# Fin-STAR: Structure-as-Semantics to Resolve Implicitness in Financial Retrieval

Anonymous ACL submission

## Abstract

Understanding financial documents is critical for high-stakes decision-making yet hindered by *systemic semantic implicitness*: key facts are rarely explicit in surface text and often determined by global structural cues. Missing these cues invites semantic misinterpretations, such as misreading what a number refers to, an outcome unacceptable in high-stakes environments. However, existing Retrieval-Augmented Generation (RAG) systems typically treat structure as a physical navigational skeleton rather than intrinsic semantic knowledge. To address this, we introduce **Fin-STAR** (**Fin**ancial **ST**ructure-**A**s-Semantics **R**etrieval), a framework redefining hierarchy as intrinsic semantics. Fin-STAR incorporates a novel *Structure-Enriched Semantic Indexing* mechanism that augments the hierarchical lineage with snippet-derived virtual nodes, and injects this enriched context via a semantic cross-attention paradigm, rendering implicit cues explicit. By grounding evidence within its structural scope, we preserve factual invariance and ensure contextual integrity. Addressing the lack of granular public datasets, we conduct experiments on **FinTierQA Gold**, a curated expert benchmark. Results show that Fin-STAR outperforms state-of-the-art hierarchical and graph-based baselines across diverse query complexities, document types, and markets. Ablations confirm that our semantic injection consistently outperforms alternative strategies. Finally, we release **FinTierQA**, comprising 3.9M pairs automatically constructed from 78k documents via our framework.

## 1 Introduction

To break through the capability ceiling of Large Language Models (LLMs), research must transition from simplified synthetic settings to noisy, high-stakes real-world domains. Financial document understanding represents a quintessential challenge in this frontier, imposing strict demands for long-context reasoning and multi-hop inference where precision is non-negotiable (Reddy et al., 2024).

Unlike standard RAG benchmarks (e.g., scientific papers (Sarthi et al., 2024), technical reports (Guinet et al., 2024)), financial documents such as annual reports, prospectuses, and ESG reports are characterized by *systemic semantic implicitness* (Loughran and McDonald, 2011; Li, 2008; Zhu et al., 2021). Here, driven by tight document coupling, key facts are rarely self-contained or explicit in surface text; instead, they are contingent upon the overarching structural hierarchy (e.g., headers) that is often physically detached from the local snippet. Therefore, expert analysis in finance extends beyond simple retrieval. Experts must routinely explicate implicit clues by anchoring isolated segments within their macro-structural framework, thereby reconstructing the complete cognitive evidence chain (as shown in Figure 1). This creates a fundamental tension: the need for rigorous factual accuracy versus the reality of structurally dispersed and implicitly defined evidence.

Current RAG systems, however, lack the mechanism to render these implicit structural cues explicit. Vanilla RAG's mechanical chunking severs snippets from their hierarchy, stripping away the structural lineage (Izacard and Grave, 2021; Karpukhin et al., 2020). While hierarchical approaches attempt to restore structure via summarization, they trade precision for abstraction, blurring the rigid constraints essential for finance (Sarthi et al., 2024; Zhang et al., 2025a). Similarly, Graph-based RAG overlooks the "container" semantics of document hierarchy, failing to distinguish when a fact is implicitly scoped by a "Risk Factors" header versus a "Financial Results" header (Edge et al., 2024; Guo et al., 2024).

This structural blindness extends to evaluation. By limiting inputs to short, self-contained segments, existing financial QA benchmarks (Chen et al., 2021; Zhu et al., 2021) strip away deep hi-

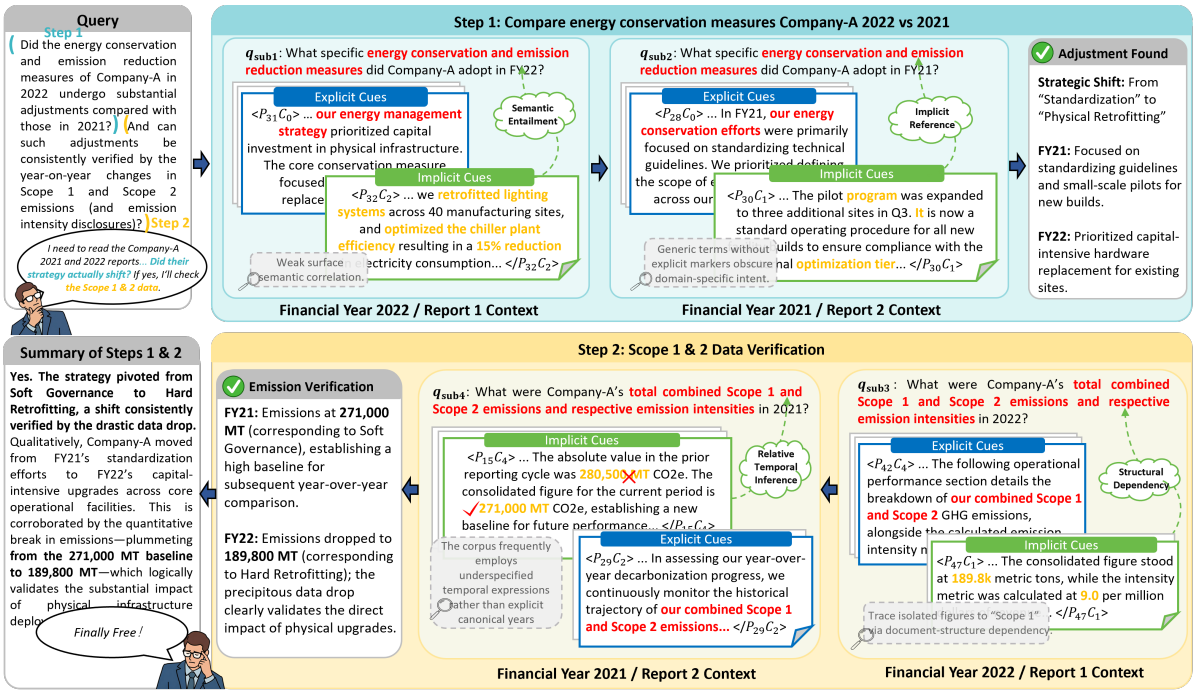



Figure 1: Cognitive challenges in financial document reasoning. Complex queries are decomposed into sub-questions ($q_{subi}$). Red/Yellow text denotes explicit/implicit clues; $< P_xC_y >$ marks the $y$-th content block on page $x$. Key challenges: (1) Semantic Entailment: Bridging macro-level queries and micro-level descriptions; (2) Implicit Reference: Disambiguating generic terms (e.g., "program") via structural paths; (3) Structural Dependency: Grounding isolated figures (e.g., "189.8") in hierarchical headers; and (4) Relative Temporal Inference: Resolving relative time expressions via metadata.

erarchy, artificially simplifying the task and masking the inability to explicate implicit cues. Compounded by data paucity and narrow coverage, they fail to benchmark real-world financial complexity.

Therefore, we propose **Fin-STAR**, which operationalizes document hierarchy as intrinsic semantic knowledge. At its core is the *Structure-Enriched Semantic Indexing* mechanism, which augments the structural lineage with snippet-derived virtual nodes and injects this enriched context via a semantic cross-attention paradigm. This process renders implicit cues explicit, ensuring that localized text carries the full weight of its global context. Crucially, we explicitly anchor facts within their native structural containers, safeguarding factual invariance against semantic noise. Finally, we release **FinTierQA Gold**, a rigorously adjudicated benchmark refined through expert consensus to preserve the deep structural dependencies required to evaluate a model's ability to explicate implicit cues in complex financial contexts.

