FinSAgent: Corpus-Aligned Multi-Agent RAG Framework for Evidence-Grounded SEC Filing Question Answering

Jijun Chi[3], Zhenghan Tai[1,3], Hanwei Wu[1,12], Tung Sum Thomas Kwok[1,4], Hailin He[1], Zixing Liao[1], Bohuai Xiao[1], Chaolong Jiang[1], Jianliang Lei[1], Jerry Huang[7,9], Peng Lu[7], Muzhi Li[5], Liheng Ma[1,2,9], Yihong Wu[7], Sicheng Lyu[1,2,9], Jingrui Tian[2], Yihan Li[8], Yanzhang Ma[1,11], Dingtao Hu[2], Yufei Cui[2], Ling Zhou[10], Lei Ding[1,6], Xinyu Wang[1,2]

[1]SimpleWay.AI [2]McGill University [3]University of Toronto [4]University of California, Los Angeles [5]The Chinese University of Hong Kong [6]University of Manitoba [7]Université de Montréal [8]Boston University [9]Mila - Quebec AI Institute [10]CG Matrix Technology Limited [11]Lakehead University [12]McMaster University

ferris.chi@alumni.utoronto.ca, xinyu.wang5@mail.mcgill.ca

Abstract

Financial question answering over U.S. Securities and Exchange Commission (SEC) filings requires retrieving and synthesizing heterogeneous evidence dispersed across long, standardized, and highly redundant disclosures. Existing retrieval-augmented and multi-agent systems typically derive retrieval queries directly from the user's question and rank candidates by semantic similarity. Together, these choices create *prior–corpus misalignment*: a mismatch between model priors and the target filings' structure, terminology, and evidence standards. As a result, query generation misses corpus-specific evidence, while semantic reranking favors topically similar but evidentially invalid "false-positive" chunks. We propose FinSAgent, an evidence-grounded multi-agent framework that reframes SEC filing QA as corpus-aligned retrieval planning and corrects both ends with a single principle: inject corpus-side conditioning wherever model priors would otherwise dominate. FinSAgent combines (1) role-specialized agents anchored to the mandated 10-K item structure, (2) database-aware query decomposition that conditions each agent's sub-queries on a lightweight, summary-level view of the local corpus, and (3) multi-path retrieval with a learned feature-gated reranker that separates evidential validity from semantic similarity. Across five offline financial QA benchmarks, FinSAgent improves retrieval coverage and answer correctness over strong single-agent and multi-agent baselines; in a three-arm randomized online experiment with 1,000 anonymous user ratings, it also receives higher scores than baselines.

Keywords

Financial NLP, Retrieval-Augmented Generation, Multi-Agent Systems, SEC Filings, Question Answering, Evidence Grounding

1 Introduction

Financial question answering (QA) over U.S. Securities and Exchange Commission (SEC) filings asks a system to answer analyst-style questions from annual and quarterly corporate disclosures such as 10-Ks and 10-Qs. The task matters because filings are among the primary public sources for understanding a firm's operations, risks, strategy, and financial condition. In practice, users rarely consult filings to recover an isolated fact. They ask whether a company's expansion appears sustainable, which risks are most material to a strategic shift, or how management's narrative aligns with reported numbers. Answering such questions requires more than fluent generation: it requires locating the right evidence in long, structured documents and synthesizing responses that remain grounded in disclosed sources [4, 5, 9, 13, 14, 31, 39, 46, 47, 52].

We focus on *evidence-grounded* financial QA over a filing database, where answering a question often requires combining numerical and narrative evidence scattered across business descriptions, management discussion, risk factors, legal proceedings, footnotes, and financial tables, while near-duplicate boilerplate recurs across issuers and years. The task is therefore not merely long-context prompting, but requires both corpus-aligned retrieval planning and validity-aware evidence selection over a large, structured, and highly redundant corpus. Recent work has strengthened individual components through domain-adapted language models [23, 41, 49], retrieval-centric preprocessing, indexing, and reranking [32, 37], and role-specialized multi-agent reasoning [2, 28, 43, 51]. Yet it leaves a central question unresolved: *how should a system discover and select valid evidence for complex filing questions?*

The core problem. We use *prior–corpus misalignment* to describe the mismatch between model priors and the target filing corpus's structure, terminology, and criteria for valid evidence. This mismatch affects both query formation and evidence selection (Figure 2). At the **front end**, prior-driven query decomposition produces sub-queries that are semantically plausible yet misaligned with how evidence is actually written and organized in the filings, so genuinely disclosed material is never retrieved. At the **back end**, a semantic reranker rewards chunks that match the same priors, so semantic relevance separates from evidential validity: near-verbatim boilerplate (e.g., generic risk-factor or compliance language that recurs across issuers) scores highly yet is invalid for the specific company, crowding out the true disclosure. Effective retrieval therefore requires corpus alignment at both stages: retrieval planning determines whether relevant evidence enters the candidate set, while evidence selection determines whether it survives the final evidence budget.

FinSAgent. To address both stages, we propose FinSAgent (Figure 2), an evidence-grounded multi-agent framework built around this problem. The central design principle is to *inject corpus-side conditioning wherever model priors would otherwise dominate*, so that evidence exploration is broadened while staying aligned with the local corpus and conservative about evidence quality. FinSAgent instantiates three coupled mechanisms: (1) **role-specialized parallel**

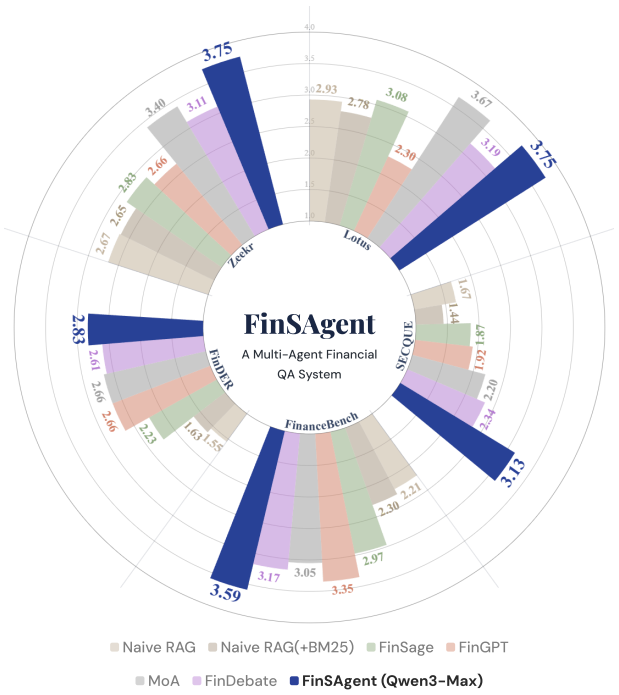

Figure 1: Answer correctness across five financial QA benchmarks. FinSAgent attains the highest correctness on every benchmark, outperforming single-agent RAG pipelines and multi-agent baselines under a matched evidence budget.

agents, anchored to the mandated 10-K item structure, that separate heterogeneous evidence needs (general, quantitative, market, legal, and company) before retrieval to broaden coverage; (2) **database-aware query decomposition**, which conditions each agent's sub-query generation on a lightweight, summary-level view of the local filing database rather than the user question alone; and (3) **multi-path retrieval with feature-gated reranking**, in which a learned gate over non-semantic features down-weights confident false positives that a semantic reranker cannot separate. The first two components support corpus-aligned retrieval planning, while the third performs validity-aware evidence selection.

Our contributions are:

- We frame SEC filing QA as requiring alignment at both retrieval planning stage and evidence selection stage, and introduce *prior-corpus misalignment* as a unifying lens for failures at both stages.
- We introduce FinSAgent, which combines role-specialized, database-aware retrieval planning with multi-path retrieval and feature-gated evidence selection for candidate selection (§3.3, §3.4).
- Across **five benchmarks**, FinSAgent improves retrieval coverage and answer correctness. Tests under production scenario further supports these results: across ~1,400 anonymous user ratings, FinSAgent receives higher scores than baselines (§4).

2 Related Work

Financial foundation models. The linguistic and numerical complexity of financial corpora has driven a line of domain-specialized language models [20]. Early efforts such as BloombergGPT [41] and open-source frameworks like FinGPT [23] and XuanYuan [49] demonstrate strong performance on domain tasks such as sentiment analysis and entity extraction. Recent work explores multimodality by post-training on textual and visual financial data [1, 38]: TAT-LLM [53] targets discrete reasoning over hybrid tabular–textual data, and Open-FinLLMs [11] extends coverage across text, tables [18, 44], time series [22], and charts [12]. These models encode deep domain knowledge but operate as standalone generators without cross-sourced reasoning [17] across heterogeneous evidence, and thus struggle with the multi-step, longitudinal reasoning that long-form SEC filings demand.

Agentic and retrieval-augmented systems. Retrieval-augmented generation (RAG) grounds model outputs in verifiable evidence and has become the prevalent approach to financial QA [7, 50]. FinSage [37] adopts a deterministic "extract-and-summarize" workflow built on multi-path retrieval and a domain-specialized reranker, but its static pipeline cannot dynamically align queries with heterogeneous filing sections [35]. A parallel line pivots to multi-agent frameworks: FinDebate [2] combines role-specialized agents with a safety-constrained debate protocol to curb model overconfidence [45], FinSight [15] coordinates specialists for data collection, context analysis, and report generation through a shared programmable variable space, and mixture-of-agents designs aggregate independent specialist drafts [10, 25, 36, 40]. On the retrieval-planning side, query decomposition methods [24, 30, 42, 48] improve coverage by splitting complex questions into sub-queries, but remain primarily *question-driven*: they refine or expand the original query without conditioning on the local structure and terminology of the target corpus, a limitation shared with classical pseudo-relevance feedback [3, 29]. Uncertainty-aware retrieval [26, 27] highlights the importance of evidence reliability but typically relies on expensive stochastic inference and is under-explored for structured filing QA.

FinSAgent differs from these systems along the axis that matters for filing QA: it treats retrieval as a corpus-aligned planning problem rather than a question-conditioned function. Concretely, it introduces database-aware decomposition to condition sub-queries on a lightweight corpus view, and augments multi-path retrieval with a learned feature gate that models evidential validity as separable from semantic similarity.

3 The FinSAgent Framework

3.1 Problem Formulation and Overview

We formulate SEC filing QA as evidence-grounded generation over a filing database $\mathcal{D}$. Given a user question q, the goal is to retrieve supporting evidence $E \subset \mathcal{D}$ and generate an answer y grounded in E. Unlike general multi-hop QA, which typically requires a sequential reasoning chain across entities, SEC filing questions demand *evidence aggregation across heterogeneous perspectives* that are scattered over fixed sections and mixed table–prose formats.

Following the two-ended view of prior–corpus misalignment introduced in §1, FinSAgent is a three-stage pipeline (Figure 2):

Table 1: Agent roles anchored to the mandated 10-K item structure.

