Uncertainty Quantification for Financial Foundation Models: A Survey

Zijie Zhao
Massachusetts Institute of Technology
Cambridge, MA, USA
zijiezha@mit.edu

Mingjun Sun
Carnegie Mellon University
Pittsburgh, PA, USA
mingjuns@andrew.cmu.edu

Shenbo Xu
Massachusetts Institute of Technology
Cambridge, MA, USA
xushenbo@mit.edu

Abstract

Financial foundation models increasingly support workflows spanning financial question answering, disclosure analysis, forecasting, and decision-making. In these settings, uncertainty may concern whether evidence is valid and available at the relevant time, whether an interpretation is justified, which outcomes remain plausible, or which action is preferred under costs and risk constraints. Because these objects arise at different points in a workflow, uncertainty cannot be reduced to confidence in a final output. We introduce a source-aware framework that distinguishes six sources of uncertainty across evidence and context, model and inference, and outcome and action, while treating agentic workflows as the execution layer through which uncertainty can propagate, change form, or be reduced. We develop a mechanism-based taxonomy grounded in how methods construct or verify uncertainty and reliability representations, and use it to assess their applicability across financial task domains. We further map finance-specific empirical studies across the taxonomy and task domains to characterize the current research landscape. Our analysis shows that most reviewed studies address uncertainty in individual parts of financial workflows, with limited work connecting evidence and reasoning reliability to downstream prediction and constrained action. By connecting where uncertainty arises, how it is represented, and how it bears on downstream decisions, this survey provides a unified foundation for studying and evaluating uncertainty in financial foundation models.

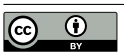
CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*.

Keywords

uncertainty quantification, financial foundation models, calibration, financial agents, decision-making under uncertainty

ACM Reference Format:
Zijie Zhao, Mingjun Sun, and Shenbo Xu. 2026. Uncertainty Quantification for Financial Foundation Models: A Survey. In *7th ACM International Conference on AI in Finance (ICAIF '26), November 14–17, 2026, Milan, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3858063.3858824



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICAIF '26, Milan, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3087-0/2026/11
https://doi.org/10.1145/3858063.3858824

1 Introduction

Financial foundation models increasingly support workflows that connect financial information to analysis, forecasting, and decisions. Their outputs may be fluent and plausible yet unsupported, numerically inconsistent, or based on information unavailable at the relevant decision time. Uncertainty quantification (UQ) can help identify unreliable cases, but uncertainty in financial workflows is not confined to the final output. Consider a system that retrieves an earnings release, interprets revised guidance, forecasts the market response, and recommends a trade. At each stage, a different question may remain unresolved: whether the evidence was available and relevant at the decision time, whether the interpretation follows from the reported figures and market context, which outcomes remain plausible, and which feasible action is preferred under stated objectives, costs, and risk constraints. Because these questions concern distinct uncertain objects, uncertainty cannot be reduced to confidence in the final output: such a score cannot indicate whether the system should retrieve additional evidence, verify its reasoning, revise its forecast, or change its action. Financial UQ must therefore identify what is uncertain, represent the remaining uncertainty, and determine whether system behavior should change.

Throughout this survey, *uncertainty* denotes the extent to which a quantity, proposition, outcome, or preferred action remains unresolved relative to a specified information set and modeling assumptions; for decisions, it is also relative to stated objectives and constraints. Its representation depends on the object and may take the form of a support judgment, confidence score, disagreement measure, predictive distribution, interval, or prediction set. *Confidence* is a score attached to a prediction, answer, or claim and need not be calibrated. *Calibration* describes the correspondence between stated probabilities or confidence levels and empirical frequencies over comparable cases [25]. We use *aleatoric uncertainty* for variability that remains conditional on the specified information set and assumptions, and *epistemic uncertainty* for uncertainty arising from incomplete evidence, data, knowledge, or model specification that is reducible in principle. The object-based framework identifies what remains unresolved, whereas the aleatoric–epistemic distinction characterizes the nature of that uncertainty; each object may involve both components.

General LLM UQ surveys synthesize work on confidence estimation, semantic consistency, calibration, conformal methods, and abstention [27, 55]. Financial UQ surveys focus primarily on probabilistic modeling and uncertainty estimation for financial time series [3, 15]. Surveys of financial language models and agents instead emphasize models, tasks, agent architectures, and deployment settings [14, 43]. These lines of work provide important founda-

tions but do not jointly organize point-in-time evidence validity, task-level capability, case-specific reasoning and interpretation, uncertain financial outcomes, and actions under stated objectives and constraints, or explain how these forms of uncertainty interact across retrieval, reasoning, tool use, memory, verification, and execution. We address this intersection by organizing uncertainty around the object that remains unresolved and tracing its role across financial foundation-model workflows, including agentic systems.

Survey scope and methodology. We focus on uncertainty in financial workflows supported by language, multimodal, and time-series foundation models, while drawing on established general and financial UQ where they provide relevant methodological or evaluative foundations. We searched ACM Digital Library, ACL Anthology, IEEE Xplore, arXiv, SSRN, and major machine-learning and finance venues for work available through July 25, 2026. The search combined terms for financial language, time-series, and agentic foundation models with uncertainty, calibration, grounding, verification, probabilistic forecasting, risk, and distribution shift. We included finance-specific studies that construct, verify, propagate, evaluate, or use operational uncertainty or reliability representations, together with directly relevant methodological and evaluation foundations. Duplicate preprint, working-paper, and published versions were merged. Section 5 defines the narrower empirical subset used for the coverage analysis.

