# PreHypo: Indexing-Time Multi-Hop Dependency Pre-Scoring for Unstructured Document QA

Anonymous ACL submission


## Abstract

RAG over SEC filings is bottlenecked by multi-hop dependency reasoning paid at query time, atop OCR that flattens tables and chunking that splits line items. We present **PreHypo**, which moves this multi-hop work to indexing time, and we show its two contributions are separable. Cross-encoder *pre-scored HOP edges* make the query path a deterministic 1-hop expansion with no per-hop LLM reasoning, delivering ≈7s/query against HopRAG's ≈163s on FinanceBench (a ∼23× latency cut); per-chunk $Q^-$/$Q^+$ *co-indexing* (with topology-preserving VLM OCR and adaptive chunking) drives retrieval quality, reaching LLM-judged correctness $0.32$ (vs. $\leq 0.19$) and page-level grounding $0.19$ (vs. $\leq 0.04$) with a small Qwen3-4B generator ($n$=150). The page-grounding gap is the robust one; the correctness gap is resolved but tight under a query-level paired bootstrap. The design generalizes across domains: on the MultiHop-RAG news benchmark PreHypo reaches correctness $0.82$ vs. $\leq 0.70$, statistically separated from every baseline.


## 1 Introduction

SEC 10-K and 10-Q filings are a natural target for Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), but resist the assumptions of most RAG stacks: a filing mixes dense numerical tables, figures cross-referenced across many pages, and tightly defined domain terminology, none of which survive the standard "flatten and chunk" pipeline cleanly.

Three structural limitations persist in RAG over such corpora: (1) *Table structure loss*: standard OCR flattens tables into plain text and destroys row-column topology, so values arrive at the LLM stripped of their headers. (2) *Context fragmentation*: fixed-length chunking splits documents without regard for semantic boundaries; the income statement and its footnotes end up in different chunks. (3) *Query-time multi-hop reasoning*: which inter-chunk connections to follow is decided at retrieval time (LLM-guided exploration, graph propagation, or per-query re-ranking), paying latency on every query and scoring each link under a tight query-time compute budget.

**PreHypo** pushes graph construction and edge scoring to indexing time; the query path traverses the pre-built graph in a single deterministic 1-hop expansion without runtime LLM reasoning. Four indexing-time designs make this work:

- **Topology-Preserving OCR**: a VLM OCR step renders pages to Markdown so that tables retain row-column alignment, preventing wrong-cell extraction at synthesis time.
- **Adaptive Context-Aware Chunking**: an adaptive semantic chunking algorithm groups semantically contiguous sentences within page clusters and attaches a rolling header (anchor, milestone summaries, prior-group summary) to each chunk embedding.
- **Predictive Knowledge Mapping**: each chunk receives dual hypothetical-query annotations—$Q^-$ for self-contained facts the chunk can answer alone, $Q^+$ for outgoing dependencies—used both as additional retrieval signal and as the seed for HOP edge construction (the end-to-end effect of the directional Stage-1/Stage-2 weighting is descriptive at $n$=150; see Appendix G).
- **Rank-Based HOP Edge Pre-Construction**: HOP edges between chunks are pre-computed and pre-scored offline by retrieving $Q^+$-candidates, scoring each pair with a cross-encoder reranker, and keeping the top-ranked links above threshold; the runtime path traverses these pre-scored edges in a deterministic 1-hop expansion, with no LLM reasoning at retrieval.

The query path is minimal: two-stage hybrid retrieve over the pre-built indices, a rerank pass, a deterministic 1-hop expansion over pre-built HOP edges, and a single LLM call. The two indexing-time contributions are separable: pre-scored HOP edges remove per-hop reasoning from the query path (the latency contribution), while $Q^-/Q^+$ co-indexing supplies retrieval quality; Appendix G isolates the two.

## 2 Related Work

**RAG and graph-based retrieval.** The foundational RAG framework (Lewis et al., 2020) combines dense passage retrieval (Karpukhin et al., 2020) with generation; extensions include iterative retrieval (Shao et al., 2023), reasoning–action loops (Yao et al., 2023), and dynamically-triggered retrieval (Su et al., 2024). A second line imposes structure on the chunk corpus at indexing time: MS-GraphRAG (Edge et al., 2024) extracts an entity-relation KG with community summaries queried via map-reduce; RAPTOR (Sarthi et al., 2024) builds a recursive summary tree; LightRAG (Guo et al., 2024) replaces entity extraction with dual-level keyword retrieval; HippoRAG (Gutiérrez et al., 2024) retrieves via personalized PageRank over an OpenIE graph. PreHypo shares the indexing-time commitment but uses single-level page-cluster chunks rather than recursive summary trees, connects chunks via cross-encoder pre-scored HOP edges rather than entity-relation or OpenIE links, and traverses them in a single deterministic 1-hop with no runtime LLM reasoning, map-reduce, or PageRank.

**Hypothetical embeddings.** Predicting at indexing time the queries a passage answers narrows the query–document gap: doc2query (Nogueira et al., 2019) appends predicted queries to each document, and HyPE (Vake et al., 2025) indexes per-chunk hypothetical questions for question–question matching—both covering only a chunk's *self-contained*, single-hop questions (PreHypo's $Q^-$ role) with no inter-chunk dependency. Query-time variants instead synthesize hypothetical text per query (HyDE Gao et al., 2023; Query2Doc Wang et al., 2023). Closest to PreHypo, HopRAG (Liu et al., 2025) also generates *directional* in-/outcoming pseudo-queries per chunk at indexing time—the analogues of $Q^-/Q^+$—so the directional split is not by itself our contribution. PreHypo differs in *use*: (1) $Q^-/Q^+$ are indexed as *separate retrieval channels* queried directionally, not merely as edge labels; and (2) each HOP edge is *pre-scored offline* by a cross-encoder, making the query path a deterministic 1-hop expansion, whereas HopRAG runs a *retrieve–reason–prune* loop with an LLM call per hop.

**When is multi-hop relevance scored?** A dimension orthogonal to whether the graph is *built* offline is when each multi-hop link's *relevance* is *scored*. Offline-graph systems still defer this decision to query time: MS-GraphRAG (Edge et al., 2024) runs map-reduce over community reports per query, HippoRAG (Gutiérrez et al., 2024) runs personalized PageRank per query, and RAPTOR (Sarthi et al., 2024) descends its summary tree per query. Query-time-reasoning systems make the cost explicit—HopRAG (Liu et al., 2025) pays an LLM reasoning call at *each* hop, and iterative or agentic retrieval (Shao et al., 2023; Yao et al., 2023; Su et al., 2024) re-scores links under a tight per-query budget. A related recent system, QCG-RAG (Wu et al., 2025), builds a doc2query graph offline but still performs multi-hop *traversal* at query time over that graph. PreHypo instead moves the link-scoring step itself—the cross-encoder relevance judgment on each candidate inter-chunk edge—entirely to indexing time, so the query path traverses pre-scored edges with no per-hop scoring, PageRank, map-reduce, traversal decision, or LLM reasoning. We are not aware of prior financial-QA RAG work that pre-scores inter-chunk dependency edges with a cross-encoder offline and reuses the same threshold at retrieval; the closest neighbors above either pay the multi-hop decision per query or score links by a cheaper heuristic (co-occurrence, PageRank) than the cross-encoder used at retrieval time.

**Financial document QA.** FinanceBench (Islam et al., 2023) provides 150 expert-curated questions over SEC 10-K/10-Q filings, with cited page-level evidence and multi-step numerical/cross-document reasoning. Earlier financial QA benchmarks (FinQA Chen et al., 2021, TAT-QA Zhu et al., 2021) supply the relevant table and surrounding text *pre-extracted* per question, so they isolate numerical reasoning from retrieval and document ingestion; they neither require end-to-end PDF parsing nor citation-grounded generation, and the answer's source passage is given rather than retrieved. This makes them a different setting from both PreHypo's full corpus-level retrieval over raw

filings and from open-domain RAG over free-text prose: their table is structured and already localized, so the open problem there is reasoning, not recovering the evidence from an unparsed document.