Agent	Focus	10-K item(s)
General	Cross-item / definitional	(always on)
Quantitative	Financial metrics, numerics	Item 8, 7A
Market	Industry context, positioning	Item 7 (MD&A)
Legal	Legal proceedings, compliance	Item 1A, 3
Company	Business, operations, segments	Item 1, 1B–2

(1) a **role-specialized multi-agent module** constructs complementary general, quantitative, market, legal, and company views of the question and broadens coverage (§3.2); (2) a **database-aware decomposition module** generates agent-specific sub-queries conditioned on a lightweight, summary-level view of the local corpus, correcting the front-end gap (§3.3); and (3) a **multi-path retrieval module with feature-gated reranking** selects reliable evidence, correcting the back-end gap (§3.4). An orchestrator routes q to the relevant agents and aggregates their outputs into the final grounded answer. An optional conversational-memory extension for multi-turn analysis is described in Appendix E.

3.2 Role-Specialized Parallel Multi-Agent System

SEC filing QA requires evidence distributed across multiple parts of a filing, which a single search trajectory captures poorly. To improve coverage, FinSAgent first decomposes the question into complementary evidence-seeking views using a role-specialized parallel multi-agent module.

Let $\mathcal{A} = \{a^g, a^q, a^m, a^l, a^c\}$ denote agents for general, quantitative, market, legal, and company evidence. Each agent $a \in \mathcal{A}$ carries a role policy $\pi_a = (\rho_a, \Omega_a, \phi_a)$, where ρ_a specifies its responsibility, Ω_a its retrieval scope, and ϕ_a its analytical focus. Given q, each agent produces a role-conditioned analytical view $z_a = g_a(q; \pi_a)$, where $g_a(\cdot)$ is the prompting function induced by π_a and z_a is a short natural-language interpretation of q that guides downstream decomposition and retrieval. All agents operate in *parallel* on the same question, disentangling heterogeneous evidence needs before retrieval and reducing the risk that one search path overlooks relevant support. Role prompts are given in Appendix F.2.

Why these roles. The agent roles are derived from recurring evidence categories reflected in the standardized 10-K item structure[1] and expert-validated (Table 1). As shown in Table 1, the general, quantitative, market, legal, and company agents cover complementary retrieval scopes across business disclosures, management discussion, risk and legal sections, and financial statements. This mapping is deliberately created to tie agents' queries to specific filing sections and distinct retrieval scopes. The general agent remains active for cross-item and definitional questions, while the specialized agents target their corresponding evidence categories. Our retrieval analysis indicates that role specialization provides complementary coverage(§4.3). We therefore adopt these five roles as a practical trade-off between evidence coverage and inference cost.

[1] https://www.sec.gov/files/form10-k.pdf

3.3 Database-Aware Query Decomposition

After obtaining the role-conditioned views $\{z_a\}_{a\in\mathcal{A}}$, each agent decomposes the question into retrieval-ready sub-queries. Role specialization alone does not make this reliable: an LLM tends to generate sub-queries from its internalized priors, which differ substantially from the target corpus and lead to weak or missed retrievals.

We therefore ground decomposition by exposing each agent to a coarse-grained local view of the filing database. During offline preprocessing we generate compact summaries for contiguous spans of filing chunks and index them. At query time, a lightweight FAISS-based retriever [8] maps q to the top-k_0 most relevant *section-level* summaries, forming the local database view $R^{\mathrm{sv}}_{k_0}(q)$. This step is intentionally lightweight: it does not answer the question, but exposes which filing regions are likely task-relevant and surfaces corpus-specific terminology. Conditioned on this view, each agent performs role-specific decomposition:

$$Q_a = d_a\Big(q,\ z_a,\ R^{\mathrm{sv}}_{k_0}(q)\Big) = \{q_{a,1}, \ldots, q_{a,n_a}\},$$

where d_a is a prompted LLM call that receives q, the analytical view z_a, and the retrieved summaries as structured input and emits a list of role-specific sub-queries.

Dynamic corpus conditioning. The agent prompt templates are fixed, but the corpus view $R^{\mathrm{sv}}_{k_0}(q)$ is retrieved anew for each question. Unlike a static terminology prompt or a hand-written overview, this query-dependent view exposes the filing regions and corpus-specific terms relevant to the current question. The resulting sub-queries can therefore reflect the filing's section headings, segment names, and disclosure terms through a lightweight summary lookup, without encoding the entire database in the prompt.

3.4 Multi-Path Retrieval and Feature-Gated Reranking

Given the sub-query sets $\{Q_a\}_{a\in\mathcal{A}}$, FinSAgent retrieves candidates within each agent's scope and filters them through a feature-gated reranker: multi-path retrieval for broad coverage, then reranking for evidence reliability.

3.4.1 Multi-Path Evidence Retrieval. A single retrieval signal is often insufficient, since relevant evidence may appear under exact lexical matches, semantic paraphrases, or coarse section-level summaries. For each agent a and sub-query $q_{a,j} \in Q_a$, we construct a candidate set $C_{a,j}$ as the union of the top-k results from three complementary paths restricted to the agent's scope Ω_a:

$$C_{a,j} = R^{\mathrm{sp}}_k(q_{a,j}; \Omega_a) \ \cup\ R^{\mathrm{de}}_k(q_{a,j}; \Omega_a) \ \cup\ R^{\mathrm{sm}}_k(q_{a,j}; \Omega_a),$$

where R^{sp} is sparse retrieval via BM25, R^{de} is dense retrieval via cosine similarity over chunk embeddings, and R^{sm} is summary-path retrieval over precomputed section-level summary embeddings. A dedicated table branch (§A) supplies structured tabular evidence. Formal definitions are given in Appendix F.1.

3.4.2 Feature-Gated Evidence Reranking. Because the candidate set $C_{a,j}$ is constructed for high recall, it can include topically relevant but evidentially invalid chunks that a cross-encoder reranker cannot

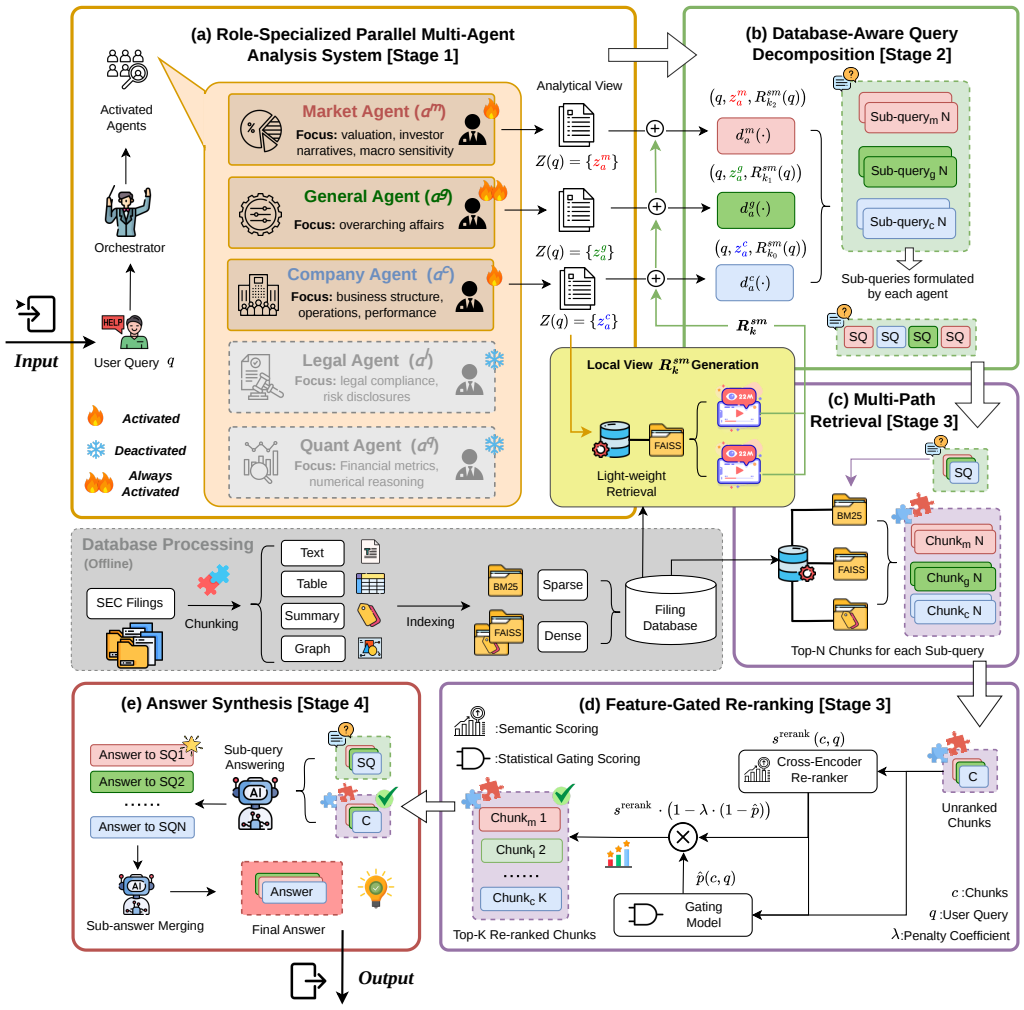

Figure 2: Overall framework of FinSAgent. Role-specialized agents first construct complementary analytical views. Database-aware decomposition then conditions each agent's sub-queries on a lightweight summary-level view of the local filing database (correcting the front-end query gap), and multi-path retrieval with a feature-gated reranker selects reliable evidence (correcting the back-end validity gap), before evidence aggregation and final answer synthesis.

reliably distinguish. FinSAgent therefore augments its semantic score with a learned feature gate.

For each candidate $c \in C_{a,j}$ we compute a base reranker score $s^{\text{rr}}(c, q_{a,j})$ and a lightweight feature vector

$$\mathbf{x}(c, q_{a,j}) = \left[\mathbf{s},\ \mathbf{r},\ \mathbf{f}_{\text{query}},\ \mathbf{f}_{\text{chunk}},\ \mathbf{f}_{\text{overlap}},\ s^{\text{rr}}\right],$$

where $\mathbf{s}$ collects the three path scores, $\mathbf{r}$ records path-specific rank and provenance (e.g., how many paths surfaced the chunk), $\mathbf{f}_{\text{query}}$ captures query-level statistics, $\mathbf{f}_{\text{chunk}}$ chunk-level metadata, and $\mathbf{f}_{\text{overlap}}$ lexical-overlap features. Critically, these are *non-semantic* signals the cross-encoder does not model. An offline-trained LightGBM model [16] $\mathcal{M}$ maps them to a validity estimate $\hat{p}(c, q_{a,j}) = \mathcal{M}(\mathbf{x}) \in [0, 1]$; we choose a gradient-boosted tree for its low inference overhead and interpretability. The estimate combines with the reranker score through a multiplicative gate:

$$\tilde{s}(c, q_{a,j}) = s^{\text{rr}}(c, q_{a,j}) \left(1 - \lambda_a \big(1 - \hat{p}(c, q_{a,j})\big)\right),$$

where $\lambda_a \in [0, 1]$ is the per-agent gating strength tuned on a validation split; performance is stable across a broad range (§C.3). The multiplicative form scales the penalty with the reranker's own confidence, so a high-scoring chunk judged risky is demoted more than a low-scoring one—confident false positives are the more damaging case. The final evidence set is $E_{a,j} = \text{TopK}^{k_e}_{c \in C_{a,j}} \tilde{s}(c, q_{a,j})$. Each agent then drafts a grounded response over $E_{a,j}$, and the orchestrator concatenates the retrieved evidence and drafts to synthesize the final answer to q.

