

REVIEW



From General-Purpose Large Language Models to Domain-Specific Financial Models: A Structured Review and Comparative Case Analysis

Fumin Niu¹, Fengshuo Liu¹, Zekang Zhang¹, Yishi Li¹ and Yufei Ren^{1,*}

¹*School of Economics and Management, North China Electric Power University, China*

Abstract: The use of large language models (LLMs) in finance has expanded rapidly, yet general-purpose systems face recurring problems in high-risk and regulated applications: Unsupported factual generation, weak reasoning across text, tables, charts, and numerical data, and limited traceability of model-assisted decisions. This study conducts a Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA)-informed structured review of 67 studies, principally published between January 2020 and March 2026, with three foundational pre-2020 studies retained through backward citation tracking. It develops a qualitative analytical and diagnostic framework with three pathways: Specialized knowledge modeling, task-constrained multimodal reasoning, and Trust-by-Design. A study-level evidence map and an explicit coding protocol show how the pathways were derived. The framework is used to compare BloombergGPT, FinGPT, FinTral, Fin-o1, Fin-R1, and Agentar-Fin-R1 through common criteria for knowledge controls, modality handling, lifecycle governance, evidence quality, and deployment maturity. The case analysis distinguishes model-paper design descriptions, developer-reported benchmark results, platform-level deployment reports, and prospective applications. The analysis indicates that financial LLMs differ materially in how they organize domain knowledge, coordinate heterogeneous inputs, and govern model use across the lifecycle. Agentar-Fin-R1 illustrates the framework; the case does not validate the framework or establish independently observed production performance. The paper provides a structured basis for assessing domain-specific financial LLMs and connects responsible model design with sustainable financial innovation through durable governance arrangements, controlled model and compliance risk, responsible scaling, financial inclusion, and institutional resilience.

Keywords: LLMs, domain-specific financial models, multimodal financial reasoning, hallucination mitigation, trustworthy artificial intelligence (AI)

1. Introduction

1.1. Financial AI under high risk and strong constraints

In recent years, general-purpose large language models (LLMs) trained on large-scale open-domain corpora have demonstrated strong capabilities in language understanding, reasoning, and knowledge generation. When these models are applied directly to finance, however, their limitations become evident. Credit approval, risk assessment, anti-money laundering (AML) screening, and compliance supervision demand rigorous reasoning, a low tolerance for error, traceable evidence, and clear responsibility. Errors in these tasks can generate substantial economic losses and regulatory violations. Financial artificial intelligence (AI) systems must therefore combine useful

outputs with reliability, explainability, auditability, compliance, and accountable human oversight [1, 2].

General-purpose models are optimized for broad language generation and do not contain all safeguards required for financial correctness, cross-modal reasoning, and compliance-sensitive decision support. These limitations have directed attention toward domain-specific financial models that combine specialized knowledge, task structure, and lifecycle governance.

1.2. Structural limitations of general-purpose large models in financial scenarios

The limitations of general-purpose large models in financial applications mainly manifest in three aspects. First, the “hallucination” problem. The model may generate content that is semantically coherent but factually incorrect, especially when dealing with complex constraints or incomplete information. Existing research indicates that hallucinations stem from the inherent probabilistic generation mechanism of large models, meaning that the hallucination problem is difficult to fundamentally eliminate

*Corresponding author: Yufei Ren, School of Economics and Management, North China Electric Power University, China. Email: 50602529@ncepu.edu.cn

through prompt engineering alone [3]. In finance, where precision and accuracy are strictly required, even a relatively low hallucination rate can lead to serious analytical biases, misjudgments of financial risks, or compliance issues.

Second, financial decision-making depends on the integrated interpretation of multimodal information, often involving text, structured financial data, chart information, and time series indicator data. Although general-purpose large models are relatively stable in single-modal tasks, they still have significant deficiencies in multimodal alignment and joint reasoning, which may lead to partial or even contradictory judgments [4].

Third, the internal decision-making process of large models is complex and difficult to explain, and the black-box nature of their algorithms conflicts with the financial industry’s requirements for transparent decision-making and traceable accountability. When a model participates in high-risk business decisions, the absence of verifiable grounds for its outputs creates both regulatory and legal uncertainty. This issue has been repeatedly emphasized in the literature on AI safety and governance [3].

Taken together, hallucination, weak multimodal coordination, and limited decision traceability restrict the use of general-purpose LLMs in core financial processes. Shallow fine-tuning or isolated engineering patches address only part of these problems, which motivates a more systematic domain-specific design approach.

1.3. Research question, research approach, and contributions

This paper addresses the following question: How can domain-specific financial LLMs be designed and assessed in ways that respond to the structural limitations of general-purpose models while meeting the professional and regulatory requirements of finance?

Existing studies have extensively explored LLM applications in finance, covering areas such as financial text analysis, market sentiment, portfolio advisory, risk assessment, and compliance monitoring [5–11]. However, three gaps remain.

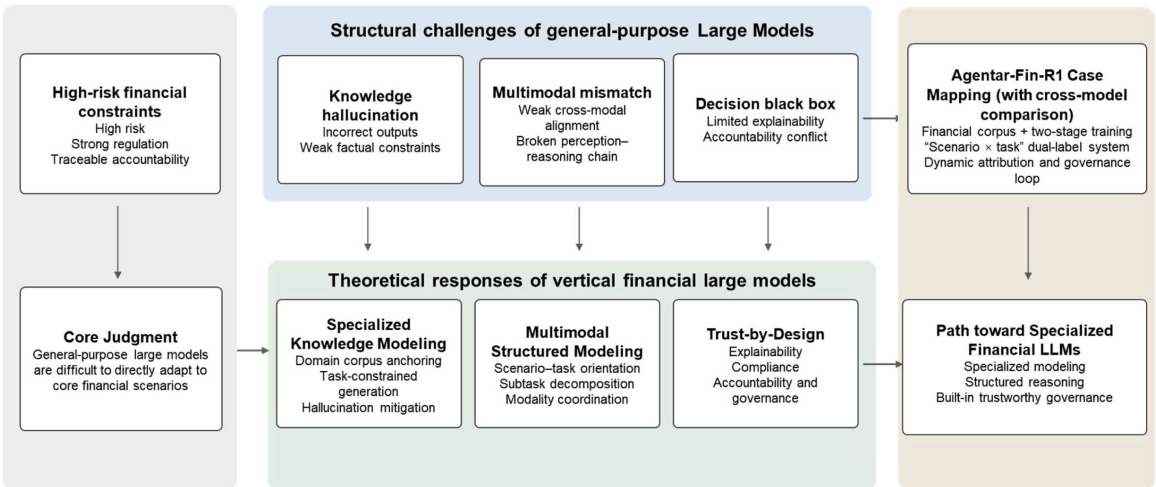
First, prior work often examines technical limitations, such as hallucination and multimodal reasoning gaps, and governance concerns, such as explainability and accountability, as separate analytical streams. Second, many model papers emphasize benchmark accuracy or application feasibility, leaving the relation between model architecture and financial accountability insufficiently specified. Third, case-based discussions of financial LLMs frequently rely on publicly reported materials, with limited comparison across representative models and limited attention to the evidentiary limits of vendor-reported claims.

This study therefore adopts a PRISMA-informed structured literature review and a comparative case analysis. The review synthesizes 67 studies, principally published between January 2020 and March 2026, with three foundational pre-2020 studies retained through backward citation tracking. It focuses on financial LLM applications, structural limitations in high-risk settings, and trustworthiness and governance. On this basis, the paper constructs a qualitative analytical and diagnostic framework centered on specialized knowledge modeling, task-constrained multimodal reasoning, and Trust-by-Design. Agentar-Fin-R1 is used as a publicly documented case, while five comparator models provide a consistent cross-model assessment.

The Agentar-Fin-R1 case illustrates the framework; it does not validate the framework or establish independently verified production performance.

The paper makes three contributions. First, it integrates research on hallucination, multimodal reasoning, and trustworthy AI into a finance-specific account of why general-purpose LLMs face heightened difficulties in regulated decision processes. Second, it converts the three pathways into a qualitative analytical and diagnostic framework with explicit indicators and an evidence hierarchy, making the derivation from the 67-study corpus visible. Third, it compares six representative financial LLMs and separates model design, developer-reported performance, platform deployment evidence, and prospective applications. This comparison supports a more disciplined assessment of responsible scaling and sustainable financial innovation. Figure 1 summarizes the analytical framework.

Figure 1
Qualitative analytical framework for domain-specific financial LLMs



Note: The framework links three recurring challenges of general-purpose LLMs—knowledge hallucination, multi-modal mismatch, and opaque decision processes—to three analytical pathways: specialized knowledge modeling, task-constrained multimodal reasoning, and Trust-by-Design. The right-hand side connects these pathways to the Agentar-Fin-R1 illustration and the cross-model comparison. The framework is conceptual and diagnostic; it does not represent an independently validated causal model.

2. Methodology

2.1. Research design

This study uses a qualitative design that combines a PRISMA-informed structured literature review, construction of an analytical framework, and comparative case application. The design follows three phases: (1) literature identification, screening, and thematic synthesis; (2) derivation and operationalization of the three-pathway framework; and (3) application of common indicators to Agentar-Fin-R1 and five comparator models. Figure 2 illustrates the research workflow.

PRISMA informs the reporting of identification, screening, and inclusion. The review is not presented as a preregistered systematic review or meta-analysis.

2.2. Historical corpus construction and prospective replication

2.2.1. Historical corpus construction

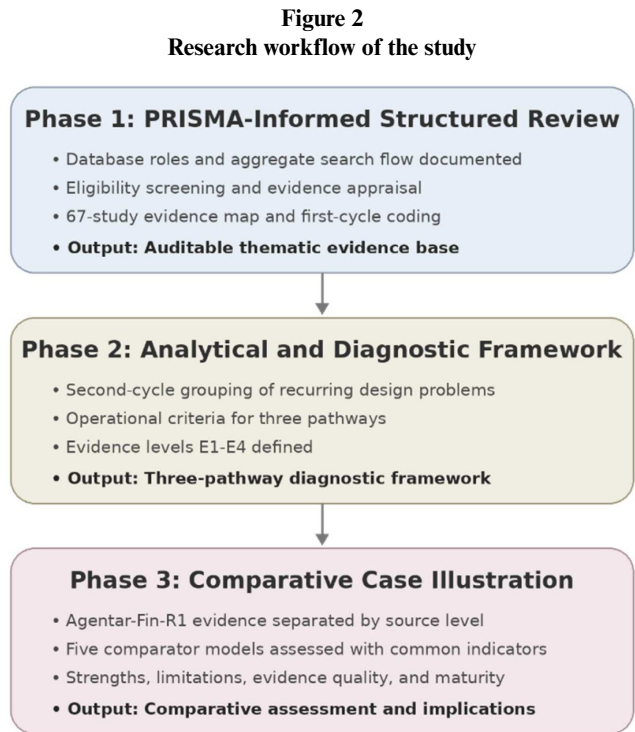
The 67-study corpus was assembled during manuscript preparation using Web of Science (WoS) as the primary bibliographic database, with Google Scholar and arXiv used for supplementary discovery. The retained historical log records the core query as (“large language model” OR “LLM” OR “foundation model”) AND (“finance” OR “financial” OR “banking”), the publication window, the thematic search families, and the aggregate screening counts. It does not preserve day-level

execution dates, source-specific yields, query-level exports, or the WoS index and field settings used in each iterative round. The historical process can therefore be documented at the aggregate level, while exact reconstruction of the original retrieval set is not supported by the retained records.

