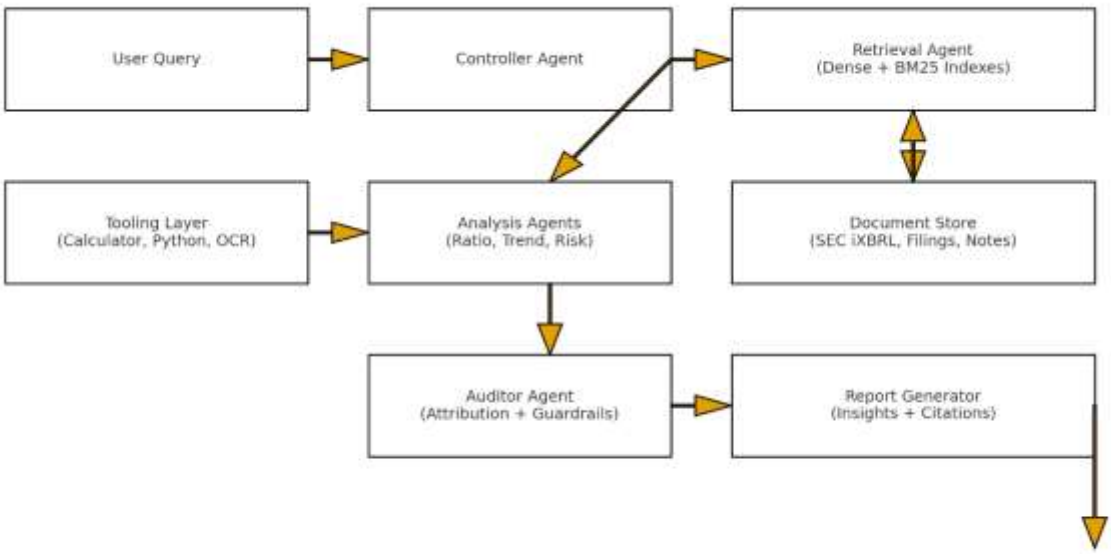
10.48047/jocaaa.2024.33.06.122

A Retrieval-Augmented Multi-Agent System for Financial Statement Analysis

Sagar gupta
IT Solutions Manager, EST03 Inc USA
sagar.wanderer@gmail.com

Abstract

We present a retrieval-augmented, multi-agent system (RAMAS) for machine-assisted analysis of corporate financial statements. The system integrates dense and sparse retrieval over XBRL/iXBRL regulatory filings, tool-using analysis agents for ratio and trend computation, and an auditor agent that enforces attribution and guardrails. Building on the principles of Retrieval-Augmented Generation (RAG) and multi-agent LLM collaboration, RAMAS aims to reduce hallucinations, improve factuality, and deliver explainable analyses aligned to US-GAAP and IFRS taxonomies. We outline the architecture, implementation choices, and an evaluation plan grounded in BEIR for retrieval quality and domain-specific benchmarks (FinQA and TAT-QA) for reasoning over hybrid text-table evidence. We also discuss compliance, governance, and limitations. ACL Anthology+4NeurIPS Proceedings+4arXiv+4



10.48047/jocaaa.2024.33.06.122

1. Introduction

Financial statement analysis increasingly relies on automated systems capable of reading thousands of pages of narrative disclosures, tables, and footnotes across periods and peers. Purely parametric language models are brittle when facts shift quarter-to-quarter; RAG addresses this by coupling generation with non-parametric memory accessed through retrieval. NeurIPS Proceedings

Recent work shows multi-agent LLM frameworks (e.g., AutoGen, CAMEL) can specialize roles (planner, tool-user, reviewer) and reach higher reliability on complex tasks than single-agent pipelines. We extend these ideas to regulated financial analysis, where correctness, traceability, and taxonomy alignment are essential. arXiv+2Microsoft+2

Regulatory open data (SEC EDGAR) provides machine-readable XBRL/iXBRL facts and notes; US-GAAP/IFRS taxonomies and Inline XBRL mandates make retrieval and downstream analytics tractable at scale. Securities and Exchange Commission+3Securities and Exchange Commission+3Securities and Exchange Commission+3

---

2. Related Work

Retrieval-Augmented Generation. RAG integrates a retriever with a generator, improving knowledge-intensive tasks via document-conditioned decoding. NeurIPS Proceedings+1