Gating model training. We construct the training set exclusively from an in-distribution training split to prevent leakage. Using targeted hard-negative mining, an LLM generates questions from random chunks; these are run through the full retrieval and reranking pipeline; within the top-k results, chunks matching the chosen ground truth are labeled positive and the rest serve as competitive

hard negatives, downsampled to an approximate 20% positive rate. Full details, the 31-feature specification, and a SHAP analysis are in Appendix C.

4 Experiments and Results

We evaluate whether FinSAgent improves evidence-grounded filing QA and, crucially, whether its gains come from *corpus alignment* rather than from added retrieval budget or extra deliberation. We report end-to-end quality (§4.2), retrieval coverage and component ablations (§4.3), a comparison to full-document long-context prompting (§4.4), evaluation-validity evidence including a blind human study and RAGAS noise sensitivity (§4.5), system overhead and per-phase latency (§4.6), a case study (§4.7), and an error analysis (§4.8). Appendices B and C report a matched-budget fairness study and a gating-strength ablation with SHAP attribution.

4.1 Experimental Setup

Datasets. We evaluate on five financial QA datasets spanning public benchmarks and proprietary in-house collections. The public benchmarks are the open-source version of `FinanceBench` [13], the `FinDER` dataset [6], and a 100-question stratified subset of `SECQUE` [47] that preserves SECQUE's native question-type distribution (Appendix A). The two proprietary datasets, `Lotus` and `Zeekr`, are drawn from real-world business data for two automotive brands and contain multi-entity QA pairs with annotated relevance labels. Corpus construction (MinerU parsing, modality-tagged chunking, table/figure branches, near-duplicate removal, coreference resolution, section-level metadata, and dense BGE-M3 plus sparse BM25 indices) is detailed in Appendix A.

Baselines and backbones. We compare against traditional and agentic systems: Naive RAG (dense-only and +BM25 variants), FinSage [37] (single agent + multi-path retrieval), FinGPT [23] (single-agent tool-augmented reasoner), and the multi-agent systems MoA [36] and FinDebate [2]. To avoid confounding gains with backbone choice, all baselines use the same generator backbone as FinSAgent within each comparison (DeepSeek-v3.1-Terminus), and we additionally report FinSAgent under Kimi-K2.5 and Qwen3-Max to show that the gains are robust to backbone choice. The base reranker is bge-reranker-v2-gemma, and the learned gate is a LightGBM model over the 31 features in Appendix C. Baseline implementation details are in Appendix A.3.

Metrics and judging protocol. We report answer **Correctness** together with five Likert-scale (1–5) [21] dimensions adopted from Fin-RATE [14]: **Information Coverage**, **Reasoning Chain**, **Factual Consistency**, **Clarity of Expression**, and **Analytical Depth**. For retrieval we report Macro-Recall against annotated relevance labels. This suite subsumes the metrics commonly used in this space: our Correctness is a finer-grained (1–5) version of SECQUE-Judge's 0/1/2 correctness and of RAGAS answer accuracy; Factual Consistency corresponds to RAGAS faithfulness; and Macro-Recall corresponds to RAGAS context recall. We add RAGAS noise sensitivity in §4.5.

Judging is **blind and order-randomized**: each candidate answer is stripped of any system identifier and presented in randomized order, and the judge receives only the question, the reference answer, and the candidate answer. Because all systems in a comparison share the same backbone that also serves as the judge, any self-preference of the judge applies equally to every system and cannot favor FinSAgent; §4.5 provides an independent human check.

4.2 End-to-End QA Performance

Table 2 reports end-to-end quality. Two patterns stand out. First, the multi-agent paradigm broadly outperforms single-agent baselines, confirming that collaborative reasoning helps on complex filing questions. Second, and more importantly, FinSAgent attains the best Correctness among all multi-agent frameworks on every dataset, and does so *regardless of backbone*: under Qwen3-Max it is best or second-best on nearly all dimensions and datasets, and its advantage is largest on the evidence-critical Factual Consistency dimension. The gains persist across DeepSeek-v3.1-Terminus, Kimi-K2.5, and Qwen3-Max, indicating they are attributable to the retrieval-planning design rather than a particular generator. The error analysis of §4.8 further shows FinSAgent suppresses the severe failure modes that dominate baseline errors. Moreover, we run a matched-budget fairness study (Appendix B) in which MoA and FinDebate receive FinSAgent's own per-role retrieval budget with our two mechanisms disabled. FinSAgent still wins Factual Consistency on all five benchmarks and Correctness on four of five, so the gains cannot be attributed to a larger number of retrieval perspectives.

4.3 Retrieval Coverage and Component Ablation

We isolate retrieval quality by holding the final context to ∼35 chunks across all configurations, so improvements reflect *smarter routing*, not more evidence. Table 3 shows the components are highly complementary: multi-role retrieval drives large gains on most corpora, feature-gated reranking filters unreliable candidates (notably lifting the hard `Lotus` set), and database-aware decomposition resolves compositional queries (peaking `FinanceBench` at 80). The full system achieves the best overall coverage. A per-agent gating-strength ablation and a global SHAP attribution (Appendix C) further show that the gate's gains are robust to the penalty strength λ_a and are driven by non-semantic features that the semantic reranker does not model.

The agent-activation analysis (Figure 3) validates every role: all agents contribute exclusive ground-truth chunks to the final pool. Even the lowest-activation *Legal* agent (8% activation) consistently retrieves exclusive ground-truth chunks when triggered, and the co-retrieval heatmap confirms that neighboring roles cannot absorb them—so dropping any role lowers recall.

On the FinanceBench full-system dip. The full system dips on `FinanceBench` (80 to 76). Under our fixed ∼35-chunk budget, components *reallocate* rather than enlarge the budget, so where added breadth helps little it can displace already-sufficient chunks. This is exactly the case for `FinanceBench`: the open-source version is corpus-shallow per entity (≈41 companies, ≈8.5 filings each, ≈73% 10-Ks), so evidence is concentrated in a single 10-K, leaving little complementary material for multi-perspective exploration to recover while the added breadth competes for the same chunks.

Table 2: End-to-end financial QA performance. Corr.=Correctness, Info.=Information Coverage, Reas.=Reasoning Chain, Fact.=Factual Consistency, Clar.=Clarity, Depth=Analytical Depth (all higher is better). Best in bold, second-best underlined, per dataset. *On FinanceBench we additionally include a full-document long-context baseline (LC) with frontier model, which reads the entire 10-K with no retrieval (§4.4).

Dataset	Type	Method	Corr.	Info.	Reas.	Fact.	Clar.	Depth
Lotus	Single	Naive RAG	2.93	2.98	3.45	3.68	4.50	3.05
	Single	Naive RAG (+BM25)	2.78	2.83	3.37	3.50	4.39	2.94
	Single	FinSage	3.08	3.26	3.75	3.99	4.53	3.48
	LC*	Long-context (full 10-K)	0	0	0	0	0	0
	Web	FinGPT	2.30	2.82	3.13	3.49	4.43	3.07
	Multi	MoA	3.67	4.03	4.14	4.11	4.62	4.21
	Multi	FinDebate	3.19	3.58	3.90	3.89	**4.69**	3.96
	Multi	FinSAgent (DS-v3.1)	3.58	3.79	3.91	**4.16**	4.58	3.83
	Multi	FinSAgent (Kimi-K2.5)	3.57	3.98	3.94	3.87	4.42	4.04
	Multi	FinSAgent (Qwen3-Max)	**3.75**	**4.20**	**4.28**	4.12	4.63	**4.32**
SECQUE	Single	Naive RAG	1.67	1.68	1.90	2.20	4.24	1.70
	Single	Naive RAG (+BM25)	1.44	1.63	1.81	1.91	4.28	1.69
	Single	FinSage	1.87	2.25	2.27	2.51	4.08	2.13
	LC*	Long-context (full 10-K)	0	0	0	0	0	0
	Web	FinGPT	1.92	2.49	2.81	3.02	4.06	2.70
	Multi	MoA	2.20	2.81	3.07	2.66	4.28	3.05
	Multi	FinDebate	2.34	2.88	3.26	2.73	4.36	3.17
	Multi	FinSAgent (DS-v3.1)	2.96	3.38	3.39	3.35	4.44	3.23
	Multi	FinSAgent (Kimi-K2.5)	2.64	3.17	3.31	2.90	4.19	3.19
	Multi	FinSAgent (Qwen3-Max)	**3.13**	**3.81**	**3.87**	**3.56**	**4.60**	**3.82**
FinanceBench	Single	Naive RAG	2.21	2.36	2.52	2.76	4.54	2.40
	Single	Naive RAG (+BM25)	2.30	2.28	2.47	2.96	4.54	2.34
	Single	FinSage	2.97	3.27	3.34	3.26	4.60	3.23
	LC*	Long-context (full 10-K)	2.98	3.29	3.52	3.37	4.60	3.42
	Web	FinGPT	3.35	**4.10**	**4.35**	3.59	**4.90**	**4.21**
	Multi	MoA	3.05	3.43	3.79	3.36	4.55	3.64
	Multi	FinDebate	3.17	3.57	4.01	3.56	4.68	3.92
	Multi	FinSAgent (DS-v3.1)	3.32	3.70	3.77	3.52	4.61	3.71
	Multi	FinSAgent (Kimi-K2.5)	3.29	3.59	3.83	3.50	4.56	3.74
	Multi	FinSAgent (Qwen3-Max)	**3.59**	3.93	4.12	**3.68**	4.74	4.07
FinDER	Single	Naive RAG	1.55	1.52	1.65	2.35	4.14	1.48
	Single	Naive RAG (+BM25)	1.63	1.56	1.69	2.46	3.97	1.51
	Single	FinSage	2.23	1.93	2.13	3.08	3.99	1.89
	LC*	Long-context (full 10-K)	0	0	0	0	0	0
	Web	FinGPT	2.66	2.56	3.04	3.32	4.35	2.96
	Multi	MoA	2.66	2.70	3.24	**3.58**	4.25	3.10
	Multi	FinDebate	2.61	2.68	3.32	3.25	**4.44**	3.24
	Multi	FinSAgent (DS-v3.1)	2.68	2.83	3.15	3.56	4.25	3.01
	Multi	FinSAgent (Kimi-K2.5)	2.38	2.82	3.04	2.93	3.90	3.00
	Multi	FinSAgent (Qwen3-Max)	**2.83**	**3.30**	**3.48**	3.41	4.32	**3.52**
Zeekr	Single	Naive RAG	2.67	2.47	2.96	3.56	4.37	2.50
	Single	Naive RAG (+BM25)	2.65	2.45	3.02	3.58	4.41	2.56
	Single	FinSage	2.83	2.76	3.32	3.96	4.47	2.84
	LC*	Long-context (full 10-K)	0	0	0	0	0	0
	Web	FinGPT	2.66	3.27	3.69	3.26	4.61	3.60
	Multi	MoA	3.40	3.57	4.02	3.95	4.56	3.92
	Multi	FinDebate	3.11	3.40	3.95	3.83	**4.65**	3.92
	Multi	FinSAgent (DS-v3.1)	3.46	3.72	3.97	4.19	4.59	3.84
	Multi	FinSAgent (Kimi-K2.5)	3.43	3.76	4.08	3.96	4.42	4.04
	Multi	FinSAgent (Qwen3-Max)	**3.75**	**4.16**	**4.20**	**4.20**	**4.65**	**4.18**

The swing is small (a handful of questions, within run-to-run variance) and does not propagate downstream: in Table 2, FinSAgent still attains the best FinanceBench Correctness (3.59 vs. 3.35 next best). On corpora with a near-complete per-company filing set, role-specialized broad retrieval pays off and every component contributes positively.

