FinReflectKG - MultiHop: Financial QA Benchmark for Reasoning with Knowledge Graph Evidence

Abhinav Arun
Domyn
New York, US
abhinav.arun@domyn.com

Reetu Raj Harsh
Domyn
India
reeturaj.harsh@domyn.com

Bhaskarjit Sarmah
Domyn
Gurugram, India
bhaskarjit.sarmah@domyn.com

Stefano Pasquali
Domyn
New York, US
stefano.pasquali@domyn.com

Abstract

Multi-hop reasoning over financial disclosures is often a retrieval problem before it becomes a reasoning or generation problem: relevant facts are dispersed across sections, filings, companies, and years, and LLMs often expend excessive tokens navigating noisy context. Without precise Knowledge Graph (KG)-guided selection of relevant context, even strong reasoning models either fail to answer or consume excessive tokens, whereas KG-linked evidence enables models to focus their reasoning on composing already retrieved facts. We present **FinReflectKG - MultiHop**, a benchmark built on FinReflectKG, a temporally indexed financial KG that links audited triples to source chunks from S&P 100 filings (2022-2024). Mining frequent 2-3 hop subgraph patterns across sectors (via GICS taxonomy), we generate financial analyst style questions with exact supporting evidence from the KG. A two-phase pipeline first creates QA pairs via pattern-specific prompts, followed by a multi-criteria quality control evaluation to ensure QA validity. We then evaluate three controlled retrieval scenarios: (S1) precise KG-linked paths; (S2) text-only page windows centered on relevant text spans; and (S3) relevant page windows with randomizations & distractors. Across both reasoning and non-reasoning models, KG-guided precise retrieval yields substantial gains on the **FinReflectKG - MultiHop** QA benchmark dataset, boosting correctness scores by $\sim 24\%$ while reducing token utilization by $\sim 84.5\%$ compared to the page-window setting, which reflects the traditional vector retrieval paradigm. Spanning intra-document, inter-year, and cross-company scopes, our work underscores the pivotal role of knowledge graphs in efficiently connecting evidence for multi-hop financial QA. We also release a curated subset of the benchmark (555 QA Pairs) to catalyze further research.
Dataset Link: **FinReflectKG - MultiHop QA Benchmark**

1 Introduction & Motivation

Despite recent advances in LLM-powered question answering, reliably handling multi-hop financial queries remains challenging due to evidence dispersed across temporal, structural, and organizational boundaries within SEC filings. Existing benchmarks either focus on single-document reasoning [1, 2], lack systematic multi-hop categorization [3], or overlook finance-specific requirements like temporal dependencies and regulatory semantics [4, 5]. To address these gaps, we present **FinReflectKG - MultiHop**, built atop FinReflectKG [6], a temporally indexed, source attributed knowledge graph from S&P 100 filings (2022-2024). We derive analyst-style, multi-hop questions from common 2-3 hop KG patterns to test our central hypothesis: KG-linked evidence paths outper-

form extended text-only context windows by enabling precise retrieval and reducing token overhead. Each question is paired with three controlled evidence protocols: (S1) exact KG-linked facts; (S2) relevant text-windowed context; and (S3) relevant windowed context with distractors. A two phase prompt driven pipeline along with defined rubrics & expert validation ensures question quality across multi-hop fidelity and temporal precision.

Key contributions:

- A KG-based multi-hop QA benchmark over 10-K filings of S&P 100 companies, spanning intra-document, inter-year, and cross-company relationships.
- Controlled evaluation contrasting KG-linked evidence against plain and distractor-augmented contexts which reflect realistic semantic search retrieval scenarios.
- Baseline results showing accuracy and efficiency gains from structured evidence retrieval.
- Empirical results highlighting that while reasoning models are effective, KG improves correctness by $\sim 24\%$ and reduces token usage by $\sim 84.5\%$ for multi-hop QA (Table 1).
- Release of a curated subset of the benchmark dataset containing top 555 QA pairs to facilitate reproducible research and community-wide evaluation.