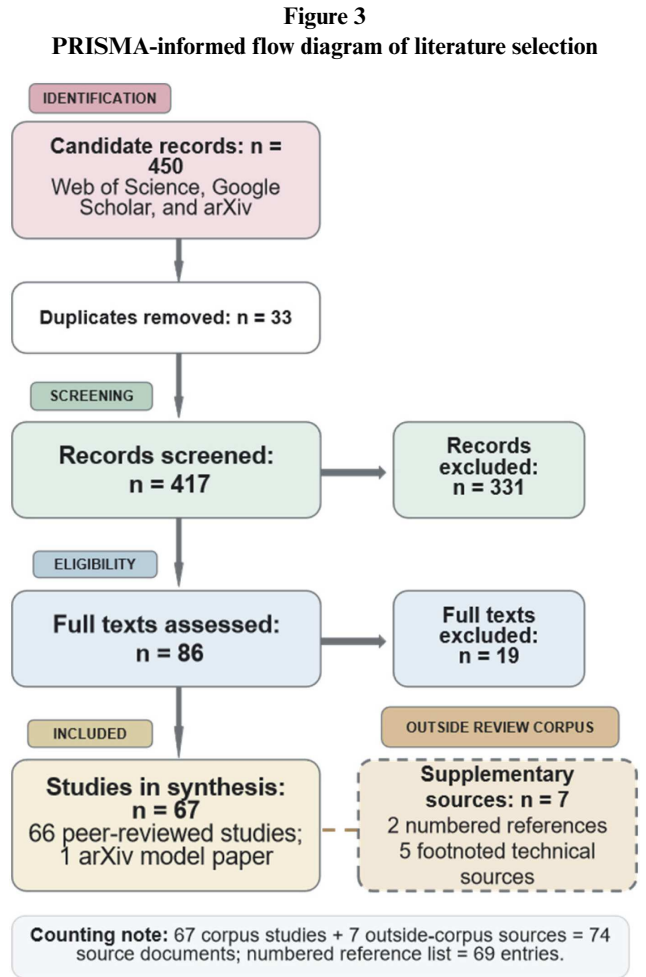
The historical search principally targeted publications from January 2020 to March 2026. Three foundational studies published before 2020 [12–14] were retained through backward citation tracking because they define multimodal learning, cloud-security evaluation, and AI accountability concepts used in the framework. Theme-specific searches covered hallucination and knowledge reliability, multimodal reasoning, explainability and explainable artificial intelligence (XAI), trustworthy AI, governance, compliance, and high-risk decision-making. The counts in Figure 3 describe this historical process.

2.2.2. Prospective replication protocol

Table S1 provides a prospective protocol for a new search. It specifies the WoS Core Collection indexes, Topic field, exact query families, equivalent arXiv searches, Google Scholar screening



Note: Phase 1 reports the retained historical review process, evidence appraisal, and thematic coding. Phase 2 derives and operationalizes a qualitative analytical and diagnostic framework. Phase 3 illustrates the framework through Agentar-Fin-R1 and applies the same indicators to five comparator models. Table S1 is a prospective replication protocol and did not generate the corpus or flow counts; Tables S2–S4 provide the study-level evidence map, code-to-pathway derivation, and register of sources outside the review corpus.



Note: The counts report the retained historical process: 450 candidate records, 33 duplicates removed, 417 title and abstract records screened, 86 full texts assessed, and 67 studies included. Seven sources outside the review corpus support model comparison and platform context: two numbered references [16, 17] and five technical sources cited in full footnotes. Table S1 presents a prospective replication protocol and did not generate these counts; Table S4 registers the seven outside-corpus sources and their use restrictions.

rules, query-level date and yield recording, independent screening, and conflict resolution. This protocol was formulated after the historical corpus had been assembled and did not produce the 450 candidate records or the 67-study corpus. A future implementation may yield a different record set because database coverage, indexing, and relevance rankings change over time.

2.3. Inclusion and exclusion criteria

Inclusion criteria were (1) peer-reviewed journal articles, conference proceedings, and technically substantive preprints; (2) studies addressing financial LLM applications, model reliability, multimodal reasoning, explainability, or governance in financial or transferable high-risk decision settings; (3) English-language publications; and (4) foundational pre-2020 studies identified through backward citation tracking when directly needed to define an analytical construct.

Exclusion criteria were (1) non-English publications, (2) duplicate records, (3) marketing-only or promotional materials without methods or analyzable mechanisms, (4) studies focused solely on nonfinancial domains without transferable relevance to high-risk financial use, and (5) records lacking sufficient detail for thematic coding. Corporate announcements and industry reports used to document Agentar platform claims were retained outside the 67-study review corpus and assigned a separate evidence category.

After 33 duplicates were removed, 417 records were screened by title and abstract. A total of 331 records were excluded for weak relevance or insufficient connection to financial LLMs, leaving 86 records for full-text assessment. Full-text assessment excluded 19 records that lacked direct financial relevance or sufficient analytical detail. The final corpus contained 67 studies: 66 peer-reviewed journal or conference studies and one arXiv model paper [15]. The retained historical log records aggregate flow counts but not source-specific yields. Two supplementary sources are retained in the numbered reference list outside the corpus: one corporate announcement [16] and one peer-reviewed comparator-model paper [17]. The numbered reference list therefore contains 69 entries. Four comparator-model preprints and one industry platform report are cited separately in full footnotes at their first occurrence. Figure 3 groups the two supplementary references and five footnoted sources as seven added sources outside the review corpus. Table S4 lists all seven sources outside the 67-study corpus and records their role, claim-evidence level, and use restriction.

Data extraction covered authors and year, context, model or system type, method, data source or benchmark, main finding, stated limitation, analytical theme, and evidence category. Table S2 presents the 67 records. First-cycle codes captured applications, hallucination and reliability, multimodal reasoning, explainability, governance, and domain-specific design. Second-cycle coding grouped these codes according to the recurring design problem addressed: knowledge reliability, heterogeneous-input reasoning, or lifecycle trustworthiness. These groups generated the three pathways. Table S3 documents the code-to-pathway rules.

The historical record does not document blinded duplicate screening, coder-by-coder decisions, or formal conflict-adjudication statistics. The study therefore reports no inter-rater coefficient and does not claim the procedures of a full systematic review. The prospective protocol in Supplementary Section S1 specifies independent screening and consensus resolution for future replication.

Because the corpus combines benchmarks, model papers, reviews, and legal or governance analyses, a single conventional risk-of-bias instrument would be inappropriate. The revision

therefore uses four evidence categories: A, peer-reviewed empirical or benchmark research with transparent methods; B, model papers or technical studies with disclosed methods but limited independent validation; C, review, conceptual, legal, or governance analysis; and D, corporate or vendor deployment material. Category D sources remain outside the 67-study synthesis and are used only as contextual evidence. This A–D study-type appraisal is distinct from the E1–E4 claim-evidence levels in Section 2.5. The former describes the evidentiary form of a reviewed study; the latter qualifies a specific mechanism or performance claim in the case and model comparison.

2.4. Case selection rationale

The financial LLM landscape has evolved rapidly since 2023. Representative models differ in design emphasis, including large-scale proprietary pre-training in BloombergGPT¹, open data-centric adaptation in FinGPT², multimodal financial training in FinTral [17]³, and reasoning-oriented supervised fine-tuning (SFT) and reinforcement-learning pipelines in Fin-o1⁴ and Fin-R1⁵. These variations provide a basis for applying common analytical indicators across models.

Agentar-Fin-R1 [15] was selected as an information-rich documentary case for three reasons. First, its model paper reports mechanisms relevant to all three framework dimensions. Second, public technical documentation and developer-reported benchmark results permit source-level examination of those claims. Third, as a 2025 model built on Qwen3 with a two-stage SFT and Group Relative Policy Optimization (GRPO) pipeline, it represents a recent financial reasoning design. Selection as the focal case does not imply representativeness or superiority. Model-paper design and benchmark claims are classified as E2 evidence; broader Agentar platform reports are classified as E3, and prospective applications are classified as E4.

A single focal case narrows the direct evidentiary base. Section 6.5 therefore applies the same operational criteria to five comparator models and reports their strengths, limitations, evidence quality, and deployment maturity. The comparison reduces case asymmetry while remaining document-based.

2.5. Analytical framework, operational criteria, and evidence levels

The framework is a qualitative analytical and diagnostic device rather than a causal theory. It links three finance-specific challenges to corresponding design pathways: (1) specialized knowledge modeling for knowledge reliability; (2) task-constrained multimodal reasoning for heterogeneous financial

¹Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kam-badur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564. <https://doi.org/10.48550/arXiv.2303.17564>, pp. 4, 15.

²Yang, H., Liu, X.-Y., & Wang, C. D. (2023). *FinGPT: Open-source financial large language models*. arXiv. <https://doi.org/10.48550/arXiv.2306.06031>, pp.4, 16.

³FinTral is the formally published comparator-model paper; its full bibliographic record is retained as reference [17].

⁴Qian, L., Zhou, W., Wang, Y., Peng, X., Yi, H., Zhao, Y., ..., & Nie, J.-Y. (2025). *Finol: On the transferability of reasoning-enhanced LLMs and reinforcement learning to finance*. arXiv. <https://doi.org/10.48550/arXiv.2502.08127>, pp. 5, 16.

⁵Liu, Z., Guo, X., Yang, Z., Lou, F., Zeng, L., Niu, J., ..., & Zhang, L. (2025). *Fin-R1: A large language model for financial reasoning through reinforcement learning*. arXiv. <https://doi.org/10.48550/arXiv.2503.16252>, pp. 5, 16.

inputs; and (3) Trust-by-Design for explainability, compliance, and accountability. Each pathway is assessed through explicit indicators and matched to an evidence level.

In the pathway comparison reported in Section 6.5, each pathway is classified as Met (M), Partial (P), or Not documented (ND). M requires all minimum components to be documented in a direct model source. P records one qualifying component or a required component supported only by platform or vendor material. ND means that the direct model source does not document a qualifying component. The status records documentary coverage and does not indicate effectiveness.

Pathway 1 requires both (i) a domain-knowledge anchoring mechanism, such as financial pre-training data, an ontology, controlled vocabulary, or retrieval, and (ii) an output-control or verification mechanism, including task-specific reward or preference alignment, structured generation, evidence binding, checking, or uncertainty handling. Financial pre-training or data curation without the second component is coded P. Meeting the documentary threshold does not establish observed hallucination reduction.