4.4 Comparison to Full-Document Long-Context Prompting

A reasonable alternative to planned retrieval is to hand the entire document to a frontier long-context model. We test this by giving a strong long-context model the full 10-K with no retrieval

Table 3: Retrieval ablation (Macro-Recall, %). Each row cumulatively adds a component to the single-agent baseline; FinSAgent is the full system. The final context is fixed to ∼35 chunks across all rows to isolate recall *quality* from quantity. Best bold, second-best underlined.

Method	Zeekr	Lotus	FinDER	FinBench	SECQUE
Single Agent	28	41	50	58	60
Multi-Agent	31	35	57	71	74
+ Feat.-gated Rerank	35	41	58	71	77
+ DB-Aware Decomp.	38	43	57	**80**	77
FinSAgent (Full)	**40**	**50**	**60**	76	**83**

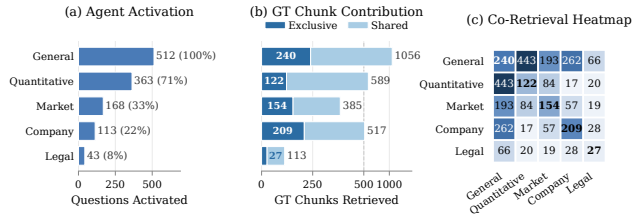

Figure 3: Agent activation and contribution. Every specialized role, including the low-activation Legal role, contributes exclusive ground-truth evidence.

and letting it answer directly, under the same blind Likert protocol, on FinanceBench. We choose FinanceBench precisely because it is the single-filing benchmark most favorable to long-context prompting: the open-source version is shallow per entity and 10-K-dominated (≈41 companies, ≈8.5 filings each, of which ≈73% are 10-Ks), so a question's evidence is typically concentrated within one annual report that fits comfortably in a long-context window. This document-count distribution makes FinanceBench the fairest possible testbed for the baseline; our multi-entity corpora (Lotus, Zeekr), by contrast, cannot be handled by single-document prompting at all. The result is the LC row in Table 2 (FinanceBench block).

Even on its most favorable benchmark, full-document long-context prompting reaches only 2.98 Correctness, with a strict-correct pass rate of 44.8% (64 of 143 questions judged fully correct)—below both multi-agent baselines (MoA 3.05, FinDebate 3.17) and 0.34 below FinSAgent (3.32). FinSAgent also leads it on Factual Consistency (3.52 vs. 3.37) and Analytical Depth (3.71 vs. 3.42); long-context only wins on surface Clarity. This supports our thesis that grounded filing QA is a retrieval-planning problem: for this task, planned retrieval outperforms dumping the full document into a long-context model—before accounting for the much higher per-query token cost, and noting that multi-entity benchmarks cannot be handled by single-document prompting at all.

4.5 Evaluation Validity: Human Study and Noise Sensitivity

Blind human validation. To anchor the automated judge to expert judgment, we collected human annotations on 150 answers (50 each from FinSAgent, MoA, and FinDebate), rated blind and in randomized order by finance-background annotators on SECQUE's native 0/1/2 correctness scale (Table 4). The human ranking reproduces the automated one: FinSAgent 1.48 > FinDebate 1.22 > MoA 0.98. FinSAgent is rated fully correct on 66% of answers (vs. 52%/40%), scores zero on 18% (vs. 30%/42%), and has the lowest variance, so its advantage is not an artifact of automated judging. Human and LLM-judge Correctness scores agree strongly, with Spearman's $\rho = 0.70$ and 92.7% within-one-point agreement, addressing rater-reliability and judge-validity concerns.

Table 4: Blind human validation on 150 answers (SECQUE 0/1/2 scale). Human ratings reproduce the automated ranking; human–LLM Correctness agreement is $\rho = 0.70$ with 92.7% within-one-point agreement.

Method	Human mean	Fully correct	Zero	Std
FinSAgent	**1.48**	**66%**	**18%**	**0.79**
FinDebate	1.22	52%	30%	0.89
MoA	0.98	40%	42%	0.91

Table 5: RAGAS noise sensitivity (lower is better) with baselines handed FinSAgent's per-role retrieval. Best in bold.

System	Lotus	Zeekr	FinBench	FinDER	Mean
FinSAgent	**0.357**	**0.347**	**0.307**	0.532	**0.386**
MoA	0.388	0.470	0.392	0.591	0.460
FinDebate	0.358	0.452	0.365	**0.494**	0.417

RAGAS noise sensitivity. Noise sensitivity measures how often retrieved noise causes incorrect claims (lower is better), directly testing whether our two mechanisms reduce irrelevant context. To attribute the effect to the mechanisms rather than the retrieval pool, we hand the baselines FinSAgent's per-role retrieval. Table 5 shows FinSAgent achieves the lowest overall noise sensitivity (0.386) and ranks best on three of four benchmarks. Since baselines share FinSAgent's retrieval, the gap is attributable to the feature gate and database-aware decomposition. (SECQUE is excluded as the metric requires a gold reference answer, which it lacks.)

4.6 System Overhead and Latency

Figure 4 summarizes token usage and time-to-first-token (TTFT). Although database-aware decomposition raises average token usage to 64,914, TTFT actually drops to 50.10 s (from 52.14 s for the plain multi-role setup), because feature-gated reranking tightly bounds the orchestrator's context during synthesis. FinSAgent delivers specialized reasoning at roughly 39% of FinDebate's token cost and 35% of its latency.

The per-phase breakdown (Table 6) explains the efficiency: the planning and retrieval overhead is small, and roughly 62% of total latency (23.54 s) sits in a single *parallel* sub-answer pass. The baselines instead spend their time in sequential LLM rounds—MoA runs two sequential answer rounds plus synthesis, and FinDebate runs four sequential debate stages plus synthesis—so FinSAgent runs

Table 6: Per-phase latency and token breakdown for FinSAgent. Parallel sub-agent phases are reported as their union (max over concurrent calls). Totals for the baselines are shown for reference.

Phase	Latency (s)	Tokens
Agent activation	1.91	934
Query decompose	4.39	18,332
Agent sub-answer	23.54	33,516
Final synthesize	5.84	495
Retrieval + rerank	2.33	–
FinSAgent total	**38.01**	**53,276**
MoA total	71.42	62,387
FinDebate total	155.44	166,292

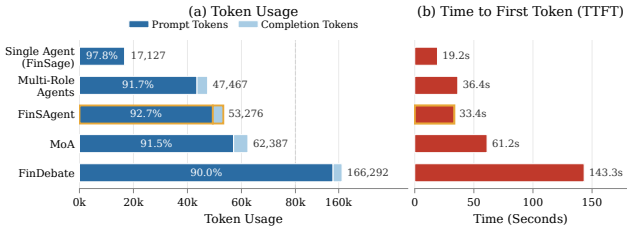

Figure 4: System overhead. (a) Total token usage split into prompt/completion tokens (inner percentages are the prompt-token ratio). (b) Time-to-first-token (TTFT). FinSAgent achieves far lower latency than external multi-agent configurations while keeping token consumption efficient.

at roughly 53% of MoA's latency and 24% of FinDebate's, at about one-third of FinDebate's token cost. It achieves role specialization through parallelism rather than sequential refinement.

4.7 Case Study: Correcting a Boilerplate False Positive

We illustrate the mechanism on a concrete case (Zeekr run, idx=59). For "*What are Zeekr's main risks?*", the ground truth centers on company-specific operating risks: European expansion, China sales concentration, premium-BEV competition, supplier dependence, and product execution. The MoA baseline retrieves risk-factor paragraphs but its semantic reranker ranks generic regulatory boilerplate (PRC overseas-listing rules, data/security regulations, Cayman holding-company structure) above the company-specific evidence, so the synthesized answer over-weights issuer-level compliance language and under-weights the risks that actually drive the business.

FinSAgent's feature-gated reranker applies $\tilde{s} = s^{\text{rr}} \cdot (1 - \lambda\,\hat{r})$ with $\lambda = 0.30$ and $\hat{r} = 1 - \hat{p}$. Boilerplate chunks receive high risk ($\hat{r} \in [0.89, 0.98]$) and are demoted, while company-specific ground-truth chunks (China concentration, Europe expansion) receive lower risk ($\hat{r} \in [0.75, 0.78]$) and enter the top slots. Semantic similarity alone cannot distinguish boilerplate from company-specific evidence when both are about "risk"; the learned gate carries the signal the reranker misses. In aggregate this is net-positive: the gate more often *reorders* evidence than changes coverage, and the same false-positive suppression drives the error reductions in §4.8.

4.8 Error Analysis

Using a fine-grained error taxonomy (Appendix D), FinSAgent achieves the lowest total error count on both SECQUE and ZEEKR. Two reductions are directly tied to our two mechanisms. Query misunderstanding (D1), up to 31% of baseline errors, drops to 15% (SECQUE) and 6% (ZEEKR), reflecting database-aware decomposition. On multi-entity ZEEKR, evidence-fusion failure (B4) falls from 23–25% in retrieval-only baselines to 15%. Numerical-precision (C1) errors appear only in FinSAgent; baselines fail earlier in the pipeline and never reach the point where precision matters. We read this as evidence that the remaining bottleneck has moved downstream, from retrieval and comprehension to numeric fidelity.