**Comparability of reported FinanceBench numbers.** Reported FinanceBench scores vary widely with the evaluation protocol: systems given the relevant filing or pre-extracted evidence reach 80–94% (Luong et al., 2024; Hashmi, 2026), whereas our setting—end-to-end retrieval from a raw-PDF corpus, scored as strict binary correctness with a small generator—is a much harder regime, so cross-paper numbers are not directly comparable to ours. We detail the protocol differences and place our absolute scores within the published end-to-end band (∼19–50%) in Appendix C; our controlled comparison is against HopRAG, MS-GraphRAG, and Naive RAG run in an identical retrieval-from-corpus environment (§4.2).

## 3 Methodology

**Task formulation.** Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be a corpus of documents (e.g., PDF filings). After preprocessing, each document is represented as text chunks forming a directed graph $G = (V, E)$, where each node $v \in V$ is a chunk with associated embeddings and edges are either NEXT (sequential adjacency) or HOP (semantic dependency). Given query $q$, the system produces an answer $a$ grounded in a cited evidence set $\mathcal{E}$ where every claim is supported by verbatim spans from $\mathcal{D}$. Figure 1 shows the full architecture: the offline indexing pipeline produces $Q^-$/$Q^+$ annotations and pre-built HOP edges, and the query pipeline is a minimal retrieve-then-generate path that reads from those pre-built structures.

### 3.1 Indexing Pipeline

Indexing proceeds in four stages: documents are normalized through VLM OCR, segmented by adaptive chunking, annotated with hypothetical queries, and connected by pre-built HOP edges.

#### 3.1.1 Document Normalization

SEC filings rely on tabular layouts—income statements, balance sheets, cash flow statements—where a numeric value's meaning depends on its row label and column header. Standard OCR flattens these tables into whitespace-aligned plain text, severing the value–header alignment and causing wrong-cell extractions at synthesis time. Raw PDFs are therefore processed by a Vision-Language Model (VLM) OCR service[1] that renders tables as Markdown with explicit header rows and column-aligned cells (an example output is given in Appendix D), keeping the alignment between column headers (e.g., fiscal years) and row values (e.g., financial metrics) intact for downstream chunking and retrieval.

#### 3.1.2 Adaptive Context-Aware Chunking

Fixed-length chunking splits documents at positions that ignore semantic structure: an income statement line item can be separated from its footnote, or a multi-page disclosure can be chopped mid-sentence. PreHypo's chunking algorithm operates at two levels to keep semantically contiguous content together.

**Page-level grouping (Level 1).** A single financial statement or disclosure often spans multiple pages; treating each page as an independent unit scatters their joint structure. Documents are parsed page-by-page, each page receives an LLM-generated summary, and the summary is embedded. Adjacent pages with cosine similarity $\geq \tau_{\text{page}}$ (default: 0.5) are grouped into a semantic *page cluster*, keeping closely related content (e.g., consecutive financial statement pages) together downstream.

**Sentence-level adaptive splitting (Level 2).** A page cluster can still cover multiple sub-topics, so a second level of splitting finds topic boundaries within the cluster. Sentences are embedded and split at positions where the cosine similarity between adjacent sentence embeddings falls below a group-adaptive threshold $\tau_{\text{chunk}}$ (default: 0.65); chunks must contain at least $M_{\min} = 2$ sentences.

**Rolling context accumulation.** A chunk viewed in isolation loses the document-level context it depends on—which filing it belongs to and what came before in the running narrative. To keep this context attached to each chunk's embedding, each chunk $c_i$ in page-group $g_k$ is enriched with a three-part context header before embedding: an *anchor* (the first page-group's summary, with the document title as fallback), *milestones* (up to two of the most recent prior-group summaries, excluding the immediately preceding one), and a *prev-summary* (the immediately preceding page-group's summary $S(g_{k-1})$).

[1] lightonai/LightOnOCR-1B-1025; https://huggingface.co/lightonai/LightOnOCR-1B-1025

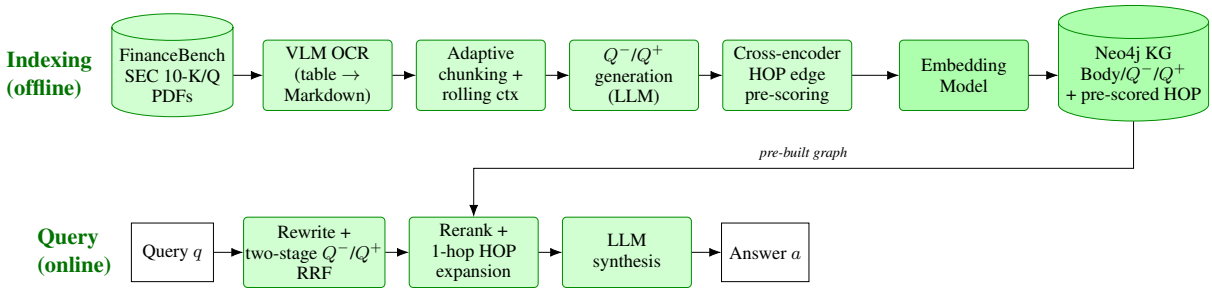


Figure 1: **PreHypo end-to-end architecture.** The offline **indexing pipeline** (top) normalizes raw SEC PDFs via VLM OCR, chunks them adaptively, generates per-chunk $Q^-/Q^+$ hypothetical-query annotations, and *pre-scores* HOP dependency edges with a cross-encoder—all written to a Neo4j knowledge graph. The **query pipeline** (bottom) reads from this pre-built graph through a two-stage hybrid retrieve, a rerank pass with a deterministic 1-hop expansion over the pre-scored edges, and a single LLM synthesis call. Central design point: all multi-hop dependency scoring is paid once at indexing time, so the query path runs no per-hop LLM reasoning (≈7s/query vs. HopRAG's ≈163s). Shaded boxes denote shared infrastructure populated at indexing time and read at query time.

The enriched representation

$$\tilde{c}_i = [\,\text{anchor; milestones; prev-summary; } c_i\,]$$

is embedded and stored in the knowledge graph, so the embedding captures how each chunk relates to the document, not just its local content.

#### 3.1.3 Predictive Knowledge Mapping

A chunk has two retrieval roles: it can be an *answer source* when its content directly satisfies a query, or a *dependency hub* when it surfaces a question that other chunks must answer. Indexing only the chunk's raw text conflates these roles. *Predictive Knowledge Mapping* separates them by annotating each chunk with a directed pair of hypothetical queries: $Q^-$ is a hypothetical question the chunk can answer alone, and $Q^+$ is a question whose answer requires content beyond the chunk. The two are indexed separately so the chunk's incoming and outgoing semantic roles can be searched independently. Following HopRAG (Liu et al., 2025), the queries are LLM-generated:

- $Q^-$ **(Incoming Projection)**: a question that is *directly and completely* answerable from $c_i$ alone, e.g., *"What was Apple's net income in FY2022?"*
- $Q^+$ **(Outgoing Projection)**: a question for which $c_i$ provides partial evidence but requires additional document context, e.g., *"How did Apple's FCF margin change from FY2021 to FY2022?"*

The $Q^-$ embedding captures the chunk's *self-contained* factual content; the $Q^+$ embedding captures its role as a *dependency node* in multi-hop chains.