Experiments on FinTierQA Gold across 15 diverse regimes (3 markets $\times$ 5 query complexity tiers) demonstrate that Fin-STAR consistently establishes robust dominance over state-of-the-art baselines (e.g., GraphRAG, LightRAG, RAPTOR). Crucially, our analysis confirms that these gains stem from rigorous evidence alignment and verifiable numerical precision rather than mere surface-level textual fluency, establishing a new baseline for reliable, cost-efficient robustness in complex multi-hop financial reasoning.

To bridge the gap between academic and real-world complexity, we leverage Fin-STAR to automatically construct **FinTierQA**, a massive-scale dataset simulating analyst reasoning on over 78k filings across the CN, HK, and US markets (2015–2024). Unlike existing corpora, FinTierQA features a five-tier difficulty structure comprising nearly 4 million samples. By unifying diverse regulatory frameworks with granular annotations, it establishes a new standard for evaluating robust reasoning over structurally implicit semantics in high-stakes financial contexts.

In summary, our main contributions are as follows: **(1)** We propose Fin-STAR, a framework that resolves systemic semantic implicitness by operationalizing document hierarchy as intrinsic semantic knowledge, ensuring precise explication of implicit cues with factual integrity. **(2)** We validate

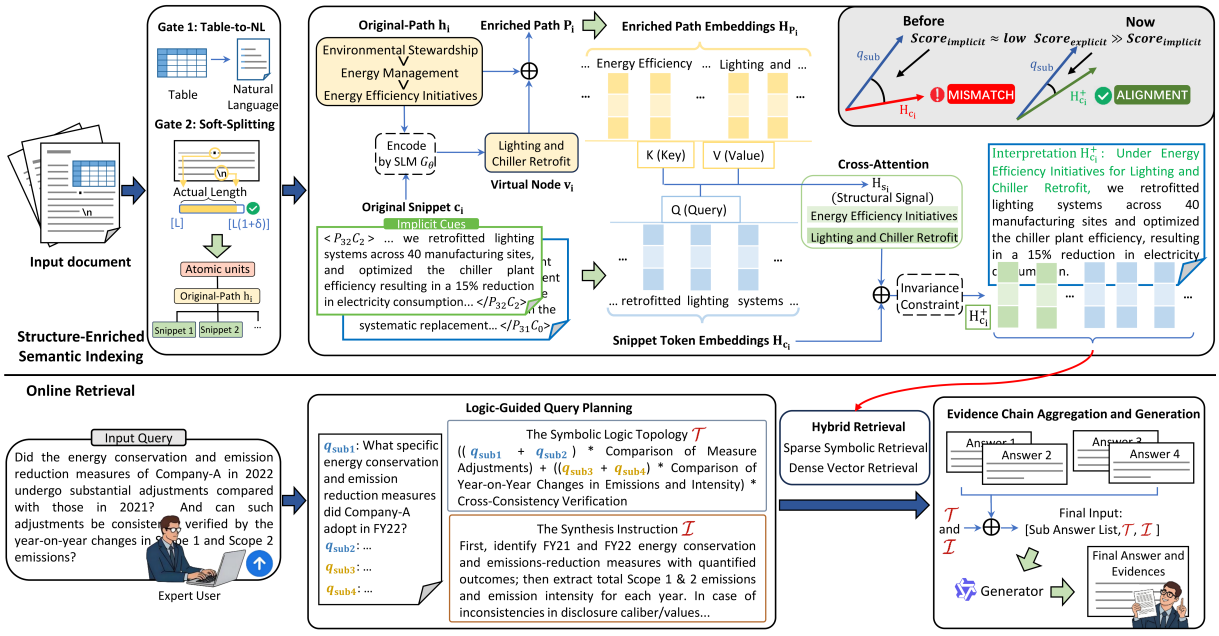



Figure 2: Overview of the Fin-STAR.

Fin-STAR across diverse markets, document types, and query complexities, demonstrating that our structure-injection strategy is a prerequisite for verifiable numerical precision in high-stakes reasoning. **(3)** We release FinTierQA Gold and FinTierQA, establishing a rigorous benchmark for evaluating structure-dependent implicit reasoning at expert-level depth and industrial scale.

## 2 Related Work

Research faces a dual barrier: RAG architectures struggle with systemic semantic implicitness, while benchmarks strip away deep hierarchy, artificially simplifying tasks and masking the inability to explicate implicit cues.

**RAG in Finance.** While RAG effectively handles explicit queries in open domains, it struggles with the systemic semantic implicitness inherent in financial documentation. First, standard retrieval methods mechanically sever snippets from their hierarchy, stripping away the defining structural lineage and rendering context-dependent facts (e.g., isolated numbers) ambiguous and prone to misinterpretation (Izacard and Grave, 2021). Second, hierarchical approaches attempt to restore context via summarization, but fundamentally trade precision for abstraction. This process inevitably blurs the rigid factual constraints essential for high-stakes decision-making, introducing semantic noise rather than preserving factual invariance (Sarthi et al., 2024; Zhang et al., 2025a). Finally, Graph-based RAG prioritizes semantic connectivity over structural containment, leading to topological dilution. By failing to enforce hierarchical scoping as a hard constraint, these systems allow retrieval paths to drift across conflicting boundaries (e.g., merging evidence from "Risk Factors" and "Financial Results"), compromising logical consistency (Edge et al., 2024; Guo et al., 2024).

**Financial QA Benchmarks.** Existing benchmarks often fail to capture the structural depth essential for real-world financial reasoning. First, seminal datasets like FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) rely on cropped segments. By stripping away deep hierarchical dependencies, they reduce complex global reasoning to simplified local reading comprehension, masking the inability to explicate implicit cues. Conversely, while recent works like FINANCEBENCH (Islam et al., 2023), DocFinQA (Reddy et al., 2024), and SEC-QA (Lai et al., 2025) expand context windows, they prioritize retrieval scale over hierarchical granularity. By failing to validate rigorous structural anchoring, these benchmarks serve as insufficient proxies for evaluating the resolution of systemic semantic implicitness within complex financial hierarchies.

## 3 Methodology

We propose Fin-STAR (Figure 2), a framework comprising two phases: *(1) Structure-Enriched Semantic Indexing*, which resolves implicitness by leveraging a semantic cross-attention paradigm to anchor snippets to their structural lineage; and *(2) Online Retrieval*, which orchestrates reasoning through logic-guided planning, hybrid evidence gathering, and verifiable answer synthesis.

### 3.1 Structure-Enriched Semantic Indexing

Fin-STAR constructs a snippet-specific structural path by combining a lightweight semantic abstraction of each snippet with its hierarchical context, and leverages a semantic cross-attention paradigm to inject this path. By linearizing the structural hierarchy into natural language, Fin-STAR forces the model to synthesize specific structural cues that are implicit when the snippet is considered in isolation, anchoring the snippet to its global context without altering its local semantics.

**Dual-Gated Semantic Unit Construction.** In financial documents, the interleaving of narrative text and tables often fragments coherent semantics. To ensure that each snippet remains semantically complete and modality-consistent, Fin-STAR employs a dual-gated normalization mechanism that constructs semantically self-contained snippets. The first gate encapsulates each table as an indivisible unit and linearizes it into a natural language representation that preserves entity–attribute–value relations, ensuring semantic compatibility with surrounding text. The second gate performs boundary-aware soft splitting, dynamically adjusting snippet boundaries to align with natural semantic boundaries and to fully include encapsulated tables.

**Explicitization of Implicit Semantics.** To surface latent meanings, Fin-STAR first constructs structural paths and then facilitates semantic injection during indexing, progressively integrating structural cues into snippet representations.

