Uncertainty-Aware Financial Risk Analysis and Investment Decision Support Using Retrieval-Augmented Large Language Models and Evidence-Based Reasoning

Tobiasz Ksel
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
tobiaszksel@ucf.edu

Ryder Strawbridge
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
ryder631@missouri.edu

Abstract

Large language models (LLMs) offer new capabilities for synthesizing financial disclosures, news, and analyst commentary, yet their deployment in risk analysis and investment decision support remains constrained by hallucination, miscalibrated confidence, and weak traceability to source evidence. This paper develops a methodological framework for uncertainty-aware financial risk analysis that couples retrieval-augmented generation (RAG) with evidence-based reasoning and explicit uncertainty quantification. The central contribution is a composite decision-support architecture in which every generated risk claim is linked to a retrievable evidence set, decomposed into aleatoric and epistemic components, and mapped to a calibrated confidence interval that feeds a risk-sensitive investment rule. We formalize the evidence graph, the uncertainty decomposition, a verifiability-weighted generation objective, and a decision-theoretic portfolio adjustment mechanism. The framework clarifies how retrieval quality, source reliability, and disagreement among evidence items propagate into downstream allocation decisions. Analytical results show that evidence-conditioned uncertainty penalties can reduce exposure to unverifiable claims, that the marginal value of additional retrieval diminishes when evidence redundancy is high, and that calibration across evidence cohorts is a necessary condition for trustworthy decision support. The framework provides a technically defensible path from document retrieval to auditable financial risk judgments without requiring unavailable proprietary data or unverified empirical claims.

Keywords

retrieval-augmented generation; financial risk analysis; uncertainty quantification; evidence-based reasoning; decision support; large language models

1. Introduction

Financial risk analysis is an evidence-intensive activity. Analysts integrate regulatory filings, earnings calls, credit rating actions, macroeconomic releases, and market microdata into judgments about exposure, liquidity, and expected return. The rise of large language models has made it possible to synthesize such heterogeneous textual evidence at scale, but the same generative flexibility that enables synthesis also introduces failure modes that are

unacceptable in risk-sensitive settings. An LLM can produce a fluent risk narrative that is not grounded in any retrieved document, can conflate stale and current evidence, and can express unwarranted confidence when the underlying evidence is sparse or contradictory [1], [2]. These problems are not cosmetic. In investment decision support, a miscalibrated risk score can translate directly into inappropriate position sizing, inadequate hedging, or delayed risk recognition.

Retrieval-augmented generation has emerged as a partial remedy by conditioning generation on external documents rather than relying solely on parametric memory [3], [4]. RAG systems retrieve relevant passages from a corpus, inject them into the model context, and generate responses that are intended to be grounded in those passages. In finance, this architecture has been explored for question answering over filings, news summarization, and compliance review. However, retrieval alone does not guarantee evidential fidelity. Retrieved documents may be irrelevant, contradictory, outdated, or of heterogeneous reliability. A model may still ignore retrieved evidence, overgeneralize from a single passage, or fail to represent the uncertainty implied by disagreement across sources. The gap is therefore not merely retrieval but evidence-aware uncertainty representation: the system must know what it knows, how well the evidence supports a claim, and how that support should modulate a financial decision.

This paper addresses that gap by developing a methodological framework for uncertainty-aware financial risk analysis and investment decision support using retrieval-augmented LLMs and evidence-based reasoning. The central contribution is a composite architecture and analytical formalism that links evidence retrieval, claim verification, uncertainty decomposition, and risk-sensitive decision rules. The framework is designed to be auditable: each generated risk claim is associated with an evidence set, a verifiability score, and a calibrated uncertainty interval. The decision layer then translates these quantities into exposure adjustments rather than treating model output as a point estimate.