We make three contributions:

- **A source-aware framework.** We distinguish six uncertainty sources across evidence and context, model and inference, and outcome and action, and treat agentic workflows as the execution layer through which these sources interact and propagate.
- **A mechanism-based taxonomy and task analysis.** We classify methods by how they construct or verify uncertainty and reliability representations and analyze their applicability across financial task domains.
- **An empirical landscape and research priorities.** We map finance-specific empirical studies across method families and task domains to identify concentrations, missing connections, and research priorities.

2 A Source-Aware Framework for Financial Uncertainty

We propose a source-aware framework that organizes financial uncertainty by the object that remains unresolved in a foundation-model workflow, rather than by a fixed pipeline stage or downstream failure. Figure 1 summarizes six sources across evidence and context, model and inference, and outcome and action, together with the agentic execution layer that connects them. Several sources may coexist in the same case; when a primary source is needed, we assign it according to what is directly uncertain.

2.1 Sources of Financial Uncertainty

Evidence and context uncertainty concerns the validity of the available information and whether previously established reliability remains applicable. **Evidence and grounding uncertainty** (U_E) asks whether a claim is supported by evidence that is correct, sufficiently complete, traceable, and available at the relevant decision time. In finance, an authentic filing cannot support a historical decision if it was released or restated only afterward, while memorized future information can contaminate retrospective analysis [33, 44]. **Distribution shift uncertainty** (U_S) asks whether reliability established during training or calibration transfers to the current environment as market regimes, disclosure practices, accounting standards, or the population of firms and instruments changes [22, 49]. Thus, U_E concerns the validity of evidence in the current case, whereas U_S concerns whether established reliability transfers to the current setting.

Model and inference uncertainty distinguishes whether a system has the knowledge and capability required for a class of tasks from whether its reasoning and interpretation are adequate in the current case. **Knowledge and capability uncertainty** (U_K) asks whether the system has the domain knowledge and skills required for a class of financial inputs, including accounting conventions, numerical reasoning over filings, structured-data understanding, and tool use [32, 61]. **Reasoning and interpretation uncertainty** (U_R) concerns whether the model's inference from the available evidence and stated assumptions to a specific conclusion is logically, numerically, and economically justified. It may arise from ambiguity in the problem formulation, faulty decomposition, numerical or tool-use errors, inappropriate comparison periods, omitted premises, or unsupported economic assumptions [12, 62]. Because internal reasoning is not always observable, U_R is assessed through available intermediate outputs, calculations, tool calls, stated assumptions, and the consistency of the conclusion with the evidence.

Outcome and action uncertainty separates uncertainty about future financial results from uncertainty about how to act on them. **Financial outcome uncertainty** (U_O) concerns the conditional distribution of a defined financial target and horizon, such as returns, earnings, volatility, default risk, or portfolio loss [24, 51]. This uncertainty does not disappear when the evidence and interpretation are sound. Market outcomes remain uncertain because prices reflect prior expectations and strategic responses, while subsequent information, liquidity conditions, and future shocks affect the realized result [45]. **Decision uncertainty under constraints** (U_D) concerns the preferred feasible action and its stability under uncertainty in outcomes, objectives, costs, liquidity, risk limits, and policy requirements. Known constraints define the feasible set; uncertainty lies in which feasible action is preferred and how sensitive that choice is to its inputs [29, 54]. Thus, U_O concerns what may happen, whereas U_D concerns what should be done.

The aleatoric–epistemic distinction complements rather than replaces the source framework. Evidence, capability, reasoning, and distribution-shift uncertainty often contain substantial epistemic components when additional information, data, computation, verification, or model improvement can reduce them. Reasoning and interpretation uncertainty may also contain aleatoric components when the task or evidence admits multiple defensible interpretations. Financial outcome uncertainty is mixed: uncertainty in the estimated conditional distribution may be reducible, while part of realized variation remains aleatoric under the stated information set and assumptions. Decision uncertainty inherits both forms through uncertain outcomes and model inputs. The distinction is therefore conditional on the information set and modeling assumptions and,

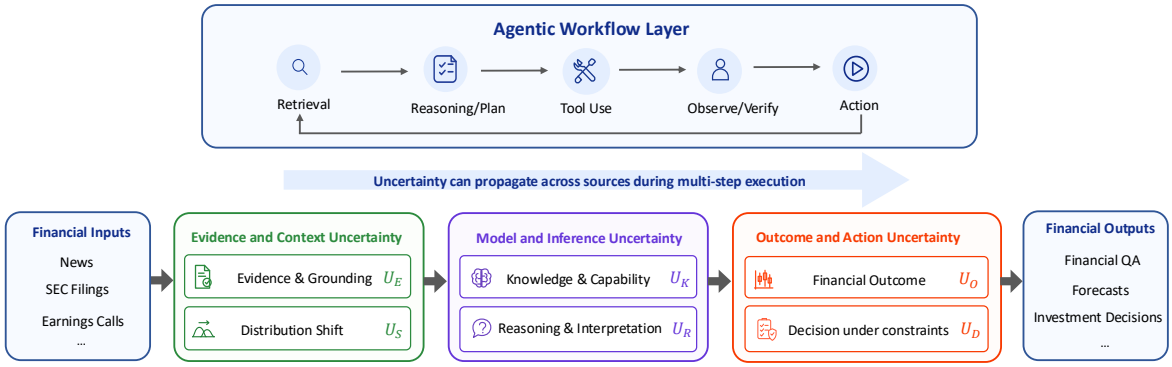


Figure 1: A source-aware framework for financial uncertainty. The agentic workflow layer connects the six sources as uncertainty propagates, changes form, or is reduced during execution.

for decisions, the stated objectives and constraints; it is not an intrinsic label attached permanently to a source.