2 Related Work

- **Financial Question Answering** FinQA [1] introduced 8,281 numerical reasoning pairs with program supervision but limited scope to single documents. ConvFinQA [2] extended this to 3,892 conversational dialogues while maintaining single-document focus. FinanceBench [3] broadened coverage with 10,231 questions but lacks systematic multi-hop evaluation and has limited public release. Recent work by Sahai et al. [5] showed RAG improvements but with limited scale (18 reports, 111 questions).
- **Multi-Hop Reasoning** General domain benchmarks include FanOutQA [4] (1,034 Wikipedia questions) and MultiHop-RAG [7] (news articles), both missing finance-specific temporal dependencies and regulatory semantics. BioHopR [8] demonstrated domain-specific multi-hop challenges in biomedicine but without financial applicability. Earlier approaches like QUEST [9] and reinforcement learning methods [10], established early foundations but predate modern LLM capabilities and QA requirements.
- **Research Gaps** Existing work lacks: (1) systematic temporal financial reasoning across periods and firms; (2) controlled evaluations contrasting KG retrieval with semantic retrieval reflecting realistic retrieval challenges; (3) a structured taxonomy of multi-hop scopes; (4) quality control and validation with audit-trail provenance using KGs. **FinReflectKG - MultiHop** addresses these gaps through a multi-phase methodology that combines pattern-based generation, systematic evaluation, and realistic retrieval simulation with similarity-based chunk ordering. We build directly on FinReflectKG [6], an agentic framework for constructing temporally indexed, source attributed financial KGs from SEC 10-K filings, which forms the foundation for multi-hop question generation and evidence grounding.

3 Dataset Design: FinReflectKG - MultiHop

3.1 Question & Answer Generation

We generate financial analyst-style QA pairs from frequent 2-3 hop subgraph patterns in FinReflectKG, capturing segmented and temporal relations. To ensure meaningfulness and domain coverage, we iteratively construct QA pairs across sectors using the GICS taxonomy (e.g., Financials, Information Technology, Energy), which encourages interconnected queries that require both retrieval and reasoning. A two-phase pipeline produces questions: (1) pattern-specific prompts transform KG paths into conversational queries with financial terminology and temporal context; (2) a quality-control rubric scores each question on five criteria (financial analyst-like, multi-hop fidelity, groundedness, relevance, expertise) out of 50, retaining items scoring above 40. We target coverage across 2-hop (52%) and 3-hop (48%) patterns, and across three evidence scopes: intra-document (48.7%), inter-year & same company (41.6%), and cross-company & same year (9.7%). Each question is paired with three evidence contexts for experiments: (a) precise KG-linked chunks, (b) text windows

centered on relevant sections simulating semantic retrieval with high precision, and (c) randomized, semantically similar windows with distractors simulating semantic retrieval with lower precision.

3.2 Annotation Schema & Process

Each QA instance includes: (i) the question; (ii) gold answer; (iii) the supporting KG path with document attribution (file, page, chunk, triplets); (iv) evidence packs for all the contexts with metadata; and (v) segmentation metadata (hop count, document relation scope).

All pairs undergo automatic screening using the quality rubric scores, with only those scoring $\geq$40/50 retained. Manual review is in progress via a labeling tool. (See Appendix, Figs. 1, 2). **Note:** The manually verified MultiHop QA benchmark release will follow soon after expert validation.

4 Experimental Design & Evaluation Protocol

We evaluate multi-hop QA over financial filings along three document relationships: *intra-document* (within a single filing), *inter-year* (same company across years), and *cross-company* (different companies in the same year). Each question is evaluated under three evidence regimes: (i) **KG-linked minimal evidence** - exact interconnected chunks from the knowledge graph; (ii) **page-window evidence** - a $\pm$5 page window around relevant chunks for each source, with deduplication and source tagging to approximate vector/semantic search with accurate retrieval; and (iii) **distractor-augmented evidence** - page windows mixed with random irrelevant pages or semantically similar but non-answering chunks, capturing retrieval noise typical in real-world RAG pipelines.

Experiments are run on two families of open-source LLMs: Qwen3 (8B, 32B) and OpenAI GPT OSS (20B, 120B) with "high" reasoning, deployed in our private cloud for cost-efficiency and reproducibility. To probe reasoning robustness, we also test non-reasoning Qwen variants augmented with *Think step-by-step* prompting. All prompts and decoding parameters are held fixed across runs.