The paper makes three specific contributions. First, it formalizes an evidence graph that represents documents, claims, and support relations, and defines a verifiability measure that penalizes unsupported generation. Second, it decomposes predictive uncertainty into aleatoric and epistemic components conditioned on evidence quality, and specifies how these components should be estimated in a RAG setting without requiring access to proprietary labeled datasets. Third, it derives a decision rule that maps uncertainty-adjusted risk forecasts into portfolio exposure, making explicit the trade-off between expected return and evidence-contingent risk. The analysis is analytical and methodological rather than empirical; it develops design choices, derivable consequences, and evidence-supported comparisons rather than reporting experimental results.

The remainder of the paper is organized as follows. Section 2 reviews related work on RAG, uncertainty quantification, and evidence-based reasoning in finance. Section 3 presents the methodological framework, including the evidence graph, uncertainty decomposition, verifiable generation objective, and decision rule. Section 4 provides a technical analysis of calibration, risk-return trade-offs, and sensitivity to retrieval quality. Section 5 discusses implementation conditions, limitations, and comparisons with alternative approaches. Section 6 concludes.

2. Background and Related Work

2.1 Retrieval-Augmented Generation and Its Financial Applications

RAG combines a retriever, which selects relevant passages from a corpus, with a generator, which conditions its output on those passages [3]. The standard formulation treats retrieval as a latent variable and trains the generator to maximize the likelihood of a target response given retrieved context. In domain-specific settings, retrieval can be sparse or dense, and the corpus may include structured filings, unstructured news, and analyst reports. Financial applications have emphasized filing question answering, earnings call summarization, and risk factor extraction. The appeal of RAG in finance is straightforward: regulatory and market documents change frequently, and parametric models cannot reliably encode the latest disclosures.

However, the financial domain imposes requirements that general-purpose RAG systems do not satisfy. First, source reliability varies enormously across document types. A audited annual report and a social media post are not equivalent evidence. Second, temporal validity matters: a risk factor disclosed two years ago may no longer describe current exposure. Third, contradictory evidence is common, and a useful system must represent disagreement rather than averaging it away. Fourth, regulatory and internal governance contexts require traceability from claim to source. These requirements motivate an evidence-based reasoning layer that goes beyond passage retrieval.

2.2 Uncertainty Quantification for Language Models

Uncertainty quantification for neural models distinguishes aleatoric uncertainty, which arises from inherent noise or ambiguity in the data-generating process, from epistemic uncertainty, which arises from limited knowledge or model ignorance [5], [6]. In LLM settings, epistemic uncertainty is often associated with regions of input space where the model lacks sufficient training signal, while aleatoric uncertainty reflects ambiguity in the prompt or evidence. Bayesian approaches, ensembles, and conformal prediction provide complementary tools for estimating these components [7], [8].

In RAG systems, uncertainty has an additional evidence-conditioned dimension. The model's confidence should depend not only on its parametric knowledge but on the quality, coverage, and consistency of retrieved evidence. Recent work on uncertainty-aware RAG has begun to address trustworthiness and risk control in generation, emphasizing the need to calibrate confidence against retrieval quality and to detect when retrieved evidence is insufficient [16]. This perspective aligns with the financial requirement that a risk claim be accompanied by an explicit statement of evidential support.

2.3 Evidence-Based Reasoning and Decision Support

Evidence-based reasoning in medicine and policy provides a useful analogy. The GRADE framework, for example, separates the quality of evidence from the strength of recommendations, explicitly acknowledging that recommendations may be weak even when evidence is moderate [9]. In finance, analogous distinctions appear in credit rating methodologies and risk factor taxonomies, where the weight assigned to a signal depends on data quality and model validation. Decision-theoretic approaches to investment under uncertainty formalize the mapping from beliefs to actions through expected utility or risk-sensitive objectives [10], [11].

The framework developed here integrates these strands. It treats retrieved evidence as a set of uncertain support relations, decomposes uncertainty into components that are interpretable to risk managers, and maps the resulting belief state into an investment decision through a risk-sensitive rule. The approach is consistent with the broader literature on

trustworthy machine learning, which emphasizes calibration, robustness, and explainability as prerequisites for high-stakes deployment [12], [13].