2.2 Agentic Workflow Layer and Uncertainty Propagation

An agentic workflow is not a seventh source but an execution layer connecting the six sources across retrieval, reasoning and planning, tool use, memory, verification, and action. Intermediate outputs become inputs: evidence shapes interpretation, interpretation informs outcome estimates, and those estimates guide actions. A filing unavailable at the decision time may support an apparently coherent interpretation and subsequently distort prediction and action. Actions can change later observations, stored state, and available choices, making uncertainty trajectory-dependent [48, 66].

Propagation need not be monotonic or irreversible. Retrieval, recomputation, provenance checks, or numerical verification may reduce epistemic uncertainty, although later steps may leave it unresolved or introduce uncertainty. When uncertainty cannot be resolved before action, abstention or escalation can limit its consequences rather than remove it. The source framework identifies what remains uncertain; the agentic layer describes how uncertainty propagates, changes form, or is mitigated across steps.

3 Methods and Evaluation for Financial Uncertainty Quantification

We focus on UQ methods used in or directly applicable to financial workflows supported by foundation models, drawing on established work on uncertainty representation and evaluation. We group these methods into five families based on how they construct or verify uncertainty or reliability representations: assessment against supporting evidence, direct estimation from a model, repeated execution of a fixed system, comparison across model or parameter realizations, or probabilistic modeling of a defined target. This taxonomy distinguishes methods by how their representations are obtained rather than by application, mathematical tradition, or downstream use. For hybrid methods, the primary family is determined by the mechanism producing the main uncertainty result.

Bayesian modeling may produce parameter variation, predictive distributions, or sequential belief updates, while calibration evaluates a representation rather than defining how it was constructed. Abstention, deferral, and action adjustment use uncertainty downstream and are therefore not separate method families.

3.1 Methods for Constructing Uncertainty Representations

Evidence and provenance verification. These methods assess a claim or result against supporting evidence and provenance information valid at the relevant decision time. Attribution methods align claims with cited passages, while financial applications must also check document provenance, version, numerical consistency, and point-in-time availability [19, 33, 44]. Recent methods decompose financial answers into atomic claims, verify them against retrieved evidence, and detect or revise unsupported statements [26, 57, 64]. Finance-specific benchmarks cover long-document claim verification, hallucination detection in grounded financial QA, and justified abstention in multi-turn financial RAG [9, 38, 68]. We include verification in this taxonomy only when it produces a claim- or instance-level judgment of support or validity; aggregate retrieval or task accuracy alone is not an uncertainty representation. This family primarily characterizes evidence and grounding uncertainty (U_E) and may also expose unsupported reasoning (U_R).

Direct uncertainty estimation. Direct methods produce an instance-level estimate without relying on repeated stochastic executions or multiple model realizations. They use token probabilities or entropy, verbalized confidence, self-evaluation, hidden-state features, or auxiliary error predictors [31, 40, 58]. Financial applications estimate confidence in monetary-policy interpretation, identify incorrect answers from model representations, and predict epistemic error in cross-sectional stock ranking [54, 59, 62]. Probability-based estimates require access to model probabilities, whereas learned probes may require internal states and labeled data; neither is necessarily calibrated or aligned with downstream decision uncertainty. This family mainly characterizes knowledge and capability uncertainty (U_K) and reasoning and interpretation uncertainty (U_R).

Sampling-based consistency estimation. These methods keep the underlying model or system fixed and compare outputs across stochastic generations, input perturbations, or repeated executions.

SelfCheckGPT measures agreement across sampled generations, while semantic uncertainty methods group meaning-equivalent answers before measuring variation [17, 37, 46]. A finance-specific hybrid combines multi-sample semantic entropy with token-level and evidence-use features [56]. Agreement does not establish correctness because repeated outputs may share the same error, and repeated execution adds latency and cost. Variation across runs mainly characterizes instability in U_K and U_R; under input perturbations, it may also reveal sensitivity to distribution shift (U_S).

Ensemble-based uncertainty estimation. Ensemble methods compare predictions from multiple models or parameter realizations. Deep ensembles use independently trained models, while Monte Carlo dropout approximates parameter sampling through repeated forward passes [18, 41]. Financial applications use disagreement across independently trained hedging networks to adjust hedge positions and agreement across LLM agents to filter unreliable information before trading decisions [42, 52]. Unlike sampling-based methods, ensembles vary the model or its parameters rather than repeatedly execute a fixed system. Disagreement is informative only when members are sufficiently diverse; highly correlated members may agree while sharing the same failure. This family can characterize capability uncertainty (U_K), reliability under distribution shift (U_S), and uncertainty in model-based financial outcomes (U_O).