Evaluation Metrics. For each generated answer, we report: (i) **LLM-as-a-Judge** (Qwen3-235B): correctness score as 0-10, with higher values indicating closer alignment with the gold answer; (ii) **BERTScore** (using *microsoft/deberta-xlarge-mnli* [11]) for semantic similarity, we report F1 score as it balances both precision & recall; (iii) **Input Tokens** for tokens utilized in prompt & context; and (iv) **Compl. Tokens** which includes output completion tokens as a measure of resource efficiency.

Results are also stratified by document relationship type in Table 3. The *KG-linked minimal evidence* setting serves as the primary baseline, with windowed and distractor contexts evaluated relative to it. For this study, we evaluate a representative subset of top **150 QA pairs** selected based on quality rubric score and spanning the two GICS sectors (Financials, Information Technology). This subset was sufficient to capture the benchmark's reasoning patterns and illustrate our main findings.

5 Experimental Results & Analysis

Model	Evidence Mode	LLM-Judge ↑	BERTScore ↑	Input Tokens ↓	Compl. Tokens ↓
GPT-OSS-120B	KG-linked	8.09	0.66	1967	1192
	Page-window	7.12	0.60	12414	1724
	Window+Distract	6.97	0.60	17873	1834
GPT-OSS-20B	KG-linked	7.75	0.54	1967	2478
	Page-window	6.46	0.52	12451	2523
	Window+Distract	6.69	0.51	17803	2389
Qwen3-32B	KG-linked	**8.23**	**0.71**	2069	703
	Page-window	6.59	0.66	13602	965
	Window+Distract	6.82	0.65	19182	999
Qwen3-8B	KG-linked	8.03	0.70	2069	814
	Page-window	5.77	0.66	13601	988
	Window+Distract	5.81	0.66	19172	1017

Table 1: Multi-hop QA performance by reasoning model and evidence retrieval mode.

Table 1 reports multi-hop QA performance across four reasoning models (GPT-OSS 120B/20B and Qwen3 32B/8B) under three evidence retrieval modes: KG-linked, Page-window, and Window+Distract. Metrics include correctness (LLM-Judge), semantic similarity (BERTScore), input tokens (retrieval efficiency), and compl. tokens (generation efficiency). **Across all models, KG-linked evidence consistently yields the highest LLM-Judge and BERTScore values, demonstrating that structured KG-linked evidence reduces retrieval noise and enables more accurate reasoning for answering financial multihop questions. On average, KG-linked improves LLM-Judge score by $\sim 24\%$ over Page-window while using $\sim 84.5\%$ fewer input tokens**, detailed summary in Appendix Tables 5, 6. The improvement is more pronounced for smaller models, emphasizing the importance of efficient retrieval for limited-capacity LLMs. Page-window and Window+Distract modes show lower correctness scores. The relatively small differences in BERTScore, indicates that these models capture semantics but struggle with precise answer generation when retrieval is noisy. Additionally, reasoning models effectively filter distractor information, demonstrating robustness to irrelevant evidence while maintaining high completion quality.

Model	Evidence Mode	LLM-Judge ↑	BERTScore ↑	Input Tokens ↓	Compl. Tokens ↓
Qwen3-32B	Reasoning	8.23	0.71	2069	703
Qwen3-32B	Non-Reasoning	7.69	0.72	2074	135
Qwen3-8B	Reasoning	8.03	0.70	2069	814
Qwen3-8B	Non-Reasoning	6.74	0.70	2073	107

Table 2: Qwen model comparison (32B vs 8B) in the KG-linked setting, with/without reasoning. Additional comparisons are shown in Table 4 (Appendix).

Comparing reasoning versus non-reasoning variants (Tables 2, 4), reasoning models generally outperform their non-reasoning counterparts. This underscores the importance of reasoning capability for multi-hop financial QA, especially when combining dispersed evidence.

Tables 1 and 2 support our central hypothesis that KG-guided retrieval enhances LLM performance on multi-hop financial QA. By filtering noise and grounding evidence in key financial disclosures, KG-linked inputs enable reasoning models to deliver higher correctness and semantic fidelity.

Doc Rel	LLM-Judge ↑	Input Tokens ↓	Compl. Tokens ↓
cross-company	7.12	13,560	1,704
inter-year	6.72	10,309	1,540
intra-document	**7.47**	11,075	1,330

Table 3: Aggregated results averaged across all reasoning models (GPT-OSS and Qwen3-Reasoning). Detailed comparisons are shown in Table 7 (Appendix).