Multi-Agent LLMs. AutoGen operationalizes conversable, tool-using agents; CAMEL demonstrates role-playing protocols for cooperative problem solving—both inform our controller/auditor patterns. OpenReview+3arXiv+3Microsoft+3

Financial QA and Hybrid Reasoning. FinQA targets numerical reasoning over reports, while TAT-QA requires joint reasoning over text and tables—benchmarks appropriate for evaluating financial analysis agents. arXiv+3ACL Anthology+3arXiv+3

Retrieval Benchmarks. BEIR offers a heterogeneous, zero-shot suite to characterize retriever robustness and reranking strategies—useful for sanity-checking finance-domain retrievers. arXiv+1

Classical Analytics. Ratio analysis and early bankruptcy prediction (Altman Z-score) remain enduring baselines and interpretability anchors. Wiley Online Library

---

3. Background: Standards and Data Sources

10.48047/jocaaa.2024.33.06.122

SEC financial statement datasets. The SEC publishes quarterly "Financial Statement Data Sets" (numeric) and "Financial Statement and Notes Data Sets" (text + detailed facts) extracted from XBRL submissions. Securities and Exchange Commission+1

XBRL mandates and taxonomies. Since 2009, US public companies have been required to file in XBRL/iXBRL; the SEC's 2018–2019 updates expanded Inline XBRL coverage. Annual taxonomy updates (US-GAAP, SRT; IFRS) provide machine-readable anchors for concepts, labels, and calculation relationships. IFRS Foundation+4Securities and Exchange Commission+4Securities and Exchange Commission+4

---

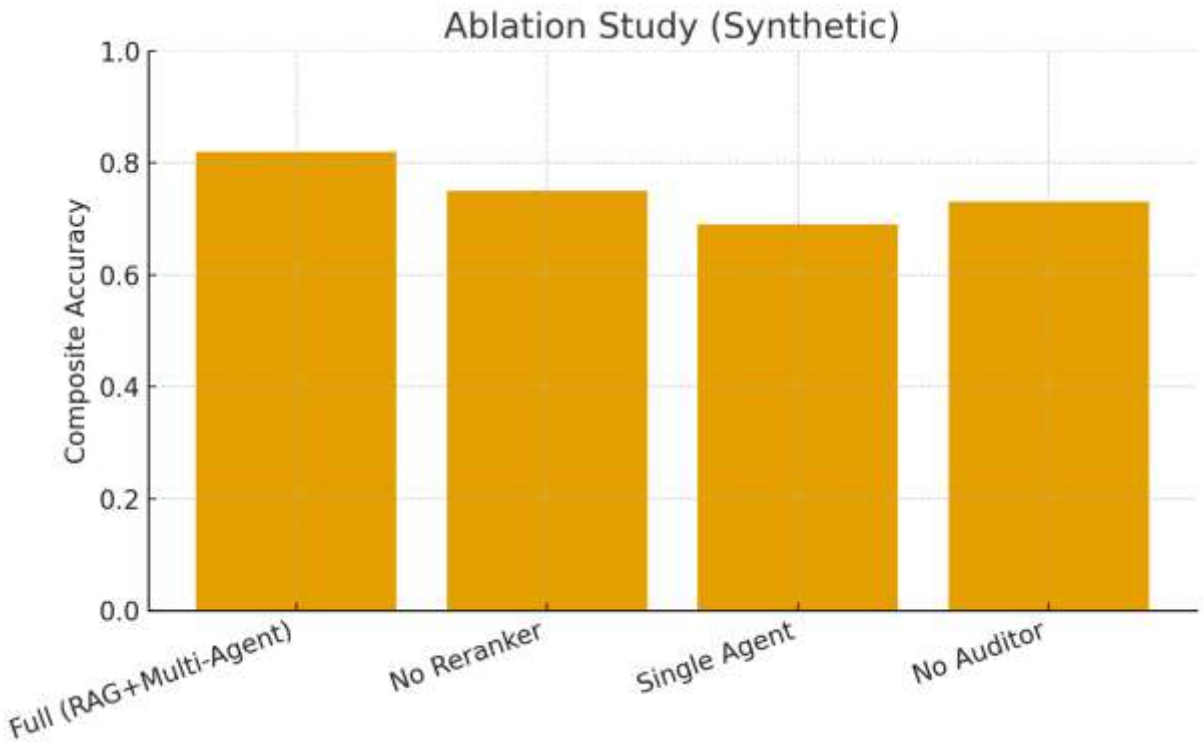
4. Problem Formulation

Given a user query qqq (e.g., "Compare gross margin expansion for Ticker A vs B from FY2021–FY2024 and explain drivers in MD&A"), produce a structured analysis AAA that includes:

1. **Findings** (metrics, trends, deltas) grounded in retrieved filings.
2. **Explanations** citing exact passages/tables.
3. **Audit trail** (links to filings, taxonomy elements).