Pathway 2 is coded M when a direct model source documents at least two substantively different financial input types, including a structured or visual modality, together with an explicit cross-input alignment, fusion, routing, or task-decomposition mechanism. P records task structuring, heterogeneous-format processing, or text-numerical reasoning without a documented native cross-modal coordination mechanism. ND records a text-only design without a qualifying coordination mechanism.

Pathway 3 is coded M when a direct model source documents controls across at least three lifecycle stages, including at least one post-training stage (deployment or monitoring), and at least two governance functions, including explainability, privacy and security, fairness, compliance, or accountability. P records narrower development-stage coverage or a required post-training component supported only by platform material. Platform-level controls are reported as E3 qualifiers and do not upgrade a model-level status.

These thresholds follow a minimum-conjunction rule derived from the second-cycle coding in Supplementary Table S3. Each pathway requires documentary evidence for the relevant problem domain and for a mechanism that acts on it. The rule prevents a single generic feature, such as financial pre-training, multiple input formats, or preference alignment, from being treated as full pathway coverage. The thresholds are author-defined qualitative decision rules for comparison; they are not statistical cutoffs or evidence of causal effectiveness.

Evidence is classified as E1 when performance is independently evaluated on shared benchmarks or in deployment, E2 when it comes from an original model paper or developer-reported benchmark, E3 when it comes from a platform or vendor deployment report, and E4 when it reflects design intention or theoretical expectation. The hierarchy qualifies the strength of a claim and does not rank overall model quality. Throughout Sections 4–6, “reported mechanism” refers to source-documented design, “developer-reported effect” to E2 benchmark evidence, “platform outcome” to E3 deployment material, and “theoretical expectation” to E4 reasoning. “Independently demonstrated” is reserved for E1 evidence. No Agentar-Fin-R1 production-effectiveness claim is classified E1.

3. Literature Review

3.1. The current state of LLM applications in finance

In recent years, the application of LLMs in finance has shifted from feasibility studies to deeper integration into business

operations. Two directions are prominent. The first focuses on workflow integration, embedding LLMs into investment research, internal operations, and customer-facing services to improve efficiency through process optimization. For instance, Fabozzi and López de Prado [5] placed foundation models within the asset management workflow and showed that model value depends on governance integration, institutional process fit, and performance.

The second direction involves the use of LLMs in core financial activities and decision-making. In this process, LLMs no longer function merely as conversation or writing tools; they serve as modeling tools for investment modeling, market forecasting, and financial sentiment analysis [6]. LLMs can extract market-related signals from massive financial texts and provide information for trading decisions [7, 8], but their effectiveness depends on data quality, market conditions, and strategy design.

Dialogue-based investment advisory services for individual investors provided by large models have also attracted research attention. Schneider and Yilmaz [9] demonstrated that LLMs can, to a certain extent, generate portfolio suggestions for investors with different risk preferences and adjust the risk-return characteristics of the portfolios. However, the stability of the model still has a significant impact on the results. This also means that once the general-purpose model is applied to actual decision-making, the issues of output stability, appropriateness, and accountability that it faces become even more urgent.

In finance, LLMs are moving from peripheral workflow support toward core institutional and retail decision processes. Research consequently places greater emphasis on reliability, auditability, and governance in high-risk financial environments [10, 11].

3.2. Limitations of general-purpose LLMs in high-risk domains

Scholars have conducted in-depth discussions on the applicability of general-purpose LLMs in high-risk fields. The core focus is whether general-purpose large models can still be safely and reliably used under conditions where the cost of errors is high, reasoning requires evidence and traceability, and responsibility cannot be left ambiguous. This issue is particularly important in the financial sector.

Research in other high-risk fields identifies three recurring concerns. The first is reliability: LLMs may amplify uncertainty in source information while obscuring their own assumptions and reasoning limits [18, 19]. The second is explanation and trust: Users may place excessive confidence in fluent outputs even when the model provides limited grounds for its conclusions [1, 2]. The third is accountability: Once a model enters a decision process, supervisory roles and responsibility for harmful outcomes become harder to specify [1, 2].

Financial decisions have low tolerance for errors, are often made under conditions of information asymmetry, and are subject to strict regulation and accountability. Under such circumstances, the plausible outputs of general-purpose large models are far from sufficient. Users need stable reasoning, transparent justification, and clear accountability for decision outcomes. Recent research indicates that general-purpose models show certain potential in information processing and decision support, but there are still deficiencies in repeatability and method consistency [1]. In more specific financial tasks, such as investor risk analysis, studies have shown that even if the model outputs are close to expert judgments in some cases, they still often fall

short of the transparency and accountability standards expected in financial advisory and risk management settings [20]. Frequent errors, inconsistencies, and insufficient explanations have greatly limited the usability of general-purpose models in high-risk decision-making environments [20, 21].

The limitations of general-purpose LLMs in high-risk environments cannot be solely attributed to performance deficiencies. Essentially, they reflect the mismatch between the structure and training goals of general-purpose large models and the requirements of high-risk applications such as finance. Specifically, there is a contradiction between flexible model outputs and the strict requirements for reliability, interpretability, and governance. This is the starting point for the discussion of vertical financial large models below.

3.3. Related research on hallucination, multimodal reasoning, and XAI

3.3.1. Hallucination

In finance, the consequences of hallucinations undermine the credibility of model outputs, complicate responsibility tracing, and pose compliance risks.

At the conceptual level, van Deemter [22] points out that the classification of hallucinations in the field of natural language generation has long mixed issues such as factual inconsistency, unsupported inferences, and information omission. This conceptual ambiguity is itself a risk. In financial decision-making scenarios, the outputs often need to be tied to verifiable data, clear institutional rules, and explainable reasoning processes. The unclear boundaries of the hallucination problem make the assessment and governance of the model outputs more difficult.

Leippold [11] shows that general-purpose models can produce fluent and persuasive responses on climate-finance questions while relying on weak factual foundations. Dahl et al. [23] similarly find substantial hallucination in legal tasks and limited self-detection of errors. These findings matter for finance because contracts, regulation, and legal interpretation form part of many financial decisions. The principal risk lies in confident answers whose supporting evidence cannot be verified.

Recent studies have also viewed the hallucination problem from the perspective of uncertainty. Vashurin et al. [24] showed that LLMs often output highly confident answers even under high uncertainty conditions. Without interpretable confidence signals, human decision-makers would have difficulty effectively supervising model behavior. In financial scenarios, post hoc auditing and responsibility attribution have involved a large number of institutional requirements. Such overconfidence will further increase the cost of control.

Regarding hallucination governance, existing studies have proposed multiple mitigation paths, including uncertainty-aware generation and automatic correction mechanisms. Pan et al. [25] reviewed feedback-based automatic correction methods, indicating that these techniques can, to some extent, reduce the frequency of hallucinations. However, the literature also shows that when models lack structured financial knowledge and specific domain constraints, these methods only alleviate superficial symptoms and are difficult to address the problem.

Therefore, the hallucination problem of general-purpose large models should be regarded as a structural risk. Simply relying on prompt engineering and post-correction is difficult to meet the reliability requirements of high-risk financial scenarios. This

is precisely why we need to design vertical models specifically for financial scenarios.

3.3.2. Cognitive and security challenges of multimodal data integration

In the financial environment, risk signals rarely appear solely in textual form. Decisions increasingly rely on multimodal inputs, including digital time series, table disclosures, charts, images, audio-visual content, and heterogeneous metadata from multiple sources. Studies on multimodal large language models (MLLMs) have shown that combining LLMs with multimodal encoders may bring about “emergent capabilities” beyond traditional pipelines, such as image-driven reasoning [26]. However, in high-risk financial environments, this capability also introduces another issue: Can the model achieve cross-modal semantic alignment, stable reasoning, and accurate evaluation and verification of results under time pressure and strict accountability requirements?

The first problem lies in cross-modal alignment. Different information forms must be coordinated in a common representation before joint reasoning can occur, and performance depends on the alignment method [4]. In finance, alignment failure may directly distort risk interpretation—for example, by misreading chart trends, confusing table footnotes, or overemphasizing salient visual cues while overlooking quantitative signals. Subsequent reasoning then rests on an unstable evidence base.

A further difficulty is that multimodal reasoning ability may vary. Visual-language models are typically good at perception-level tasks but are insufficient in scientific reasoning [27]. In financial environments, this gap is significant because multimodal inputs often only present the visible surface of the problem, while the true task lies at a deeper level—causal attribution, scenario stress testing, liquidity spirals, or the assessment of risk contagion. Therefore, although the large models appear reliable at the perceptual level, their reasoning mechanisms can be fragile, potentially leading to unreasonable judgments.

The third problem lies in the evaluation of multimodal systems. Existing benchmarks have broadened multimodal evaluation, but the gap between deployment and verification is still significant. For example, TinyLVLM-eHub assesses the model from multiple dimensions (including perception, knowledge, reasoning, common sense, and object hallucination), and the results show that even powerful closed-source models are still vulnerable in important dimensions [28]. For financial institutions, this means that model performance cannot be simplified to a single indicator. It is important for the evaluation framework to be able to capture changes in distribution under real decision-making conditions and compliance error patterns.

The black-box nature of multimodal reasoning further exacerbates the accountability challenge. Some studies have attempted to enhance the interpretability of models by translating the reasoning process of multimodal models into interpretable rule structures [29], but this method is still in its early stages and has not yet become a standard for deploying systems. In the financial domain, decisions often require traceable reasons, and this black-box attribute adds to the governance burden.

Therefore, in the framework of general-purpose large models in finance, multimodal integration functions as both a capability enhancement and a source of structural risk. As the input space becomes richer, the risks of semantic misalignment, weak reasoning, and limited accountability also increase [4, 26]. This explains why financial large models need task-constrained

multimodal structured approaches and Trust-by-Design mechanisms, with capability expansion placed under explicit constraints.

3.3.3. Research on XAI

In finance, explainability is closely related to whether large models can be applied in practice. Having only one seemingly correct result is not enough; institutions also need to know the core basis for obtaining this result, as well as whether the reasons behind it can be verified and audited. This means that explainability is no longer an optional add-on and gradually becomes a core requirement for trustworthy AI [30, 31].

Current research on LLMs has proposed different ways of thinking about explainability. In text-level explanation studies, Zhao et al. [32] distinguished explanation methods through training paradigms and explanation goals. They also demonstrated that explainability has cognitive, communication, debugging, error-diagnosis, and system-improvement functions. Recent research has also focused on the role of visualization in enhancing the explainability of foundation models. Studies have shown that visualization analysis can help understand, evaluate, and improve complex models, especially in enhancing transparency, fairness, and robustness [33]. For financial applications, the explainability of large models helps users understand results, identify reasoning failures, and reduce operational risks during deployment.

In multimodal foundation models, explainability becomes more complex. Kazmierczak et al. [34] noted that multimodal foundation models are difficult to explain because of their structural complexity, while they are also used to build systems with explanation capabilities. Therefore, explainability should be treated as a model-system attribute that encompasses model architecture design, explanation interfaces, and evaluation mechanisms.