Table 3 shows that intra-document questions achieve the highest correctness, with inter-year settings proving most challenging for models. Cross-company queries show moderate accuracy, likely due to aligned financial contexts across firms. Table 7 (Appendix) includes a detailed analysis.

6 Conclusion & Future Work

We introduce **FinReflectKG - MultiHop**, a benchmark for multihop financial question answering and show that KG-guided, minimally noisy evidence consistently outperforms page-window style retrieval in both correctness and token efficiency across open-source reasoning/non-reasoning models (enhancing correctness scores by $\sim 24\%$ while reducing token utilization by $\sim 84.5\%$). By simulating realistic scenarios (including cross-company and inter-year queries) and releasing a curated subset of 555 multi-hop QA pairs, we provide a testbed for grounded financial reasoning. While our current evaluation uses Qwen3-235B as the LLM-as-a-Judge and may introduce model-family bias, we plan to diversify & calibrate the LLM Judge (including proprietary models), broaden the model coverage to test the prowess of closed-source reasoning models, expand dataset coverage by increasing cross-company questions, and expand expert audits to deliver a larger, manually validated benchmark as part of future work. We hope **FinReflectKG - MultiHop** catalyzes research toward trustworthy, interpretable, and cost-efficient financial QA grounded in structured knowledge.

References

[1] 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. 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, 2021.

[2] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, 2022.

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

[4] Zhiyu Zhu, Yifeng Wang, Chao-Chun Hsu, Yuan-Kuei Wu, and Qingyu Zhou. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–10, 2024.

[5] Siddharth Sahai, Akshay Chaturvedi, and Rushang Karia. Multi-document financial question answering using llms. *arXiv preprint arXiv:2411.07264*, 2024.

[6] Abhinav Arun, Fabrizio Dimino, Tejas Prakash Agarwal, Bhaskarjit Sarmah, and Stefano Pasquali. Finreflectkg: Agentic construction and evaluation of financial knowledge graphs, 2025. URL `https://arxiv.org/abs/2508.17906`. Dataset available at: `https://huggingface.co/datasets/iGeniusAI/FinReflectKG`.

[7] Yixuan Tang, Yi Yang, Ahmed Masry, Mor Geva, Dipanjan Das, Avijit Mitra, Dragomir Radev, and Ben Zhou. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

[8] Yash Sharma, Shubham Mehta, Kartik Sachdeva, and Srinivasan Parthasarathy. Biohopr: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[9] Xiaolu Lu, Soumajit Pramanik, Rishiraj Saha Roy, Abdalghani Abujabal, Yafang Wang, and Gerhard Weikum. Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114, 2019.

[10] Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3243–3253, 2018.

[11] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL `https://arxiv.org/abs/2006.03654`.

7 Appendix

We provide additional comparison between reasoning & non reasoning Qwen models for the Page-window evidence mode in Table 4.

Model	Evidence Mode	LLM-Judge ↑	BERTScore ↑	Input Tokens ↓	Compl. Tokens ↓
Qwen3-32B	Reasoning	6.59	0.66	13602	965
Qwen3-32B	Non-Reasoning	5.76	0.66	13605	147
Qwen3-8B	Reasoning	5.77	0.65	13601	988
Qwen3-8B	Non-Reasoning	4.68	0.62	13606	76

Table 4: Qwen model comparison (32B vs 8B) in the Page-window setting, with/without reasoning.

Tables 5, 6 below summarize the average improvements in LLM-Judge correctness score and input token savings when using KG-linked evidence compared to Page-window during multihopQA (inference), derived from the detailed results in Table 1

Model	KG-linked	Page-window	% Improvement
GPT-OSS-120B	8.09	7.12	13.6%
GPT-OSS-20B	7.75	6.46	20.0%
Qwen3-32B	8.23	6.59	24.9%
Qwen3-8B	8.03	5.77	39.2%

Table 5: Average correctness improvement of KG-linked over Page-window mode across models.

Model	Page-window	KG-linked	% Savings
GPT-OSS-120B	12414	1967	84.2%
GPT-OSS-20B	12451	1967	84.2%
Qwen3-32B	13602	2069	84.8%
Qwen3-8B	13601	2069	84.8%

Table 6: Input token savings of KG-linked over Page-window mode across models.