3. Methodological Framework

3.1 Problem Definition and Notation

Let q denote a financial risk query, such as an assessment of a firm's liquidity risk or sector exposure. Let D be a document corpus containing financial disclosures, news articles, and analyst reports. A retriever selects a set of passages E = {e1, ..., en} from D that are relevant to q. A generator produces a risk claim c and a supporting rationale conditioned on q and E. The decision layer maps c and its associated uncertainty to a portfolio exposure vector w.

The evidence graph G = (V, R) represents documents, claims, and support relations. Vertices V include document nodes, passage nodes, and claim nodes. Edges R encode support, contradiction, and temporal precedence relations. Each document node carries a reliability weight r(d) in [0, 1], which may be assigned by source type, editorial review, or regulatory status. Each support edge carries a strength s(e, c) in [0, 1], representing the degree to which passage e supports claim c. The evidence graph provides the structural basis for verifiability assessment.

3.2 Verifiability-Weighted Generation

A central design choice is to make generation explicitly accountable to the evidence graph. Rather than optimizing only for fluency and relevance, the generator is trained or prompted to produce claims whose support in G exceeds a threshold. We define the verifiability score of a claim c as a reliability-weighted aggregate of its incoming support edges:

$$V(c) = (\sum_\{e \in S(c)\} r(d(e)) \cdot s(e, c)) / (1 + \sum_\{e \in S(c)\} r(d(e)))$$

Here S(c) is the set of passages that support claim c, d(e) is the document containing passage e, r(d(e)) is the reliability weight of that document, and s(e, c) is the support strength. The denominator includes a smoothing term that prevents a single weak support edge from producing a high verifiability score. The score V(c) lies in [0, 1] and increases with both the reliability of supporting documents and the strength of support.

The verifiability score contributes to the generation objective by penalizing claims with low V(c). If L_gen is a standard generation loss, such as negative log-likelihood of the target response, the evidence-aware objective is:

$$L_total = L_gen + \lambda \cdot max(0, \tau - V(c))$$

where $\lambda > 0$ controls the penalty strength and τ is a minimum verifiability threshold. This objective is analytically motivated rather than empirically fitted; it encodes the requirement that a claim below the evidence threshold should incur additional loss. The threshold τ can be set by governance policy, reflecting the organization's tolerance for unsupported claims.

3.3 Uncertainty Decomposition

The uncertainty associated with a risk claim c has two primary components. Aleatoric uncertainty U_a reflects ambiguity in the evidence itself, such as contradictory passages or imprecise language. Epistemic uncertainty U_e reflects the model's lack of knowledge about the claim, which may arise from sparse retrieval, weak support, or restricted coverage of the corpus. We propose the following decomposition:

$$U(c) = \alpha \cdot U_a(c) + \beta \cdot U_e(c) + \gamma \cdot \big(1 - V(c)\big)$$

where α, β, and γ are nonnegative weights that sum to one. The term (1 - V(c)) is an explicit evidence-deficit penalty; it captures the intuition that a claim with low verifiability is inherently uncertain regardless of the model's internal confidence. Aleatoric uncertainty can be estimated from disagreement among retrieved passages, for example by measuring the variance of sentiment or risk-factor polarity across passages that mention the same entity. Epistemic uncertainty can be estimated from retrieval coverage, such as the fraction of query-relevant document types represented in E, or from ensemble disagreement across multiple generator configurations.

The decomposition is useful because it supports differentiated risk responses. High aleatoric uncertainty suggests that the evidence is ambiguous and that the claim should be treated with caution but not necessarily discarded. High epistemic uncertainty suggests that the system lacks sufficient knowledge and that additional retrieval or human review is warranted. The weighted combination U(c) provides a scalar that can be compared across claims and used in the decision layer.