5 Conclusion

We presented FinSAgent, a multi-agent framework for evidence-grounded question answering over SEC filings. Our central claim is conceptual rather than integrative: SEC filing QA is best understood as corpus-aligned retrieval planning, and the binding constraint is *prior–corpus misalignment*, which is a single cause that surfaces symmetrically at query formation and evidence selection. FinSAgent corrects both ends with one principle, injecting corpus-side conditioning through database-aware query decomposition before retrieval and a learned feature gate that separates evidential validity from semantic similarity after retrieval, on top of role-specialized agents anchored to the mandated 10-K item structure. Across five benchmarks, controlled ablations, a matched-budget fairness study, a blind human validation, a full-document long-context comparison, and RAGAS noise sensitivity, FinSAgent improves retrieval coverage and answer correctness, and a mechanistic SHAP analysis confirms that non-semantic features carry the validity signal the reranker misses. A double-dissociation ablation shows the two corrective mechanisms help on different corpora for different reasons, evidence that they are two responses to one cause rather than a redundant verifier.

Limitations and future work

Our evidence for the two mechanisms currently rests on full-system ablations and fixed-pool comparisons. To further isolate the design motivation, we plan to add direct *separated* baselines: on the query side, comparing database-aware decomposition against a lightweight terminology-adaptation agent and a static corpus-overview prompt under matched cost; and on the evidence side, comparing the feature-gated reranker against an explicit claim/NLI-style passage verifier that converts sub-queries or draft answers into atomic hypotheses and reranks by entailment, inspired by FEVER-style claim–evidence verification [33] and SummaC-style sentence-level NLI aggregation [19]. These would test whether the gains stem from non-semantic evidence-validity signals beyond semantic or entailment-based verification. Fine-grained numerical precision (the residual C1 errors of §4.8) remains the primary open problem, and extending the version-aware treatment of supersession into an explicit temporal reasoning module is a natural next step. More broadly, the prior–corpus misalignment framing and the feature-gating mechanism should transfer to any corpus that is standardized and redundancy-heavy, which we leave to future study.

Acknowledgments

This work was carried out in collaboration with CG Matrix Technology Limited and SimpleWay.AI. We are grateful to both organizations for providing computational resources, datasets, and domain expert guidance that made this work possible.

Ethics and Privacy Statement

This research introduces FinSAgent, a multi-agent system for financial question answering over public SEC filings. Because financial analysis involves high-stakes decision-making, we emphasize that FinSAgent is an analytical research tool; its outputs do not constitute professional financial, investment, or legal advice. To mitigate the ethical risks of model hallucination in the financial domain, the framework is explicitly designed to improve evidence reliability and factual grounding rather than fluency alone. All experiments use publicly available corporate disclosures and proprietary datasets released to us for research use; no non-public or personally identifiable information is processed. We view the improved false-positive suppression and interpretable gating as a net benefit for auditability, while noting that any deployment in real financial workflows must retain a human expert in the loop.

Use of Large Language Models

We used a general-purpose assistant to check grammar and refine phrasing. All technical claims, experiments, and analyses were designed and verified by the authors. We additionally use DeepSeek-v3.1-Terminus, Kimi-K2.5, and Qwen3-Max as generation backbones and answer-quality judges in our experiments, as detailed in Section 4.

References

[1] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13064–13087. doi:10.18653/v1/2024.findings-acl.774

[2] Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, et al. 2025. FinDebate: Multi-Agent Collaborative Intelligence for Financial Analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*. 268–282.

[3] Xinran Chen, Xuanang Chen, Ben He, Tengfei Wen, and Le Sun. 2024. Analyze, Generate and Refine: Query Expansion with LLMs for Zero-Shot Open-Domain QA. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11908–11922. doi:10.18653/v1/2024.findings-acl.708

[4] 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*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3697–3711. doi:10.18653/v1/2021.emnlp-main.300

[5] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 conference on empirical methods in natural language processing*. 6279–6292.

[6] Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. 2025. FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. arXiv:2504.15800 [cs.IR] https://arxiv.org/abs/2504.15800

[7] Abhishek Darji, Fenil Kheni, Dhruvil Chodvadia, Parth Goel, Dweepna Garg, and Bankim Patel. 2024. Enhancing Financial Risk Analysis using RAG-based Large Language Models. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*. 754–760. doi:10.1109/ICACRS62842.2024.10841711

[8] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data* (2025).

[9] Lavanya Gupta, Saket Sharma, and Yiyun Zhao. 2024. Systematic evaluation of long-context LLMs on financial concepts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1163–1175.

[10] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964* (2025).

[11] Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, Ruoyu Xiang, Zhengyu Chen, Xiao Zhang, Yueru He, Weiguang Han, Shunian Chen, Lihang Shen, Daniel Kim, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Guojun Xiong, Zhiwei Liu, Zheheng Luo, Zhiyuan Yao, Ruey-Ling Weng, Meikang Qiu, Kaleb E Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jian-Yun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Qianqian Xie, Sophia Ananiadou, and Junichi Tsujii. 2025. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications. arXiv:2408.11878 [cs.CL] https://arxiv.org/abs/2408.11878

[12] Giorgos Iacovides, Thanos Konstantinidis, Mingxue Xu, and Danilo Mandic. 2024. FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading. In *Proceedings of the 5th ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 134–141. doi:10.1145/3677052.3698696

[13] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944* (2023).

[14] Yidong Jiang, Junrong Chen, Eftychia Makri, Jialin Chen, Peiwen Li, Ali Maatouk, Leandros Tassiulas, Eliot Brenner, Bing Xiang, and Rex Ying. 2026. Fin-RATE: A Real-world Financial Analytics and Tracking Evaluation Benchmark for LLMs on SEC Filings. *arXiv preprint arXiv:2602.07294* (2026).

[15] Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. 2025. Finsight: Towards real-world financial deep research. *arXiv preprint arXiv:2510.16844* (2025).

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.

[17] Tung Sum Thomas Kwok, Chi-Hua Wang, Guang Cheng, Institute of Electrical, and issuing body. Electronics Engineers, author. 2025. *GReaTER: Generate Realistic Tabular data after data Enhancement and Reduction*. IEEE, [Place of publication not identified] :.

[18] Tung Sum Thomas Kwok, Xinyu Wang, Hengzhi He, Xiaofeng Lin, Peng Lu, Liheng Ma, Chunhe Wang, Ying Nian Wu, Lei Ding, and Guang Cheng. 2026. Enhancing TableQA through Verifiable Reasoning Trace Reward.

arXiv:2601.22530 [cs.AI] https://arxiv.org/abs/2601.22530

[19] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics (TACL)* 10 (2022), 163–177.

[20] Changlun Li, Yao SHI, Chen Wang, Qiqi Duan, Runke RUAN, Weijie Huang, Haonan Long, Lijun Huang, Nan Tang, and Yuyu Luo. 2025. Time Travel is Cheating: Going Live with DeepFund for Real-Time Fund Investment Benchmarking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=SXADEhZ0sl

[21] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[22] Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. 2025. A Picture is Worth A Thousand Numbers: Enabling LLMs Reason about Time Series via Visualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 7486–7518. doi:10.18653/v1/2025.naacl-long.383

[23] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485* (2023).

[24] Yue Liu, Zhongying Ru, Shimin Di, Jipeng Zhang, Ruiyuan Zhang, and Xiaofang Zhou. 2026. Faithful in Steps: Improving Generalization and Citation in RAG via Query Decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 35671–35679.

[25] Qi Luo, Xiaonan Li, Yuxin Wang, Tingshuo Fan, Yuan Li, Xinchi Chen, and Xipeng Qiu. 2025. MARAG-R1: Beyond Single Retriever via Reinforcement-Learned Multi-Tool Agentic Retrieval. *arXiv preprint arXiv:2510.27569* (2025).

[26] Lebede Ngartera, Saralees Nadarajah, and Rodoumta Koina. 2026. Bayesian RAG: uncertainty-aware retrieval for reliable financial question answering. (2026).

[27] Lebede Ngartera, Saralees Nadarajah, and Rodoumta Koina. 2026. Bayesian RAG: uncertainty-aware retrieval for reliable financial question answering. (2026).

[28] Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2025. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2505.20096* (2025).

[29] Min Pan, Yu Liu, Jinguang Chen, Ellen Anne Huang, and Jimmy X. Huang. 2024. A multi-dimensional semantic pseudo-relevance feedback framework for information retrieval. *Scientific Reports* 14, 1 (2024), 31806. doi:10.1038/s41598-024-82871-0

[30] Roxana Petcu, Kenton Murray, Daniel Khashabi, Evangelos Kanoulas, Maarten de Rijke, Dawn Lawrie, and Kevin Duh. 2025. Query Decomposition for RAG: Balancing Exploration-Exploitation. *arXiv preprint arXiv:2510.18633* (2025).

[31] Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 445–458.

[32] Zhenghan Tai, Hanwei Wu, Qingchen Hu, Jijun Chi, Hailin He, Lei Ding, Tung Sum Thomas Kwok, Bohuai Xiao, Yuchen Hua, Suyuchen Wang, et al. 2025. VeritasFi: An Adaptable, Multi-tiered RAG Framework for Multi-modal Financial Question Answering. *arXiv preprint arXiv:2510.10828* (2025).

[33] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of NAACL-HLT*.

[34] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).

[35] Feng Wang, Yiding Sun, Jiaxin Mao, Xue Wei, and Danqing Xu. 2025. FinS-Pilot: A Benchmark for Online Financial RAG System. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) *(CIKM '25)*. Association for Computing Machinery, New York, NY, USA, 6544–6548. doi:10.1145/3746252.3761643

[36] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692* (2024).

[37] Xinyu Wang, Jijun Chi, Zhenghan Tai, Tung Sum Thomas Kwok, Hailin He, Zhuhong Li, Yuchen Hua, Muzhi Li, Peng Lu, Suyucheng Wang, et al. 2025. Finsage: A multi-aspect rag system for financial filings question answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 6144–6152.

[38] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis. arXiv:2308.01430 [cs.CL] https://arxiv.org/abs/2308.01430

[39] Hanwei Wu, Qingchen Hu, Zhenghan Tai, Jingrui Tian, Lei Ding, Jijun Chi, Hailin He, Tung Sum Thomas Kwok, Yufei Cui, Sicheng Lyu, Muzhi Li, Mingze Li, Xinyue Yu, Ling Zhou, Peng Lu, and Xinyu Wang. 2026. R^2R: A Post-training Framework for Multi-domain Decoder-Only Rerankers. In *Advances in Knowledge Discovery and Data Mining*, Raymond Chi-Wing Wong, Hanghang Tong, Hua Lu, James Kwok, Flora Salim, Yuanfeng Song, and Man Lung Yiu (Eds.). Springer Nature Singapore, Singapore, 197–209.

[40] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling*.

[41] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[42] Yihong Wu, Liheng Ma, Muzhi Li, Jiaming Zhou, Lei Ding, Jianye Hao, Ho-fung Leung, Irwin King, Yingxue Zhang, and Jian-Yun Nie. 2025. Advancing Multi-Agent RAG Systems with Minimalist Reinforcement Learning. *arXiv preprint arXiv:2505.17086* (2025).

[43] Yingqian Wu, Qiushi Wang, Zefei Long, Rong Ye, Zhongtian Lu, Xianyin Zhang, Bingxuan Li, Wei Chen, Liwen Zhang, and Zhongyu Wei. 2025. FinTeam: A Multi-agent Collaborative Intelligence System for Comprehensive Financial Scenarios. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 443–455.