#### 3.1.4 Rank-Based HOP Edge Pre-Construction

Multi-hop financial questions (e.g., margin computations spanning multiple line items, YoY comparisons across prior-period chunks) require following dependencies across chunks. Whether a candidate inter-chunk link is truly relevant must be scored somewhere in the pipeline; existing systems pay this scoring cost per query, while PreHypo performs it offline at indexing time through a *retrieve–rerank–prune* procedure. For each source chunk $c_i$ with $Q^+$ embedding $\mathbf{q}_i^+$, we (1) retrieve the top-$K_{\text{hop}}$ candidates ($K_{\text{hop}} = 15$) from the $Q^+$ vector index by ANN search on $\mathbf{q}_i^+$, restricted to chunks from the same company but a different source document so edges stay within-issuer yet cross-document (falling back to the $Q^-$ index when $Q^+$ returns nothing); (2) score each candidate $c_j$ with a cross-encoder reranker $r$ on $(Q_i^+, c_j)$; and (3) keep a HOP edge $c_i \to c_j$ if $r(Q_i^+, c_j) \geq \tau_r$ ($\tau_r = 0.4$), retaining the top $L_{\text{hop}} = 5$ edges by reranker score. This separates candidate generation from dependency verification: ANN search over the $Q^+$ index retrieves chunks with related outgoing dependencies, while the cross-encoder reranker $r(Q_i^+, c_j)$ performs the dependency check itself by scoring whether $c_j$ supplies the content that $c_i$'s outgoing question asks for. The same $\tau_r$ is used during indexing and at retrieval, so HOP edges follow the same scoring criterion the system applies when reading them; multi-hop discovery happens once, at indexing time, fixing which edges to follow in advance rather than re-deciding per query—the source of PreHypo's query-time latency advantage (§3.3).

The resulting graph $G$ stores three vector indices per chunk—body, $Q^-$, $Q^+$—and three corresponding BM25-style fulltext indices.

### 3.2 Query Pipeline

The query path is kept minimal: two-stage hybrid retrieve, rerank with top-up, and deterministic 1-hop expansion over the pre-built graph.

**Query rewriting.** Financial filings use a constrained but variable surface vocabulary, so a single phrasing often misses chunks that use an equivalent term. To bridge this lexical gap without drifting from intent, an LLM generates up to $N_r$=2 retrieval variants under fixed rules—company/period tokens preserved verbatim, metric and line-item terms unchanged except for a fixed synonym list, and statement anchors inserted when implied (full rule set and synonym list in Appendix B). Both the original and the variants are issued to the retrieval stage below.

**Two-stage hybrid retrieve with RRF.** Retrieval must surface two kinds of content—chunks that directly answer the query, and chunks that supply the dependencies the answer needs—so PreHypo routes the query through different chunk projections in two stages. Within each channel, vector and fulltext search are combined via reciprocal rank fusion (RRF) (Cormack et al., 2009): $\mathrm{RRF}(d) = w_v/(k + r_v(d)) + w_t/(k + r_t(d))$ with $k = 60$, $w_v = 1.3$, $w_t = 1.0$—vector captures semantic similarity, fulltext anchors on exact entity and period tokens. The original query and its $N_r$ variants are issued through each channel; variant contributions are down-weighted to $w_r = 0.85$ of the original when results are accumulated across queries. Stage 1 fetches candidates from the $Q^-$ (channel weight 0.7) and body (0.3) indices for grounded, self-contained content; if Stage 1 yields fewer candidates than the slot budget *or its top rerank score falls below* $\tau_r + 0.08$, Stage 2 expands by querying the $Q^+$ channel (0.6) together with a $Q^-$ support pool (0.4) to introduce outgoing-dependency chunks.

**Rerank with top-up.** Bi-encoder retrieval (vector + fulltext via RRF) is fast but coarse; a cross-encoder re-scores query–document pairs at finer granularity. Candidates are reranked at threshold $\tau_r = 0.4$—the same threshold used during HOP edge construction, so retrieval and indexing apply the same relevance bar. If fewer chunks cross the threshold than the requested top-$k$, the pool is topped up with the next-best ungated candidates so the downstream context never runs short of candidates due to an over-strict gate.

**Deterministic 1-hop expansion.** Multi-hop dependencies must be followed at query time, but per-query decisions about which neighbor to follow are expensive: HopRAG pays an LLM reasoning call at each hop. Because PreHypo's HOP edges are pre-scored by the same cross-encoder used at retrieval, no such per-query decision is needed. Starting from the reranked seed nodes, the system expands the frontier exactly once along pre-built NEXT and HOP edges. There is no LLM continuation check, no multi-turn loop, and no runtime hypothetical-query expansion—the offline edges fix the connectivity used at query time.

**Synthesis.** The expanded context is concatenated—each chunk preceded by a `[[Title, Page, Chunk]]` marker so the LLM can see the source location of every span—and passed to a single LLM call with the instruction to answer using only the provided context, or to say it does not know if the context is insufficient. The same synthesis prompt is used for PreHypo and every baseline so any score gap traces back to retrieval, not synthesis-prompt asymmetry. There is no reflection stage, no refinement stage, and no agent.

### 3.3 Query-Time Cost

PreHypo's query path runs in ≈7 seconds per question on FinanceBench (mean over 150 queries) against ≈163 seconds for HopRAG; the asymmetry follows from where LLM reasoning is paid—PreHypo issues at most three LLM calls per query (query rewrite, reranker-query simplification when the input exceeds 80 characters, and a single synthesis call), while HopRAG performs $O(K_{\text{hop}})$ LLM reasoning calls during its retrieve–reason–prune multi-hop exploration over pre-built edges.

## 4 Experiments

### 4.1 Setup

**Datasets.** We evaluate on two benchmarks spanning two document domains. **FinanceBench** (Islam et al., 2023): 150 expert-curated questions over SEC 10-K/10-Q filings from publicly traded companies, requiring numeric extraction, multi-step arithmetic (e.g., free cash flow margin, YoY

CAPEX change), and cross-document reasoning; ground-truth answers include expert-annotated supporting evidence and computation formulas. **MultiHop-RAG** (Tang and Yang, 2024): a news-article benchmark whose queries require retrieving and reasoning over evidence spread across multiple documents; we use a balanced sample of $n$=200 queries (50 each of comparison / inference / temporal / null types; details in §4.3). FinanceBench is our primary benchmark (it carries the page-grounding and ablation analyses); MultiHop-RAG tests cross-domain generalization.

**Baselines.**

- **Naive RAG**: sentence-level chunking with vector similarity search (no graph, no adaptive chunking, no hypothetical queries).
- **HopRAG** (Liu et al., 2025): pre-built passage graph with LLM-generated pseudo-query edges, traversed at retrieval time via the *retrieve–reason–prune* multi-hop mechanism with LLM reasoning at each hop.
- **MS-GraphRAG** (Edge et al., 2024): entity-relation knowledge graph with community-based retrieval.

The four systems share the upstream PDF and OCR output and then each runs its own published indexing and retrieval pipeline on that shared OCR markdown. Baseline code provenance and serving details are in Appendix B.

**LLM models.** All systems use **Qwen3-4B-Instruct-2507** (Qwen Team, 2025) as the generation model: PreHypo targets small, locally-deployable LLMs, so a 4B generator is the intended operating point rather than a constraint. Embeddings use **Qwen3-Embedding-0.6B** and candidates are reranked by **Qwen3-Reranker-0.6B**, which is also the cross-encoder used for offline HOP edge pre-scoring (§3.1.4). Hyperparameters are in Appendix A; serving stack and storage details are in Appendix B.

**Metrics.** We report the following metrics (higher is better unless noted):

- **Correctness** (J, 0–1): correctness against ground truth scored by an LLM judge (`gpt-5.5-2026-04-23`); abstentions score 0. For brevity we refer to this LLM-judged correctness score as the *Judge score* (J) throughout.
- **Halluc.** (H, 0–1, lower better): substantive responses with factually incorrect content, flagged by the same judge; abstentions score 0.
- **DocM** (0–1): fraction of queries where the gold document is in the retrieved set.
- **PgM** (0–1): fraction of queries where the gold page is in the retrieved set (FinanceBench only; news articles are not paginated).
- **Att** (0–1): fraction of queries answered substantively (vs. abstaining).
- **Latency** (s, lower better): mean per-query response time (query-time only, indexing excluded).

For MultiHop-RAG we additionally report the benchmark's official fact-level retrieval metrics—MRR@10, MAP@10, Hits@10—and evidence-document recall (fraction of gold evidence documents retrieved). We always report H jointly with Att so that low H from heavy abstention is not misread as faithfulness.

**Reporting protocol.** FinanceBench numbers are 5-fold mean ± std (folds of 30; per-fold plots in Appendix H). MultiHop-RAG numbers are 5-fold mean ± std (folds of 40, same seed). Significance throughout is established by query-level paired bootstrap CIs (Appendix I for FinanceBench; §4.3 for MultiHop-RAG), which—unlike fold std—are not subject to partition variance.