Financial decision-makers need to assess an output, understand how it was formed, identify its supporting evidence, and review the process when disputes, losses, or regulatory questions arise. Explainability links model outputs with verifiable evidence, supports human supervision, and facilitates audit and governance. It should therefore be treated as a system-design attribute in financial AI.

3.3.4. Developments in the literature on financial regulation and trustworthy AI

The financial industry is applying AI to high-risk decision-making tasks, including credit scoring, fraud detection, AML monitoring, and risk management. The use of AI has changed the risk exposure and compliance responsibilities of financial institutions, so regulatory attention increasingly centers on system trustworthiness: Whether the system is safe, robust, fair, privacy-preserving, explainable, auditable, and accountable. When models enter high-risk scenarios such as finance, trustworthiness must be a fundamental requirement throughout [1, 30].

A major advance in the literature is the conversion of broad ethical principles into lifecycle requirements. From data collection and pre-training through deployment, monitoring, and updating, trustworthy-AI research specifies requirements and corresponding verification methods [1]. Huang et al. [3] classify these methods as evaluation and verification, formal verification, runtime monitoring, and normative governance. This sequence aligns with regulatory practice: Requirements are defined, evidence is generated, drift and new vulnerabilities are monitored, and accountable records are retained.

The literature also asks how trustworthiness can be measured. Giudici and Raffinetti [35] propose the Sustainability,

Accuracy, Fairness, and Explainability (SAFE) AI framework for financial applications, while Giudici et al. [36] develop key AI risk indicators that translate regulatory principles into measurable indicators. These approaches connect governance requirements with evidence and continuous monitoring. They also place sustainability within financial AI assessment by directing attention to durable controls, proportional resource use, inclusion, and institutional resilience.

Trustworthy AI depends on model attributes, governance structures, and institutional control. Bartsch et al. [37] pointed out the direction of high-risk AI governance: responsibility allocation, transparency, auditability, and human supervision. McIntosh et al. [38] compared the adaptability of various cybersecurity frameworks in the commercialization of LLMs and clarified that existing frameworks still need to evolve to cover the multi-faceted risks of LLMs. Trustworthy AI is embedded in enterprise controls, process documents, continuous verification and auditing, and assurance processes.

The issues of law and liability attribution have emerged. Hacker [39] emphasizes that high-risk regulation, interpretability, and compensation mechanisms should be more closely linked, and discusses the regulatory boundaries of general/basic models in specific high-risk domains. Malgieri and Pasquale [40] advocate strengthening ex-ante obligations before model deployment and clarifying the specific responsibilities of developers. For finance, trustworthy AI must function within the framework, and when disputes or losses occur, responsibilities can be divided and borne.

Taken together, this literature defines trustworthy financial AI through reliable operation, fairness and privacy protection, necessary transparency, and lifecycle accountability. These attributes also support sustainable financial innovation: Durable governance arrangements can reduce recurring model and compliance risk, responsible scaling can align computational resources with decision value, and inclusive, resilient institutions can maintain service quality as AI use expands.

3.4. Critical synthesis

The reviewed literature provides substantial knowledge on individual components of financial LLM design. Application studies document uses ranging from sentiment analysis to portfolio support [5, 7, 9]. Hallucination and multimodal research identify recurring failure mechanisms and partial remedies [4, 22, 26, 41]. XAI and trustworthy AI studies develop taxonomies, indicators, and lifecycle requirements [1, 3, 32, 35]. The study-level evidence map shows that these lines of research converge on three recurring design problems: controlling financial knowledge and evidence, coordinating heterogeneous inputs within a task, and governing model use throughout the lifecycle. Second-cycle coding converts these problem groups into the three analytical pathways. Tables S2 and S3 make the derivation visible. The same synthesis links technical design to sustainable financial innovation through durable governance, responsible scaling, inclusion, and institutional resilience. Table 1 summarizes this critical synthesis and the framework derivation that follows.

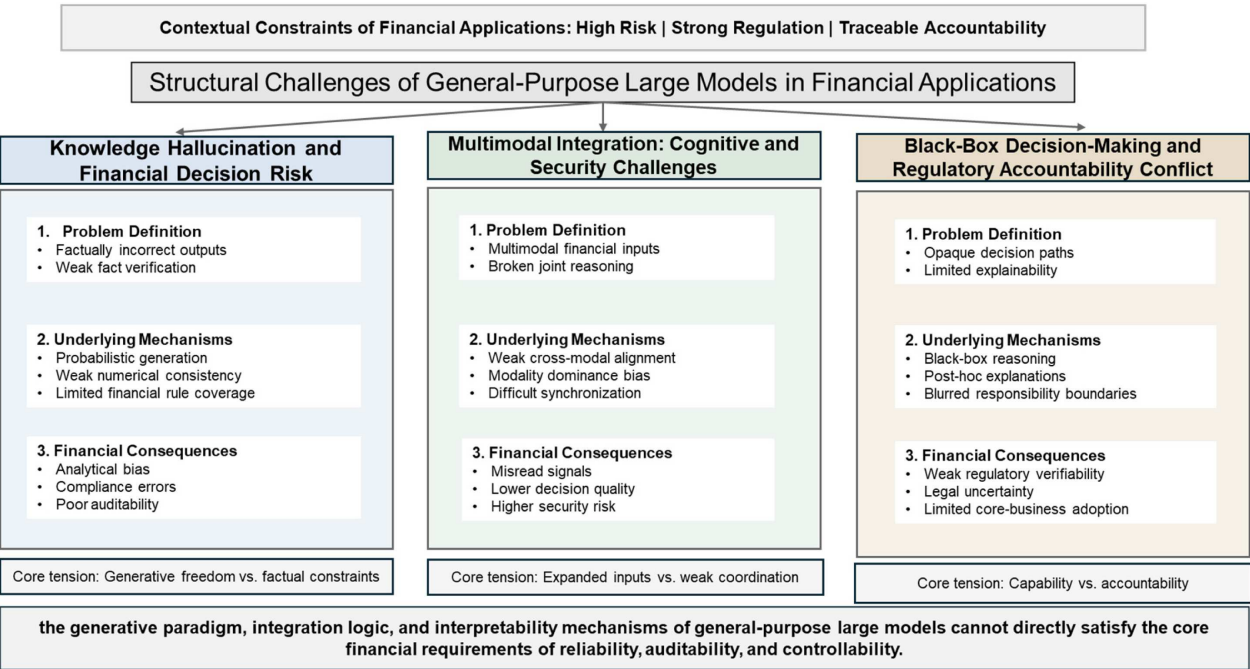
4. Structural Challenges of General-Purpose Large Models in Financial Applications

Figure 4 summarizes the three challenges that arise when general-purpose large models are applied to financial scenarios: knowledge hallucination, multimodal integration and security challenges, and black-box decision-making under accountability

Table 1
Critical synthesis and framework derivation

Theme	Critical synthesis and remaining gap
Applications	Studies on workflow integration, sentiment analysis, and portfolio advisory show practical promise but give limited attention to auditability and accountability in high-risk financial use [5, 7, 9].
Hallucination	Research clarifies factual error types, retrieval-augmented generation-based mitigation, and citation risks, while finance-specific constraints and evidence requirements remain insufficiently specified [22, 41, 42].
Multimodal	MLLM studies identify alignment and reasoning challenges across text, tables, charts, and images. Finance-specific benchmark coverage remains limited [4, 26, 43].
XAI	XAI work provides taxonomies and system-level explainability concepts, while regulatory traceability in financial decision processes remains underdeveloped [32, 34, 44].
Governance	Trustworthy-AI frameworks define lifecycle requirements and measurable indicators, while links from governance requirements to model architecture remain weak [1, 3, 35].
This study	The 67-study evidence map and second-cycle coding connect knowledge-reliability studies to Pathway 1, multimodal studies to Pathway 2, and XAI/governance studies to Pathway 3. Sustainability implications are assessed through durable governance, responsible scaling, inclusion, and institutional resilience (Tables S2–S3). Sources outside the 67-study corpus and their use restrictions are listed in Table S4.

Figure 4
Structural challenges of general-purpose large models in financial applications



Note: This figure identifies three challenges that limit reliable use of general-purpose LLMs in high-risk financial applications: Knowledge hallucination, multimodal integration and security, and opaque decision-making under regulatory accountability. Each challenge is examined through its definition, underlying mechanisms, and financial consequences. The bottom box synthesizes these findings into the analytical motivation for Section 5.

pressure. The figure explains the problems, mechanisms, and financial consequences associated with these issues. These three challenges reflect a deeper mismatch between general-purpose modeling approaches and the requirements of financial scenarios. The following sections examine these challenges in turn.

4.1. Knowledge hallucination and financial decision risk

Hallucination denotes fluent model output that lacks factual support or conflicts with available evidence. One source of risk is probabilistic next-token generation combined with the absence of

built-in factual verification; data coverage, prompting, and task conditions also affect error rates [3, 32]. In finance, the consequences can be serious because decisions often depend on precise numerical information, rule interpretation, and compliance judgments. Task-specific studies report fictitious citations and other factual errors in finance-related generation tasks [45]. These findings establish risk in the evaluated tasks without implying a uniform hallucination rate across all general-purpose models. The risk of unreliable output becomes more salient when models work with structured financial information, especially tables and mixed text–table documents [46]. Financial statements, disclosure reports, and regulatory materials require precise

information extraction, numerical alignment, and cross-table calculation. Many general-purpose models lack dedicated mechanisms for these operations and for rule-based verification. Evaluations on financial document question answering report errors in numerical consistency and multi-step reasoning [46]. In operational use, such errors could contribute to valuation, risk, or compliance mistakes; the magnitude depends on task design and human oversight.

Finance also contains small-sample, high-heterogeneity tasks, including credit assessment in specialized industries and pricing of structured products. A domain-adaptation study using financial terminology reports limited transfer from general corpora under specialized conditions [47]. This evidence supports the expectation that domain knowledge and business-rule constraints may be needed in some financial tasks; it does not quantify their effect on hallucination across applications.

Taken together, hallucination risk and limited transfer to specialized financial tasks undermine accuracy, auditability, and risk control. These problems motivate the knowledge-modeling pathway developed in Section 5.

4.2. Cognitive and security challenges of multimodal data integration

Financial decision-making often requires joint interpretation of text, structured financial data, market charts, and qualitative signals such as management tone [48, 49]. General-purpose models perform strongly on many single-modal tasks, yet their perception-to-inference process remains unstable in multimodal financial tasks [43].

Cross-modal integration creates risks of semantic misalignment and modality bias [3, 50]. A model may rely on one dominant modality, overlook signals from another, or form inconsistent associations between text, charts, and numerical data. In finance, these failures could produce inconsistent use of risk warnings, trend reversals, or qualitative management signals. The extent of this problem remains task- and benchmark-dependent.

