

Retrieval-Augmented Generation for Financial Intelligence: A Framework for Evidence-Based Decision-Making

Kennedy Ayanga

Department of Computer Science, Maseno University, P.O. Box 333 40105, Maseno, Kenya
kennedy.ayangal@maseno.ac.ke

Abstract

Financial decisions must rest on evidence: what a filing says, what a policy requires, what a contract permits, and what the market has reported. Large language models (LLMs) can read and synthesize such material quickly, but on their own they answer from memory, which may be outdated, incomplete, or invented. Retrieval-augmented generation (RAG) addresses this by retrieving relevant documents and grounding the model's output in them. In finance, however, grounding is not enough. The evidence must be current, authoritative, permitted for the user, numerically accurate, cited, and traceable into the decision it supports. This article proposes a framework for RAG-based financial intelligence built around evidence quality. It combines event-driven knowledge ingestion, an evidence-aware index, entitlement-aware and time-aware retrieval, deliberate context assembly, citation-bound generation guided by prompt patterns, a verification layer with guardrails and numerical recomputation, workflow integration through evidence packs, and continuous evaluation. Using a design-oriented approach grounded in ten studies published between 2020 and 2024, the article maps the evidence base and presents the framework and its evidence chain. It defines evidence quality criteria, failure modes with their controls, and evaluation metrics. It illustrates the framework with an asset manager's investment committee reviewing a holding after an earnings release and a credit downgrade. It argues that RAG supports sound financial decisions only when every claim can be traced to permitted, current, and verified evidence.

Keywords: retrieval-augmented generation, financial intelligence, large language models, evidence-based decision-making, citation, verification, prompt engineering, guardrails, event-driven architecture, enterprise workflow

1. Introduction

Financial professionals work through documents. An investment analyst reads earnings releases and regulatory filings. A credit officer reads loan agreements and financial statements.

A compliance officer reads regulations and internal policies. A risk manager reads incident reports and market news. The quality of their decisions depends on finding the right evidence, interpreting it correctly, and being able to show later what the decision was based on.

LLMs can read, summarize, and answer questions about such documents far faster than people. Used directly, however, they answer from patterns learned during training, which may be outdated or wrong, and they cannot show where an answer came from. Lewis et al. (2020) introduced retrieval-augmented generation, in which a retriever selects relevant documents and a generator conditions its output on them. This improves factual accuracy and allows knowledge to be updated without retraining. Dense passage retrieval (Karpukhin et al., 2020) made retrieval of relevant passages accurate enough to support this approach at scale.

Finance places particular demands on RAG. Information changes constantly, so knowledge must be updated as events occur (Sannidhanam, 2021). Much financial information is numerical and tabular, and numerical reasoning over reports remains hard for language models (Chen et al., 2021). Models do not use all retrieved context equally, and information in the middle of long contexts tends to be overlooked (Liu et al., 2023). Decision-makers and auditors need citations they can check (Gao et al., 2023). Outputs must be reliable (Sannidhanam, 2024a) and safe (Sannidhanam, 2024b), and consequential decisions must be governed (Basireddy, 2022). RAG systems also need to be evaluated continuously (Es et al., 2023).

This article asks: *How should RAG be designed so that financial decisions based on its outputs are grounded in evidence that is current, authoritative, permitted, accurate, and traceable?* It contributes a mapping of the evidence base (Figures 1 and 2), a framework architecture (Figure 3), an evidence chain from question to decision (Figure 4), evidence quality criteria, a catalog of failure modes and controls, evaluation metrics, and an illustrative application.

2. Related Work

2.1 Retrieval foundations

Lewis et al. (2020) combined a pre-trained retriever with a sequence-to-sequence generator and trained them together. The approach outperformed purely parametric models on knowledge-intensive tasks and produced more specific and factual text. Karpukhin et al. (2020) showed that dense passage retrieval, using learned vector representations of questions and passages, clearly outperformed traditional keyword-based retrieval on open-domain question answering. Keyword methods remain useful for exact terms such as ticker symbols, clause numbers, and regulatory references. Financial RAG systems therefore benefit from combining both.

2.2 How models use retrieved context

Liu et al. (2023) found that language models use information best when it appears at the beginning or end of a long context and considerably worse when it sits in the middle. Adding more retrieved documents does not guarantee better answers. The selection, ordering, and amount of context all matter. Gao et al. (2023) built a benchmark for LLMs that generate text with citations and evaluated both whether citations actually support the statements they accompany and how complete the supporting citations are. They found that even strong models often produce citations that do not fully support their claims, which shows that citations themselves need checking.