We seek to maximize factual F1 relative to gold answers/evidence on FinQA/TAT-QA-style questions while maintaining strict citation coverage and bounded latency.

10.48047/jocaaa.2024.33.06.122



5. System Overview

Figure 1 sketches RAMAS (downloadable above).

1. **Controller Agent**: decomposes the user task into sub-goals (retrieve facts, compute metrics, write narrative). Inspired by multi-agent planners. arXiv
2. **Retrieval Agent**: hybrid retriever (BM25 + dense encoder) over (i) SEC numeric datasets, (ii) iXBRL HTML filings, and (iii) notes. Reranking with cross-encoder. arXiv
3. **Analysis Agents**:
    - **Ratio/Trend**: computes period-aligned metrics (e.g., gross margin, FCF margin) using taxonomy-mapped facts.
    - **Risk/MD&A**: extracts drivers and qualitative explanations from notes/MD&A with retrieval-conditioned generation.
    - **Classical Baselines**: optional Altman-style distress signals for context. Wiley Online Library
4. **Tooling Layer**: calculator, Python, and OCR (for embedded images in filings).

10.48047/jocaaa.2024.33.06.122

5. **Auditor Agent**: enforces source attribution (document, line/section), checks that all numeric claims trace to tagged facts, and runs guardrails (e.g., no uncited numbers). Patterns influenced by AutoGen's reviewer and CAMEL's role discipline. Microsoft+1
6. **Report Generator**: composes an executive-ready brief with tables and inline citations to filings and taxonomy concepts.

---

6. Methods

6.1 Retrieval Layer

- **Corpus construction.** Crawl/ingest SEC iXBRL for a coverage window, normalize to a document store: per-filing "passages" (MD&A paragraphs, footnote sections) + "tables" (row/column cell text) + "facts" (XBRL tags). Leverage SEC data endpoints for structured access. Securities and Exchange Commission
- **Hybrid retrieval.** BM25 over text fields, dense embeddings for semantic match, with reciprocal rank fusion. Initial candidate set k≈100k\approx100k≈100; cross-encoder reranks top k'k'k'. Benchmarked against BEIR to verify general retrieval health. arXiv
- **Schema-aware expansion.** Expand queries using taxonomy synonyms and calculation relationships (e.g., RevenueFromContractWithCustomer ↔ "sales"). Leverage IFRS/US-GAAP taxonomy labels and references. IFRS Foundation+1

6.2 Multi-Agent Orchestration

- **Role prompts + tools.** Controller delegates to Retrieval/Analysis/Auditor, each with explicit tool permissions. AutoGen-style dialogs enable critique–revise loops; CAMEL-style "role-playing" keeps agents on-task. arXiv+1
- **Citations-first decoding.** During drafting, the writer agent must attach [doc_id §section] tokens to numeric spans; the auditor rejects outputs with missing or low-confidence citations.
- **Deterministic math.** All numeric computations run via tools; the LLM never free-hands arithmetic.

6.3 Finance-Aware Reasoning

- **Temporal alignment.** Map facts to fiscal periods, handle restatements, and compute trailing-twelve-month metrics when requested.