The technical challenge arises from heterogeneous data representations and the need to maintain alignment over time. Different modalities use different encoding mechanisms, while financial judgments require consistency in timing, semantics, and decision weights. In models using static or weakly adaptive fusion, reasoning quality may decline when modality relevance changes [12, 51]. This relationship remains a theoretical transfer to finance until tested on financial multimodal benchmarks. Cross-modal errors can also be difficult to trace, increasing the burden of review and accountability [52].

Data privacy creates an additional requirement. Financial information is sensitive and subject to privacy and compliance rules. Multimodal processing expands the types of information handled and the number of processing steps. Sending business secrets or customer data to a cloud-based general-purpose model can increase exposure to leakage and noncompliance; the realized risk depends on deployment architecture and control measures [13, 53].

Cognitive alignment and security issues limit the application of multimodal large models in finance.

4.3. Black-box decision-making and regulatory accountability

The operation of the financial system is highly dependent on explainability, traceability, and accountability in the decision-making process. Whether the task involves credit approval, risk rating, AML, or compliance review, financial institutions need

to explain why a decision was made. However, the algorithmic black-box characteristic of general-purpose large models makes output explanation and traceability difficult. When the model is involved in high-risk decisions such as loan approval, risk rating, or compliance review, the absence of a verifiable decision basis may infringe consumers' right to know and expose financial institutions to significant compliance and legal risks.

XAI seeks to address this problem, yet many existing methods provide post hoc approximations through feature attribution or attention visualization. Such explanations do not automatically create a reproducible and auditable decision record [54]. The literature also notes that larger and more capable models frequently become harder to interpret [32, 34, 44], which limits the direct regulatory value of post hoc explanation alone.

Even more seriously, the problem of algorithmic black boxes has further triggered a systemic accountability dilemma. If the training data contains historical biases or structural unfairness, then in the absence of clear rules for constraints, large models may replicate or even amplify it, thereby infringing upon fairness and financial inclusiveness [55]. Once such decisions suffer losses, the responsibility is even more difficult to handle. Who should bear the responsibility—the model developers, the deployment institutions, or the business users—is often difficult to clearly define, forming what is called an “accountability vacuum” [14, 56].

The interaction between opaque decision processes and regulatory accountability therefore remains a structural obstacle to deeper deployment of general-purpose LLMs in finance.

5. Analytical Design Pathways for Domain-Specific Financial LLMs

5.1. Specialized knowledge modeling and hallucination mitigation mechanisms

In finance, hallucination risk reflects the gap between general training objectives and specialized, rule-based financial knowledge. The literature identifies three design principles for managing this risk: domain-specific corpus anchoring, task-level generation controls, and verification mechanisms that continue from training into deployment [41, 57–59]. These mechanisms provide a design rationale; their presence does not by itself establish a lower hallucination rate in production.

Task constraints narrow the model's generation space. Even a specialized model can produce plausible but unverifiable conclusions when its reasoning conditions are unclear. Structured outputs, evidence binding, terminology controls, and retrieval-augmented generation can improve grounding when evidence sources are identifiable and locatable [41, 60]. The resulting reliability gain remains an empirical question for the financial task and benchmark under study.

Training and verification practices support longer-term control of hallucination risk. Domain-adaptive pre-training and SFT can strengthen financial terminology and reasoning patterns [61], while annotated reasoning data may improve output consistency [62]. Deployment still requires citation checking, error feedback, and updating procedures [42]. A complete pathway therefore combines domain learning with output verification and ongoing monitoring.

5.2. Multimodal structured modeling for financial tasks

Financial decisions draw on text, structured tables, charts, and unstructured signals. The central challenge for multimodal

models in finance is to organize and jointly reason over these inputs within a defined task while preserving evidence consistency.

At the task-context level, multimodal modeling requires a clear definition of the financial decision and the inputs relevant to it. Song et al. [4] identify a persistent semantic gap across modalities and show that alignment choices depend on the application. Investment research, risk assessment, and compliance review place different demands on text, numbers, tables, and charts. Selective modeling around a defined task can therefore reduce irrelevant inputs and make the reasoning process more transparent.

At the task-decomposition level, explicit sub-tasks can make multimodal reasoning easier to inspect. Wang et al. [63] report more robust stepwise reasoning in visual question answering through a knowledge-guided mixture-of-experts framework. The result provides a transferable design hypothesis for finance. Financial-statement interpretation and macroeconomic analysis would still require task-specific tests of extraction, local judgment, and synthesis steps.

At the modality-coordination level, Li et al. [64] report reduced cross-modal bias and improved generalization in their Uni-MoE benchmark through modality-specific encoders and sparse expert routing. Transfer to financial data remains a theoretical expectation. Xia et al. [65] also report failures by advanced multimodal models on structured chart-reasoning tasks. These findings support evaluating dedicated chart and table pathways; they do not demonstrate improved financial decisions.

The resulting pathway is a structured multimodal design organized around the application context, explicit task decomposition, and coordinated modality roles. It provides a basis for evaluating whether a financial LLM can combine heterogeneous evidence without allowing one modality or one reasoning step to dominate the decision.

5.3. Trust-by-Design for lifecycle governance

Trust-by-Design is treated here as a design proposition in which trustworthiness requirements are specified across the system lifecycle. Reliability, security, fairness, privacy, transparency, and accountability are translated into engineering requirements, evaluation procedures, and governance responsibilities [1, 2]. The existence of these controls documents governance design; outcome effectiveness requires separate evaluation.

Song et al. [1] systematically reviewed the trust requirements of LLMs in practical deployment and decomposed them into multiple dimensions, including reliability, security, fairness, privacy protection, transparency, and accountability. The study pointed out that trust requirements exist continuously across stages such as data collection, model training, system integration, and operation maintenance. Lu et al. [2] further argued from a system-architecture perspective that foundation models have gradually become the core of deployed generative-AI systems. Trust therefore needs to be embedded into model invocation, tool coordination, log recording, and human-computer interaction through architectural decisions, with post-audit and manual intervention functioning as additional controls.

5.3.1. Explainability

In high-risk fields such as finance, explainability should mean that the decision-making process has systematic traceability, auditability, and verifiability. Therefore, responsible base model systems should support the recording of reasoning paths, the association of evidence, and the replay of the decision-making process at the architectural level, so that the model output can be understood and verified by external actors [2].

From the perspective of verification and validation (V&V), explainability is one of the key prerequisites for achieving trustworthy AI. Only when the model behavior can be systematically analyzed, replicated, and refuted can the assessment of security, fairness, and compliance have a realistic basis [3]. Therefore, in the Trust-by-Design framework, explainability should be regarded as a systematic engineering capability.

5.3.2. Compliance

Compliance is a necessary condition for trustworthy AI. Song et al. [1] identify compliance and privacy protection as core requirements of foundation-model systems, with data governance and output control serving as principal mechanisms. Zhang [66], using China's Personal Information Protection Law as an example, explains why training and application stages require differentiated controls. This lifecycle perspective is consistent with Trust-by-Design.

5.3.3. Accountability and governance

When foundational models deeply participate in the decision-making process, the traditional human-centered responsibility allocation model will face new challenges. Liu et al. [67], from the perspective of governance structure, summarized the core governance issues of the foundational model system into three dimensions: decision-making power, incentive mechanism, and responsibility attribution, and believed that technology architecture should be utilized to enhance the traceability of responsibility. Lu et al. [2] also emphasized that the responsibility boundaries between models, tools, and human decision-makers should be clearly defined at the system level to avoid a responsibility vacuum.

At the operational level, continuous risk assessment and adversarial testing form part of governance. Jabbar et al. [68] argue that red-teaming can reveal vulnerabilities in security, bias, and manipulation, while Dong et al. [69] emphasize guardrails, runtime monitoring, and incident response across the model lifecycle.

Trust-by-Design therefore converts trust requirements into architecture, evaluation, operating controls, and responsibility allocation. In financial applications, it provides a common basis for examining how explainability, compliance, and accountability are specified before and after deployment.

6. Comparative Case Illustration: Agentar-Fin-R1

Section 6 applies the framework to a documentary case and a comparator set. Four evidence categories are kept separate: model-paper design characteristics and developer-reported benchmarks (E2), platform deployment reports (E3), and prospective applications inferred from the framework (E4). No E1 independent production evaluation of Agentar-Fin-R1 was identified.

The Agentar-Fin-R1 case illustrates the framework; it does not validate the framework or establish independently observed production performance.

Ant Digital Technologies released Agentar-Fin-R1 in 2025 [15, 16]. Its model paper describes a domain-specific knowledge system, task-oriented reasoning mechanisms, and a trustworthy-AI engineering approach. The following analysis reports what the source documents specify and labels every claim by evidence level.

Agentar-Fin-R1 can be assessed through the three analytical pathways: specialized knowledge modeling, task-constrained multimodal reasoning, and Trust-by-Design. The assessment distinguishes reported mechanisms from observed outcomes and keeps model-level and platform-level evidence separate.

Figure 5 presents a conceptual case representation based on public design descriptions. It shows where model-paper mechanisms, platform functions, and prospective financial applications enter the analytical framework.

6.1. Specialized knowledge modeling and hallucination mitigation

As discussed in Section 4.1, financial hallucination risk can arise when general-purpose corpora and generation objectives do not capture specialized financial rules and evidence requirements. The Agentar-Fin-R1 model paper reports a specialized financial corpus and a multi-stage training strategy intended to strengthen domain knowledge [15]. These descriptions are classified as E2 evidence.

At the corpus level, the model paper reports the use of regulatory documents, brokerage research, financial news, and academic literature. This controlled corpus is intended to stabilize terminology and improve domain coverage. It does not eliminate the possibility of confident factual error, which requires direct benchmark and deployment testing.

At the training level, the model paper reports a two-stage SFT and GRPO process. SFT develops financial terminology and rule knowledge, while the subsequent stage targets complex reasoning. A difficulty-aware weighting mechanism is designed to allocate training effort across tasks of different complexity [15].

These mechanisms indicate a combined strategy of domain learning and reasoning control. Their effect on hallucination is supported only by developer-reported design and benchmark evidence in the available source; independent financial-task and production evaluation remains necessary.

6.2. Task-structured financial reasoning and partial multimodal coverage

The Agentar-Fin-R1 model paper reports a dual-dimensional “scenario × task” labeling system that organizes training and

evaluation records by financial context and task type [15]. At the context level, banking, securities, insurance, trusts, and mutual funds are treated as distinct domains. At the task level, operations such as named entity recognition, intent classification, slot filling, and consultation-style question answering are represented as explicit categories. The paper also reports heterogeneous-format normalization and task-specific difficulty weighting. These mechanisms provide E2 evidence of task structuring and data organization.

The direct source does not specify a native multimodal architecture or an explicit alignment mechanism spanning text, tables, charts, and images. Scenario–task labeling organizes financial domains and operations, while data normalization and weighting operate on records and task categories. Agentar-Fin-R1 is therefore coded P for Pathway 2. A Met classification would require direct evidence of multiple financial input modalities and a documented cross-input coordination mechanism, followed by finance-specific multimodal evaluation.