2.3 Financial reasoning

Chen et al. (2021) introduced FinQA, which requires numerical reasoning over financial reports that combine text and tables. Models of the time performed well below human experts. Financial RAG must therefore treat tables and figures carefully and should not rely on the model's own arithmetic.

2.4 Evaluating RAG

Es et al. (2023) proposed RAGAS, a framework for evaluating RAG systems without human-written reference answers. It measures faithfulness (whether the answer is supported by the retrieved context), answer relevance (whether it addresses the question), and context relevance (whether the retrieved material is focused on what was asked). Such metrics allow RAG quality to be monitored continuously rather than assessed once.

2.5 Real-time knowledge ingestion

Sannidhanam (2021) examines real-time claims processing using event-driven architectures, in which business events are published to a stream and processed by independent services as they occur. The same pattern suits financial knowledge. A new filing, price move, rating action, policy update, or news story is an event that should update the retrieval index within minutes, not at the next batch run.

2.6 Prompt patterns and guardrails

Sannidhanam (2024a) sets out prompt engineering patterns for reliable LLM outputs in production environments. These include clear role and task instructions, structured output formats, grounding in supplied context, explicit constraints, and validation of outputs. Sannidhanam (2024b) examines guardrails and safety mechanisms for LLM-powered enterprise applications, including input filtering, output validation, restrictions on topics and actions, protection of sensitive data, and escalation to humans. Together, they define how a RAG system should instruct the model and what checks its outputs must pass.

2.7 Decision workflows

Basireddy (2022) proposes configurable enterprise workflow architecture for digitizing risk management, with defined steps for assessment, approval, escalation, and remediation and complete audit trails. For evidence-based decision-making, RAG outputs should enter such workflows together with their evidence, so that decisions can be reviewed and later reconstructed.

2.8 Research gap

These strands have developed separately, leaving limited guidance on combining them so that RAG supports evidence-based financial decisions.

3. Research Methodology

The study follows a design-oriented research approach. The problem, that grounding alone does not guarantee that RAG outputs meet the evidence standards of financial decision-making, was identified from the literature. Requirements were derived from the reviewed studies: timely knowledge, suitable retrieval, deliberate context assembly, verifiable citations, numerical accuracy, reliable and safe outputs, governed decisions, and continuous evaluation. The framework was designed from ten studies and demonstrated through an illustrative investment committee case.

To make the evidence base transparent, each study was assigned to the theme it primarily supports. Figure 1 shows the distribution across themes, and Figure 2 shows the studies by publication year. These figures describe the literature used in this article. They are not performance measurements. The framework has not been evaluated empirically, which is left to future work.

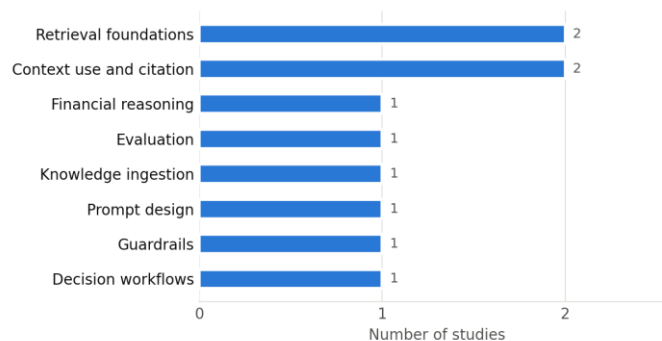


Figure 1. Number of reviewed studies primarily supporting each theme (n = 10)

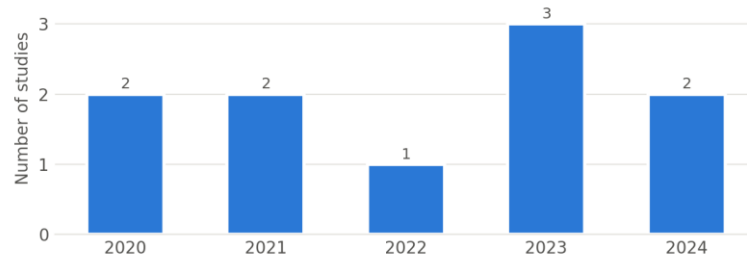


Figure 2. Reviewed studies by publication year (n = 10)

The evidence is spread across the full RAG pipeline, with no single dominant theme, and most studies date from 2020 onward, following the introduction of RAG. Evaluation and financial reasoning each rest on one study, which points to a need for evaluation methods designed specifically for financial RAG.