### 4.2 Main Results

Table 1 reports main results. We treat two findings as co-primary: page-level grounding (PgM) reaches 0.19 against $\leq$ 0.04 for every baseline—the largest gap relative to fold variance, though 0.19 itself is a low absolute level—and the Judge gap is resolved but tight under a query-level paired bootstrap (Appendix I); we lead with PgM as the more robust of the two. Defensive analyses are in Appendix J.

**Page-level grounding: largest *relative* gap, low absolute floor.** PreHypo reaches PgM 0.19±0.07 vs. the best baseline's 0.04 (Figure 2)—the largest gap in Table 1 relative to fold variance, though the absolute level is low (∼19% of queries). We treat PgM as a cleanly comparable proxy rather than a destination metric. Cross-page semantic grouping (Level-1 chunking, §3.1.2) is the largest

| System | Judge↑ All | Judge↑ \| Att | Halluc.↓ All | Halluc.↓ \| Att | PgM↑ | Att↑ | Lat. (s)↓ |
|---|---|---|---|---|---|---|---|
| **PreHypo (ours)** | **0.32 ± 0.11** | **0.57 ± 0.09** | 0.17 ± 0.04 | 0.31 ± 0.08 | **0.19 ± 0.07** | 0.55 ± 0.10 | 6.81 |
| MS-GraphRAG | 0.19 ± 0.10 | 0.23 ± 0.10 | 0.32 ± 0.05 | 0.40 ± 0.11 | 0.03 ± 0.04 | 0.80 ± 0.08 | 26.4 |
| HopRAG | 0.19 ± 0.05 | 0.39 ± 0.05 | 0.13 ± 0.06 | 0.26 ± 0.04 | 0.04 ± 0.04 | 0.50 ± 0.11 | 163.2 |
| Naive | 0.05 ± 0.04 | 0.10 ± 0.05 | 0.04 ± 0.04 | 0.06 ± 0.04 | 0.00 ± 0.00 | 0.59 ± 0.08 | 2.1 |

Table 1: FinanceBench (n=150) main results, 5-fold mean ± std. **Bold** marks columns where PreHypo's lead is unambiguous (gap to runner-up exceeds fold std or is corroborated by paired bootstrap, §4.2). **Judge | Att** and **Halluc. | Att** are conditional metrics over attempted queries (substantive response, not abstention) and expose retrieval/synthesis quality independent of abstention rate. PreHypo leads on Judge in both views; Naive's nominally low Halluc. is not bolded because it reflects an answered set of which only $\sim$10% are correct rather than faithfulness. The PgM gap (0.19 vs. $\leq 0.04$) is the largest *relative* gap; the absolute level 0.19 is low. Document Match (DocM) is $\geq 0.89$ for all four systems (PreHypo/HopRAG 0.99, MS-GraphRAG 0.95, Naive 0.89) and is omitted from the table.

| System | Judge↑ | Halluc.↓ | MRR@10↑ | MAP@10↑ | Hits@10↑ | DocRec↑ | Att↑ | Lat. (s)↓ |
|---|---|---|---|---|---|---|---|---|
| **PreHypo (ours)** | **0.82 ± 0.06** | 0.12 ± 0.05 | 0.66 ± 0.07 | 0.41 ± 0.08 | 0.56 ± 0.08 | **0.84 ± 0.03** | 0.95 ± 0.04 | 7.8 |
| MS-GraphRAG | 0.70 ± 0.07 | 0.12 ± 0.06 | 0.62 ± 0.05 | 0.35 ± 0.02 | 0.60 ± 0.06 | 0.53 ± 0.03 | 1.00 ± 0.00 | 35.9 |
| HopRAG | 0.64 ± 0.07 | 0.09 ± 0.03 | 0.53 ± 0.08 | 0.31 ± 0.06 | 0.52 ± 0.09 | 0.77 ± 0.03 | 0.81 ± 0.05 | 18.7 |
| Naive | 0.65 ± 0.06 | 0.17 ± 0.05 | 0.35 ± 0.08 | 0.16 ± 0.04 | 0.25 ± 0.04 | 0.46 ± 0.04 | 0.93 ± 0.05 | 3.2 |

Table 2: **MultiHop-RAG results** (balanced $n$=200 sample; 5-fold mean ± std, folds of 40, seed 42). **Judge** (LLM-judged correctness), Halluc., and Att are over all 200 queries; retrieval metrics (MRR@10, MAP@10, Hits@10) and **DocRec** (evidence-document recall) are over the gold-bearing (non-null) queries. **Bold** marks columns where PreHypo's lead is significant against *every* baseline under a query-level paired bootstrap ($B$=10,000, seed 42): Judge (vs. MS-GraphRAG $+0.12$ $[+0.06, +0.19]$, vs. HopRAG $+0.19$ $[+0.12, +0.26]$, vs. Naive $+0.17$ $[+0.10, +0.24]$) and evidence-document recall (vs. MS-GraphRAG $+0.31$). On MRR/MAP/Hits, PreHypo is significant vs. Naive and HopRAG but ties MS-GraphRAG (CIs include zero); MS-GraphRAG's nominal Hits@10 edge is not significant. Page-level grounding does not apply (news articles are unpaginated).

single ablated contributor: disabling it drops PgM $0.19 \rightarrow 0.11$ ($\sim$42% of the gap); the residual $0.11 > 0.04$ stays unattributed. The cross-page-synthesis slice (DocM= 1, PgM= 0, answer still judged correct), where evidence must be synthesized across sibling pages rather than the single gold page, is discussed in Limitations—a pattern also reported for financial QA (Choe et al., 2025; Wang et al., 2025).

**Judge and latency.** PreHypo's Judge (0.32) is above every baseline ($\leq 0.19$, $\sim$13pp); fold std on its Judge cell ($\sim$0.11) overlaps baselines on individual folds, but query-level paired bootstrap CIs exclude zero on every pairing (Appendix I). The absolute level is low but in line with published end-to-end FinanceBench systems, which report 19–50% accuracy in this retrieve-from-corpus regime (Appendix C). Latency is $\approx$ 7s vs. HopRAG's $\approx$ 163s.

PreHypo also sits on the J–H Pareto frontier—no other system dominates it on both axes (Figure 4, Appendix H).

| Variant | J↑ | H↓ | PgM↑ | Att↑ | Ctx↓ |
|---|---|---|---|---|---|
| **Full** (PreHypo) | 0.32 | 0.17 | 0.19 | 0.55 | 5.5 |
| $Q^-$-only | 0.30 | 0.17 | 0.16 | 0.55 | 5.5 |
| $Q^+$-only | 0.22 | 0.13 | 0.18 | 0.44 | 5.5 |
| Combined | 0.28 | 0.17 | 0.19 | 0.50 | 5.4 |

Table 3: **Channel ablation** (FinanceBench, 5-fold mean, $n$=150): query-side mechanism check of the $Q^-$/$Q^+$ channels, indexing held fixed. **Full** = PreHypo; **Combined** = HopRAG-style single non-directional channel. Detail in Appendix G.

**Channel ablation ($Q^-$/$Q^+$).** A query-side mechanism check isolates each channel's role (Table 3). $Q^+$-only drops Att $0.55 \rightarrow 0.44$ ($Q^+$ indexes *outgoing-dependency* questions) while $Q^-$-only matches Full, locating the answer-bearing signal in $Q^-$. A HopRAG-style non-directional channel (Combined) already beats HopRAG by $\sim$9pp (0.28 vs. 0.19), placing most of the advantage in annotation quality; the directional weighting adds $+0.04$ but is not resolved at $n$=150 (Appendix G). This makes the two contributions cleanly separable: the quality comes from $Q^-$/$Q^+$ co-indexing, while the

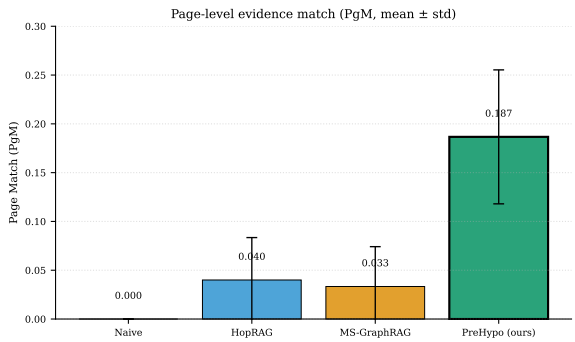


Figure 2: **Page-level evidence match (PgM, 0–1, higher is better) on FinanceBench** ($n$=150, 5-fold mean ± std). PgM is the fraction of queries for which the gold expert-cited page appears in the retrieved set. PreHypo reaches 0.19 while every baseline is at $\leq$ 0.04—the largest gap in the paper *relative* to fold variance (exceeds it on every pair), though 0.19 is a low absolute floor (§4.2). Ablations attribute most of this gain to cross-page semantic grouping (Table 5).

pre-scored HOP edges contribute the latency (§3.3) rather than the quality.