***Structural Path Construction.*** Given a normalized snippet $c_i$ and its hierarchical path $h_i$, Fin-STAR employs a Small Language Model (SLM) generator $G_\theta$ to synthesize a discriminative virtual node $v_i = G_\theta(c_i, h_i)$, yielding the enriched path $P_i = h_i \oplus v_i$. To ensure robustness, path construction is constrained by two principles: (1) *discriminativeness*, requiring $v_i$ to capture essential semantics absent from $h_i$, and (2) *depth control*, enforcing $\mathrm{Depth}(P_i) \leq 5$.

***Semantic Injection via Cross-Attention.*** Fin-STAR operationalizes semantic injection via a query-driven extraction mechanism. Let $\mathbf{H}_{c_i}$ and $\mathbf{H}_{P_i}$ denote the dense representations of the snippet and the enriched path, respectively. To bridge the semantic gap without modifying the underlying architecture, we conceptualize the snippet $\mathbf{H}_{c_i}$ as the Query ($\mathbf{Q}$) and the path $\mathbf{H}_{P_i}$ as the Key ($\mathbf{K}$) and Value ($\mathbf{V}$). Rather than treating hierarchy as a static label, this interaction forces the model to dynamically weigh and aggregate only the structural cues relevant to the specific ambiguity of the snippet, synthesizing a tailored structural context vector $\mathbf{H}_{s_i}$:

$$\mathbf{H}_{s_i} = \mathrm{Attn}(\mathbf{Q} = \mathbf{H}_{c_i}, \mathbf{K} = \mathbf{V} = \mathbf{H}_{P_i}). \quad (1)$$

The final structure-injected representation is then constructed via concatenation:

$$\mathbf{H}^+_{c_i} = \mathbf{H}_{s_i} \oplus \mathbf{H}_{c_i}. \quad (2)$$

This architecture enforces a principle of ***factual invariance*** through a structural orthogonality constraint. Unlike summarization or additive injection which risks mixing semantics and blurring numerical precision, concatenation isolates the original snippet $\mathbf{H}_{c_i}$ in an independent subspace. This ensures that explicit entities and critical numerals remain mathematically unadulterated by structural priors, allowing the model to attend to hierarchy without corrupting the facts. The structural signal $\mathbf{H}_{s_i}$ acts exclusively as a contextual augmentation, resolving implicitness without overwriting high-precision evidence.

Formally, let $\mathcal{E}(\cdot)$ denote a factual extraction function that identifies key information elements (e.g., named entities and numerical data). Our design guarantees that this core set remains strictly recoverable ($\mathcal{E}(c_i) \subseteq \mathcal{E}(\mathbf{H}^+_{c_i})$), preserving the factual integrity required for financial analysis.

### 3.2 Online Retrieval

**Logic-Guided Query Planning.** Complex analytical queries within the financial domain often involve multiple constraints (e.g., specific entities, time windows, and logical operations) that necessitate structured decomposition. Fin-STAR addresses this by mapping the user intent to an explicit execution plan composed of atomic subqueries. The query processor parses the composite intent $\mathcal{Q}$ into a sequence of atomic units $\{q_{\text{sub}}\}$. To

Table 1: Summary of difficulty levels.

| Level | Type | Scope | Logic | Count (%) |
|---|---|---|---|---|
| $T_1$ | Atomic | Local Span | Direct Retrieval | 4,050 (30%) |
| $T_2$ | Intra-topic | Intra-H1 | Chain ($\geq$2 hops) | 3,240 (24%) |
| $T_3$ | Inter-topic | Cross-H1 | Chain ($\geq$3 hops) | 2,700 (20%) |
| $T_4$ | Global | Full Doc | Aggregation | 2,160 (16%) |
| $T_5$ | Multi-doc | Cross-Doc | Align & Compare | 1,350 (10%) |

handle multi-hop dependencies, we apply semantic completion and coreference resolution, ensuring that each $q_{\text{sub}}$ explicitly specifies its target entity. The refined sub-query is then encoded into a dense vector representation $\mathbf{q}_{\text{sub}} \in \mathbb{R}^d$ for retrieval. Following Yao et al. (2022), Fin-STAR further builds a *Symbolic Logic Topology* $\mathcal{T}$ that specifies how sub-query results are combined (e.g., $\cap$, $\setminus$, aggregation) and in what order, yielding a concrete evidence-composition pathway. In parallel, a *Synthesis Instruction* $\mathcal{I}$ is produced to guide the downstream generation module, describing how retrieved facts should be organized and composed into a coherent explanation.

**Hybrid Retrieval.** To ensure the precision of micro-facts, we execute independent retrieval for each atomic sub-query $q_{\text{sub}}$, targeting the Top-$K_{\text{sub}}$ evidence fragments. Specifically, we adopt a dual-stream hybrid strategy that integrates semantic matching (between $\mathbf{H}_{c_i}^{+}$ and $\mathbf{q}_{\text{sub}}$) with sparse symbolic constraints (BM25). These signals are fused to safeguard exact lexical recall while preserving semantic coverage.

**Evidence Chain Aggregation and Generation.** Fin-STAR employs a two-stage synthesis process to guarantee that responses are structurally coherent and verifiably grounded:

***Atomic Fact Verification.*** For each atomic query $q_{\text{sub}}$, the model executes isolated verification against its specific Top-$K_{\text{sub}}$ evidence. This decoupled processing yields verified atomic sub-answers, minimizing noise interference across distinct intents and ensuring local factual precision.

***Logic-Controlled Synthesis.*** The final response is constructed via constrained aggregation. Guided by the *Symbolic Logic Topology* $\mathcal{T}$ and *Synthesis Instruction* $\mathcal{I}$, the generator synthesizes these verified sub-answers into a cohesive narrative, enforcing global logical consistency while maintaining strict faithfulness to the evidence chain.

# 4 Experimental Setup

## 4.1 FinTierQA Gold

We present FinTierQA Gold, an expert-adjudicated benchmark curated through a rigorous, difficulty-stratified annotation pipeline.

**Corpus Preparation.** We compiled a corpus of 270 long-form documents spanning 17 industries and nine categories, covering China, Hong Kong, and U.S. markets with a mix of Annual, IPO, and ESG reports. This collection totals approximately 60K pages of source material. Leveraging Chen et al. (2025), we parsed the raw documents into hierarchical JSON blocks aligned with section headers (H1–H4, where H1 denotes the top level).

**Difficulty-Aware Annotation Pipeline.** Guided by inquiry principles distilled from authentic financial analyst workflows, we employed a two-stage AI-Human pipeline to generate high-fidelity QA samples across five complexity tiers ($T_1$–$T_5$), as detailed in Table 1. *Stage 1: Multi-Model Synthesis.* We utilized an ensemble of three SOTA reasoning models (Gemini 3 Pro, Grok 4.1-Thinking, GPT 5.2-Thinking) (Chiang et al., 2024) to generate standardized JSON tuples (`question`, `answer`, `reasoning`, `position_index`) from identical document contexts, strictly following the prompting protocol illustrated in Appendix A. *Stage 2: Consensus Verification.* Eighteen domain experts independently validated each instance, strictly assessing both the factual correctness of the answer and the precise alignment of the `position_index` with supporting evidence. Instances achieving unanimous consensus were automatically retained, while discordant cases necessitated expert adjudication, followed by a senior expert meta-review. This rigorous cross-validation procedure achieved a Fleiss' Kappa of 0.84 (Fleiss, 1971) and a strict retention rate of 74.3%, yielding a robust gold standard.

## 4.2 Baselines

We benchmark Fin-STAR against six structured RAG systems across three paradigms:

**Global Graph-Index RAG.** Targeting macro-narratives, *GraphRAG* (Edge et al., 2024) (Global Search) synthesizes hierarchical summaries via community detection, while *LightRAG* (Guo et al., 2024) unifies entity-specific and concept-abstract retrieval to overcome single-granularity limits.