4. Framework

The framework has two pipelines (Figure 3). The knowledge pipeline keeps evidence current and well organized. The decision pipeline turns a question into a verified, cited answer and then into a governed decision. Table 1 describes the components.

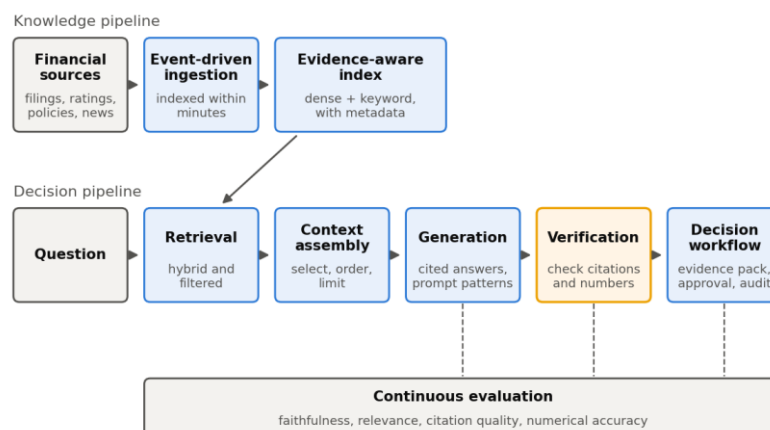


Figure 3. Framework for RAG-based financial intelligence

Table 1. Components of the RAG financial intelligence framework

Component	Function	Key mechanisms	Supporting studies
Event-driven ingestion	Keep knowledge current	New filings, prices, ratings, policies, and news processed as events	Sannidhanam (2021)

Component	Function	Key mechanisms	Supporting studies
Evidence-aware index	Organize evidence for retrieval	Dense and keyword indexes, table-aware chunking, metadata for source, date, authority, and entitlement	Karpukhin et al. (2020); Chen et al. (2021)
Retrieval	Find relevant, permitted, current evidence	Hybrid search, entitlement filters, time filters	Lewis et al. (2020); Karpukhin et al. (2020)
Context assembly	Present evidence effectively to the model	Selection, deduplication, ordering, size limits	Liu et al. (2023)
Citation-bound generation	Produce answers tied to evidence	Prompt patterns, required citations, structured output	Sannidhanam (2024a); Gao et al. (2023)
Verification	Check answers before use	Citation support checks, numerical recomputation, guardrails	Gao et al. (2023); Chen et al. (2021); Sannidhanam (2024b)
Decision workflow	Route answers into governed decisions	Evidence packs, approvals, audit trail	Basireddy (2022)
Evaluation	Monitor quality over time	Faithfulness, answer relevance, context relevance, citation quality	Es et al. (2023); Gao et al. (2023)

Event-driven ingestion. Each new document or data point is published as an event and indexed within minutes (Sannidhanam, 2021). Examples include an earnings release, a rating change, a regulatory notice, or a revised internal policy. Superseded documents are marked as such rather than deleted, so past decisions can still be reconstructed.

Evidence-aware index. Documents are split into passages, and tables are kept intact with their headings so that figures are not separated from their labels (Chen et al., 2021). Each passage carries metadata: source, publication date, authority level (for example, an audited filing versus a news report), and entitlement, meaning who may see it. Dense and keyword indexes are maintained side by side (Karpukhin et al., 2020).

Retrieval and context assembly. Queries search both indexes, filtered by the user's entitlements and by date (Lewis et al., 2020). The results are deduplicated, limited to the most relevant, and ordered so that the most important evidence is not buried in the middle of the context (Liu et al., 2023).

Citation-bound generation. The model is instructed, using structured prompt patterns (Sannidhanam, 2024a), to answer only from the supplied evidence, to cite a passage for every factual claim, to state when the evidence is insufficient, and to return its output in a fixed structure.

Verification. Each citation is checked to confirm that the cited passage supports the claim (Gao et al., 2023). Any figure the model reports or derives is recomputed from the source table by a calculation tool (Chen et al., 2021). Guardrails check for disclosure of restricted information,

unsupported recommendations, and prohibited content, and escalate failures (Sannidhanam, 2024b).

Decision workflow. Verified answers enter configurable workflows as evidence packs containing the question, answer, citations, source documents, verification results, and model and index versions (Basireddy, 2022). Decision-makers review the pack, decide, and record their reasoning.

Evaluation. Samples of answers are scored continuously and falling scores trigger investigation (Section 5.3).

5. The Evidence Chain

Figure 4 shows the chain from question to decision. Each link must hold for the decision to be evidence-based.

