction data and evaluation benchmark for finance* (arXiv:2306.05443). arXiv.
- Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., & Chua, T.-S. (2021). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 3277–3287).