**Fine-Grained KG Reasoning RAG.** Focusing on path navigation, *HippoRAG2* (Gutiérrez et al., 2025) employs Personalized PageRank on associative graphs to dynamically activate multi-hop paths, and *ArchRAG* (Wang et al., 2025) guides retrieval by parsing the document's architectural skeleton.

**Hierarchical Tree RAG.** Using recursive structures, *RAPTOR* (Sarthi et al., 2024) constructs recursive cluster trees for summarization-augmented retrieval, whereas *PageIndex* (Zhang et al., 2025a) employs a vector-less, Table-of-Contents-based mechanism for routing-augmented retrieval.

### 4.3 Metrics

We evaluate model performance across four dimensions: textual overlap, retrieval precision, logical consistency, and numerical rigor.

**Answer F1.** Following Rajpurkar et al. (2016), we compute the Answer F1 score ($F1_a$) based on the token overlap between the normalized gold reference ($\mathcal{T}_G$) and prediction ($\mathcal{T}_P$):

$$F1_a = \frac{2 \cdot |\mathcal{T}_G \cap \mathcal{T}_P|}{|\mathcal{T}_G| + |\mathcal{T}_P|} \tag{3}$$

**Evidence & Joint F1.** For high-order reasoning ($T_2$–$T_5$), Answer F1 is prone to rewarding "lucky guesses" (spurious correctness). To enforce consistency, following Yang et al. (2018), we compute Evidence F1 ($F1_e$) based on the sentence-level overlap between gold ($\mathcal{E}_G$) and retrieved ($\mathcal{E}_P$) sets, and derive Joint F1 ($F1_j$):

$$F1_e = \frac{2|\mathcal{E}_G \cap \mathcal{E}_P|}{|\mathcal{E}_G| + |\mathcal{E}_P|}, \quad F1_j = \frac{2F1_aF1_e}{F1_a + F1_e} \tag{4}$$

**Numeric Accuracy.** Addressing format heterogeneity (e.g., "25%" vs. "0.25"), we follow Chen et al. (2021) in verifying the coverage of $V_G$ by $V_P$. Given absolute ($\epsilon_{atol}$) and relative ($\epsilon_{rtol}$) tolerances, we define the matching criterion $\mathcal{M} \iff |v_g - v_p| \leq \epsilon_{atol} + \epsilon_{rtol}|v_p|$ to calculate Numeric Accuracy ($Acc_n$):

$$Acc_n = \mathbb{I}\left(\forall v_g \in V_G, \exists v_p \in V_P : \mathcal{M}\right) \tag{5}$$

### 4.4 Implementation Details

For all methods, we employ `Qwen3-Embedding-8B` as the embedding model to compute dense vector representations for indexing and retrieval, and `Qwen3-32B` as the answer generator for reasoning and final response generation (Zhang et al., 2025b).

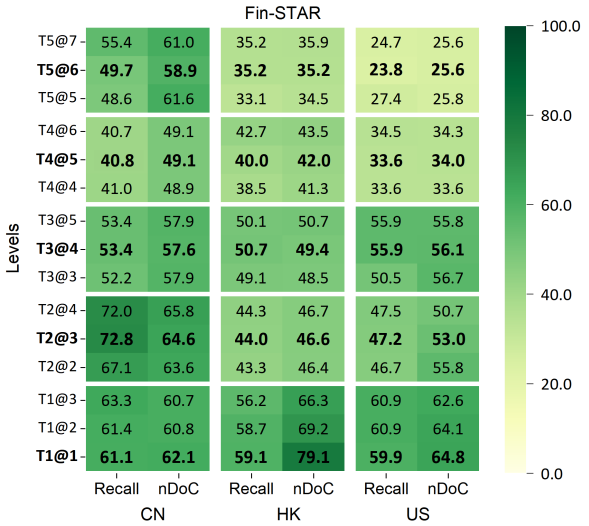



Figure 3: Retrieval sensitivity (%) of Fin-STAR ($T_x@k$: Tier $x$ at depth $k$). Performance is evaluated using two metrics: Recall@K, which measures retrieval completeness by capturing the proportion of gold evidence, and nDoC@K, which assesses ranking quality by prioritizing evidence appearing earlier in the list.

For both the main experiments and all ablation studies, the small language model (SLM) $G_\theta$ component consistently uses Qwen3-8B (Zhang et al., 2025b). For cross-market generalization experiments, we only change the output language instruction in the prompt (to match the target market language), while keeping all other configurations unchanged. All experiments are conducted on 4 NVIDIA Tesla V100 (32GB) GPUs.

### 4.5 Retrieval Sensitivity & Adaptive Protocol

Benchmarking multi-tier reasoning necessitates a calibrated retrieval depth $k$ (i.e., the count of top-ranked snippets), as static thresholds introduce bias due to varying evidence demands.

**Sensitivity Analysis.** We analyzed performance across $k \in \{1, \ldots, 10\}$ for all baselines (see Figure 3 for Fin-STAR; others in Appendix B). Results reveal distinct structural heterogeneity: atomic queries ($T_1$) favor precise retrieval ($k = 1$), whereas high-level aggregation ($T_5$) requires broader scope ($k = 6$) to encompass dispersed financial indicators.

**Unified Tier-Adaptive Protocol.** To mitigate selection bias, we derive a unified Tier-Adaptive Protocol based on observed stability regions, fixing $k$ to $\{1, 3, 4, 5, 6\}$ for $T_1$–$T_5$, respectively (highlighted in Figure 3). This consensus setting ensures that metrics reflect intrinsic reasoning capa-

Table 2: Main experimental results across three markets (%). Bold indicates the best performance in each column.