Distributional and statistical uncertainty estimation. These methods represent uncertainty over a defined target and horizon through predictive distributions, conditional quantiles, intervals, prediction sets, volatility estimates, or tail-risk measures. Established statistical methods such as ARCH/GARCH and quantile regression provide foundations for predictive uncertainty in financial forecasting and risk modeling [4, 16, 34]. Conformal and evidential methods provide prediction sets, intervals, and uncertainty decompositions [1, 22, 53]. Recent work applies these ideas to time-series foundation models, financial forecasting, and risk estimation [2, 24, 51, 60]. ProbFM uses evidential uncertainty decomposition for financial forecasting, while Bayesian state filters maintain posterior regime distributions and risk measures in agentic settings [8, 13]. This family primarily characterizes financial outcome uncertainty (U_O); adaptive procedures may also address reliability under distribution shift (U_S).

No single representation covers the full framework: evidence verification cannot describe future outcome distributions, while a calibrated distribution does not establish valid evidence or reasoning. Evidence-, model-, and outcome-level representations should remain distinct rather than be collapsed into generic confidence. In agentic workflows, these representations may also be updated or used across dependent stages.

3.2 Uncertainty Quantification in Financial Agentic Workflows

In financial agentic workflows, local uncertainty estimates may be propagated, aggregated, or used to guide later steps [48, 66]. Repeated executions of trading-agent systems reveal run-to-run instability [36]; agreement across multiple agents provides an ensemble-based reliability measure [42]; Bayesian filters update beliefs and risk measures over time [13]; and trajectory-level probes audit failures across agent roles and stages [6]. These approaches combine the five method families above rather than define a separate agent-specific family. The financial studies reviewed here do not yet provide a general method for preserving the meaning of heterogeneous uncertainty representations across a workflow. Verification, abstention, deferral, position adjustment, and escalation govern how an agentic workflow responds to uncertainty.

3.3 Evaluating Uncertainty Representations

An uncertainty representation should be evaluated against the object it characterizes and, when used downstream, the decision or workflow it is intended to support. For example, among comparable forecasts assigned 90% probability to a positive market reaction, the event should occur approximately 90% of the time. The estimate should also identify unreliable cases and, when used for decision support, improve decisions after costs and constraints.

We distinguish three complementary evaluation targets. (1) Task performance measures the underlying output against a reference outcome, such as answer correctness, forecast realization, or financial utility. (2) Uncertainty quality measures whether a score, distribution, interval, or prediction set corresponds to observed errors or outcomes; relevant criteria include calibration, error detection, risk–coverage curves, proper scoring rules, coverage, and sharpness [5, 20, 23, 25]. (3) Decision value measures whether acting on the representation improves the workflow or financial action after accounting for verification, abstention, implementation costs, and constraints. Strong task performance does not imply reliable uncertainty, and reliable uncertainty does not by itself establish decision value.

The appropriate test depends on the representation. Evidence judgments require labeled supported and unsupported claims, as provided by atomic factuality and financial verification benchmarks [47, 50, 65]. Predictive intervals require empirical coverage at useful width, while tail-risk estimates require exceedance and dependence backtests [11, 39]. Evaluation must also preserve temporal order because a representation calibrated in one regime may fail after a shift [49]. Recent financial work extends these requirements to time-gated calibration, agent trajectories, financial biases, market frictions, and deployment assumptions [6, 21, 35, 63].

A financial UQ study should report the target and horizon, data vintage and leakage controls, construction and evaluation of the uncertainty representation, robustness across time or regimes, and downstream costs and constraints [29, 63].

4 Uncertainty Quantification Across Financial Tasks

Building on Section 3, we examine which uncertainty sources arise in representative financial tasks supported by foundation models and which methods are most relevant. Table 1 summarizes this mapping, which reflects conceptual relevance rather than the frequency of methods in the current literature.

4.1 Information and Document Intelligence

Information and Document Intelligence covers retrieval, extraction, financial QA, document understanding, and grounded summarization; economic interpretation belongs to Analysis, Valuation, and Advisory. Reliability depends on whether the output is supported by

Table 1: Financial task domains, representative tasks, dominant uncertainty sources, and relevant method families.

Task domain	Representative tasks	Dominant sources[a]	Most relevant method families
Information and Document Intelligence	Retrieval and RAG; financial QA; information extraction; table, chart, and multimodal document understanding; grounded summarization	U_E, U_K, U_R	Evidence and Provenance Verification
Analysis, Valuation, and Advisory	Disclosure and event interpretation; financial-statement analysis; valuation; investment research and due diligence; investment and wealth advisory	U_R, U_K, U_D	Direct Uncertainty Estimation; Sampling-Based Consistency Estimation
Forecasting and Risk Modeling	Return, earnings, cash-flow, and volatility forecasting; credit and default risk; VaR/ES; stress testing	U_O, U_S	Distributional and Statistical Uncertainty Estimation; Ensemble-Based Uncertainty Estimation
Financial Decision-Making	Security selection and trading; portfolio allocation and rebalancing; hedging and position sizing; execution and market making; lending and underwriting	U_D, U_O, U_S	Distributional and Statistical Uncertainty Estimation; Direct Uncertainty Estimation; Ensemble-Based Uncertainty Estimation
Accounting, Compliance, and Financial Operations	Accounting and reporting; regulatory interpretation and compliance review; KYC/AML investigation; audit and internal controls; payments, settlement, reconciliation, and claims review	U_E, U_R, U_D	Evidence and Provenance Verification; Direct Uncertainty Estimation

[a] U_E: Evidence and Grounding Uncertainty; U_S: Distribution Shift Uncertainty; U_K: Knowledge and Capability Uncertainty; U_R: Reasoning and Interpretation Uncertainty; U_O: Financial Outcome Uncertainty; and U_D: Decision Uncertainty under Constraints.

the correct document, fiscal period, unit, version, and information available at the decision time. Evidence and Provenance Verification is therefore the primary method family. Direct or sampling-based estimates may flag unreliable cases, but they cannot establish that the supporting evidence is valid. Evaluation should examine both output correctness and supporting evidence. FinRAGBench-V and PHANTOM provide evidence- or hallucination-oriented labels, while FinQA and TAT-QA provide correctness labels that can support UQ evaluation [7, 30, 67, 69].