### 4.3 Generalization to MultiHop-RAG

To test whether PreHypo's indexing-time design generalizes beyond financial filings, we evaluate on MultiHop-RAG (Tang and Yang, 2024), a news-article benchmark whose queries require retrieving and reasoning over evidence spread across multiple documents (question types: comparison, inference, temporal, and null). We use a balanced sample of $n$=200 queries (50 per type, seed 42); all four systems run the identical sample with the same Qwen3-4B generator and `gpt-5.5` judge. PreHypo's core indexing and query design is unchanged from FinanceBench; only domain-gated configuration differs—most notably the strict company filter (specific to single-issuer financial queries) is disabled for news, and the per-chunk context carries each article's publication date and source (Appendix B).[2]

Table 2 reports results. PreHypo reaches LLM-judged correctness 0.82 against $\leq$ 0.70 for every baseline, and a query-level paired bootstrap ($B$=10,000) separates the lead from all three (every CI excludes zero), including MS-GraphRAG (+0.12 [+0.06, +0.19])—a cleaner separation than on FinanceBench, where the Judge lead was tight. The win holds across all four question types (comparison 0.66, inference 0.94, temporal 0.68, null 1.00; each above every baseline). Excluding null (included by design to test hallucination avoidance, where all systems score near-perfectly), PreHypo's mean over the answerable types is 0.76 vs. $\leq$ 0.61, so the lead widens rather than resting on the null slice. PreHypo also leads evidence-document recall (0.84 vs. $\leq$ 0.77, significant vs. all), at ≈7.8s/query versus MS-GraphRAG's ≈36s. Here the gain concentrates in correctness and evidence-document recall—consistent with pre-scored cross-document HOP edges surfacing the right evidence—while rank-quality metrics (MRR/MAP/Hits@10) tie MS-GraphRAG within bootstrap CIs.

[2]These domain gates leave the FinanceBench configuration unchanged; they adapt the same pipeline to the news domain rather than altering the method.

## 5 Failure Analysis

PreHypo's 0.55 attempt rate reflects three abstention drivers (strict company filter, $\tau_r$=0.4 rerank gate, synthesis prompt's "insufficient evidence" default) whose individual contributions we do not separately quantify. Two failure modes dominate the incorrect attempts, both LLM-side rather than retrieval-side, and we do not separately quantify them. *(i) Non-answer despite correct retrieval*: a large share of failures occur when DocM= 1 but the answer is "insufficient evidence," over-representing multi-step numerical computations (e.g., CAPEX change rate)—the primitives are retrieved but the retrieval-only path has no arithmetic-audit stage, and at PgM= 0.19 the exact gold page is frequently absent. *(ii) Formula misapplication*: the LLM retrieves the right values but applies the wrong formula (e.g., ROA as Net Income/Equity).

## 6 Conclusion

PreHypo moves multi-hop dependency scoring to indexing time, separating two contributions: pre-scored HOP edges make the query path a deterministic 1-hop expansion with no per-hop LLM reasoning—the latency win, ≈7s vs. HopRAG's ≈163s—while $Q^-/Q^+$ co-indexing (with VLM OCR and adaptive chunking) supplies retrieval quality. On FinanceBench (Qwen3-4B) it reaches PgM 0.19 vs. $\leq$ 0.04 (largest relative gap) and Judge 0.32 (13pp, bootstrap CIs above zero but tight), with grounding tracing to cross-page semantic grouping. The design generalizes: on MultiHop-RAG it reaches correctness 0.82 vs. $\leq$ 0.70, separated from every baseline. Code: `https://anonymous.4open.science/r/PreHypo-FB60`.

## Limitations

**Statistical power.** Evaluation uses 5-fold contiguous random splits ($n = 30$ per fold) over FinanceBench's 150 queries; fold std on PreHypo's Judge cell ($\sigma = 0.11$) is the largest among the four systems and overlaps with baselines on individual folds. We verified seed robustness by repeating the partition under 10 random seeds: PreHypo–baseline Judge gaps stay positive on every seed and remain invariant in mean (vs. HopRAG $+0.127$, vs. MS-GraphRAG $+0.133$, vs. Naive $+0.267$). The query-level paired bootstrap (§4.2, Appendix I) is not subject to fold-partition variance and excludes zero on all three pairings, including under Bonferroni correction (HopRAG lower bound shrinks to 0.7pp). The $Q^-/Q^+$ directional weighting (Appendix G, Combined row) is *not* statistically resolved at this sample size (95% CI $[-0.053, +0.133]$, $p$=0.35); we report it descriptively.

**Benchmark and domain coverage.** We evaluate on two benchmarks spanning two domains—financial filings (FinanceBench) and multi-hop news QA (MultiHop-RAG)—but do not cover legal, medical, or the adjacent pre-extracted table-and-text benchmarks (FinQA, TAT-QA). On MultiHop-RAG we use a balanced $n$=200 sample rather than the full 2,556 queries (the graph baselines are slow—HopRAG alone is ≈28h on the full set), with significance established by query-level paired bootstrap rather than 5-fold splits; the per-type counts (50 each) are small, so per-category numbers are indicative rather than precise. Page-level grounding is FinanceBench-specific and may behave differently on single-page-evidence (TAT-QA) or longer numerical-reasoning (FinQA) benchmarks, and the indexing modules' relative weights may shift across domains. Absolute margins should be read as descriptive of these two benchmarks rather than as calibrated estimates of broader generalization.

**LLM-based evaluation.** Judge and hallucination labels are produced by a closed-model judge (`gpt-5.5-2026-04-23`) in a single combined call. We do not report human–LLM agreement or inter-rater reliability for these labels; both metrics should be read as approximations whose absolute calibration depends on judge behavior. Honest "insufficient evidence" abstentions are deterministically assigned $H = 0$, so abstaining is never punished as hallucination, but this means H must always be read jointly with Att.

## Ethical Considerations

This work develops a question-answering system over publicly available financial filings (SEC 10-K/10-Q documents), which are legally mandated disclosures in the public domain. No private or personally identifiable information was collected or used. The LLM-based evaluation (judge score, hallucination rate) introduces potential model bias, and scores should be interpreted alongside the deterministic retrieval metrics (DocM, PgM). The system is intended to assist analysts, not replace professional financial judgment; outputs should be independently verified before use in financial decisions.

## References


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.

Jaeyoung Choe, Jihoon Kim, and Woohwan Jung. 2025. Hierarchical retrieval with evidence curation for open-domain financial question answering on standardized documents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16663–16681.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

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*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

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, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*.

Afshan Hashmi. 2026. Adaptive query routing: A tier-based framework for hybrid retrieval across financial, legal, and medical documents. *arXiv preprint arXiv:2604.14222*.

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*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. *arXiv preprint arXiv:2502.12442*.