| Method | $T_1$@1 | | | | $T_2$@3 | | | | $T_3$@4 | | | | $T_4$@5 | | | | $T_5$@6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ |
| **Panel A: United States (US) Market** | | | | | | | | | | | | | | | | | | | | |
| GraphRAG | 63.6 | 71.6 | 46.3 | 53.6 | **45.3** | 55.6 | 26.3 | 33.3 | **43.0** | 62.5 | 33.9 | 37.9 | 27.8 | 34.0 | 20.0 | 23.3 | 22.9 | 18.8 | 10.3 | 14.2 |
| LightRAG | **72.1** | 76.0 | 54.9 | **62.3** | 27.6 | 57.4 | 33.5 | 30.2 | 26.5 | 60.7 | 41.3 | 32.3 | 19.1 | 40.0 | 21.6 | 20.3 | 16.4 | 25.0 | 14.1 | 15.1 |
| ArchRAG | 56.6 | 59.6 | 44.2 | 49.6 | 32.1 | 47.2 | 25.6 | 28.5 | 34.3 | 51.8 | 34.6 | 34.4 | 21.7 | 31.0 | 18.6 | 20.0 | 24.5 | 18.8 | 17.3 | 20.3 |
| HippoRAG2 | 26.2 | 35.6 | 27.7 | 26.9 | 28.1 | 38.9 | 17.2 | 21.4 | 39.4 | 47.3 | 27.8 | 32.6 | **28.5** | 25.0 | 10.1 | 14.9 | 22.5 | 12.5 | 4.0 | 6.8 |
| RAPTOR | 50.0 | 50.0 | 58.3 | 53.9 | 25.8 | 43.8 | 31.8 | 28.4 | 34.4 | 53.1 | 30.1 | 32.1 | 27.6 | 45.0 | 27.6 | 27.6 | 13.8 | 12.5 | **21.8** | 16.9 |
| PageIndex | 49.9 | 61.1 | 52.3 | 51.1 | 27.6 | 57.4 | 33.5 | 30.2 | 25.3 | 56.3 | 37.6 | 30.3 | 15.5 | 31.3 | 24.5 | 19.0 | **37.7** | 15.2 | 15.2 | **21.6** |
| **Fin-STAR** | 66.7 | **76.4** | **55.9** | 60.9 | 40.1 | **61.1** | **43.5** | **41.7** | 38.8 | **68.8** | **43.4** | **41.0** | 27.6 | **49.0** | **30.6** | **29.0** | 22.7 | **37.5** | 18.4 | 20.3 |
| **Panel B: China (CN) Market** | | | | | | | | | | | | | | | | | | | | |
| GraphRAG | 55.5 | 63.3 | 44.0 | 49.1 | 48.9 | 21.3 | 35.3 | 41.0 | **52.6** | 51.6 | 30.9 | 38.9 | **46.0** | 42.5 | 25.5 | 32.8 | **45.5** | 43.8 | 44.5 | 45.0 |
| LightRAG | **69.9** | 67.2 | 44.0 | 54.0 | 45.2 | 31.5 | 51.6 | 46.1 | 30.6 | 41.9 | 34.4 | 32.4 | 26.9 | 56.6 | 28.8 | 27.9 | 16.3 | 25.0 | 21.3 | 18.5 |
| ArchRAG | 43.6 | 52.2 | 28.5 | 34.5 | 37.5 | 9.8 | 39.0 | 38.3 | 37.0 | 41.9 | 22.5 | 28.0 | 30.7 | 54.7 | 29.1 | 29.9 | 38.8 | 37.5 | 39.6 | 39.2 |
| HippoRAG2 | 26.2 | 34.4 | 17.8 | 21.2 | 41.1 | 37.7 | 32.1 | 36.0 | 42.8 | 50.0 | 30.0 | 35.2 | 31.4 | 10.4 | 16.5 | 21.6 | 22.3 | 21.4 | 26.7 | 24.3 |
| RAPTOR | 50.6 | 62.2 | 30.0 | 37.7 | 54.9 | 29.7 | 36.9 | 44.1 | 51.4 | 64.5 | 28.0 | 36.2 | 20.6 | 18.9 | 15.2 | 17.5 | 25.8 | 12.5 | 18.2 | 21.3 |
| PageIndex | 53.3 | 58.4 | 43.2 | 47.7 | 42.8 | 27.9 | 35.8 | 39.0 | 52.3 | 54.9 | 29.7 | 37.9 | 31.5 | 25.9 | 28.7 | 30.0 | 44.9 | 41.0 | 23.5 | 30.9 |
| **Fin-STAR** | 67.9 | **74.4** | **56.5** | **61.7** | **60.2** | **56.6** | **52.3** | **56.0** | 46.8 | **65.3** | **46.8** | **46.8** | 39.2 | **58.5** | **43.0** | **41.0** | **45.5** | **57.1** | **51.4** | **48.3** |
| **Panel C: Hong Kong (HK) Market** | | | | | | | | | | | | | | | | | | | | |
| GraphRAG | 44.2 | 60.5 | 28.2 | 34.4 | **46.3** | 45.3 | 31.9 | 37.8 | 36.7 | 48.2 | 33.5 | 35.1 | **45.5** | 36.7 | 25.1 | **32.4** | 44.5 | 23.8 | 18.9 | 26.5 |
| LightRAG | **65.7** | 60.7 | 34.9 | 45.6 | 28.8 | 51.9 | 34.9 | 31.6 | 24.6 | 50.5 | **37.7** | 29.7 | 21.0 | 45.0 | 24.7 | 22.7 | 43.4 | 31.3 | 24.2 | 31.1 |
| ArchRAG | 50.0 | 60.5 | 28.5 | 36.3 | 39.1 | 55.7 | 36.1 | 37.5 | 30.7 | 40.6 | 24.2 | 27.1 | 30.2 | 41.7 | 22.6 | 25.9 | 43.4 | 31.3 | 24.2 | 31.1 |
| HippoRAG2 | 17.5 | 32.6 | 20.1 | 18.7 | 28.4 | 34.9 | 23.7 | 25.8 | 36.2 | 42.0 | 25.9 | 30.2 | 27.3 | 23.3 | 19.2 | 22.5 | 26.0 | 6.3 | 18.9 | 21.9 |
| RAPTOR | 17.4 | 35.7 | 24.2 | 20.2 | 29.1 | 30.2 | 8.4 | 13.0 | 39.1 | 37.5 | 14.5 | 21.2 | 23.7 | 34.1 | 16.8 | 10.6 | 6.1 | 12.5 | 12.7 | 8.3 |
| PageIndex | 64.8 | 53.6 | 36.5 | 46.6 | 28.9 | 56.3 | 27.3 | 28.1 | 22.4 | 31.3 | 26.0 | 24.1 | 28.3 | 25.0 | **32.3** | 30.2 | 38.4 | 32.5 | 20.7 | 26.9 |
| **Fin-STAR** | 65.0 | **76.7** | **56.5** | **60.4** | 44.8 | **56.6** | **40.5** | **42.6** | **53.5** | **56.3** | 36.1 | **43.1** | 30.4 | **48.3** | 31.2 | 30.8 | **51.4** | **43.8** | **32.6** | **39.9** |

bilities rather than model-specific hyperparameter artifacts.

## 5 Experimental Results & Analysis

We evaluate Fin-STAR across 15 regimes (3 markets $\times$ 5 difficulty tiers). The main experimental results are summarized in Table 2.

**Holistic Performance across Markets and Tiers.** Fin-STAR establishes a robust dominance, achieving the highest Joint F1 score in 12 out of 15 scenarios. While methods like LightRAG and GraphRAG achieve marginal leads in isolated scenarios (e.g., US-$T_1$, HK-$T_4$) with negligible gaps ($\Delta < 0.016$), Fin-STAR maintains consistent superiority across diverse regulatory frameworks and escalating difficulty levels. Crucially, this effectiveness is underpinned by a superior performance-efficiency balance (see Appendix C). This establishes a baseline of uniform reliability and cost-efficiency essential for high-stakes finance, prioritizing scalable stability over occasional, narrow-domain peaks.

**Source of Gain: Verification over Fluency.** A granular decomposition of metrics reveals a distinct "asymmetric advantage." Fin-STAR's superiority is not driven by surface-level textual similarity—it leads Answer F1 in only 4/15 cases and Semantic Answer Similarity (SAS, see Appendix C) in just 2/15. Conversely, it dominates in verification-centric metrics, securing best-in-class performance for Numeric Accuracy (14/15) and Evidence F1 (11/15). This divergence confirms that the improvements in Joint F1 stem from the rigorous alignment of evidence chains rather than mere stylistic fluency. By enforcing structural constraints to mitigate scope drift" and numerical misattribution, Fin-STAR prioritizes the precise, verifiable grounding essential for financial integrity.