4.2 Analysis, Valuation, and Advisory

Analysis, Valuation, and Advisory covers financial interpretation, valuation, research, and advice rather than fact retrieval, future-outcome forecasting, or execution. The key question is whether a conclusion follows from the evidence, comparison basis, and financial assumptions. Direct Uncertainty Estimation can estimate confidence or error risk, while Sampling-Based Consistency Estimation can reveal sensitivity to alternative generations or formulations. For valuation, uncertainty should be tied to the assumptions and the resulting range of implied values. In advisory settings, it should support suitability assessment, abstention, or escalation under client-specific objectives and constraints. RealFin and FinBen provide relevant reasoning cases, but dedicated UQ protocols for valuation and personalized advice remain limited [12, 61].

4.3 Forecasting and Risk Modeling

Forecasting and risk modeling concern conditional distributions of future returns, earnings, cash flows, volatility, defaults, and losses. The main sources are financial outcome uncertainty and reliability under changing regimes. Distributional and Statistical Uncertainty Estimation is the primary method family because the target is usually a conditional distribution, quantile, interval, or tail-risk measure; ensembles can additionally characterize variation across model realizations. Average forecast error is insufficient: forecasts require calibration, coverage, and sharpness, while risk estimates require tail and stress-period validation. FinBench provides a time-gated calibration setting, complementing established backtests for interval and tail-risk forecasts [11, 21, 39].

4.4 Financial Decision-Making

Financial decision-making converts forecasts, risk estimates, or analytical conclusions into feasible actions: predicting default is a risk-modeling task, whereas approving a loan is a decision. The uncertainty concerns which feasible action is preferred and how stable that choice is to uncertain outcomes and model inputs. Distributional methods represent outcome uncertainty, while direct and ensemble-based estimates can assess model reliability or support selective action and position adjustment. A decision-level UQ claim should show that uncertainty changes and improves the selected action, not merely that it predicts errors. Comparisons should use matched information, costs, execution timing, constraints, and intervention budgets [29, 63].

4.5 Accounting, Compliance, and Financial Operations

Accounting, Compliance, and Financial Operations covers model-supported judgments and exception handling in accounting, regulatory review, KYC/AML, audit, payments, settlement, reconciliation, and claims. A fraud or default probability belongs to Forecasting and Risk Modeling; investigation, rule application, approval, reporting, and escalation belong here. The main uncertainties concern the validity of records and rules, the interpretation of accounting or regulatory requirements, and whether the resulting action is permissible. Evidence and Provenance Verification is central, while Direct Uncertainty Estimation can support selective review and escalation. Evaluation should test whether rules are applied correctly, conclusions can be traced to the relevant records and requirements, and asymmetric errors and regulatory changes are handled appropriately. FAITH, FinVerBench, and FinTrust cover selected aspects of statement verification and trustworthiness, but not the full range of these processes [28, 50, 65].

Because an agentic workflow describes how a system executes a financial task rather than a separate task domain, we classify agents by the evaluated output; tool-call validity and trajectory stability remain cross-cutting concerns [6, 10, 70].



Figure 2: Method–task coverage in the included finance-specific studies. Cells report unique-paper counts across 24 papers and 26 assignments; gray denotes zero, orange one, and red more than one study.

5 Method–Task Landscape and Research Priorities

We code the reviewed finance studies by their primary method family and financial task label(s). Figure 2 reports aggregate method–task coverage; the paper-level coding table is omitted due to space limits.

Selection and coding criteria. We retain a study only if it satisfies two criteria. First, it must construct, use, or directly evaluate an uncertainty or reliability representation covered by Section 3, such as evidence support, confidence or error estimates, disagreement measures, predictive distributions or sets, risk measures, or uncertainty-conditioned actions. Studies reporting only aggregate task performance, e.g., accuracy, return, or Sharpe ratio, are excluded. Second, it must report substantive quantitative evaluation on financial data or workflows in at least one task domain from Section 4. General methods without financial experiments, surveys, and conceptual studies are excluded. Each paper receives one primary method family according to the mechanism producing its main uncertainty result. Multiple task labels require distinct empirical results for those domains; passing through several workflow stages alone does not create multiple assignments. Duplicate versions are merged. This yields 24 papers and 26 method–task assignments.

Coverage patterns. Three patterns stand out in the included corpus. First, method–task specialization is strong: Evidence and Provenance Verification accounts for nine of eleven assignments in Information and Document Intelligence, while all six Distributional and Statistical assignments occur in Forecasting and Risk Modeling. Direct Uncertainty Estimation is the only family represented across all five task domains. Second, task coverage is uneven. Information and Document Intelligence and Forecasting and Risk Modeling contain most assignments, whereas Analysis, Valuation, and Advisory and Accounting, Compliance, and Financial Operations are sparsely represented. Financial Decision-Making is represented mainly by trading, hedging, and agent-based settings, with limited evidence in lending, underwriting, and execution. Third, cross-stage UQ is rare in the included corpus: only two studies span more than one task domain, and we did not identify an included study that carries evidence or reasoning reliability through predictive uncertainty into constrained financial action. Most included work studies uncertainty within parts of the workflow rather than across the path from evidence to action. A zero indicates no included study met our criteria for that method–task pair; it does not imply that no related work exists.