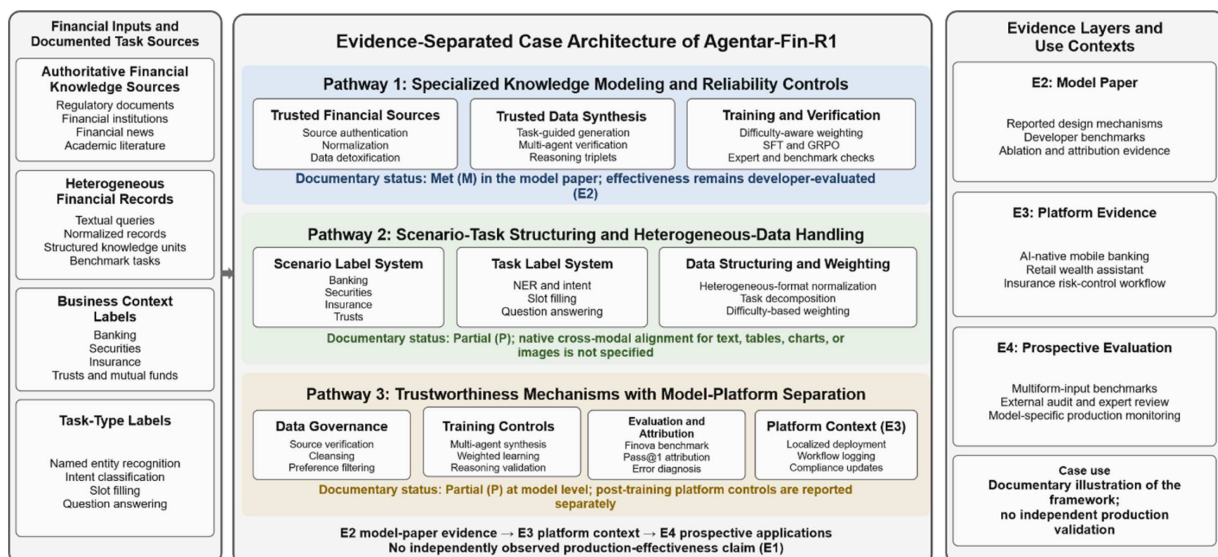
6.3. Trust-by-Design architecture

Trust-by-Design requires model and platform controls to be specified across data processing, training, evaluation, deployment, and monitoring. The relevant Agentar sources describe mechanisms at different system levels, which must be reported separately.

Zheng et al. [15] report model-level mechanisms such as data cleaning, multi-agent synthesis or checking, and dynamic attribution. Corporate and platform materials [16] and the footnoted 2025 financial-agent report describe localized deployment, workflow logging, compliance checking, regulatory updates, and continuous feedback at the broader Agentar platform level. Table 2 distinguishes these sources and evidence levels.

Table 2 classifies each mechanism by source and evidence level. It separates model-paper characteristics, developer-reported benchmarks, platform-level deployment functions, and prospective effects so that platform capabilities are not treated as verified model outcomes.

Figure 5
Conceptual case representation of Agentar-Fin-R1 and the Agentar platform



Note: Conceptual representation based on publicly reported design. Model-paper mechanisms, platform-level mechanisms, and prospective applications are shown as separate evidence layers. The figure does not constitute independent architectural or performance validation.

Table 2
Evidence-level classification of Agentar-Fin-R1 and Agentar platform mechanisms

Dimension	Evidence source and level	Documented mechanism	Interpretation
Reliability and robustness	Agentar-Fin-R1 model paper [15]; E2	Two-stage SFT/GRPO, difficulty-aware weighting, multi-agent synthesis or checking, and expert-annotated reasoning data.	Developer-reported design and benchmark evidence; no independent production validation identified.
Transparency and explainability	Model paper [15], E2; Agentar platform report, E3	Dynamic attribution in model documentation; white-box workflow logic and logs at platform level.	Model and platform mechanisms are reported separately; audit effectiveness has not been externally verified.
Fairness and inclusion	Model paper [15], E2; prospective application, E4	Data detoxification and debiasing; diversified financial corpus.	The design may support bias control, while subgroup fairness and inclusion outcomes remain unevaluated.
Privacy and security	Model paper [15], E2; platform materials [16] and the footnoted 2025 financial-agent report, E3	Safety and compliance evaluation; private or localized deployment at platform level.	Security design and deployment options are documented; no independent production security audit is reported.
Accountability and governance	Platform report, E3; prospective model use, E4	Compliance checking, regulatory-update integration, feedback, and attribution functions at platform level.	Platform governance provides deployment context; model-specific accountability effects remain prospective.

The documentary record therefore indicates a governance-oriented design intention and an established platform context. It does not establish independent production effectiveness for Agentar-Fin-R1.

6.4. Evidence levels in model and platform deployment

Financial LLM capabilities enter business processes through platform interfaces, data connections, agents, and workflow orchestration. The Agentar platform supplies this deployment context, but platform outcomes do not isolate the contribution of Agentar-Fin-R1. The three platform cases presented in Sections 6.4.1–6.4.3 are therefore classified as E3 contextual evidence and are followed by separately labeled E4 prospective model applications.

Each subsection first reports the platform claim contained in the 2025 financial-agent report⁶ and then states how the Agentar-Fin-R1 design might be evaluated in a related task. No causal link from the model to the reported platform outcome is asserted.

6.4.1. Platform evidence on knowledge-intensive financial assistance

Platform-level evidence (E3). The 2025 financial-agent report⁶ describes Shanghai Bank's AI-native mobile-banking project and reports a 10-percentage-point increase in business conversion after the Agentar platform introduced conversational service. The source attributes the outcome to the platform-level project; it does not isolate Agentar-Fin-R1 and is not treated as model validation.

Prospective model application (E4). The reported two-stage training strategy and difficulty-aware weighting could be evaluated on long-tail, ambiguous, colloquial, or error-prone banking instructions. Relevant tests would compare factual accuracy,

evidence use, abstention, and expert review outcomes against general-purpose and financial-domain baselines.

6.4.2. Platform evidence on heterogeneous data integration

Platform-level evidence (E3). The report describes a retail wealth-management assistant that integrated customer profiles, market information, and product data and reports a 30% increase in service coverage per relationship manager. This metric documents platform-level workflow support and remains vendor-reported.

Prospective model application (E4). Agentar-Fin-R1's scenario–task labeling and reported modality weighting could be tested on combined financial reports, tables, charts, and policy documents. Evaluation should examine cross-modal consistency, numerical accuracy, source attribution, and whether modality-specific errors are visible to human reviewers.

6.4.3. Platform evidence on traceable risk-control workflows

Platform-level evidence (E3). The report describes an insurance risk-control application in which the Agentar platform parsed historical reports and supported model construction and strategy optimization, with a reported 30% improvement in modeling efficiency. The metric concerns the platform workflow and has not been independently verified.

Prospective model application (E4). Reported mechanisms such as data detoxification, multi-model checking, and dynamic attribution could support source-linked explanations and review logs in risk-control tasks. Their practical value should be tested through external audit, error analysis, and production monitoring rather than inferred from their existence.

Together, the three platform cases document deployment context and business-process maturity for the Agentar platform. Agentar-Fin-R1-specific production maturity remains unverified. The cases identify possible integration points without providing an independent estimate of model effect.

⁶Zero One Think Tank Research Group. (2025). *2025 report on deep applications of financial agents: Deployment practices across 100+ in-depth business scenarios*. <https://hulianhutongshequ.cn/upload/tank/report/2025/202509/1/152fda22743a4c6f890c93c108d81667.pdf>

The model–platform–business relationship is therefore treated as deployment context rather than evidence that the model improves a business outcome.

6.5. Comparative assessment of representative financial LLMs

The comparison applies the M/P/ND criteria in Section 2.5 to BloombergGPT, FinGPT, FinTral [17], Fin-o1, Fin-R1, and Agentar-Fin-R1 [15]. Each model is assessed on the three pathways, principal strength, principal limitation, evidence quality, and deployment maturity. Claims come from original model papers and are classified as E2 unless an additional source is identified. Because the models have not been tested on one

common benchmark and production dataset, the comparison does not establish an overall ranking. Table 3 reports the resulting cross-model assessment.

The revised comparison explicitly reports strengths, weaknesses, evidence quality, and deployment maturity in addition to architectural characteristics.

1) Pathway 1: Specialized knowledge modeling

BloombergGPT and FinGPT are coded P for Pathway 1 because their direct sources document financial data specialization without a separate output-control or verification mechanism. FinTral [17], Fin-o1, Fin-R1, and Agentar-Fin-R1 [15] are coded M because they combine domain data or financial reasoning data with preference or reward alignment, structured task control, or

Table 3
Standardized comparative assessment of representative financial LLMs

Model and direct source	Pathway status (M/P/ND), evidence, and maturity
BloombergGPT (Wu et al., 2023; arXiv:2303.17564)	P1—P: Proprietary financial pre-training corpus; the direct source does not document a separate output-control or verification mechanism. P2—ND: Text-centric. P3—ND: Lifecycle governance not systematically specified. Strength: Scale and domain coverage. Limitation: Closed data/model and no independent deployment evaluation. Evidence: E2 model paper. Maturity: Research model.
FinGPT (Yang et al., 2023; arXiv:2306.06031)	P1—P: Open financial data curation and Low-Rank Adaptation (LoRA) adaptation; no separate output-control or verification mechanism is documented. P2—ND: Text-centric. P3—ND: Lifecycle governance not systematic. Strength: Openness and adaptability. Limitation: Reliability depends on the data pipeline, and governance remains underspecified. Evidence: E2 model paper. Maturity: Research ecosystem; production maturity not established.
FinTral [17]	P1—M: Domain pre-training plus instruction and preference alignment. P2—M: Text, numerical, table, and image inputs within a native multimodal design. P3—P: Preference alignment covers one output stage without the required lifecycle breadth. Strength: Broad native modality coverage. Limitation: Lifecycle governance and independent deployment evidence remain limited. Evidence: E2 peer-reviewed model paper. Maturity: Research model.
Fin-o1 (paper title: Fino1; Qian et al., 2025; arXiv:2502.08127)	P1—M: Financial reasoning data plus SFT and GRPO provide domain anchoring and task-specific reward alignment. P2—P: Text and structured numerical inputs are used, while a distinct cross-modal coordination mechanism is not documented. P3—ND: Reasoning evaluation without lifecycle governance. Strength: Reasoning-focused training. Limitation: Modality scope and deployment governance are limited. Evidence: E2 preprint and developer benchmark. Maturity: Research model.
Fin-R1 (Liu et al., 2025; arXiv:2503.16252)	P1—M: Financial chain-of-thought data plus SFT and GRPO provide domain anchoring and task-specific reward alignment. P2—P: Text and structured numerical inputs are used, while a distinct cross-modal coordination mechanism is not documented. P3—ND: Reasoning evaluation without lifecycle governance. Strength: Open reasoning pipeline. Limitation: Production controls and lifecycle governance are not established. Evidence: E2 preprint. Maturity: Research model.
Agentar-Fin-R1 [15]	P1—M: Domain training and label-guided weighting are combined with multi-agent data synthesis and verification. P2—P: Scenario–task structuring, heterogeneous-format normalization, and task weighting are documented, while native cross-modal alignment is not. P3—P at model level: Data, training, evaluation, and attribution mechanisms are documented, but post-training deployment or monitoring coverage depends on E3 platform materials and does not upgrade model-level status. Strength: Detailed model-development and benchmark documentation. Limitation: Native multimodal architecture and independent production validation are not established. Evidence: E2 model paper plus E3 platform context. Maturity: Platform deployed; model-specific production maturity unverified.