**Tier-wise Sensitivity and Multi-hop Reasoning.** Performance breakdown by difficulty tier validates our design motivation. We calculate gains by comparing the cross-market average of Fin-STAR against the strongest tier-specific baseline. The

Table 3: Ablation on the US market across difficulty tiers (%). Bold denotes the best score in each column. "w/o Virt. Node" removes virtual-scope generation and uses only the original hierarchical paths; "w/o Sem. Inj." replaces our cross-attention semantic injection with simple prefix-style path concatenation; "w/o Invariance" drops the factual invariance constraint during injection, allowing factual elements (e.g., entities/numbers/time) to be unconstrained.

| Config. | $T_1$@1 | | | | $T_2$@3 | | | | $T_3$@4 | | | | $T_4$@5 | | | | $T_5$@6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ |
| w/o Virt. Node | 54.5 | 49.5 | 39.9 | 46.1 | 33.7 | 58.0 | 33.4 | 33.5 | 32.9 | 39.7 | 24.2 | 27.9 | 21.3 | 40.3 | 22.8 | 22.0 | 15.9 | 16.7 | 4.2 | 6.6 |
| w/o Sem. Inj. | 62.8 | 68.7 | 45.0 | 52.4 | 36.3 | 46.8 | 37.5 | 36.9 | 34.7 | 56.3 | 39.3 | 36.9 | 25.0 | 40.7 | 18.5 | 21.3 | 19.7 | 33.3 | 6.3 | 9.5 |
| w/o Invariance | 57.8 | 61.0 | 45.3 | 50.8 | 31.8 | 54.0 | 31.1 | 31.5 | 37.3 | 56.3 | 35.9 | 36.6 | 24.4 | 42.9 | 18.9 | 21.3 | 19.0 | 19.9 | 13.2 | 15.6 |
| **Fin-STAR** | **66.7** | **76.4** | **55.9** | **60.9** | **40.1** | **61.1** | **43.5** | **41.7** | **38.8** | **68.8** | **43.4** | **41.0** | **27.6** | **49.0** | **30.6** | **29.0** | **22.7** | **37.5** | **18.4** | **20.3** |

results reveal a dynamic advantage profile: starting with solid gains in simple tasks ($T_1$, +13.0%), performance peaks at $T_2$ (+25.2%) where explicit structural navigation is most critical. The advantage narrows slightly at $T_4$ (+14.0%) as global aggregation becomes dominant, but notably rebounds at $T_5$ (+19.8%). This trajectory confirms that explicating implicit structural cues confers a resilient edge, enabling Fin-STAR to outperform specialized architectural baselines even in high-entropy, complex reasoning scenarios.

## 6 Ablation Studies

We focus our ablation study on the US market (Table 3), while results for the CN and HK markets (which exhibit consistent trends) are detailed in Appendix D. A critical pattern emerges: removing structural components triggers a catastrophic drop in *Evidence F1* (e.g., $-77\%$ in US-$T_5$ for w/o Virt. Node) whereas *Answer F1* declines more moderately. This "asymmetric degradation" attributes Fin-STAR's gains to verifiable numerical binding rather than superficial answer fluency.

**Component Validity.** A detailed breakdown validates our design choices: (1) *Virtual Nodes as Anchors*: The collapse of w/o Virt. Node in complex tiers identifies virtual nodes as essential anchors that prevent "scope drift" by grounding implicit constraints. (2) *Semantic Injection*: The superiority over w/o Sem. Inj. underscores that naïve integration of raw paths fails to resolve implicitness; our cross-attention paradigm is essential to dynamically extract structural alignment, effectively filtering noise before integration. (3) *Invariance as Prerequisite*: Crucially, the drop in w/o Invariance reveals that unconstrained injection acts as adversarial noise. This confirms that gains derive from the rigorous factual correctness of structural constraints, not merely the injection mechanism itself.

## 7 FinTierQA

To bridge the reality gap, we leverage Fin-STAR to construct FinTierQA by simulating analyst reasoning on real-world filings, establishing a new standard for scope and difficulty.

In terms of breadth, FinTierQA aggregates 78,564 Annual, IPO, and ESG filings from 8,657 companies across China, Hong Kong, and the US (2015–2024), covering diverse regulatory and linguistic frameworks. For depth, the dataset comprises 3,928,500 samples, systematically balanced across five difficulty tiers through predefined sampling ratios ($T_1$: 1,178,550, $T_2$: 942,840, $T_3$: 785,700, $T_4$: 628,560, $T_5$: 392,850), where each instance is programmatically annotated with the `question`, `answer`, `evidence`, and `tier label`.

This dataset propels dual-domain advancements. In computer science, its trilateral market coverage and tiered complexity facilitate cross-lingual alignment and robust reasoning over structurally implicit semantics. In finance, by scaling systematic cross-market analysis, it promotes a paradigm shift from manual information retrieval to automated, reasoning-centric analytical workflows.

## 8 Conclusion

We introduce Fin-STAR to resolve *systemic semantic implicitness* by transforming document hierarchy from a physical skeleton into intrinsic semantic knowledge. Through virtual node injection and cross-attention, our framework enforces rigorous evidence anchoring, a capability validated by its superiority over SOTA baselines on FinTierQA Gold. These findings establish explicit structural injection as a prerequisite for precise numerical binding. To further bridge the academic-industrial gap, we release FinTierQA (3.9M pairs), a massive-scale resource catalyzing the shift toward robust, structure-aware financial intelligence.

## Limitations

While Fin-STAR establishes robust dominance, achieving the highest Joint F1 score in 12 out of 15 scenarios, baseline methods like LightRAG and GraphRAG achieve marginal leads in isolated scenarios (e.g., US-$T_1$, HK-$T_4$). This reflects our experimental isolation of indexing-stage semantic injection: we employed a standard hybrid retrieval backbone without the specialized runtime routing or global navigation used by GraphRAG or ArchRAG. Consequently, Fin-STAR faces performance ceilings in scenarios heavily dependent on such runtime heuristics. However, this limitation validates our core hypothesis: Fin-STAR's consistent superiority in Numeric Accuracy and Evidence F1 proves that indexing-level injection independently secures factual precision. This establishes a high-fidelity foundation, offering a clear path to further elevate performance by integrating task-specific routing mechanisms in future work.

## Ethics Statement

**Data Sourcing and Privacy.** The dataset constructed in this work, FinTierQA, is derived exclusively from publicly available financial disclosures (e.g., Annual Reports, IPO prospectuses, ESG reports) filed with regulatory bodies in the United States (SEC), Hong Kong (HKEX), and China (SSE/SZSE). These documents are in the public domain and intended for investor scrutiny. We have strictly adhered to the terms of use of the respective data repositories. Given the corporate nature of these documents, they contain minimal personally identifiable information (PII) regarding private individuals; any incidental PII (e.g., names of corporate officers) appears in their public professional capacity.

**Annotator Recruitment & Compensation.** Our dataset construction involved a human-in-the-loop verification process, specifically during the "Consensus Verification" stage described in Section 4.1. We recruited 18 financial experts through professional industry channels. All participants were explicitly informed about the purpose of the study and the intended usage of the annotated data. To ensure fair labor practices, annotators were compensated at an hourly rate that significantly exceeds the local minimum wage and aligns with the competitive market rate for financial analysts in their respective regions.

**Intended Use and Potential Risks.** Fin-STAR is designed to assist financial professionals in information retrieval and complex reasoning over long documents. However, financial decision-making is a high-stakes domain. While our method improves numerical accuracy and factual grounding compared to baselines, it is not immune to errors or hallucinations inherent in language models. Therefore, this system should be deployed as a "copilot" to augment human analysts, not as a fully autonomous financial advisor. We explicitly warn against using this model for automated trading or investment decisions without rigorous human oversight.

**AI-Assisted Data Synthesis.** We utilized large language models (as detailed in Section 4.1) to assist in the initial generation of QA pairs. We acknowledge that synthetic data may carry biases present in the underlying models. The subsequent human verification stage was implemented specifically to mitigate such biases and ensure logical soundness.

## References


Yan Chen, Yu Zou, Jialei Zeng, Haoran You, Xiaorui Zhou, and Aixi Zhong. 2025. Pharos-esg: A framework for multimodal parsing, contextual narration, and hierarchical labeling of esg report. *arXiv preprint arXiv:2511.16417*.

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*, pages 3697–3711.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated

evaluation of retrieval-augmented language models with task-specific exam generation. *arXiv preprint arXiv:2405.13622*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

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*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Viet Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. 2025. Secqa: A systematic evaluation corpus for financial qa. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 221–236.