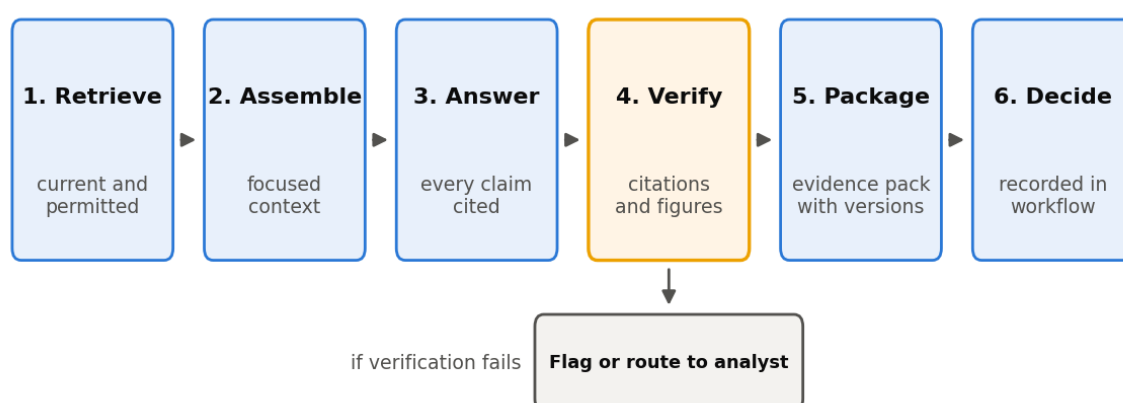


Figure 4. Evidence chain from question to decision

A question is retrieved against current, permitted evidence. The evidence is assembled into a focused context. The model answers with citations. The answer is verified by citation checks, recomputation, and guardrails. The verified answer and its evidence are packaged, and the decision-maker decides within a workflow that records the evidence and the reasoning. If verification fails, the answer is returned with its problems flagged or passed to an analyst instead.

5.1 Evidence quality criteria

Not all retrieved evidence is equally good. Table 2 sets out the criteria the framework applies.

Table 2. Evidence quality criteria for financial RAG

Criterion	Question	Mechanism
Relevance	Does the evidence address the question?	Hybrid retrieval, context relevance scoring
Timeliness	Is it the latest applicable version?	Event-driven ingestion, date filters, superseded flags
Authority	How reliable is the source?	Authority metadata, preference for audited and official sources
Entitlement	Is the user permitted to see it?	Entitlement metadata and filtering
Accuracy	Are the figures correctly reported and derived?	Table-aware indexing, numerical recomputation
Traceability	Can each claim be traced to its source?	Required citations, citation support checks, evidence packs

5.2 Failure modes and controls

Table 3 lists common failure modes of RAG in finance and the controls that address them.

Table 3. Failure modes of financial RAG and their controls

Failure mode	Example	Control	Supporting studies
Stale evidence	Answer uses last quarter's figures after new results	Event-driven ingestion, date filters	Sannidhanam (2021)
Missed evidence	Relevant clause not retrieved because of different wording	Hybrid dense and keyword retrieval	Karpukhin et al. (2020)
Ignored evidence	Key passage in the middle of a long context overlooked	Context limits and ordering	Liu et al. (2023)
Unsupported citation	Citation attached to a passage that does not support the claim	Citation support checks	Gao et al. (2023)
Calculation error	Wrong ratio derived from a table	Recomputation by calculation tools	Chen et al. (2021)
Restricted disclosure	Confidential deal information shown to an unauthorized user	Entitlement filters, output guardrails	Sannidhanam (2024b)
Inconsistent output	Answers vary in format and completeness	Structured prompt patterns	Sannidhanam (2024a)
Ungoverned decision	Action taken on an answer without review	Workflow approval with evidence pack	Basireddy (2022)

5.3 Evaluation metrics

Following Es et al. (2023) and Gao et al. (2023), the framework monitors four metrics. Faithfulness is the share of claims supported by the retrieved evidence. Answer relevance is how well the answer addresses the question. Context relevance is how focused the retrieved evidence is. Citation quality is whether citations support their claims and whether all claims are cited. For financial use, a fifth metric is added: numerical accuracy, the share of reported figures that match recomputation from source.

6. Illustrative Case: Investment Committee Review

Consider an asset management firm whose investment committee must decide whether to keep, reduce, or sell a large holding in an industrial company. On the same morning, the company has released quarterly results and a rating agency has downgraded its debt.

Ingestion. The earnings release, the rating agency announcement, and several news reports are published as events and indexed within minutes (Sannidhanam, 2021). The previous quarter's figures are marked as superseded. The firm's internal investment policy, including its limits on holdings below a certain credit rating, is already in the index.