Vinh Luong, Sang Dinh, Shruti Raghavan, William Nguyen, Zooey Nguyen, Quynh Le, Hung Vo, Kentaro Maegaito, Loc Nguyen, Thao Nguyen, Anh Hai Ha, and Christopher Nguyen. 2024. DANA: Domain-aware neurosymbolic agents for consistency and accuracy. *arXiv preprint arXiv:2410.02823*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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 (ICLR)*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Domen Vake, Jernej Vičič, and Aleksandar Tošić. 2025. Bridging the question–answer gap in retrieval-augmented generation: Hypothetical prompt embeddings. *IEEE Access*. IEEE Xplore document 11080443.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Xinyu Wang, Jijun Chi, Zhenghan Tai, Tung Sum Thomas Kwok, Muzhi Li, Zhuhong Li, Hailin He, Yuchen Hua, Peng Lu, Suyuchen Wang, Yihong Wu, Jerry Huang, Jingrui Tian, Fengran Mo, Yufei Cui, and Ling Zhou. 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 (CIKM)*.

Yaxiong Wu, Jianyuan Bo, Yongyue Zhang, Sheng Liang, and Yong Liu. 2025. Query-centric graph retrieval augmented generation. *arXiv preprint arXiv:2509.21237*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287.

## A Hyperparameter Summary

| Parameter | Value | Description |
|---|---|---|
| $\tau_{\text{page}}$ | 0.5 | Page similarity threshold |
| $\tau_{\text{chunk}}$ | 0.65 | Sentence cohesion threshold |
| $M_{\min}$ | 2 | Min sentences/chunk |
| $L_{\text{hop}}$ | 5 | Max outgoing HOP edges |
| $K_{\text{hop}}$ | 15 | HOP candidate pool size |
| $\tau_r$ | 0.4 | Reranker score threshold |
| $\delta_{\text{stage2}}$ | 0.08 | Stage-2 trigger slack |
| $k$ (RRF) | 60 | RRF constant |
| $w_v$ | 1.3 | Vector RRF weight |
| $w_t$ | 1.0 | Fulltext RRF weight |
| $w_r$ | 0.85 | Query-variant down-weight |
| $N_r$ | 2 | Query-rewrite variants |
| Stage 1 weights | 0.7/0.3 | $Q^-$/body |
| Stage 2 weights | 0.6/0.4 | $Q^+$/$Q^-$ support |
| $d_{\text{emb}}$ | 1024 | Embedding dimensions |

Table 4: Key hyperparameters of PreHypo. The query path has no agent loop, so there are no $T_{\max}$ or $R_{\max}$ parameters. $\delta_{\text{stage2}}$, $N_r$, $K_{\text{hop}}$, and $L_{\text{hop}}$ were fixed before the main experiments based on small-scale pilots and we do not report a formal sensitivity sweep; the values shown reflect the configuration used for every Table 1 and Table 5 cell.

## B Implementation Details

**Serving stack.** The generation, embedding, and reranker models are served via vLLM on separate GPUs. The knowledge graph—chunks, NEXT/HOP edges, vector indices, and BM25-style fulltext indices—is stored in Neo4j. Each system writes to its own Neo4j namespace so retrieval is fully isolated across baselines.

**Embedding configuration.** Qwen3-Embedding-0.6B is served with embedding dimensionality $d_{\text{emb}} = 1024$ and a context window of $d_{\max} = 8192$ tokens. Inputs longer than $d_{\max}$ are truncated client-side before embedding.

**Baseline code provenance.** HopRAG runs the authors' official HopRetriever implementation, adapted to share our embedding service and Neo4j backend. MS-GraphRAG uses the official `graphrag` package. Naive RAG is a minimal top-$k$ vector-similarity baseline implemented in this repository.

**Judge call.** The Judge and Halluc. labels (§4) are produced by a single combined LLM call per query, scoring correctness and flagging hallucination in one pass to keep the two metrics self-consistent and amortize evaluation cost.

**Query rewriting rules.** The up-to-$N_r$=2 retrieval variants (§3.2) are generated under fixed rules: company and period tokens are preserved verbatim; metric and line-item terms are kept unchanged except for a fixed synonym list (revenue↔net sales, capex↔purchases of PP&E, net PP&E↔property, plant and equipment—net, net AR↔trade accounts receivable, net); and financial statement anchors (e.g., income statement) are inserted when the query implies them.

**Strict company filter.** FinanceBench queries are anchored to a single company (e.g., "Apple's FY2022 net income"), and cross-company chunks that score highly on reranker similarity—structurally similar finance tables from unrelated tickers—can produce hallucinated answers if they enter the synthesis context. For single-issuer queries the rerank pool is therefore restricted to the queried company between rerank and 1-hop expansion, with a fallback to a no-filter pass if strict filtering would empty the pool. This is a domain configuration for single-company queries (disabled for the news domain, §4.3), not a tuning of the retrieval scorer.

**Canonical 5-fold protocol.** Every per-cell mean ± std in this paper is computed by Python's `random.Random(42).shuffle` on the list `[0, 1, . . . , 149]` (indices into the per-run `details` array, in file order), partitioned into five contiguous slices of 30. For each fold we average the per-query metric, and over the five fold means we report the population mean and population standard deviation (denominator $N$, not $N-1$). This is the protocol used throughout Tables 1, 5, and 3, and reproduced verbatim by the `analyze_context_size.py` script shipped in the supplementary archive.

## C Comparability of External FinanceBench Numbers

Reported FinanceBench scores vary widely with the evaluation protocol, so cross-paper numbers are not directly comparable to ours (§2). DANA (Luong et al., 2024), a domain-aware neurosymbolic agent, reports ∼90% with each system given access to the relevant SEC filing; the tier-based router of Hashmi (2026) reports 0.82–0.94 but evaluates on FinanceBench's *pre-extracted evidence passages* ("without raw PDF access") and scores a composite *accuracy+completeness+relevance* quality by

an LLM judge rather than strict binary correctness. Both hand the system the gold document or evidence, so they isolate reasoning from retrieval. PreHypo instead retrieves from scratch over a raw-PDF corpus, scores strict binary correctness with abstentions counted as wrong, and uses a deliberately small generator; we therefore do not place these reported figures in our results table. For absolute calibration, the closest published condition to ours—FinanceBench's shared-vector-store retrieval with no gold evidence—reaches 19% correct even with a far larger generator (9% closed-book) (Islam et al., 2023), and end-to-end accuracy approaches ∼50% only with a heavily engineered system and a 72B generator (Wang et al., 2025); the oracle and long-context settings that report 79–85% hand over the evidence and are not the same task. PreHypo's 0.32 with a small generator therefore sits above its nearest-condition peer and within this published end-to-end band.

## D Topology-Preserving OCR Example

A representative output of the VLM OCR step (§3.1) on a financial statement table, illustrating the row/column alignment preserved by the Markdown rendering:

```
| Metric     | FY2022 | FY2021 |
|------------|--------|--------|
| Revenue    | $394B  | $366B  |
| Net Income | $100B  | $95B   |
```

## E $Q^-$/$Q^+$ Generation Prompt

The offline prompt used to generate dual hypothetical-query annotations per chunk (§3.1.3). The model is the same Qwen3-4B generator used at synthesis; `{global_context}` below contains the rolling-context header described in §3.1.2.

**Definitions (kept strictly distinct).** $Q^-$ **(Incoming, self-contained):** questions the chunk alone answers verbatim. $Q^+$ **(Outgoing, Bridge):** questions grounded in this chunk's entity/period/metric whose answer requires content from another chunk or document; $Q^+$ builds the multi-hop graph and is not a paraphrase of $Q^-$.

**Rules.**

1. $Q^-$: up to 3 self-contained questions directly answerable from this chunk; empty list if no concrete facts.
2. $Q^+$: up to 3 outgoing questions, each satisfying at least one of: (a) same metric, different period; (b) derived metric (margin, ratio, growth, YoY, FCF, ROA, ROE) whose operands are not all present here; (c) related line item in a different statement; (d) multi-doc bridge (prior-year filing, segment breakdown). Empty if none apply.
3. Each question is specific, answerable in finite SEC-filing context, and $\leq 22$ words.
4. Each question includes grounding tokens (entity, metric, period, source anchor) from this chunk.
5. Include a period token in every $Q^-$; a different period token is allowed (and preferred) in $Q^+$.
6. No placeholders or meta phrases; never fabricate unseen values.
7. TOC, exhibits, signatures, and boilerplate → shorter or empty lists.
8. Computation cues (change, ratio, margin, prior-period) → at least one $Q^+$ of type (a) or (b).
9. Dense Summary: exactly one sentence, $\leq 35$ words, grounded only in this chunk; preserve numeric strings.