Note: M = All minimum components documented in a direct model source; P = Partial coverage or dependence on platform/vendor material for a required component; ND = Qualifying mechanism not documented in the direct source. Status records documentary coverage and does not indicate effectiveness. The first column provides a direct author–year and reference or arXiv identifier for every model. E2 denotes an original model paper or developer benchmark; E3 denotes platform or vendor deployment evidence. E3 evidence is reported as context and does not upgrade a model-level status. Deployment maturity refers to evidence documented for the named model, not to the commercial maturity of an affiliated platform.

reported checking. These statuses record documentary mechanism coverage. None of the source papers alone establishes a general reduction in production hallucination.

2) Pathway 2: Multimodal reasoning

BloombergGPT and FinGPT are coded ND for Pathway 2 because their source papers are text-centric. Fin-o1 and Fin-R1 are coded P because they combine text with structured numerical inputs without documenting a distinct cross-modal coordination mechanism. FinTral is coded M for native processing of text, numbers, tables, and images [17]. Agentar-Fin-R1 is coded P because its model paper documents scenario-task structuring, heterogeneous-format normalization, and task weighting, but does not specify native cross-modal alignment across text, tables, charts, or images [15].

3) Pathway 3: Trust-by-Design

BloombergGPT, FinGPT, Fin-o1, and Fin-R1 are coded ND for Pathway 3 because their original papers do not specify the required lifecycle governance coverage. FinTral is coded P because preference alignment addresses one output stage [17]. Agentar-Fin-R1 is also coded P at model level: The model paper documents several governance functions, while fuller lifecycle coverage depends partly on E3 platform materials. The additional platform evidence is reported as context and does not upgrade the model-level status.

The comparison reveals distinct areas of specialization rather than a linear progression. BloombergGPT and FinGPT are strongest in data strategy, FinTral in native multimodality, and Fin-o1 and Fin-R1 in reasoning-oriented training. Agentar-Fin-R1 meets the documentary threshold for Pathway 1 and has partial Pathway 2 and Pathway 3 coverage at model level; the broader Agentar platform supplies additional E3 governance context. Evidence quality remains model paper based for all six models, and model-specific production maturity is generally unverified. The framework is therefore most useful for locating design emphasis, partial coverage, and evidentiary gaps.

7. Discussion and Conclusion

7.1. Theoretical and practical implications

This paper asked how financial LLMs can move from general-purpose architectures toward domain-specific designs that meet professional and regulatory requirements. The 67-study synthesis and six-model comparison organize the documentary evidence through knowledge controls, heterogeneous-input reasoning, and lifecycle governance.

Agentar-Fin-R1 serves as a documentary illustration of the framework. It does not validate the framework or establish independently observed production performance.

For financial institutions, the framework can inform model procurement, internal evaluation, and governance design. Benchmark accuracy can be considered together with task fit, evidence grounding, modality handling, audit records, and the source level of performance claims. Domain specialization and task constraints provide design mechanisms; measurable improvements require independent benchmarks, expert review, and production monitoring. Organizational responsibility for model-assisted decisions should be assigned before deployment.

From a technological perspective, the three pathways form a qualitative analytical and diagnostic lens. The M/P/ND status

indicates where direct source documentation meets the minimum criteria, where coverage is partial, and where qualifying mechanisms are not documented. It does not measure effectiveness. The E1–E4 hierarchy separately records the strength and source of performance or mechanism claims.

From a regulatory perspective, the analysis is consistent with lifecycle governance from data collection through deployment monitoring [1, 35, 36]. It also clarifies the sustainability dimension of financial AI. Sustainable financial innovation depends on governance arrangements that remain effective as systems scale, control recurring model and compliance risk, align computational resources with decision value, preserve access and fairness, and strengthen institutional resilience. The framework identifies where these objectives require model-level mechanisms, platform controls, and independent measurement.

7.2. Contributions, evidence limits, and research directions

The paper contributes in three ways. First, it provides a study-level evidence map and code-to-pathway account that makes the derivation of the framework transparent. Second, it operationalizes specialized knowledge modeling, task-constrained multimodal reasoning, and Trust-by-Design as a qualitative analytical and diagnostic framework with common indicators. Third, it compares six financial LLMs through an evidence hierarchy that separates model design, developer benchmarks, platform deployment reports, and prospective applications, while connecting these choices to sustainable financial innovation.

The Agentar-Fin-R1 case illustrates, rather than validates, the framework.

The evidence base remains documentary. Agentar-Fin-R1-specific claims originate from an original model paper and company or platform materials, and the study does not include shared experimental benchmarks or production data. The revised evidence hierarchy and comparative tables make this condition explicit and prevent claims from being transferred across model, platform, and prospective application levels.

Future research can implement the prospective protocol in Supplementary Section S1, apply the indicators to a larger model set, and test the pathways on shared financial benchmarks, expert review, and independently observed deployment data. Quantitative indicators can then be linked with financial AI governance standards and with measured outcomes in resource use, inclusion, compliance performance, and institutional resilience.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Tables S1–S4 in the Supplementary Materials provide the prospective replication protocol, the study-level evidence map for the 67-study corpus, the code-to-pathway derivation and threshold rationale, and the register of seven sources outside the review corpus. The original query-level exports, source-specific yields, and coder-level screening logs were not retained. Data sharing

is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Fumin Niu: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Fengshuo Liu:** Conceptualization, Investigation, Writing – original draft. **Zekang Zhang:** Methodology, Writing – original draft. **Yishi Li:** Writing – original draft. **Yufei Ren:** Supervision, Project administration.

Supplementary Materials

The supplementary materials are available from the article webpage at <https://doi.org/10.47852/bonviewFSI62029857>.

References

- [1] Song, P., Ojo, A., & Curry, E. (2026). Trustworthy requirements for foundation models—A comprehensive survey and roadmap. *Engineering Applications of Artificial Intelligence*, 163, 113111. <https://doi.org/10.1016/j.engappai.2025.113111>
- [2] Lu, Q., Zhu, L., Xu, X., Xing, Z., & Whittle, J. (2024). Toward responsible AI in the era of generative AI: A reference architecture for designing foundation model-based systems. *IEEE Software*, 41(6), 91–100. <https://doi.org/10.1109/MS.2024.3406333>
- [3] Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., . . . , & Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7), 175. <https://doi.org/10.1007/s10462-024-10824-0>
- [4] Song, S., Li, X., Li, S., Zhao, S., Yu, J., Ma, J., . . . , & Mao, X. (2025). How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9), 5311–5329. <https://doi.org/10.1109/TKDE.2025.3527978>
- [5] Fabozzi, F. A., & López de Prado, M. (2025). Implementing AI foundation models in asset management: A practical guide. *The Journal of Portfolio Management*, 52(2), 11–22. <https://doi.org/10.3905/jpm.2025.1.778>
- [6] Guo, J., & Shum, H.-Y. (2025). Large investment model. *Frontiers of Information Technology & Electronic Engineering*, 26(10), 1771–1792. <https://doi.org/10.1631/FITEE.2500268>
- [7] Liu, C., Arulappan, A., Naha, R., Mahanti, A., Kamruzzaman, J., & Ra, I.-H. (2024). Large language models and sentiment analysis in financial markets: A review, datasets, and case study. *IEEE Access*, 12, 134041–134061. <https://doi.org/10.1109/ACCESS.2024.3445413>
- [8] Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62, 105227. <https://doi.org/10.1016/j.frl.2024.105227>
- [9] Schneider, C. J., & Yilmaz, Y. (2025). Stock portfolio selection based on risk appetite: Evidence from ChatGPT. *Finance Research Letters*, 82, 107517. <https://doi.org/10.1016/j.frl.2025.107517>
- [10] Wang, C., Liu, Y., Guo, T., Li, D., He, T., Li, Z., . . . , & Wen, Y. (2024). Systems engineering issues for industry applications of large language model. *Applied Soft Computing*, 151, 111165. <https://doi.org/10.1016/j.asoc.2023.111165>
- [11] Leippold, M. (2023). Thus spoke GPT-3: Interviewing a large-language model on climate finance. *Finance Research Letters*, 53, 103617. <https://doi.org/10.1016/j.frl.2022.103617>
- [12] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>
- [13] Shirvani, M. H., Rahmani, A. M., & Sahafi, A. (2018). An iterative mathematical decision model for cloud migration: A cost and security risk approach. *Software: Practice and Experience*, 48(3), 449–485. <https://doi.org/10.1002/spe.2528>
- [14] Crawford, K., & Schultz, J. (2019). AI systems as state actors. *Columbia Law Review*, 119(7), 1941–1972.
- [15] Zheng, Y., Du, X., Liao, L., Zhao, X., Zhou, Z., Song, J., & Zhang, P. (2025). Agentar-Fin-R1: Enhancing financial intelligence through domain expertise, training efficiency, and advanced reasoning. *arXiv Preprint: 2507.16802*.
- [16] Ant Digital Technologies. (2025). *Ant Digital Technologies released the financial reasoning large model Agentar-Fin-R1*. Retrieved from: <https://antdigital.com/community/articles/31702>
- [17] Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024). FinTral: A family of GPT-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 13064–13087). <https://doi.org/10.18653/v1/2024.findings-acl.774>
- [18] Ray, P. P. (2025). Should LLMs be over empowered for high-stake regulatory research? *Briefings in Bioinformatics*, 26(3), bbaf299. <https://doi.org/10.1093/bib/bbaf299>
- [19] Church, K. W., & Yue, R. (2023). Emerging trends: Smooth-talking machines. *Natural Language Engineering*, 29(5), 1402–1410. <https://doi.org/10.1017/S1351324923000463>
- [20] Fazlija, B., Ibraimi, M., Forouzandeh, A., & Fazlija, A. (2025). Reasoning with financial regulatory texts via large language models. *Journal of Behavioral and Experimental Finance*, 47, 101067. <https://doi.org/10.1016/j.jbef.2025.101067>
- [21] Marasco, E., & Bourlai, T. (2025). Enhancing trust in Large Language Models for streamlined decision-making in military operations. *Image and Vision Computing*, 158, 105489. <https://doi.org/10.1016/j.imavis.2025.105489>
- [22] van Deemter, K. (2024). The pitfalls of defining hallucination. *Computational Linguistics*, 50(2), 807–816. https://doi.org/10.1162/coli_a_00509
- [23] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64–93. <https://doi.org/10.1093/jla/lae003>
- [24] Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Vasilev, D., Tsvigun, A., . . . , & Shelmanov, A. (2025). Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13, 220–248. https://doi.org/10.1162/tacl_a_00737
- [25] Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12, 484–506. https://doi.org/10.1162/tacl_a_00660
- [26] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models.