3.4 Calibrated Confidence Intervals

A risk claim is more useful when accompanied by an interval rather than a point estimate. Let μ(c) be the central estimate of a risk quantity, such as a probability of default or a volatility forecast, derived from the generated claim and any structured data available in the evidence set. We define a calibrated interval:

$$I(c) = \left[\mu(c) - z_{1-\frac{\delta}{2}} \cdot \sigma(c), \mu(c) + z_{1-\frac{\delta}{2}} \cdot \sigma(c)\right]$$

where σ(c) is a scale parameter derived from U(c) and δ is the desired coverage level. The multiplier z_{1-δ/2} is the standard normal quantile for a two-sided interval. This formulation treats U(c) as a variance-like quantity after appropriate transformation; in practice, σ(c) can be set as a monotone function of U(c), such as σ(c) = σ_0 · exp(κ · U(c)), where σ_0 is a baseline scale and κ > 0 controls the sensitivity of the interval width to uncertainty.

The interval I(c) is not claimed to be empirically calibrated in this paper. Rather, it provides a formal mechanism for propagating evidence-conditioned uncertainty into downstream decisions. Calibration would require out-of-sample evaluation, which is outside the scope of this methodological treatment. The framework specifies the conditions under which calibration should be assessed, namely across evidence cohorts defined by document type, retrieval coverage, and claim verifiability.

3.5 Risk-Sensitive Decision Rule

The decision layer maps the uncertain risk estimate into a portfolio exposure. Let w_0 be a baseline exposure derived from a standard risk model, such as a mean-variance or factor-model allocation. The evidence-aware adjustment reduces exposure when uncertainty is high and increases it when evidence strongly supports the risk estimate. A simple and interpretable rule is:

$$w(c) = w_0 \cdot (1 - \eta \cdot U(c)) \cdot 1\{V(c) \geq \tau\}$$

where η in [0, 1] controls the sensitivity of exposure to uncertainty, and 1{·} is an indicator function that sets exposure to zero when verifiability falls below the governance threshold τ. This rule has three desirable properties. First, it is monotone decreasing in

uncertainty, so higher uncertainty always implies lower exposure, all else equal. Second, it enforces a hard constraint on unverifiable claims, which is consistent with auditability requirements. Third, it separates the evidence threshold from the continuous uncertainty penalty, allowing governance policy to set τ independently of the risk model's uncertainty sensitivity η.

The decision rule can be extended to a portfolio context by aggregating adjusted exposures across claims and applying a budget constraint. If claims are indexed by i and baseline exposures are w_{0,i}, the portfolio-level exposure is:

$$W = \sum_i w_{0,i} \cdot (1 - \eta \cdot U(c_i)) \cdot 1\{V(c_i) \geq \tau\}$$

subject to $\sum_i |w_{0,i}| \leq B$, where B is a risk budget. This aggregation preserves the evidence-contingent structure at the portfolio level and makes explicit that uncertainty penalties compound across claims.

4. Technical Analysis

4.1 Calibration Across Evidence Cohorts

The decision rule in Equation (5) depends on the calibration of U(c) and V(c). If U(c) is systematically understated for a particular evidence cohort, such as claims supported only by news articles, the uncertainty penalty will be too small and exposure will be too high. Conversely, if U(c) is overstated for high-reliability cohorts, the system will be overly conservative. Calibration should therefore be assessed conditional on evidence cohort, where a cohort is defined by document type, temporal recency, and retrieval coverage.

A useful analytical condition is that the expected coverage of I(c) should approximate 1 - δ within each cohort. If coverage is below the nominal level, the interval is too narrow and the uncertainty scale σ(c) should be increased. If coverage is above the nominal level, the interval is conservative and exposure is unnecessarily reduced. This condition is standard in conformal prediction and interval calibration [7], [8]. The framework does not assume that calibration holds automatically; it specifies the diagnostic that must be performed before deployment.