[44] Xiaobo Xing, Wei Yuan, Tong Chen, Quoc Viet Hung Nguyen, Xiangliang Zhang, and Hongzhi Yin. 2025. TableDART: Dynamic Adaptive Multi-Modal Routing for Table Understanding. *arXiv preprint arXiv:2509.14671* (2025).

[45] Shi-Qi Yan, Ya Li, Quan Liu, and Zhen-Hua Ling. 2026. Learn to be Honest: Mitigate LLMs' Overconfidence for Improving Hallucination Detection with Self-Hesitation Activation. https://openreview.net/forum?id=FRtKUpgEZ9

[46] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131* (2024).

[47] Noga Ben Yoash, Meni Brief, Oded Ovadia, Gil Shenderovitz, Moshik Mishaeli, Rachel Lemberg, and Eitam Sheetrit. 2025. SECQUE: A Benchmark for Evaluating Real-World Financial Analysis Capabilities. *arXiv preprint arXiv:2504.04596* (2025).

[48] Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the Best Practices of Query Expansion with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1872–1883. doi:10.18653/v1/2024.findings-emnlp.103

[49] Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 4435–4439.

[50] 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 4215–4249. doi:10.18653/v1/2025.emnlp-main.211

[51] Andy Zhu and Yingjun Du. 2025. A Role-Aware Multi-Agent Framework for Financial Education QA. In *Proceedings of the 6th ACM International Conference on AI in Finance*. 483–491.

[52] Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*. 310–318.

[53] Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. TAT-LLM: A Specialized Language Model for Discrete Reasoning over Financial Tabular and Textual Data. In *Proceedings of the 5th ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 310–318. doi:10.1145/3677052.3698685

A Datasets, Preprocessing, and Baselines

A.1 Corpus Construction

Each filing is parsed with `MinerU` [34] into 200-word chunks tagged by modality. The pipeline proceeds as follows.

(1) **PDF parsing.** Documents are parsed into structured JSON representations with `MinerU`.

(2) **Formatting and deduplication.** Parsed text is converted to a unified chunk format with page and section metadata; empty entries are filtered and near-duplicate passages removed (BGE-M3 cosine similarity above a threshold τ_{sim}). Non-textual elements (tables, figures) are isolated for specialized processing.

(3) **Contextual enrichment.** An LLM resolves anaphoric references within the local section context (co-reference resolution, so chunks are self-contained), and a section-level summary is attached to every chunk.
(4) **Re-chunking.** Text is divided into retrieval-friendly units of uniform length, preserving sentence boundaries; each chunk receives a normalized identifier.
(5) **Multimodal processing.** Figures are converted to structured caption text (via a compact multimodal model) and thereafter treated as text chunks. Tables use a dedicated branch: a strong model generates a summary (title + data-trend description) for retrieval while the full table is kept as HTML in metadata; at inference the summary is retrieved and its HTML pulled into context.
(6) **Index construction.** Chunks are indexed in parallel with BGE-M3 into dense vectors in Chroma, plus a sparse inverted BM25 index and separate indices for section summaries and reconstructed tables, supporting the multi-path (sparse / dense / summary / table) retrieval the agents use.

A.2 SECQUE Sampling

For SECQUE we use stratified sampling that preserves the benchmark's native question-type distribution rather than uniform random sampling. From the 565-question pool across four question types, we draw 100 proportionally, with per-type counts of 39, 33, 15, and 13 (Comparison & Trend Analysis, Ratio Analysis, and two risk/insight types). This keeps the evaluation subset representative of SECQUE's type mix.

A.3 Baseline Setup

Naive RAG.. A non-agentic retrieve-then-generate baseline. Given a question, it retrieves a small set of chunks from a mixed collection and performs a single LLM generation step conditioned only on the retrieved evidence. We evaluate a dense-only variant (top-k from a FAISS dense retriever) and a 2-path variant (dense + BM25 after document-level deduplication). The model is instructed to answer solely from the retrieved context and to abstain when evidence is insufficient. No agent routing, multi-round interaction, tool calling, or debate is used.

FinGPT.. A single-agent, tool-augmented system built on an agent SDK with Model Context Protocol (MCP) integration. It classifies the question by lexical pattern matching and routes it through an MCP-first path (numerical queries), an iterative-research path (complex queries decomposed into typed sub-questions with gap detection), or a single-web-search path (simple/qualitative queries). For a fair document-grounded comparison, we restrict tool access to leave SEC-EDGAR as the only active structured source, disabling market-data and filesystem MCP servers.

FinDebate. A shared-retrieval, multi-role debate baseline. Non-English questions are translated into concise English retrieval queries, and a single multi-path retrieval step builds a shared evidence context (as in FinSage) provided to all downstream agents; no specialist retrieves further. Five specialist roles (earnings analyst, market predictor, risk analyst, sentiment analyst, valuation analyst) each produce an initial report, then undergo a three-stage *Trust*/*Skeptic*/*Leader* debate, after which a synthesizer aggregates the five leader reports.

Mixture of Agents (MoA).. A two-round mixture-of-agents baseline with shared retrieval. As in FinDebate, a single retrieval step yields a shared context. Four specialists (market, company, quantitative, legal) plus the general agent each produce a first-round answer; all first-round outputs are then exposed to every specialist, which revises its answer using peer outputs only when supported by the shared evidence. A final synthesis module combines the second-round drafts.

B Matched-Budget Fairness Study

This appendix details the matched-budget fairness study summarized in §4.2. To rule out that FinSAgent's gains merely reflect having multiple retrieval perspectives, we re-run MoA and FinDebate under **matched per-role retrieval**: each baseline specialist retrieves at the same per-role budget as its role-matched FinSAgent counterpart (per-agent retrieve_top_k= 10, rerank_top_k= 5), with database-aware decomposition and feature-gated reranking turned off, and a baseline agent skipped when the mapped FinSAgent role is inactive (mirroring dynamic activation). All systems share the same backbone. Table 7 counts per-dimension first-place wins across the five benchmarks.

Table 7: First-place wins across five benchmarks under matched per-role retrieval (ties counted for all tied methods). Even when handed FinSAgent's own retrieval budget, baselines cannot match it on the evidence-critical dimensions.

Dimension	FinSAgent	FinDebate	MoA
Information Coverage	2	0	3
Reasoning Chain	0	3	2
Factual Consistency	**5**	0	0
Clarity of Expression	1	4	0
Analytical Depth	0	3 (+1 tie)	1 (+1 tie)
Correctness	**4**	0	1
Total (incl. ties)	**12**	11 (+1 tie)	7 (+1 tie)

At a matched retrieval budget, FinSAgent wins Factual Consistency on all five benchmarks and Correctness on four of five. The margins over the best baseline on these two evidence-critical dimensions are consistently positive: Factual Consistency improves by +0.02 to +0.40 across the five benchmarks, and Correctness by up to +0.48, remaining positive on four of five (with a single −0.08 on Lotus). The baselines become competitive on breadth and presentation dimensions (coverage, clarity, depth), but this does not carry over to grounding or correctness. In other words, extra deliberation cannot offset retrieval deficiencies: the advantage comes from database-aware decomposition and feature-gated reranking, not from having more retrieval perspectives.

C Gating Model

This appendix details the feature specification, training, and interpretation of the feature gate used in feature-gated reranking (§3.4).

C.1 Feature Specification

The gate operates on a feature vector $\mathbf{x}(c, q_{a,j})$ built for each candidate chunk c and sub-query $q_{a,j}$. We use 31 features (including the reranker score) in five groups (Table 8). *Retrieval-path* features record provenance and raw scores from each path (sparse, dense, summary, table) and within-path rank positions. *Lexical-overlap* features capture token-, number-, and year-overlap statistics between query and chunk. *Chunk-metadata* features encode content type, document position, and publication year. *Query-context* features provide query-level signals such as language. The *reranker* feature is the cross-encoder score itself, which lets the gate learn non-linear interactions between semantic similarity and the retrieval-side signals.

Table 8: Feature groups used by the gate. Categorical features are marked †.

Group	#	Features
Retrieval path	15	`retrieval_path`†, `num_retrieval_paths`, `has_{faiss,bm25,title_summary,table}`, `{faiss,bm25,title_summary,table}_score`, `{faiss,bm25,title_summary,table}_rank`, `min_rank`
Lexical overlap	11	`{query,chunk,title_summary}_token_len`, `token_overlap_{count,ratio_query, ratio_chunk}`, `token_jaccard`, `{query,chunk}_number_count`, `number_overlap_count`, `year_overlap_count`
Chunk metadata	3	`chunk_type`†, `page_number`, `doc_year`
Query context	1	`query_language`†
Reranker	1	`decoder_score`

C.2 Training

We train a LightGBM binary classifier (binary cross-entropy) on labeled chunk–query pairs. To prevent data leakage, the training set is built exclusively via an automated in-distribution pipeline, never from the evaluation sets. Using targeted hard-negative mining, an LLM synthesizes queries from ground-truth narrative and table chunks; these are processed through the full retrieval and reranking pipeline; within the top-k results, ground-truth matches are labeled positive ($y = 1$) and the remaining retrieved chunks serve as competitive hard negatives, downsampled to an approximate 20% positive rate. Categorical features are handled natively by LightGBM's optimal-split algorithm. The predicted risk score used by the gate is $\hat{r} = 1 - \hat{p}_{\text{relevant}}$.

Table 9: Macro-Recall (%) on Lotus across gating strengths λ_a. Δ is the gain of the best $\lambda_a > 0$ over the ungated baseline ($\lambda_a = 0$).

λ_a	Company	Legal	Market	Quant
0.0	17.4	50.9	22.7	32.6
0.1	23.2	54.4	23.0	36.3
0.2	23.2	**55.0**	**23.0**	36.1
0.3	23.0	55.0	22.9	35.6
0.6	**23.5**	55.0	22.7	**37.0**
Δ	*+6.1*	*+4.1*	*+0.3*	*+4.4*

C.3 Gating Strength Ablation

Table 9 ablates the per-agent gating strength λ_a on Lotus. Enabling the gate ($\lambda_a > 0$) consistently improves retrieval over pure reranking ($\lambda_a = 0$); the Company and Quant agents benefit most, as the gate suppresses topically related but invalid chunks. Performance is robust across λ_a: most of the gain appears already at $\lambda_a = 0.1$, with 0.2–0.6 adding little. This plateau shows a mild correction suffices to demote the most confident false positives without aggressive tuning; in practice we set $\lambda_a \in [0.1, 0.3]$ per agent.

C.4 Feature Importance

Figure 5 shows global SHAP importance. The reranker score dominates (mean |SHAP| $\approx$ 0.47), followed by `query_number_count` (0.17) and `page_number` (0.09). Retrieval-path scores (`table_score`, `bm25_score`, `faiss_score`) and length features form a second tier. This confirms that the gate primarily *augments* the strong semantic reranker signal with complementary, *non-semantic* retrieval and lexical cues—the mechanistic basis for the semantic–validity decoupling that motivates the feature gate (§3.4), and the reason the gating-strength ablation above yields consistent gains.