**Output schema.** JSON only: `{"summary": "...", "q_minus": [...], "q_plus": [...]}`.

## F Indexing-Module Ablation

This appendix reports the one-at-a-time removal of each upstream indexing-time module (referenced from §4.2 and from the channel ablation in §G). Each module shows a distinct dominant metric.

**Adaptive chunking determines page-level grounding.** The −Chunking variant disables Level-1 page-cluster grouping while keeping Level-2 splitting and the rolling context header intact. PgM drops from 0.19 to 0.11 (−8pp) while Judge moves by only 1pp: when pages no longer co-cluster, multi-page statements split into per-page units, the chunk with the gold row drops out of top-$k$, and the answer cites a neighboring narrative chunk (correct value, wrong-page citation). The retrieved context also shrinks (5.5k → 4.6k chars, −18% across all five folds): single-page chunks are inherently shorter, yet PgM nearly halves—Full's longer chunks are not wasteful padding but denser, page-aligned

| Variant | J↑ | H↓ | PgM↑ | Att↑ | Ctx↓ |
|---|---|---|---|---|---|
| **Full** | $0.32 \pm 0.11$ | $0.17 \pm 0.04$ | $0.19 \pm 0.07$ | $0.55 \pm 0.10$ | $5.5 \pm 0.6$ |
| −Chunking | $0.31 \pm 0.07$ | $0.14 \pm 0.04$ | $0.11 \pm 0.04$ | $0.51 \pm 0.08$ | $4.6 \pm 0.5$ |
| −Summary | $0.29 \pm 0.04$ | $0.16 \pm 0.03$ | $0.13 \pm 0.06$ | $0.49 \pm 0.06$ | $5.8 \pm 0.6$ |
| −Table-to-text | $0.30 \pm 0.14$ | $0.19 \pm 0.06$ | $0.20 \pm 0.05$ | $0.57 \pm 0.14$ | $9.5 \pm 2.2$ |

Table 5: PreHypo upstream-module ablations on FinanceBench (5-fold mean ± std): one-at-a-time removal of each indexing-time module. **Ctx** is the mean number of characters in the retrieved context fed to the synthesis LLM (top-$k$=8 chunks per query, reported in thousands of chars; lower is better at matched quality). DocM is $\geq 0.97$ across all rows and is omitted. The complementary query-side ablation of the $Q^-$/$Q^+$ directional channels is in Appendix G (Table 3).

evidence. Cross-page semantic grouping is the PgM mechanism; J is largely unaffected by its removal.

**Table-to-text compresses the LLM input.** The strongest signal here is on context size, not hallucination: disabling table-to-text inflates the retrieved context from 5.5k to 9.5k chars per query (+70% in mean, +38 to +126% across all five folds; Ctx in Table 5)—raw whitespace-aligned tables expand into long verbatim spans, while VLM-rendered Markdown delivers the same content in much shorter form. Quality moves only marginally (H $0.17 \rightarrow 0.19$, Judge +2pp, both within fold variance $\pm 0.04$), so we treat the quality effect as suggestive; the compactness effect is direction-consistent across all five folds.

**Bundle attribution.** Of the 13pp PreHypo–HopRAG Judge gap (0.32 vs. 0.19), the upstream indexing modules contribute modestly—chunking (∼1pp) and rolling summary (∼3pp)—and the remaining ∼9pp is *jointly* attributable to dual $Q^-$+$Q^+$ co-indexing and pre-scored HOP edges; the pre-scoring fixes the multi-hop edges offline, which is what gives the query path its latency advantage. The query-side channel ablation (Appendix G, Table 3) corroborates this: a HopRAG-style single-channel use of the same $Q^-$/$Q^+$ annotations (Combined) already beats HopRAG by ∼9pp (0.28 vs. 0.19), placing most of the gap in annotation quality rather than the directional architecture; it also confirms the channel semantics ($Q^+$-only drops Att $0.55 \rightarrow 0.44$; $Q^-$ holds the answer-bearing signal).

## G Channel Ablation: $Q^-$/$Q^+$ Direction Split

This appendix expands the channel ablation summarized in §4.2 (Table 3, promoted to the main body). It is a query-side *mechanism check* for what each $Q^-$/$Q^+$ channel encodes; the indexing pipeline is held fixed across all variants (confirmed by near-identical Ctx values 5.4–5.5k), so differences isolate the retrieval channels rather than indexing. Relative to **Full** (PreHypo: Stage-1 $Q^-$ 0.7 + body 0.3, Stage-2 $Q^+$ 0.6 + $Q^-$ 0.4 on demand): $Q^-$-only and $Q^+$-only use a single hypothetical channel at weight 1.0 (no body, no Stage-2); **Combined** is a HopRAG-style single non-directional hypothetical channel ($Q^-$/$Q^+$ at $0.5/0.5$, no body, no Stage-2), isolating annotation quality from PreHypo's directional architecture.

**What each channel does.** The deterministic signature is on attempt rate: $Q^+$-only drops Att from 0.55 to 0.44 at DocM 0.99, with all five folds below Full's mean. $Q^+$ embeddings index *outgoing-dependency* questions, so a pool drawn exclusively from $Q^+$ forces the synthesis path into "insufficient evidence." $Q^-$-only matches Full on Att ($0.55 \pm 0.03$) and is within fold variance on Judge (0.30 vs. 0.32); paired with the $Q^+$-only drop in Att, this locates the answer-bearing retrieval signal in $Q^-$ (visualized per-metric in Figure 3).

**Annotation quality vs. directional architecture.** The Combined control uses the same $Q^-$/$Q^+$ annotations as HopRAG uses its pseudo-queries—a single non-directional hypothetical channel—yet beats HopRAG by ∼9pp (0.28 vs. 0.19), locating most of PreHypo's advantage in $Q^-$/$Q^+$ *annotation quality* rather than the directional architecture. The directional Stage-1/Stage-2 weighting adds the remaining margin to Full (0.32, +0.04 over Combined), but that increment is *not* statistically resolved at $n$=150 (paired bootstrap 95% CI $[-0.053, +0.133]$, $p$=0.35), so we treat it as a refinement, not a core module.

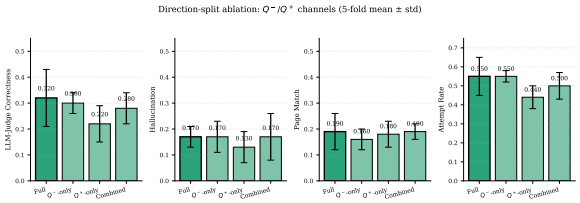


Figure 3: **Channel ablation as a mechanism check** (5-fold mean ± std, $n$=150; per-metric view of Table 3). The cleanly separable signal is on attempt rate (rightmost panel): $Q^+$-only drops Att from 0.55 to 0.44, consistent with $Q^+$ indexing outgoing-dependency questions; $Q^-$-only matches Full, consistent with $Q^-$ holding the answer-bearing signal. Judge (LLM-judged correctness) and PgM overlap within fold variance across the three variants—the end-to-end effect of the directional Stage-1/Stage-2 weighting is not resolved at $n$=150.

## H Per-Fold Variance and Pareto View

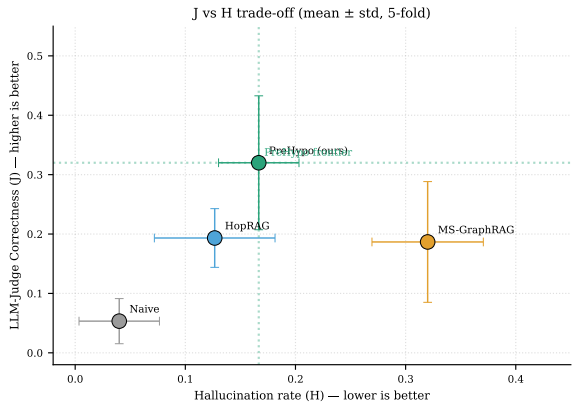


Figure 4: **Quality–faithfulness trade-off on FinanceBench** ($n$=150, 5-fold mean ± std). $x$-axis: hallucination rate (H, fraction of substantive answers flagged factually incorrect by the LLM judge, lower is better); $y$-axis: LLM-judged correctness (Judge score J, higher is better). Both labels are produced by the same judge (`gpt-5.5`); abstentions score $H$=0 and $J$=0. Up-and-left is better. PreHypo sits on the Pareto frontier—no other system has both higher J and lower H. Naive's low H reflects near-total failure to answer correctly (J=0.05), not faithfulness.