4.2 Marginal Value of Retrieval

Retrieval is costly, both computationally and in terms of latency. It is therefore useful to characterize when additional retrieval improves decision quality. Let $U_e(n)$ be the epistemic uncertainty as a function of the number of retrieved passages n. In many retrieval settings, $U_e(n)$ decreases with n but exhibits diminishing returns once the evidence set covers the relevant document types. A simple model is:

$$U_e(n) = U_\infty + (U_0 - U_\infty) \cdot exp(-\rho \cdot n)$$

where U_0 is the epistemic uncertainty with no retrieval, U_∞ is the irreducible floor, and $\rho > 0$ is a retrieval efficiency parameter. This model implies that the marginal reduction in epistemic uncertainty is proportional to the remaining gap $U_e(n) - U_\infty$. When evidence redundancy is high, ρ is small and additional retrieval yields little benefit. When the corpus contains diverse and complementary evidence, ρ is larger.

The decision-theoretic implication is that retrieval should continue until the marginal reduction in U(c) no longer justifies its cost. If C_{ret} is the cost per retrieved passage and L(w, c) is the loss associated with a decision error, the optimal retrieval depth satisfies:

$$\partial L/\partial U \cdot |\partial U/\partial n| \leq C_ret$$

at the margin. This condition is analytical rather than empirical; it clarifies the trade-off that any deployed system must resolve. In practice, the threshold can be set by policy or learned from operational data, but the framework makes the trade-off explicit.

4.3 Sensitivity to Source Reliability

The verifiability score V(c) is sensitive to the reliability weights r(d). If a claim is supported only by low-reliability documents, V(c) will be low even if the support strength s(e, c) is high. This is intentional: a claim that is strongly supported by an unreliable source should not be treated as well-evidenced. The sensitivity of V(c) to r(d) can be analyzed by considering the elasticity of V with respect to r. For a claim with a single support edge, the elasticity is:

$$E_r = (\partial V/\partial r) \cdot (r/V) = (1 - V) / (1 + r)$$

which is positive and decreasing in V. This means that reliability improvements have the largest effect on claims with low verifiability, which are precisely the claims that pose the greatest risk. The implication for system design is that effort spent on improving source reliability classification is most valuable for marginal claims, not for claims that are already well-supported.

4.4 Failure Modes and Boundary Conditions

The framework has several boundary conditions that must be respected. First, the evidence graph must be complete enough to represent the relevant support relations. If the retriever misses a crucial document, V(c) will be understated and the decision rule will be overly conservative. Second, the reliability weights must be meaningfully assigned. If all documents are assigned the same weight, the verifiability score reduces to a support-strength average and loses its ability to discriminate by source quality. Third, the uncertainty decomposition assumes that aleatoric and epistemic components can be separately estimated; in practice, the estimates may be correlated, and the weights α, β, γ may require recalibration.

A further failure mode is adversarial evidence. If a malicious actor inserts a fabricated document into the corpus, the retriever may surface it and the generator may treat it as support. The evidence graph mitigates this risk only if reliability weights are assigned by provenance and if the system can detect inconsistencies across sources. Temporal precedence relations can help: a claim that contradicts a more recent high-reliability document should be flagged. These mechanisms are not foolproof, and the framework should be understood as a structured approach to risk reduction rather than a guarantee of correctness.

5. Discussion

5.1 Comparison with Alternative Approaches

The proposed framework differs from standard RAG in its explicit treatment of evidence quality and uncertainty. A standard RAG system retrieves passages and generates a response, but it does not necessarily represent the reliability of those passages or the disagreement among them. The evidence graph and verifiability score add a structural layer that makes these factors explicit. Compared with fine-tuning a domain-specific LLM, the framework has the advantage of traceability: each claim can be linked to a retrievable source, which is essential for audit and governance. Fine-tuned models may achieve higher fluency and

domain accuracy, but they are less transparent and more difficult to update when new evidence arrives.

Compared with traditional quantitative risk models, the framework is complementary rather than substitutive. Quantitative models provide calibrated estimates of market risk, credit risk, and liquidity risk under specific assumptions. The RAG-based framework is best suited for synthesizing textual evidence that is not easily quantifiable, such as management guidance, regulatory commentary, and emerging risk narratives. The decision rule in Equation (5) is designed to adjust exposures derived from quantitative models, not to replace them.