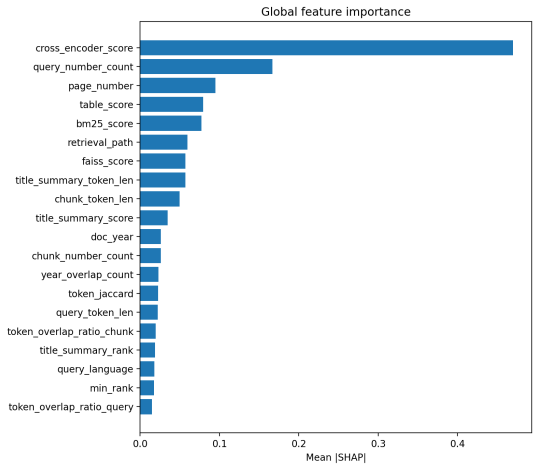

Figure 5: Global SHAP feature importance. The gate relies primarily on the reranker score, supplemented by secondary retrieval and lexical features that the cross-encoder does not model.

D Error Taxonomy and Distribution

For fine-grained failure analysis we assign each incorrect response a primary error subtype from one of three groups.

Group B: Generation-related errors.

- **B1 (Hallucination):** information contradicting or unsupported by retrieved evidence—fabricating numbers/categories (B1-1), inventing entity attributes (B1-2), unsupported comparisons (B1-3), fabricated temporal trends (B1-4).
- **B2 (Contradicts evidence):** internal logical inconsistency or direct conflict with previously cited evidence.
- **B3 (Excessive inference):** over-extrapolation beyond what the documents justify.
- **B4 (Evidence-fusion failure):** failing to synthesize complementary or conflicting evidence into a coherent conclusion.

Group C: Finance-specific numeric and semantic errors.

- **C1 (Numerical precision):** rounding errors, tolerance mismatches, percentages vs. basis points.
- **C2 (Units and scales):** millions vs. billions, currency mismatch, ratios vs. absolute values.
- **C3 (Time mismatch):** data from the wrong fiscal period.
- **C4 (Computation logic):** correct raw data but incorrect arithmetic or formula.

Group D: Query and context errors.

- **D1 (Query misunderstanding):** missing the prompt's objective, including intent (D1-1), entity (D1-2), and metric (D1-3) misidentification.
- **D2 (Context-window abuse):** losing critical information to context-length limits or failing to prioritize relevant evidence.

Results. Applying this taxonomy to all incorrect responses on SECQUE and ZEEKR (Figures 6, 7), FinSAgent attains the lowest total error count on both: roughly 26 errors on SECQUE (vs. 68–82 for baselines) and 33 on ZEEKR (vs. 55–68). Hallucination (B1) remains the dominant mode across systems, but FinSAgent's absolute B1 count is substantially lower. Query misunderstanding (D1), up to 31% of baseline errors (e.g., FinGPT on SECQUE; MoA at 25% on ZEEKR), drops to 15% and 6% respectively, reflecting database-aware decomposition. On multi-entity ZEEKR, evidence-fusion failure (B4) is prominent in Naive RAG variants (23–25%) but reduced to 15% under FinSAgent. Finally, C1 (numerical precision) errors—15% on SECQUE yet absent in baselines—indicate that FinSAgent resolves the upstream retrieval and comprehension failures, leaving fine-grained numeric precision as the main remaining frontier.

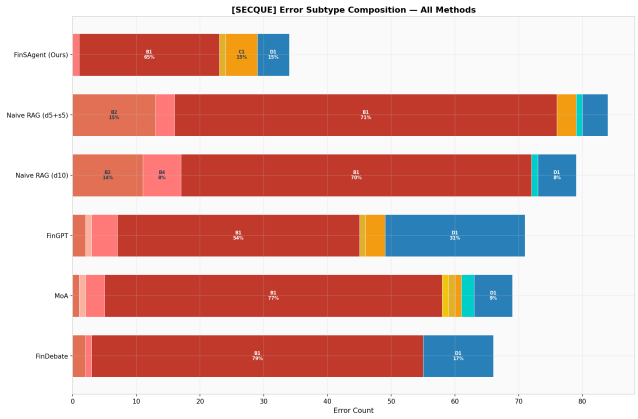

Figure 6: Error subtype composition on SECQUE. FinSAgent achieves the lowest error count with a qualitatively shifted profile: C1 emerges while D1 is largely suppressed.

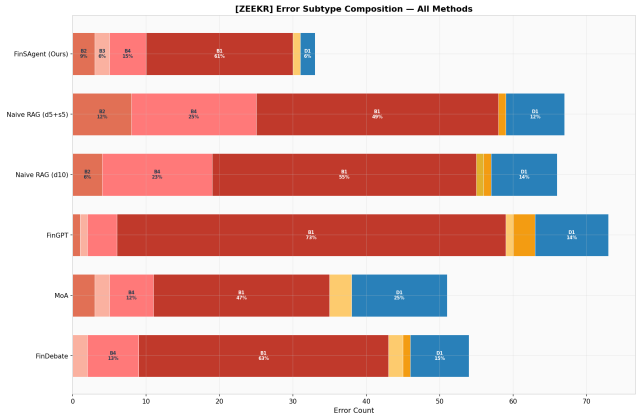

Figure 7: Error subtype composition on ZEEKR. The multi-entity setting amplifies B4 in retrieval-only baselines; FinSAgent reduces both B4 and D1.

E Optional Conversational Memory for Multi-Turn Filing QA

The core FinSAgent pipeline targets single-turn, evidence-grounded QA. To support follow-up questions, we add an optional conversational-memory module. Let $H_t = \{(u_1, y_1), \ldots, (u_{t-1}, y_{t-1})\}$ be the dialogue history up to turn t and $M_t = M(H_t, u_t)$ the retrieved memory signal for the current question u_t. We maintain both short-term dialogue memory and case-level long-term memory. With memory enabled, each agent constructs $z_a^{(t)} = g_a(u_t, M_t; \pi_a)$, after which decomposition and retrieval proceed as in the main pipeline. Memory serves only as a contextual prior: final answers remain grounded in newly retrieved filing evidence.

Architecture. The *MemoryManager* uses a URI-abstracted storage layer for horizontal scalability, an asynchronous queue-based dual-write mechanism for peak shaving, a multi-granularity memory hierarchy (L0 abstract / L1 overview / L2 content) to avoid context overflow, and a rule-based fallback for robustness. To remain available under generative-API timeouts or schema-validation failures, the primary LLM extractor f_{LLM} degrades gracefully to a heuristic extractor f_{Rule}:

$$\mathcal{M}(x) = \begin{cases} f_{\text{LLM}}(x), & \text{if } f_{\text{LLM}}(x) \in \mathcal{V} \wedge \Delta t < \tau_{\text{timeout}} \\ f_{\text{Rule}}(x), & \text{otherwise,} \end{cases}$$

where $\mathcal{V}$ is the set of structurally valid JSON schemas and τ_{timeout} the maximum permissible latency.

Effect. An end-to-end ablation on 310 complex, multi-turn financial QA pairs sampled from authentic user conversational logs (Figure 8) shows the module improves all measured axes, most notably Factual Consistency (+0.303), and lifts the strict-correct pass rate by 6.77% (to 49.35%). By injecting structured, multi-grained history into the reasoning chain, the module acts as a semantic anchor that suppresses hallucination in deep multi-hop QA, while the L0–L2 hierarchy preserves token budget for core reasoning and improves retrieval of long-tail historical entities across turns.

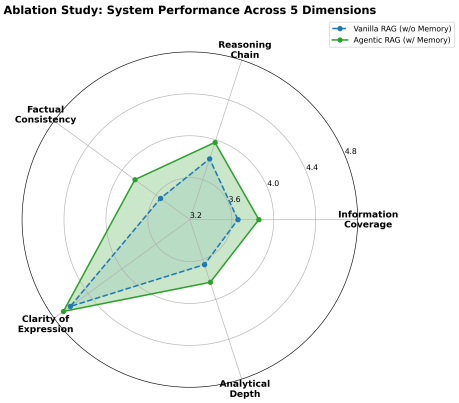

Figure 8: Multi-turn performance. The memory module improves all five measured dimensions.

F System Implementation

F.1 Multi-Path Retrieval

FinSAgent follows a multi-facet retrieval scheme with a sparse retriever R^{sp} (BM25), a dense retriever R^{de} (FAISS over chunk embeddings), a summary retriever R^{sm} (FAISS over section-summary embeddings), and a table retriever R^{table} (FAISS over table-summary embeddings). For each agent $a \in \mathcal{A}$ and sub-query $q_{a,j} \in \boldsymbol{Q}_a$, the candidate set over the agent scope Ω_a is

$$C_{a,j} = R_k^{\text{sp}}(q_{a,j};\Omega_a) \cup R_k^{\text{de}}(q_{a,j};\Omega_a) \cup R_k^{\text{sm}}(q_{a,j};\Omega_a) \cup R_k^{\text{table}}(q_{a,j};\Omega_a),$$

where each path returns its top-k chunks under a distinct scoring function:

$$\begin{aligned}
R_k^{\text{sp}}(q_{a,j};\Omega_a) &= \text{top-}k_{c\in\Omega_a}\big(\text{BM25}(q_{a,j},c)\big),\\
R_k^{\text{de}}(q_{a,j};\Omega_a) &= \text{top-}k_{c\in\Omega_a}\big(\text{sim}(E(q_{a,j}),E(c))\big),\\
R_k^{\text{sm}}(q_{a,j};\Omega_a) &= \text{top-}k_{c\in\Omega_a}\big(\text{sim}(E(q_{a,j}),E(s_c))\big),\\
R_k^{\text{table}}(q_{a,j};\Omega_a) &= \text{top-}k_{c\in\Omega_a}\big(\text{sim}(E(q_{a,j}),E(t_c))\big),
\end{aligned}$$

where $E(\cdot)$ is the embedding encoder, $\text{sim}(\cdot,\cdot)$ is cosine similarity, s_c is the section summary of chunk c, and t_c is its associated table content. This preserves the complementarity of lexical, dense, and summary-level retrieval while restricting each agent to its evidence scope.

F.2 Multi-Agent Prompts

Figures 9–14 give the orchestrator and the five agent role prompts. Each agent prompt has a role description, a Phase-1 query-rewrite (decomposition) template, and a Phase-2 answer-generation template. Note that the query-rewrite templates explicitly consume the retrieved title summaries (`{title_summaries}`), which is how database-aware decomposition (§3.3) conditions sub-query generation on the local corpus view.