Question and retrieval. A portfolio manager asks: "What changed in the company's results and credit position, and does the holding still comply with our investment policy?" Retrieval combines dense and keyword search (Karpukhin et al., 2020), filtered to the current quarter and to documents the manager is permitted to see. It returns passages from the earnings release, the rating announcement, and the policy clause on credit rating limits.

Assembly and answer. The context places the rating announcement and the policy clause first, then the key results table (Liu et al., 2023). Following the prompt pattern (Sannidhanam, 2024a), the model produces a structured answer. Revenue fell compared with the same quarter last year, net debt rose, and the rating was lowered by one notch. Under the policy, holdings at the new rating are still permitted but capped at a lower share of the portfolio than the fund currently holds. Every statement is cited.

Verification. The citation check confirms each claim is supported (Gao et al., 2023). The calculation tool recomputes the revenue change and the fund's current weight from the source table and the portfolio data (Chen et al., 2021). The recomputed weight differs slightly from the model's figure because of an intraday price move, so the answer is updated with the recomputed value. Guardrails confirm that no restricted information from another client's mandate has been included (Sannidhanam, 2024b).

Decision. The answer, its citations, the source documents, and the verification results form an evidence pack routed to the investment committee through the decision workflow (Basireddy, 2022). The committee decides to reduce the holding to within the policy cap and records its reasoning. Months later, when the firm's auditors review compliance with investment limits, the evidence pack shows exactly what was known, when, and on what basis the decision was taken.

Evaluation. The discrepancy corrected by recomputation is recorded in the evaluation sample (Es et al., 2023), prompting a change to include live portfolio data in future answers about weights.

The case shows how each part of the framework contributes to a decision that is quick, grounded, and fully traceable.

7. Discussion

7.1 Implications

For research, the framework reframes financial RAG around evidence quality rather than answer quality alone. It links retrieval, context, citation, and verification research to the requirements of timeliness, authority, entitlement, accuracy, and traceability that financial decisions impose. The evidence mapping suggests that evaluation methods designed for financial RAG are an important gap. For practitioners, the main lessons are to update knowledge as events occur, require and check citations, recompute every figure, and deliver answers as evidence packs within governed workflows.

7.2 Limitations and future work

The framework was demonstrated through an illustrative case and draws on a focused set of ten studies, and the evidence mapping reflects only these studies. Future work should implement and evaluate it, measuring faithfulness, citation quality, numerical accuracy, retrieval of current versus superseded evidence, analyst time saved, and auditor assessment of evidence packs. Further research is needed on evaluation benchmarks for financial RAG, on retrieval over complex tables and linked documents, on balancing context size against the lost-in-the-middle effect, and on regulatory expectations for AI-assisted evidence in financial decisions.

8. Conclusion

RAG can give financial professionals fast, grounded answers from the documents that matter. Grounding alone is not enough for decisions that must be defensible. This article proposed a framework that keeps knowledge current through event-driven ingestion, organizes evidence with authority and entitlement metadata, retrieves and assembles it carefully, and generates citation-bound answers with reliable prompt patterns. The framework also verifies citations and figures, applies guardrails, and delivers answers as evidence packs within governed workflows, with continuous evaluation. The investment committee case showed the framework producing a fast, verified, and auditable decision. The central conclusion is that RAG supports evidence-based financial decisions when every claim is traceable to evidence that is current, authoritative, permitted, and verified.

References

1. Basireddy, S. R. (2022). Digitizing risk management through configurable enterprise workflow architecture. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 14(4), 231–239.
2. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., ... Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3697–3711).
3. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). *RAGAS: Automated evaluation of retrieval augmented generation* (arXiv:2309.15217). arXiv.
4. Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488).
5. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781).
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
7. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). *Lost in the middle: How language models use long contexts* (arXiv:2307.03172). arXiv.
8. Sannidhanam, A. H. (2021). Real-time claims processing using event-driven architectures. *International Journal of Technology, Management and Humanities*, 7(01), 36–50. <https://doi.org/10.21590/07.01.02>

9. Sannidhanam, A. H. (2024a). Prompt engineering patterns for reliable LLM outputs in production environments. *Journal of Multidisciplinary Knowledge*, 4(1), 29–39. <https://jmk.datatablets.com/index.php/j/article/view/168>
10. Sannidhanam, A. H. (2024b). Guardrails and safety mechanisms for LLM-powered enterprise applications. *International Journal of Engineering Science & Humanities*, 14(3), 273–285. <https://www.ijesh.com/j/article/view/933>