Research priorities. Three priorities follow from these patterns. First, future work should connect uncertainty across stages while preserving the meaning of evidence, reasoning, outcome, and decision uncertainty rather than collapsing them into a single confidence score. Second, UQ should extend to under-studied settings, including valuation and advisory, accounting and compliance, and financial decisions beyond trading and hedging, with evaluation matched to task-relevant uncertainty. Third, financial decision and agent studies should establish decision value by testing whether uncertainty changes and improves the action under matched information, timing, implementation costs, constraints, and intervention budgets. Retrieval or verification may reduce epistemic uncertainty; when it cannot be resolved before action, abstention, escalation, or position adjustment can limit its consequences.

6 Conclusion

This survey develops a source-aware framework for uncertainty in financial foundation-model workflows, distinguishing what remains unresolved across evidence and context, model and inference, and outcome and action. It organizes UQ methods by how they construct or verify uncertainty and reliability representations, then examines their applicability across representative financial tasks. The resulting method–task landscape shows how current empirical work is distributed and supports research priorities for connecting uncertainty across stages, extending it to under-studied financial settings, and establishing its value for downstream decisions.

The review is bounded by its focus on foundation-model and agentic financial workflows and by the databases, search terms, and

literature cutoff described in Section 1; it does not attempt to cover the full literature on classical financial uncertainty modeling. The field is also developing quickly, and several recent studies remain preprints whose findings may change with broader evaluation or subsequent publication.

Taken together, the framework and empirical synthesis shift the focus from confidence in a final output to the different forms of uncertainty that arise across a financial workflow. The literature remains largely separated across tasks and stages, with limited evidence connecting evidence and reasoning reliability to predictive uncertainty and constrained action. This survey provides a unified foundation for understanding, evaluating, and using uncertainty in financial foundation models.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep Evidential Regression. In *Advances in Neural Information Processing Systems*, Vol. 33. 14927–14937. https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=gerNCVqqtR

[3] Txus Blasco, Josep Salvador Sánchez, and Vicente García. 2024. A Survey on Uncertainty Quantification in Deep Learning for Financial Time Series Prediction. *Neurocomputing* 576, Article 127339 (2024). doi:10.1016/j.neucom.2024.127339

[4] Tim Bollerslev. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31, 3 (1986), 307–327. doi:10.1016/0304-4076(86)90063-1

[5] Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78, 1 (1950), 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[6] Zichen Chen, Jianda Chen, Jiaao Chen, and Misha Sra. 2026. From Tasks to Teams: A Risk-First Evaluation Framework for Multi-Agent LLM Systems in Finance. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, San Diego, California, United States, 38819–38857. doi:10.18653/v1/2026.findings-acl.1934

[7] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3697–3711. doi:10.18653/v1/2021.emnlp-main.300

[8] Arundeep Chinta, Lucas Vinh Tran, and Jay Katukuri. 2026. ProbFM: Probabilistic Time Series Foundation Model with Uncertainty Decomposition. arXiv:2601.10591 https://arxiv.org/abs/2601.10591

[9] Eunbyeol Cho, Yunseung Lee, Mirae Kim, Jeewon Yang, Youngjun Kwak, and Edward Choi. 2026. K-FinHallu: A Hallucination Detection Benchmark for Multi-Turn RAG in Korean Finance. arXiv:2605.29523 [cs.LG] doi:10.48550/arXiv.2605.29523

[10] Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. 2025. FinAgentBench: A Benchmark Dataset for Agentic Retrieval in Financial Question Answering. In *Proceedings of the 6th ACM International Conference on AI in Finance*. Association for Computing Machinery, 632–637. doi:10.1145/3768292.3770362

[11] Peter F. Christoffersen. 1998. Evaluating Interval Forecasts. *International Economic Review* 39, 4 (1998), 841–862. doi:10.2307/2527341

[12] Yuyang Dai, Yan Lin, Zhuohan Xie, and Yuxia Wang. 2026. RealFin: How Well Do LLMs Reason About Finance When Users Leave Things Unsaid?. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, San Diego, California, United States, 25050–25080. doi:10.18653/v1/2026.findings-acl.1255

[13] Matthew Francis Dixon. 2026. Belief at Risk: Quantifying Agentic AI Model Risk with LLM-Inferred Bayesian State Filters. arXiv:2606.15473 https://arxiv.org/abs/2606.15473

[14] Yifei Dong, Fengyi Wu, Kunlin Zhang, Yilong Dai, Sanjian Zhang, Wanghao Ye, Sihan Chen, and Zhi-Qi Cheng. 2025. Large Language Model Agents in Finance: A Survey Bridging Research, Practice, and Real-World Deployment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, Suzhou, China, 17889–17907. doi:10.18653/v1/2025.findings-emnlp.972

[15] Sivert Eggen, Tord Johan Espe, Kristoffer Grude, Morten Risstad, and Rickard Sandberg. 2026. Financial Time Series Uncertainty: A Review of Probabilistic AI Applications. *Journal of Economic Surveys* 40, 2 (2026), 915–953. doi:10.1111/joes.70018

[16] Robert F. Engle. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50, 4 (1982), 987–1007. doi:10.2307/1912773

[17] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* 630, 8017 (2024), 625–630. doi:10.1038/s41586-024-07421-0

[18] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*. PMLR, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

[19] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398

[20] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 30.