- National Science Review*, 11(12), nwae403. <https://doi.org/10.1093/nsr/nwae403>
- [27] Jablonka, K., & Krishnan, N. M. A. (2025). Vision language models excel at perception but struggles with scientific reasoning. *Nature Computational Science*, 5(10), 852–853. <https://doi.org/10.1038/s43588-025-00871-0>
- [28] Shao, W., Lei, M., Hu, Y., Gao, P., Xu, P., Zhang, K., . . . , & Luo, P. (2025). TinyLVLm-eHub: Towards comprehensive and efficient evaluation for large vision-language models. *IEEE Transactions on Big Data*, 11(3), 933–947. <https://doi.org/10.1109/TBDATA.2025.3536930>
- [29] Yang, X., Liu, F., & Lin, G. (2024). Neural logic vision language explainer. *IEEE Transactions on Multimedia*, 26, 3331–3340. <https://doi.org/10.1109/TMM.2023.3310277>
- [30] McGraw, G., Bonett, R., Figueroa, H., & McMahon, K. (2024). 23 security risks in black-box large language model foundation models. *Computer*, 57(4), 160–164. <https://doi.org/10.1109/MC.2024.3363250>
- [31] Han, B. (2025). Trustworthy machine learning in the era of foundation models. *IEEE Intelligent Systems*, 40(6), 73–79. <https://doi.org/10.1109/MIS.2025.3621938>
- [32] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., . . . , & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38. <https://doi.org/10.1145/3639372>
- [33] Yang, W., Liu, M., Wang, Z., & Liu, S. (2024). Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, 10(3), 399–424. <https://doi.org/10.1007/s41095-023-0393-x>
- [34] Kazmierczak, R., Berthier, E., Frehse, G., & Franchi, G. (2025). Explainability and vision foundation models: A survey. *Information Fusion*, 122, 103184. <https://doi.org/10.1016/j.inffus.2025.103184>
- [35] Giudici, P., & Raffinetti, E. (2023). Safe artificial intelligence in finance. *Finance Research Letters*, 56, 104088. <https://doi.org/10.1016/j.frl.2023.104088>
- [36] Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial Intelligence risk measurement. *Expert Systems with Applications*, 235, 121220. <https://doi.org/10.1016/j.eswa.2023.121220>
- [37] Bartsch, S., Behn, O., Benlian, A., Brownsword, R., Bucker, S., Düwell, M., . . . , & Will-Zocholl, M. (2025). Governance of high-risk AI systems in healthcare and credit scoring. *Business & Information Systems Engineering*, 67(4), 563–581. <https://doi.org/10.1007/s12599-025-00944-4>
- [38] McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., Xu, D., Liu, D., . . . , & Halgamuge, M. N. (2024). From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *Computers & Security*, 144, 103964. <https://doi.org/10.1016/j.cose.2024.103964>
- [39] Hacker, P. (2023). The European AI liability directives – Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51, 105871. <https://doi.org/10.1016/j.clsr.2023.105871>
- [40] Malgieri, G., & Pasquale, F. (2024). Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, 52, 105899. <https://doi.org/10.1016/j.clsr.2023.105899>
- [41] Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5), 856. <https://doi.org/10.3390/math13050856>
- [42] Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research*, 26, e52935. <https://doi.org/10.2196/52935>
- [43] Gan, Z., Zhang, D., Li, H., Wu, Y., Lin, X., Liu, J., . . . , & Dai, Y. (2025). MME-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 12867–12874. <https://doi.org/10.1145/3746027.3758230>
- [44] Khan, F. S., Mazhar, S. S., Mazhar, K., AlSaleh, D. A., & Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: A systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, 58(8), 232. <https://doi.org/10.1007/s10462-025-11215-9>
- [45] Erdem, O., Hassett, K., & Egriboyun, F. (2025). Hallucination in AI-generated financial literature reviews: Evaluating bibliographic accuracy. *International Journal of Data Science and Analytics*, 20(5), 4501–4510. <https://doi.org/10.1007/s41060-025-00731-0>
- [46] Srivastava, P., Malik, M., Gupta, V., Ganu, T., & Roth, D. (2024). Evaluating LLMs’ mathematical reasoning in financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 3853–3878). <https://doi.org/10.18653/v1/2024.findings-acl.231>
- [47] Alghamdi, E. A., Zakraoui, J., & Abanmy, F. A. (2024). Domain adaptation for Arabic machine translation: Financial texts as a case study. *Applied Sciences*, 14(16), 7088. <https://doi.org/10.3390/app14167088>
- [48] Zhi, Q. (2025). Research on the application of multimodal information fusion in financial data prediction. *Journal of Computer Science and Artificial Intelligence*, 4(2), 1–4. <https://doi.org/10.54097/c6jmgg90>
- [49] Yan, Y. (2025). Application of multimodal financial data fusion analysis in corporate strategic management. *Applied Mathematics and Nonlinear Sciences*, 10(1), 20250842. <https://doi.org/10.2478/amns-2025-0842>
- [50] Jiang, S., Liang, J., Wang, J., Dong, X., Chang, H., Yu, W., . . . , & Qin, B. (2025). From specific-MLLMs to omni-MLLMs: A survey on MLLMs aligned with multi-modalities. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 8617–8652). <https://doi.org/10.18653/v1/2025.findings-acl.453>
- [51] Liang, P. P., Zadeh, A., & Morency, L.-P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10), 264. <https://doi.org/10.1145/3656580>
- [52] Aamir, M., Rahman, Z., Bhatti, U. A., Bhutto, J. A., Choudhry, N., & Abro, W. A. (2025). A comprehensive review of language models: Evolution, applications, and future directions. In *2025 IEEE 3rd International Conference on Computer, Vision and Intelligent Technology*, 1–8. <https://doi.org/10.1109/ICCVIT67848.2025.11391358>
- [53] Maniah, Soewito, Gaol, B., L, F., & Abdurachman, E. (2022). A systematic literature review: Risk analysis in cloud migration. *Journal of King Saud University Computer and Information Sciences*, 34(6), 3111–3120. <https://doi.org/10.1016/j.jksuci.2021.01.008>

- [54] Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, 2(4), 815–826. <https://doi.org/10.1007/s43681-022-00142-y>
- [55] Wang, Z., Wu, Z., Zhang, Y., Guan, X., Jain, N., Lu, Q., . . . , & Gupta, S. (2025). Bias amplification: Large language models as increasingly biased media. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 115–132. <https://doi.org/10.18653/v1/2025.ijcnlp-long.8>
- [56] Carnat, I. (2024). Addressing the risks of generative AI for the judiciary: The accountability framework(s) under the EU AI Act. *Computer Law & Security Review*, 55, 106067. <https://doi.org/10.1016/j.clsr.2024.106067>
- [57] DeBellis, M., Dutta, N., Gino, J., & Balaji, A. (2024). Integrating ontologies and large language models to implement retrieval augmented generation. *Applied Ontology*, 19(4), 389–407. <https://doi.org/10.1177/15705838241296446>
- [58] Luo, P., & Liu, Z. (2025). Automatic subject indexing with large language model fine-tuning and controlled vocabulary retrieval. *Aslib Journal of Information Management*, 1–20. <https://doi.org/10.1108/AJIM-05-2025-0253>
- [59] Hiriyanna, S., & Zhao, W. (2025). Multi-layered framework for LLM hallucination mitigation in high-stakes applications: A tutorial. *Computers*, 14(8), 332. <https://doi.org/10.3390/computers14080332>
- [60] Tozuka, R., Johno, H., Amakawa, A., Sato, J., Muto, M., Seki, S., . . . , & Onishi, H. (2025). Application of NotebookLM, a large language model with retrieval-augmented generation, for lung cancer staging. *Japanese Journal of Radiology*, 43(4), 706–712. <https://doi.org/10.1007/s11604-024-01705-1>
- [61] Zhou, S., Liu, X., Li, D., Gu, T., Liu, K., Yang, Y., & Wong, M. O. (2025). Integrating domain-specific knowledge and fine-tuned general-purpose large language models for question-answering in construction engineering management. *Automation in Construction*, 175, 106206. <https://doi.org/10.1016/j.autcon.2025.106206>
- [62] Huang, Y., Ma, T., Yang, K., & Zhang, Z. (2026). FinSent-DistillQ: A distilled large language model with chain-of-thought fine-tuning for financial sentiment analysis. *Journal of Intelligent Information Systems*, 64(2), 735–771. <https://doi.org/10.1007/s10844-025-01020-9>
- [63] Wang, C., Yang, J., Zhou, Y., & Yue, X. (2025). CooKie: Commonsense knowledge-guided mixture-of-experts framework for fine-grained visual question answering. *Information Sciences*, 695, 121742. <https://doi.org/10.1016/j.ins.2024.121742>
- [64] Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., . . . , & Zhang, M. (2025). Uni-MoE: Scaling unified multimodal LLMs with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5), 3424–3439. <https://doi.org/10.1109/TPAMI.2025.3532688>
- [65] Xia, R., Ye, H., Yan, X., Liu, Q., Zhou, H., Chen, Z., Shi, B., Zhang, B., . . . Zhang, B. (2025). ChartX and ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 34, 7436–7447. <https://doi.org/10.1109/TIP.2025.3607618>
- [66] Zhang, L. (2025). Lawfulness of mass processing personal data to train large language models in China. *Telecommunications Policy*, 49(8), 103023. <https://doi.org/10.1016/j.telpol.2025.103023>
- [67] Liu, Y., Lu, Q., Zhu, L., & Paik, H.-Y. (2024). Decentralized governance-driven architecture for designing foundation-model-based systems: Exploring the role of blockchain in responsible AI. *IEEE Software*, 41(5), 34–42. <https://doi.org/10.1109/MS.2024.3369551>
- [68] Jabbar, M., Al-Azani, S., Alotaibi, A., & Ahmed, M. (2025). Red teaming large language models: A comprehensive review and critical analysis. *Information Processing & Management*, 62(6), 104239. <https://doi.org/10.1016/j.ipm.2025.104239>
- [69] Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., . . . , & Huang, X. (2025). Safeguarding large language models: A survey. *Artificial Intelligence Review*, 58(12), 382. <https://doi.org/10.1007/s10462-025-11389-2>

How to Cite: Niu, F., Liu, F., Zhang, Z., Li, Y., & Ren, Y. (2026). From General-Purpose Large Language Models to Domain-Specific Financial Models: A Structured Review and Comparative Case Analysis. *FinTech and Sustainable Innovation*. <https://doi.org/10.47852/bonviewFSI62029857>