Feng Li. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2-3):221–247.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. 2025. Archrag: Attributed community-based hierarchical retrieval-augmented generation. *arXiv preprint arXiv:2502.09891*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Mingtian Zhang, Yu Tang, and PageIndex Team. 2025a. Pageindex: Next-generation vectorless, reasoning-based rag. *PageIndex Blog*. Https://pageindex.ai/blog/pageindex-intro.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

# A FinTierQA Gold Construction Prompt and Supplemental Materials

```
# Role: High-Precision Multi-hop QA Annotator.
# Use ONLY the provided JSON document. No external knowledge.
# If evidence is missing, state "Not found in the document."

Steps:
  1) List all required data points (multi-hop if needed).
  2) Locate supporting snippets and record EVERY Positional_Index (e.g., [28,0]).
  3) Answer concisely and provide step-by-step reasoning grounded in those indices.

Output:
  EXACTLY one JSON object:
    {"question":"<query>","answer":"<fact>","position_index":[[x1,y1],[x2,y2],...],
    "reasoning":"<derivation using only indexed snippets>"}
Inputs:
  Document={{attachment}}, Query1={{query1}}, Query2={{query2}}.
```

Figure 4: The standardized prompt used for the AI expert panel to generate annotations for the FinTierQA Gold benchmark.

We adhere to the principle of full transparency in the creation of the FinTierQA Gold benchmark. Figure 4 presents the exact prompt utilized by our AI expert panel to generate the high-fidelity QA annotations. This prompt design ensures that the resulting dataset maintains strict factual grounding and supports complex multi-hop reasoning.

In addition, the complete RAG implementation code, detailed configurations, demo data, raw RAG outputs, and indexing demos for our ablation settings are provided in the **supplementary materials** (attachments).

# B Sensitivity Analysis of Retrieval Depth

To validate the rationale behind the retrieval settings in our main evaluation, we present the sensitivity heatmaps of retrieval depth ($k$) for all baseline models (excluding Fin-STAR) in Figure 5. We report performance across varying tiers ($T_1$–$T_5$) and markets using two complementary metrics: Recall@K, which measures evidence completeness by capturing the proportion of retrieved gold evidence, and nDoC@K, which assesses ranking quality by prioritizing evidence appearing earlier in the list.

Consistent with the observations in the main text, the results exhibit a pronounced tier-dependent preference for optimal $k$: *(1) Low-Tier Precision:* Simpler queries (e.g., $T_1$) consistently favor smaller $k$ values. This restriction helps maintain a high degree of evidence concentration (nDoC) by filtering out irrelevant context noise. *(2) High-Tier Coverage:* Conversely, complex queries (e.g., $T_4$–$T_5$) necessitate larger retrieval scopes to capture dispersed evidence.

These findings underscore that enforcing a static $k$ introduces comparative bias—either overwhelming simple queries or starving complex ones. Consequently, we derived a unified tier-adaptive configuration (i.e., $k \in \{1, 3, 4, 5, 6\}$ for $T_1$–$T_5$) based on the aggregate stability regions across all baselines. Applying this consensus schedule uniformly ensures that retrieval budgets dynamically align with reasoning complexity, guaranteeing that performance metrics reflect intrinsic algorithmic capabilities unskewed by model-specific hyperparameter artifacts.

# C Efficiency and Supplementary Semantic Analysis

In this section, we expand our evaluation to include computational efficiency (Token Consumption) and supplementary surface-level metrics (Semantic Answer Similarity, SAS), visualized in the dual-axis plots of Figure 6.

**Token Efficiency.** From a practical engineering perspective, Fin-STAR demonstrates superior cost-effectiveness. As observed in the high-tier subplots (e.g., $T_3$–$T_5$), global graph-based methods like LightRAG incur prohibitive computational costs, represented by elevated bars significantly exceeding the average. In contrast, Fin-STAR maintains a lean token footprint comparable to summarization-based baselines (e.g., RAPTOR), validating that utilizing document structure allows for targeted retrieval rather than the brute-force context flooding observed in global graph approaches.

**Source of Gain: Verification over Surface Fluency.** Decomposing the performance metrics reveals a distinct "asymmetric advantage." While Fin-STAR demonstrates competitive stability in the SAS line charts, it leads surface similarity metrics in limited scenarios—topping Answer F1 (in main text) in only 4/15 cases and SAS (Figure 6) in just 2/15. However, it decisively dominates in Numeric Accuracy (14/15) and Evidence F1 (11/15). This pattern suggests that the Joint F1 improvement reported in the main experiments does not stem from generating superficially fluent answers, but from the rigorous alignment of reasoning chains. By injecting structural scope to mitigate "scope drift" and numerical misattribution, Fin-STAR ensures the mathematical precision and verifiable grounding essential for financial integrity, prioritizing correctness over mere semantic similarity.


Figure 5: Tier-Dependent Sensitivity of Retrieval Depth ($k$) across Baseline Models (%).

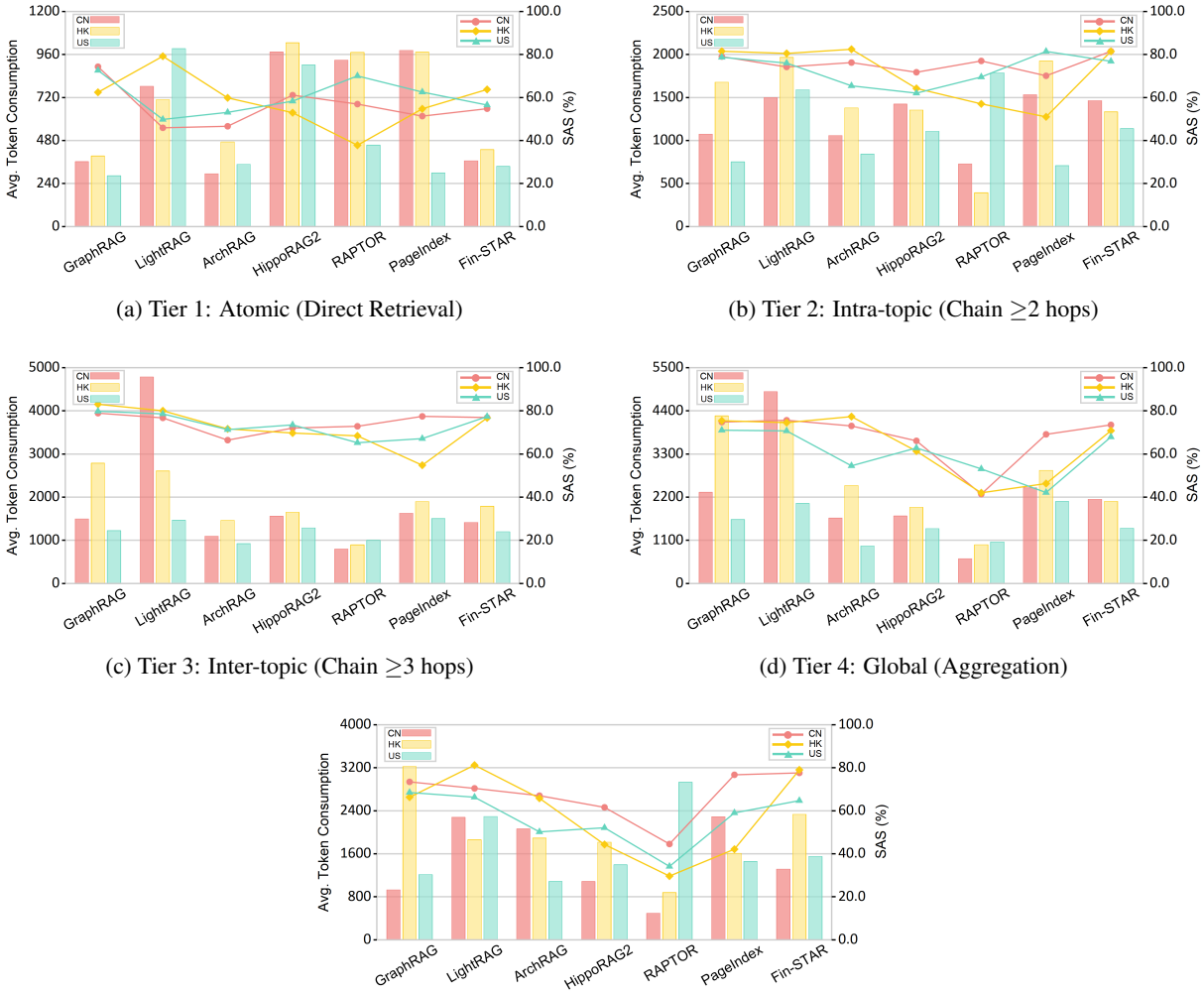
Table 4: Ablation study across difficulty tiers (%). Bold indicates the best performance within each market. "w/o Virt. Node" removes virtual-scope generation and uses only the original hierarchical paths; "w/o Sem. Inj." replaces our cross-attention semantic injection with simple prefix-style path concatenation; "w/o Invariance" drops the factual invariance constraint during injection, allowing factual elements (e.g., entities/numbers/time) to be unconstrained.