[21] Rishab Ghosh and Vinay Devarakonda. 2026. FinBench: Time-Gated Calibration and Uncertainty Benchmarking for Agentic Financial Forecasting. arXiv:2607.16229 https://arxiv.org/abs/2607.16229

[22] Isaac Gibbs and Emmanuel J. Candès. 2021. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*, Vol. 34. 1660–1672.

[23] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378. doi:10.1198/016214506000001437

[24] Anubha Goel, Puneet Pasricha, and Juho Kanniainen. 2024. Time-Series Foundation AI Model for Value-at-Risk Forecasting. arXiv:2410.11773 https://arxiv.org/abs/2410.11773

[25] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[26] Dongxin Guo, Jikun Wu, and Siu Ming Yiu. 2026. FinGround: Detecting and Grounding Financial Hallucinations via Atomic Claim Verification. arXiv:2604.23588 https://arxiv.org/abs/2604.23588

[27] Jianfeng He, Linlin Yu, Changbin Li, Runing Yang, Fanglan Chen, Kangshuo Li, Min Zhang, Shuo Lei, Xuchao Zhang, Mohammad Beigi, Kaize Ding, Bei Xiao, Lifu Huang, Feng Chen, Ming Jin, and Chang-Tien Lu. 2026. Survey of Uncertainty Estimation in LLMs: Sources, Methods, Applications, and Challenges. *Information Fusion* 130, Article 104057 (2026). doi:10.1016/j.inffus.2025.104057

[28] Tiansheng Hu, Tongyan Hu, Liuyang Bai, Yilun Zhao, Arman Cohan, and Chen Zhao. 2025. FinTrust: A Comprehensive Benchmark of Trustworthiness Evaluation in Finance Domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 10099–10128. doi:10.18653/v1/2025.emnlp-main.512

[29] Theis Ingerslev Jensen, Bryan Kelly, Semyon Malamud, and Lasse Heje Pedersen. 2026. Machine Learning and the Implementable Efficient Frontier. *The Review of Financial Studies*, Article hhag022 (2026). doi:10.1093/rfs/hhag022

[30] Lanlan Ji, Dominic Seyler, Gunkirat Kaur, Manjunath Hegde, Koustuv Dasgupta, and Bing Xiang. 2025. PHANTOM: A Benchmark for Hallucination Detection in Financial Long-Context QA. In *Advances in Neural Information Processing Systems*, Vol. 38. https://openreview.net/forum?id=5YQAo0S3Hm Datasets and Benchmarks Track.

[31] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. arXiv:2207.05221 https://arxiv.org/abs/2207.05221

[32] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. arXiv:2311.15548 https://arxiv.org/abs/2311.15548

[33] Bryan T. Kelly, Semyon Malamud, Johannes Schwab, and Teng Andrea Xu. 2026. *Scaling Point-in-Time Language Models*. Working Paper 35247. National Bureau of Economic Research. doi:10.3386/w35247

[34] Roger Koenker and Gilbert Bassett. 1978. Regression Quantiles. *Econometrica* 46, 1 (1978), 33–50. doi:10.2307/1913643

[35] Yaxuan Kong, Hoyoung Lee, Yoontae Hwang, Alejandro Lopez-Lira, Bradford Levy, Dhagash Mehta, Qingsong Wen, Chanyeol Choi, Yongjae Lee, and Stefan

Zohren. 2026. Evaluating LLMs in Finance Requires Explicit Bias Consideration. arXiv:2602.14233 https://arxiv.org/abs/2602.14233

[36] Aleksandr Koviazin, Timur Mudarisov, Evgeny Polyachenko, and Radu State. 2026. Reproducibility in the TradingAgents Framework. In *Proceedings of the 2026 International Conference on Artificial Intelligence and Fintech*. Association for Computing Machinery, 365–370. doi:10.1145/3800973.3801029

[37] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=VD-AYtP0dve

[38] Mahesh Kumar, Bhaskarjit Sarmah, and Stefano Pasquali. 2026. FinReflectKG – HalluBench: GraphRAG Hallucination Benchmark for Financial Question Answering Systems. arXiv:2603.20252 [cs.CL] doi:10.48550/arXiv.2603.20252

[39] Paul H. Kupiec. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives* 3, 2 (1995), 73–84. doi:10.3905/jod.1995.407942

[40] Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2023. DEUP: Direct Epistemic Uncertainty Prediction. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=eGLdVRvvfQ

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, Vol. 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[42] Minghan Li, Rachel Gonsalves, Weiyue Li, Sunghoon Yoon, and Mengyu Wang. 2026. TrustTrade: Human-Inspired Selective Consensus Reduces Decision Uncertainty in LLM Trading Agents. arXiv:2603.22567 https://arxiv.org/abs/2603.22567

[43] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. Association for Computing Machinery, 374–382. doi:10.1145/3604237.3626869

[44] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts? arXiv:2504.14765 doi:10.2139/ssrn.5217505

[45] A. Craig MacKinlay. 1997. Event Studies in Economics and Finance. *Journal of Economic Literature* 35, 1 (1997), 13–39.