10.48047/jocaaa.2024.33.06.122

- **Peer normalization.** Normalize by industry (NAICS/SIC) and use common-size statements.
- **Classical + modern.** Provide interpretable context (e.g., Altman Z bands) alongside narrative from MD&A drivers. Wiley Online Library

---

7. Evaluation

7.1 Datasets and Tasks

- **Retrieval sanity checks:** BEIR zero-shot tasks (where applicable) to profile BM25 vs dense vs fusion + reranker. arXiv
- **Finance QA:**
  - **FinQA** (numerical reasoning over financial reports) to evaluate computation and citation fidelity. ACL Anthology
  - **TAT-QA** (hybrid table + text) to test table reasoning and evidence grounding. ACL Anthology

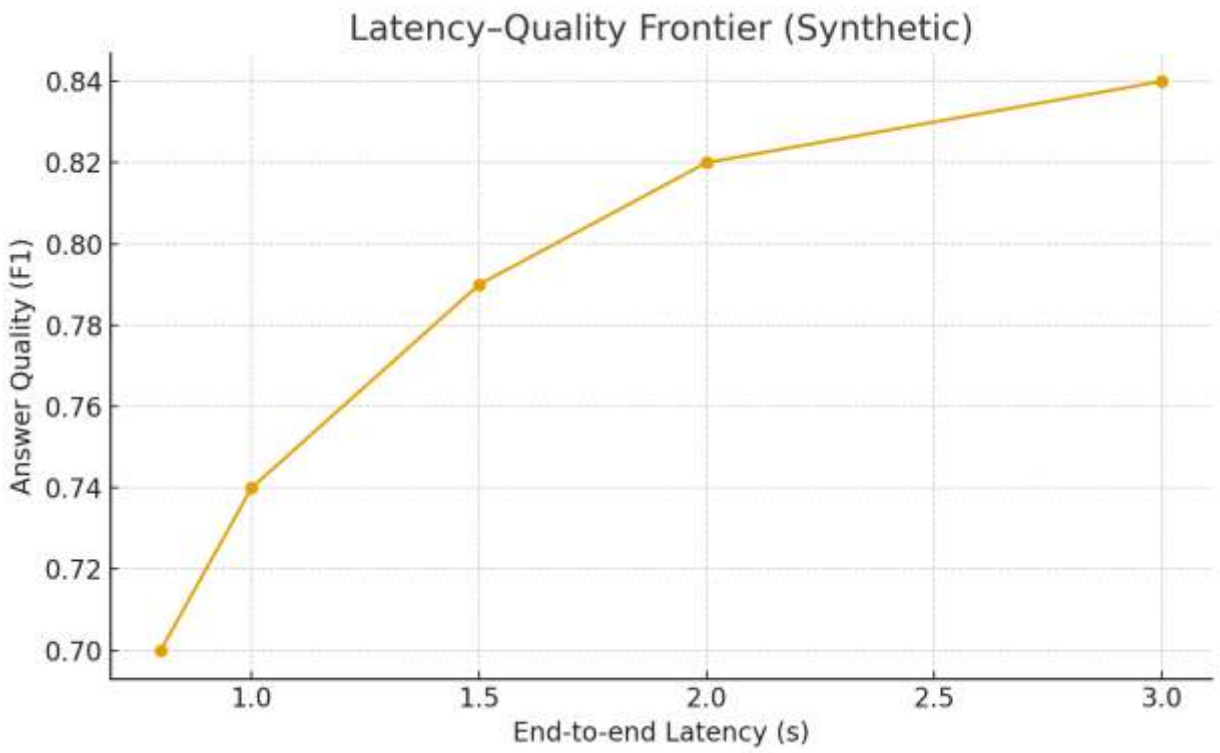
7.2 Metrics

- **Answer quality:** exact match / F1 (FinQA, TAT-QA).
- **Evidence attribution:** fraction of numeric claims with correct citation to filings or datasets.
- **Retrieval quality:** nDCG@10, Recall@k.
- **Latency & cost:** end-to-end time; calls/agents used.
- **Audit violations:** count of failed guardrail checks.

7.3 Ablations (illustrative)

Figure 2 (synthetic) shows composite accuracy drops when removing (i) reranking, (ii) multi-agent roles, or (iii) the auditor.