Figure 5 shows per-fold judge scores for the four systems.

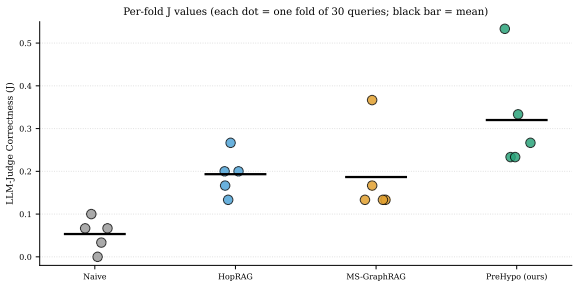


Figure 5: Per-fold judge scores. Each dot is one fold of 30 queries; the bar is the cell mean. Within-cell fold range is comparable to between-cell mean gaps for some pairings (e.g., MS-GraphRAG vs. PreHypo on a single fold), which is why we report 5-fold mean ± std throughout and treat PgM (Figure 2) and Judge as co-primary.

## I Supplementary Data Tables

Tables 6 and 7 support the defensive analyses in Appendix J: paired bootstrap CIs (Judge, conditional-Judge, and Bonferroni-corrected) and cross-document coverage of the retrieved set.

| Comparison | J | J \| Att | Bonf. |
|---|---|---|---|
| PreHypo − HopRAG | .027/.227 | .041/.355 | .007/.247 |
| PreHypo − MS-GR | .047/.220 | .230/.472 | .027/.240 |
| PreHypo − Naive | .180/.353 | .368/.619 | .160/.373 |

Table 6: Query-level paired bootstrap CIs ($B$=10,000, $n$=150, seed=42), reported as *lower/upper* bounds. J: unadjusted Judge-gap 95% CI; J | Att: conditional-Judge bootstrap (attempted queries only); Bonf.: Bonferroni-corrected 98.33% CI on the Judge gap ($\alpha$=0.05, three-comparison family). All intervals exclude zero. MS-GR = MS-GraphRAG.

| Metric | PreHypo | HopRAG | MS-GR | Naive |
|---|---|---|---|---|
| Avg retrieved-set size | 7.97 | 10.00 | 4.01 | 5.00 |
| Avg unique docs | 4.25 | 5.76 | 3.34 | 3.77 |
| $\geq$ 2 docs rate | 92.0% | 97.3% | 96.0% | 93.3% |
| $\geq$ 3 docs rate | 80.0% | 94.0% | 83.9% | 84.0% |
| Multi-doc\|DocM=1 | 91.9% | 97.3% | 96.5% | 93.3% |

Table 7: Cross-document coverage of the retrieved set across all 150 queries. All four systems reach multi-document spread without per-query graph traversal; the differentiator (§4.2) is which documents are retrieved (page-level grounding, Table 1).

## J Main Results: Detailed Analysis

This appendix expands §4.2 with the defensive analyses, the PgM residual attribution, and the per-baseline characterizations omitted from the main body.

**Paired bootstrap CIs on the Judge gap.** Fold std reflects partition variance over $n$=30 subsamples, not the standard error of the mean. A query-level paired bootstrap on Judge ($B$=10,000, $n$=150, `seed=42`) yields 95% CIs that exclude zero on all three PreHypo–baseline pairings; the HopRAG and MS-GraphRAG lower bounds (2.7 and 4.7pp) sit close to zero, giving resolved-but-tight separations. Full CIs and the Bonferroni-corrected family-wise intervals are in Appendix I.

**Conditional precision rules out abstention games.** Among attempted answers, PreHypo's Judge score is 0.57 vs. $\leq$ 0.39 for every baseline (Table 1); a query-level bootstrap on the conditional gap places all three 95% CIs above zero (Appendix I). PreHypo and HopRAG abstain at close rates (0.55 vs. 0.50), so the ∼20pp gap is over comparable answer budgets, not differing abstention thresholds.

**Multi-hop dependency is resolved offline, not at query time.** A natural concern for PreHypo's deterministic 1-hop path is whether one hop recovers what runtime multi-hop expansion provides; multi-document spread itself is not the distinguishing axis. All four systems reach two or more distinct documents on $\geq$ 92% of queries and three or more on $\geq$ 80% (per-system breakdown in Appendix I); multi-document retrieval is achievable without any graph traversal, including by Naive top-$k$ embedding lookup. What differs is *which* sub-document evidence is recovered: DocM is $\geq$ 0.89 for every system, but at the page level (§4.2) PreHypo reaches 0.19 while every baseline is at $\leq$ 0.04—the largest *relative* gap in Table 1, though the absolute level remains low. HopRAG's per-query multi-hop expansion—LLM-guided exploration over pre-built pseudo-query edges—becomes in PreHypo a cross-encoder pre-scoring of edges at indexing time; the query path then traverses pre-scored edges with no LLM reasoning at retrieval (latency ≈7s vs. ≈163s).

**PgM residual attributable to retrieval-side modules.** The PgM gap is not merely a chunking effect: even with PreHypo's cross-page semantic grouping disabled (the −Chunking variant of Table 5: Level-1 grouping off, Level-2 sentence-adaptive splitting retained), PreHypo still reaches PgM 0.11, exceeding HopRAG's 0.04 by ∼7pp on the same metric where this grouping has been removed. This residual is attributable to retrieval-and-indexing-side choices (dual hypothetical-query co-indexing, $Q^-$/$Q^+$ directional weighting, cross-encoder pre-scored HOP edges, two-stage hybrid retrieve); §F attributes most of the 13pp PreHypo–HopRAG Judge gap to per-chunk dual co-indexing and pre-scored HOP edges (a HopRAG-style single-channel use of the same annotations already covers ∼9pp; Appendix G) rather than to the directional weighting.

**Judge gap is resolvable under query-level pairing despite fold-level overlap.** Fold std reaches 0.11 on PreHypo's Judge cell—the largest of the four—and the per-fold strip plot (Appendix H) shows individual folds where the gap narrows considerably. The paired bootstrap resolves the fold-level overlap: at the query level the 95% CIs for the PreHypo–baseline Judge gap exclude zero on every pairing. We treat PgM and Judge as co-primary: PgM has the largest gap relative to fold variance, and the Judge gap is statistically resolvable under query-level paired bootstrap.

**HopRAG's hallucination rate is competitive.** HopRAG reaches H 0.13, within fold variance of PreHypo's 0.17. The two systems use related hypothetical-query mechanisms; the differences sit in (a) PreHypo's $Q^-$/$Q^+$ direction split, (b) PreHypo's offline HOP edge construction, and (c) PreHypo's chunking algorithm. The HopRAG gap on Judge (0.19 vs. 0.32) is the part our ablation traces primarily to chunking and dual co-indexing (§F); table-to-text's measured effect on this benchmark is on context compactness rather than on quality.

**MS-GraphRAG over-attempts.** MS-GraphRAG's Att 0.80 is the highest of the four systems and its H 0.32 is the highest as well. The two go together: the community-report retrieval path produces narrative completions even when the underlying context is sparse, which the judge flags as hallucinated. This is a property of the baseline architecture we did not modify.

**Rolling-summary ablation: effect is mixed.** Disabling rolling summary (Table 5, −Summary) drops PgM by 6pp and Judge by 3pp; H moves by 1pp (within fold variance). No single metric dominates, and the J and PgM drops sit at the boundary of fold variance on individual folds. We report the row for completeness rather than as a primary finding; rolling summary plausibly helps cross-chunk context bridging in some cases and adds noise in others.