| **Config.** | $T_1$@1 | | | | $T_2$@3 | | | | $T_3$@4 | | | | $T_4$@5 | | | | $T_5$@6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ | $F1_a$ | $Acc_n$ | $F1_e$ | $F1_j$ |
| **China (CN) Market** | | | | | | | | | | | | | | | | | | | | |
| w/o Virt. Node | 49.2 | 59.7 | 42.6 | 45.7 | 46.0 | 21.9 | 20.0 | 27.8 | 25.8 | 40.3 | 29.6 | 27.6 | 38.0 | 56.0 | 20.8 | 26.9 | 32.5 | 25.5 | 40.0 | 38.0 |
| w/o Sem. Inj. | 54.9 | 38.7 | 30.1 | 38.9 | 51.7 | 25.0 | 18.4 | 27.1 | 36.9 | 57.0 | 37.9 | 37.4 | 39.4 | 46.0 | 39.4 | 39.4 | 42.6 | 10.9 | 42.2 | 42.4 |
| w/o Invariance | 56.6 | 38.7 | 28.5 | 37.9 | 53.4 | 20.9 | 18.6 | 27.6 | 25.8 | 40.3 | 19.6 | 22.3 | 39.5 | 28.3 | 39.3 | 39.4 | 39.6 | 46.4 | 20.5 | 27.0 |
| **Fin-STAR** | **67.9** | **74.4** | **56.5** | **61.7** | **60.2** | **56.6** | **52.3** | **56.0** | **46.8** | **65.3** | **46.8** | **46.8** | **39.2** | **58.5** | **43.0** | **41.0** | **45.5** | **57.1** | **51.4** | **48.3** |
| **Hong Kong (HK) Market** | | | | | | | | | | | | | | | | | | | | |
| w/o Virt. Node | 57.5 | 66.0 | 45.2 | 50.6 | 41.2 | 35.2 | 15.6 | 22.6 | 51.7 | 49.3 | 27.3 | 35.7 | 24.6 | 19.6 | 25.2 | 24.9 | 46.3 | 25.0 | 25.3 | 32.7 |
| w/o Sem. Inj. | 53.4 | 56.4 | 52.2 | 52.8 | 30.8 | 49.5 | 31.6 | 31.2 | 48.5 | 37.5 | 17.7 | 25.9 | 18.1 | 27.5 | 19.0 | 18.5 | 48.9 | 10.2 | 12.0 | 19.2 |
| w/o Invariance | 35.7 | 41.0 | 35.8 | 35.8 | 41.9 | 22.6 | 17.9 | 25.1 | 48.1 | 50.0 | 19.8 | 28.1 | 21.7 | 23.7 | 18.4 | 19.9 | 46.8 | 6.3 | 22.9 | 30.7 |
| **Fin-STAR** | **65.0** | **76.7** | **56.5** | **60.4** | **44.8** | **56.6** | **40.5** | **42.6** | **53.5** | **56.3** | **36.1** | **43.1** | **30.4** | **48.3** | **31.2** | **30.8** | **51.4** | **43.8** | **32.6** | **39.9** |

## D Extended Ablation Analysis

To verify the robustness of Fin-STAR across diverse linguistic and regulatory environments, we extended the component-wise ablation study to the China (CN) and Hong Kong (HK) datasets. Detailed performance comparisons across varying difficulty tiers are presented in Table 4.

**Ablation Analysis on CN Market.** Experimental results on the China (CN) market consistently mirror the findings observed in the US market, particularly emphasizing the sensitivity of evidence-side metrics. We observe a distinct "asymmetric degradation" where the removal or low-quality injection of structural components primarily leads




(a) Tier 1: Atomic (Direct Retrieval)

(b) Tier 2: Intra-topic (Chain ≥2 hops)

(c) Tier 3: Inter-topic (Chain ≥3 hops)

(d) Tier 4: Global (Aggregation)

(e) Tier 5: Multi-doc (Align & Compare)

Figure 6: Efficiency-Performance Trade-off across Difficulty Tiers ($T_1$–$T_5$). We visualize the Average Token Consumption (Bars, Left Axis) versus Semantic Answer Similarity (SAS, Lines, Right Axis).

to a collapse in $F1_e$ and $F1_j$, while $F1_a$ remains relatively stable. For instance, in CN–$T_4$, skipping the virtual node (Config 1) causes $F1_e$ to plummet from 43.0 to 20.8 (approx. $-51.6\%$), yet $F1_a$ only decreases slightly from 39.2 to 38.0 ($-3.1\%$). Similarly, on CN–$T_2$, replacing semantic injection with simple concatenation (Config 2) reduces $F1_e$ from 52.3 to 18.4 ($-64.9\%$) and $F1_j$ from 56.0 to 27.1 ($-51.6\%$), contrasted with a moderate $14.2\%$ drop in $F1_a$. On the most complex tier CN–$T_5$, utilizing weak paths (Config 3) results in a $60.1\%$ reduction in $F1_e$ (51.4 to 20.5), while $F1_a$ only declines by $12.8\%$. These results confirm that Fin-STAR's core utility lies in transforming implicit structural cues into verifiable evidence paths rather than merely improving surface textual fluency.

**Ablation Analysis on HK Market.** The Hong Kong (HK) market further validates the necessity of explicit structural modeling for maintaining factual integrity in financial RAG. Similar to other regions, removing the virtual node on HK–$T_2$ causes $F1_e$ to drop from 40.5 to 15.6 ($-61.6\%$) and $F1_j$ from 42.6 to 22.6 ($-47.0\%$), whereas $F1_a$ only shows an $8.0\%$ decline. In multi-hop and aggregation scenarios like HK–$T_3$, simple path fusion reduces $F1_e$ from 36.1 to 17.7 ($-51.0\%$) and $F1_j$ from 43.1 to 25.9 ($-39.9\%$). Notably, on the challenging HK–$T_5$ tier, inadequate injection significantly impairs numerical precision, dragging $Acc_n$ from 43.8 down to 10.2 ($-76.7\%$) and reducing $F1_j$ by half (39.9 to 19.2). These findings consistently support that providing structural cues via simple concatenation or insufficient injection leads to severe mismatches between numerals and evidence chains. Collectively, the cross-market results emphasize that making implicit structural cues explicit and injecting them through constrained mechanisms

is vital for the robustness of financial analytical systems.