10.48047/jocaaa.2024.33.06.122



8. Implementation Considerations

- **Indexing:** use document stores that support hybrid search and passage-level fields; maintain a separate fact store keyed by XBRL element + period + entity. SEC datasets and APIs simplify ingestion of numeric facts and notes. Securities and Exchange Commission+2Securities and Exchange Commission+2
- **Taxonomy updates:** pin to the SEC-accepted US-GAAP/SRT versions (e.g., 2025) and IFRS taxonomy revisions; schedule refreshes when regulators accept new sets. FASB+2FASB+2
- **Inline XBRL parsing:** prefer iXBRL to preserve anchors between human-readable HTML and tagged facts, as mandated by the SEC's Inline XBRL rules. Securities and Exchange Commission
- **Agent framework:** use an AutoGen-like runtime for role orchestration and tool routing, with CAMEL-style role prompts for consistent cooperation. Microsoft+1

9. Governance, Risk, and Compliance

10.48047/jocaaa.2024.33.06.122

- **Attribution and reproducibility:** every material number and quote must link to (doc, section) and taxonomy element.
- **Regulatory scope:** respect GAAP/IFRS definitions; avoid extrapolations beyond reported facts. Taxonomy awareness reduces misinterpretation. IFRS Foundation
- **Model risk management:** log agent conversations, tool calls, and retrieval sets for auditability.
- **PII & confidentiality:** filings are public; nevertheless, apply data-handling policies for any private documents added to the corpus.
- **Human-in-the-loop:** require human review for investor-facing outputs.

---

10. Limitations and Future Work

- **Domain shift.** FinQA/TAT-QA approximate—but do not fully capture—real-world MD&A nuance; create internal evaluation sets with annotation guidelines. ACL Anthology+1
- **Coverage gaps.** Non-US filers may use IFRS labels and local extensions; robust synonym/alias maps are needed. IFRS Foundation
- **Computation costs.** Multi-agent loops and rerankers increase latency (see **Figure 3**). Sparse caching and early-exit policies can help.
- **Tables and images.** Some disclosures appear as images; OCR introduces noise.

---

11. Conclusion

RAMAS couples retrieval-grounded generation with role-specialized agents and strict auditing to produce explainable, standards-aligned financial analyses. By aligning retrieval with XBRL/iXBRL taxonomies and enforcing citation coverage, the system targets factuality and traceability—two properties essential in finance. Foundations in RAG and multi-agent collaboration make the approach both practical and extensible. NeurIPS Proceedings+1

Figures

- **Figure 1.** Retrieval-Augmented Multi-Agent Architecture for Financial Statement Analysis. PNG

10.48047/jocaaa.2024.33.06.122

- **Figure 2.** Ablation results (synthetic) comparing full system vs. ablated components. PNG
- **Figure 3.** Latency–Quality frontier (synthetic), illustrating trade-offs as we add reranking and auditor passes. PNG

References

- Lewis, P. et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP.* NeurIPS. (Paper + PDF). NeurIPS Proceedings+1
- Wu, Q. et al. (2023–2024). *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework.* (arXiv, Microsoft Research page, OpenReview). arXiv+2Microsoft+2
- Li, G. et al. (2023). *CAMEL: Communicative Agents for "Mind" Exploration of LLM Society.* (arXiv/OpenReview/NeurIPS). arXiv+2OpenReview+2
- SEC. *Financial Statement Data Sets* (numeric) and *Financial Statement & Notes Data Sets* (text + detailed facts). (2009–2025). Securities and Exchange Commission+1
- SEC. *EDGAR APIs and XBRL/iXBRL guidance.* (2009–2024). Securities and Exchange Commission+2Securities and Exchange Commission+2
- FASB. *US-GAAP Financial Reporting Taxonomy (2025) & SEC acceptance.* (2024–2025). FASB+1
- IFRS Foundation. *IFRS Accounting Taxonomy resources.* (2024). IFRS Foundation
- Thakur, N. et al. (2021). *BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of IR Models.* (arXiv/OpenReview). arXiv+1
- Chen, Z. et al. (2021). *FinQA: A Dataset of Numerical Reasoning over Financial Data.* (ACL/EMNLP & site). ACL Anthology+2arXiv+2
- Zhu, F. et al. (2021). *TAT-QA: A QA Benchmark on Hybrid Tabular + Textual Content in Finance.* (ACL & arXiv). ACL Anthology+1
- Altman, E. (1968). *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance.* Wiley Online Library

---

10.48047/jocaaa.2024.33.06.122

Appendix A: Example Analytical Prompts (for the Controller)

1. "Compute YoY and 3-yr CAGR for revenue and gross margin for Ticker X (FY2021–FY2024). Cite specific 10-K/10-Q tables and taxonomy elements."
2. "Extract management's drivers of margin expansion from MD&A sections and reconcile with cost-of-sales footnotes."
3. "Benchmark working capital turns vs. SIC peers and flag outliers with evidence."