5.2 Implementation Conditions

Successful implementation requires several organizational and technical conditions. The corpus must be curated and maintained, with clear provenance metadata and reliability labels. The retriever must be evaluated for recall and precision on financial queries, and the generator must be tested for faithfulness to retrieved evidence. The uncertainty estimates must be calibrated on historical data or expert judgments, and the decision rule must be parameterized in consultation with risk management. Governance must define the verifiability threshold τ and the uncertainty sensitivity η, and these parameters should be reviewed periodically.

The framework also has implications for model selection. Smaller, more controllable models may be preferable to larger, more fluent models when verifiability is the primary requirement. Retrieval quality may matter more than generator size, because the evidence graph and uncertainty decomposition depend on the retrieved set. This is consistent with the broader observation that RAG performance is often limited by retrieval rather than generation [14], [15].

5.3 Limitations

The framework is analytical and does not report empirical validation. Its claims are conditional: if the evidence graph is well-constructed and the uncertainty estimates are calibrated, then the decision rule will produce auditable and risk-sensitive exposures. The framework does not specify how to estimate aleatoric and epistemic uncertainty in all cases, and it does not provide a complete solution to adversarial evidence. These are open problems that require further research and operational experience.

The uncertainty decomposition in Equation (3) is a modeling choice, not a uniquely correct representation. Alternative decompositions may be appropriate for different risk domains. The decision rule in Equation (5) is intentionally simple to preserve interpretability; more complex rules, such as those based on expected utility or coherent risk measures, could be substituted without changing the overall architecture. The framework's value lies in the integration of evidence, uncertainty, and decision, not in the specific functional forms.

6. Conclusion

This paper has developed a methodological framework for uncertainty-aware financial risk analysis and investment decision support using retrieval-augmented large language models and evidence-based reasoning. The framework addresses a concrete problem: LLM-generated risk claims are often fluent but poorly grounded, and their confidence is not systematically tied to the quality of retrieved evidence. The proposed architecture links each claim to an evidence graph, computes a verifiability score, decomposes uncertainty into aleatoric and epistemic components, and maps the resulting belief state into a risk-sensitive exposure rule.

The central insight is that evidence-based reasoning and uncertainty quantification are not optional add-ons to RAG in finance; they are prerequisites for trustworthy decision support. The verifiability-weighted generation objective penalizes unsupported claims, the uncertainty decomposition supports differentiated risk responses, and the decision rule enforces a hard constraint on unverifiable claims while continuously adjusting exposure to evidence-conditioned uncertainty. Analytical results clarify the conditions under which the framework improves decision quality, including the importance of calibration across evidence cohorts, the diminishing marginal value of additional retrieval, and the sensitivity of verifiability to source reliability.

The framework is intended to be auditable and implementable. It does not require proprietary datasets or unavailable model internals, and it can be integrated with existing quantitative risk models. Future work should focus on empirical calibration, adversarial robustness, and the extension of the evidence graph to multi-modal financial data. The broader implication is that the value of LLMs in finance will depend less on generative fluency than on the discipline with which evidence, uncertainty, and decisions are connected.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

2. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

3. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J. B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. Proceedings of the 39th International Conference on Machine Learning, 162, 2206–2240.

4. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

5. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

6. Gal, Y. (2016). Uncertainty in deep learning (Doctoral dissertation). University of Cambridge.

7. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

8. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

9. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ, 336(7650), 924–926.

10. Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77–91.

11. Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. Mathematical Finance, 9(3), 203–228.

12. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

13. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330.

14. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781.

15. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 874–880.

16. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 526-531). IEEE.

17. Thorp, E. O. (1969). Optimal gambling systems for favorable games. Revue de l'Institut International de Statistique, 37(3), 273–293.

18. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

19. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168.

20. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

21. Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. Management Science, 64(6), 2833–2855.