[46] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 9004–9017. doi:10.18653/v1/2023.emnlp-main.557

[47] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long-Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[48] Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Sean Du, Hamed Hassani, Paul Bogdan, Dawn Song, and Sharon Li. 2026. Uncertainty Quantification in LLM Agents: Foundations, Emerging Challenges, and Opportunities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, San Diego, California, United States, 16219–16250. doi:10.18653/v1/2026.acl-long.738

[49] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems*, Vol. 32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[50] Silu Panda. 2026. FinVerBench: Benchmark Validity and Calibration in Large Language Model Financial Statement Verification. arXiv:2605.29586 https://arxiv.org/abs/2605.29586

[51] Daniel Traian Pele, Vlad Bolovăneanu, Min-Bin Lin, Rui Ren, Andrei Theodor Ginavar, Bruno Spilak, Alexandru-Victor Andrei, Filip-Mihai Toma, Stefan Lessmann, and Wolfgang Karl Härdle. 2026. In the Beginning Was the Word: LLM-VaR and LLM-ES. *Expert Systems with Applications* 295, Article 128676 (2026). doi:10.1016/j.eswa.2025.128676

[52] Manan Poddar. 2026. Uncertainty-Aware Deep Hedging. arXiv:2603.10137 https://arxiv.org/abs/2603.10137

[53] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. 2019. Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*, Vol. 32. 3543–3553. https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html

[54] Ursina Sanderink. 2026. When Alpha Breaks: Two-Level Uncertainty for Safe Deployment of Cross-Sectional Stock Rankers. arXiv:2603.13252 https://arxiv.org/abs/2603.13252

[55] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. *Comput. Surveys* 58, 3, Article 63 (2025), 38 pages. doi:10.1145/3744238

[56] Mainak Singha. 2025. Detecting AI Hallucinations in Finance: An Information-Theoretic Method Cuts Hallucination Rate by 92%. arXiv:2512.03107 https://arxiv.org/abs/2512.03107

[57] Likun Tan, Kuan-Wei Huang, and Kevin Wu. 2025. FRED: Financial Retrieval-Enhanced Detection and Editing of Hallucinations in Language Models. arXiv:2507.20930 https://arxiv.org/abs/2507.20930

[58] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 5433–5442. doi:10.18653/v1/2023.emnlp-main.330

[59] Richard Zhe Wang. 2026. Confidently Wrong: Detecting Hallucinations in Financial Question Answering from LLM Internal States. arXiv:2607.11414 https://arxiv.org/abs/2607.11414

[60] Steven Y. K. Wong, Jennifer S. K. Chan, and Lamiae Azizi. 2025. Quantifying Neural Network Uncertainty under Volatility Clustering. *Neurocomputing* 614, Article 128816 (2025). doi:10.1016/j.neucom.2024.128816

[61] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/adb1d9fa8be4576d28703b396b82ba1b-Abstract-Datasets_and_Benchmarks_Track.html Datasets and Benchmarks Track.

[62] Rui Yao, Qi Chai, Jinhai Yao, Siyuan Li, Junhao Chen, Qi Zhang, and Hao Wang. 2026. Interpreting Fedspeak with Confidence: A LLM-Based Uncertainty-Aware Framework Guided by Monetary Policy Transmission Paths. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, Vol. 40. 34414–34422. doi:10.1609/aaai.v40i40.40739

[63] Yuxuan Ye, Jun Han, Ao Hu, Juncheng Bu, Yiyi Chen, Liangjian Wen, Danilo Mandic, Danny Dongning Sun, Yinghui Xu, and Zenglin Xu. 2026. The Alpha Illusion: Reported Alpha from LLM Trading Agents Should Not Be Treated as Deployment Evidence. arXiv:2605.16895 https://arxiv.org/abs/2605.16895

[64] Taoye Yin, Haoyuan Hu, Yaxin Fan, Xinhao Chen, Xinya Wu, Kai Deng, Kezun Zhang, and Feng Wang. 2026. Mitigating Hallucination in Financial Retrieval-Augmented Generation via Fine-Grained Knowledge Verification. arXiv:2602.05723 https://arxiv.org/abs/2602.05723

[65] Mengao Zhang, Jiayu Fu, Tanya Warrier, Yuwen Wang, Tianhui Tan, and Ke-wei Huang. 2025. FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in Finance. In *Proceedings of the 6th ACM International Conference on AI in Finance*. Association for Computing Machinery, 159–167. doi:10.1145/3768292.3770433

[66] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. 2025. Uncertainty Propagation on LLM Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 6064–6073. doi:10.18653/v1/2025.acl-long.302

[67] Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. FinRAGBench-V: A Benchmark for Multimodal RAG with Visual Citation in the Financial Domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 4215–4249. doi:10.18653/v1/2025.emnlp-main.211

[68] Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. FinDVer: Explainable Claim Verification over Long and Hybrid-content Financial Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 14739–14752. doi:10.18653/v1/2024.emnlp-main.818

[69] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 3277–3287. doi:10.18653/v1/2021.acl-long.254

[70] Jie Zhu, Yimin Tian, Boyang Li, Kehao Wu, Zhongzhi Liang, Junhui Li, Xianyin Zhang, Lifan Guo, Feng Chen, Yong Liu, and Chi Zhang. 2026. FinMCP-Bench: Benchmarking LLM Agents for Real-World Financial Tool Use under the Model Context Protocol. arXiv:2603.24943 https://arxiv.org/abs/2603.24943