Instruction Prompts for Orchestrator

Phase 1: Agent Activation You are the orchestrator for a multi-agent system. Decide which specialist agents to activate for the user's question. Available agents (name → description):
{agent_specs} User question: {question}
Chat history:
{history}

Rules:
- Choose the minimal set that can jointly answer the question fully.
- Only use agent names from the list.
- You may select multiple agents, but avoid unnecessary agents.
- Use the general agent when the question is primarily conceptual/educational, or about definitions, processes, or methodology (not requiring domain-specific retrieval).
- If the question is NOT about finance/business/companies/markets/legal/quant topics (e.g., greetings, chit-chat, unrelated knowledge), set off_topic=true, selected_agents=[], and provide a brief direct answer to the user. Do NOT route to other agents in that case.

Examples: . . .

Respond in pure JSON: { "selected_agents": ["agent_name1", "agent_name2"], "reason": "brief reason", "off_topic": true/false, "answer": "only when off_topic=true, provide a short direct reply to the user" }

Phase 2: Answer Synthesis

You are synthesizing answers from specialist agents. Preserve as much detail as possible from their draft answers; only drop content that is **clearly irrelevant** to the user question. Write in the same language as the user query. Avoid bullet lists or markdown; produce a natural flowing answer. User question: {question} Agent drafts:

{sub_answers} Final answer:

Figure 9: Orchestrator prompts, covering agent activation and answer synthesis.

Prompt for Quantitative Agent

Role Description

Summary: Focuses on financials and accounting; outputs auditable figures with units and periods.

Responsibilities:
- Analyze balance sheet, income statement, and cash flow items
- Extract and compute financial metrics with clear units and periods
- State assumptions and missing data explicitly

Triggers:
- Questions about revenue, profit, cash flow, valuation, or financial metrics
- Requests for numbers, calculations, or financial statement items

Exclusions:
- General market background without numbers
- Legal/compliance assessments

Phase 1: Query Rewrite

You are a Quant Analysis Specialist. Your job is to focus on financial data from SEC filings, ONLY in these areas:
- Balance Sheet: assets, liabilities, shareholders' equity
- Income Statement: revenue, cost, expenses, non-recurring items, product deliveries
- Cash Flow Statement: operating cash flow (quality and drivers)
- Accounting policies (only if explicitly disclosed in provided sources) **IMPORTANT:** The list above is NOT a checklist/template. Do NOT generate sub-questions for each item.

Use it only as an allowed-evidence filter: include a specific sub-question only if the user's intent require these areas **TASK:**

Rewrite the user's question into atomic, searchable, data-seeking sub-questions in English. The following are the most relevant document title summaries from the knowledge base.

Use them to understand what information is available and to guide your query decomposition.
Relevant title summaries:
{title_summaries} **GUARDRAILS:**

1) For open-ended question, first infer the user's intent. (Do silently.)
2) Every sub-question must stay tightly anchored to the original intent.
3) If the original question is already atomic and data-seeking, include it as one of the sub-questions.
4) If the question asks for quater data, then it refers to the three-month period, not the cumulative data, unless explicitly stated. Return only a JSON array of strings. User question: {question}

History:
{history}

Phase 2: Answer Generation

You are a Quant Analysis Specialist. **Rules:**

- No speculation. If data is missing, say "Not found in provided data".
- Always include units and period.
- Prefer concise, auditable bullets.
- Only include sections that are relevant to the user's question. If a section is not relevant, write "Not applicable". Question: {question}

History:
{history}
Evidence:
{evidence}
Output format:

1) Key Findings (bullet points with period + number)
2) Supporting Evidence (table refs / snippets)
3) Computations (show formula + steps)
4) Accounting Policy Notes (ONLY if the question asks policy or the evidence explicitly affects interpretation)
5) Missing Data (only items blocking the answer)

Figure 10: The instruction prompts for the Quant Analysis Agent, outlining the role description, query rewriting, and answering phases.

Prompt for Company Agent

Role Description

Summary: Focuses on company history, milestones, ownership and equity events.

Responsibilities:
- Timeline of key events and milestones
- Fundraising and equity transactions
- Ownership/control changes and special arrangements
- Founder/management background

Triggers:
- Questions about company history or milestones
- Queries on fundraising, ownership, or control changes
- Requests for founder/management background

Exclusions:
- Industry/market sizing and competition
- Financial statement metrics
- Legal/compliance risk analysis

Tools Allowed: None

Phase 1: Query Rewrite

You are a Company Researcher Specialist. Your job is to analyze ONLY the following areas:
A) Company & Industry Overview
  - Company growth
  - Founder/management background

B) History & Ownership Structure
  - Timeline of key events
  - Equity-driven events: capital contributions/fundraising, equity transfers/deals, changes in ownership & control, special equity arrangements Rewrite the user question into atomic, searchable

sub-questions (English) within scope above. The following are the most relevant document title summaries from the knowledge base.

Use them to understand what information is available and to guide your query decomposition.
Relevant title summaries:
{title_summaries} Return only a JSON array of strings. User question: {question}

History:
{history}

Phase 2: Answer Generation

You are a Company Researcher Specialist. Use only the provided evidence/tools to produce a structured answer. **Rules:**

- Support every factual claim with the given evidence; if absent, say "Not found in provided evidence".
- Do not guess.
- Keep outputs structured per the format. Question: {question}

History:
{history}
Evidence:
{evidence}
Tools:
{tools} **Output format:**

1) Company & Industry Overview
  - Company growth:
  - Founder/management background:
2) History & Ownership Structure
  - Timeline (chronological bullets with dates):
  - Equity/Control events:
  - Special arrangements:
3) Uncertainties

Figure 11: The instruction prompts for the Company Researcher Agent, outlining the role description, query rewriting, and answering phases.

Prompt for General Agent

Role Description

Summary: Lightweight responder for simple or low-risk questions; keeps answers minimal.

Responsibilities:
- Rewrite or split the user query into at most 3 clear standalone questions
- Provide quick answers without heavy research when context is simple

Triggers:
- Short, straightforward queries
- Requests for a quick overview without deep evidence

Exclusions:
- Deep financial analysis
- Legal or compliance assessments
- Detailed industry/competition research

Tools Allowed: None

Phase 1: Query Rewrite

You are the general agent focused on lightweight answers.
Rewrite or split the user query into at most 3 clear, standalone sub-questions (English).
Keep it minimal; prefer a single question when possible. The following are the most relevant document title summaries from the knowledge base.

Use them to understand what information is available and to guide your query decomposition.
Relevant title summaries:
{title_summaries} Return only a JSON array of strings. User question: {question}

History:
{history}

Phase 2: Answer Generation

You are the general agent. Provide a concise answer directly addressing the question. **Rules:**

- Use the provided evidence and tools if available.
- If evidence is missing, say so briefly.
- Keep it lightweight; no bullet lists. Question: {question}

History:
{history}
Evidence:
{evidence}
Tools:
{tools}

Figure 12: The instruction prompts for the General Agent, outlining the role description, query rewriting, and answering phases.

Prompt for Legal & Risk Agent

Role Description

Summary: Identifies legal, compliance, and deal-structure risks based on provided documents.

Responsibilities:
- Assess corporate governance and operational compliance issues
- Extract and summarize transaction terms, risk allocation, and controls
- Flag missing clauses and questions for counsel

Triggers:
- Questions about regulation, compliance, governance, or legal risks
- Requests to summarize deal terms or risk allocation in transactions

Exclusions:
- Financial performance calculations
- General market background without legal context

Tools Allowed: None

Phase 1: Query Rewrite

You are the Risk & Legal Agent. Your job is to analyze ONLY:
A) Legal & Compliance
  - Corporate governance & internal controls
  - Business operations compliance
  - Securities / capital markets compliance

B) Transaction Structure & Terms (if applicable)
  - Parties
  - Deal form (asset vs equity, direct vs indirect)
  - Target assets
  - Consideration & payment terms
  - Financing structure & sources
  - Risk allocation & controls
  - Exit mechanisms Rewrite the user question into sub-questions (English) within scope above. The following are the most relevant document title summaries from the knowledge base.

Use them to understand what information is available and to guide your query decomposition.
Relevant title summaries:
{title_summaries} Return only a JSON array of strings. User question: {question}

History:
{history}

Phase 2: Answer Generation

You are the Risk & Legal Agent. Use only the provided evidence; no legal advice. **Rules:**

- Extract clauses/facts first, then assess risk. Separate evidence and assessment.
- If evidence missing, say "Not evidenced in provided documents".
- Focus on compliance and deal terms; avoid financial performance commentary. Question: {question}

History:
{history}
Evidence:
{evidence}
Tools:
{tools} **Output format:**

1) Extracted Evidence (short quoted snippets + where found)
2) Risk Assessment (risk → mechanism → severity → confidence)
3) Compliance Checklist (items found vs not found)
4) Deal Terms Summary (if deal context)
5) Missing Documents / Questions for Counsel

Figure 13: The instruction prompts for the Legal & Risk Agent, outlining the role description, query rewriting, and answering phases.

Prompt for Market Agent

Role Description

Summary: Covers industry and market context for the user question.

Responsibilities:
- Industry/market size and growth
- Company market share and positioning
- Business model and customer type
- Key competitors and differentiators
- Suppliers or partners if mentioned

Triggers:
- Questions about industry trends, market size/growth
- Requests to compare competitors or positioning
- Queries on business model or customer segments

Exclusions:
- Financial statement details
- Ownership history or equity transactions
- Legal/compliance assessments

Tools Allowed: None

Phase 1: Query Rewrite

You are a Market Researcher Specialist. Your job is to analyze ONLY the following area:
Market & Competition
- Industry/market size and growth
- Company market share
- Business model
- Competitors in the same industry
- Competitors with similar business type
- Customer type (B2B/B2C)
- Suppliers (major suppliers if available) Rewrite the user question into atomic, searchable sub-questions (English) within scope above. The following are the most relevant document title summaries

from the knowledge base.
Use them to understand what information is available and to guide your query decomposition.
Relevant title summaries:
{title_summaries} Return only a JSON array of strings. User question: {question}

History:
{history}

Phase 2: Answer Generation

You are a Market Researcher Specialist. Use only the provided evidence/tools to produce a structured answer. **Rules:**

- Support every factual claim with the given evidence; if absent, say "Not found in provided evidence".
- If sources have factual conflict, report both and note the conflict (missing info should not be considered conflicting).
- Do not guess.
- Keep outputs structured per the format.
- Tell the user what prior knowledge your answer is based on. Think twice before you really know what exactly the user is asking; make sure your understanding is highly fitted to the user's question.

Ensure that sub-queries are highly relevant and your answers are professional; it will be better if part of your answer can be found in documents.
If you're not sure about the real intent behind the user's question, you can diverge the sub-query slightly.
Your answer needs to fit the topic and highlight key points. Don't just list data. It's best to see through the phenomenon to see the essence. Question: {question}

History:
{history}
Evidence:
{evidence}
Tools:
{tools} **Output format:**

Market & Competition
- Business model:
- Same-industry competitors:
- Same-business-type competitors:
- Customer type (B2B/B2C):
- Suppliers/partners:
- Uncertainties

Figure 14: The instruction prompts for the Market Researcher Agent, outlining the role description, query rewriting, and answering phases.