Model	Doc Rel	LLM-Judge ↑	Input Tokens ↓	Compl. Tokens ↓
GPT-OSS-120B-Reasoning	cross-company	7.71	13440	2083
	inter-year	7.16	9810	1635
	intra-document	7.89	10721	1372
GPT-OSS-20B-Reasoning	cross-company	7.03	13132	2855
	inter-year	6.53	9873	2543
	intra-document	7.49	10709	2268
Qwen3-32B-Reasoning	cross-company	7.12	14305	904
	inter-year	7.02	10824	950
	intra-document	7.55	11467	835
Qwen3-8B-Reasoning	cross-company	6.62	14365	972
	inter-year	6.17	10729	1030
	intra-document	6.93	11401	845
Qwen3-32B-Non-Reasoning	cross-company	6.45	9995	146
	inter-year	6.50	7238	146
	intra-document	7.10	7751	135
Qwen3-8B-Non-Reasoning	cross-company	5.64	9995	99
	inter-year	5.27	7207	95
	intra-document	6.19	7753	87

Table 7: Detailed results across document relationship categories.

From Table 7, we observe a clear trend across all model families, both reasoning and non-reasoning. Intra-document questions consistently achieve the highest correctness scores, which is intuitive since all relevant evidence resides within a single document, reducing the need for complex reasoning or long-range context tracking. In contrast, inter-document settings pose greater difficulty. Between the two inter-document cases, cross-company questions yield higher correctness scores than inter-year questions. This may be explained by the fact that cross-company comparisons often involve semantically aligned contexts (e.g., the same financial metric reported across different firms in the same year), whereas inter-year comparisons require temporal alignment and reasoning over potentially evolving terminology, financial structures, or reporting practices, making them harder for models to resolve reliably. Additionally, improved coverage of cross-company questions in the future dataset will provide greater clarity.

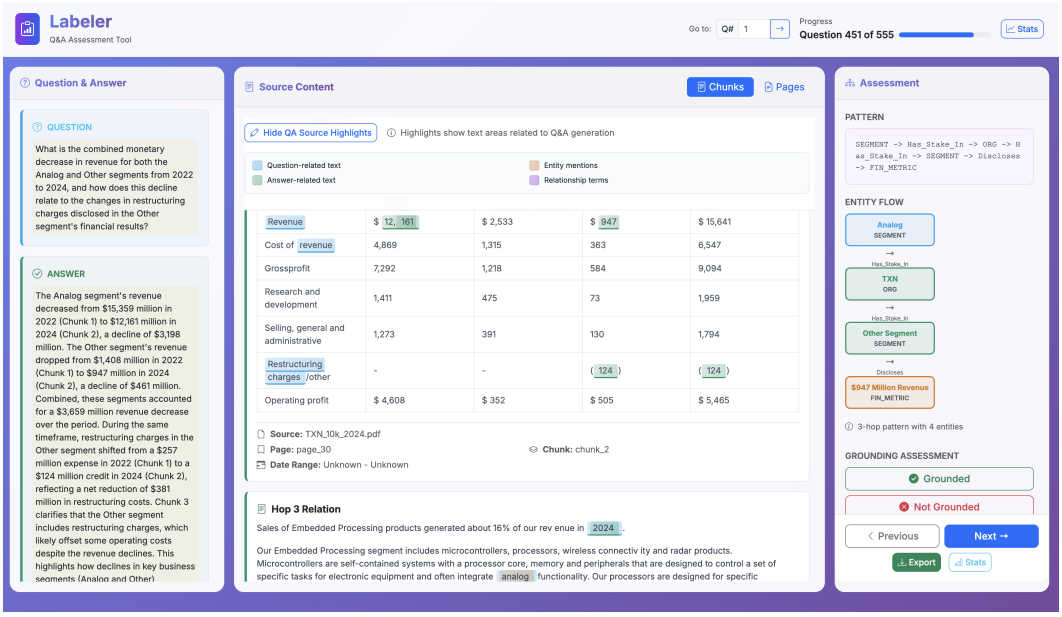

Figure 1: Interactive labeling interface used for manual verification of reliability, groundedness, and relevance in multi-hop financial QA. The interface highlights how evidence is linked across multiple segments and disclosures, also requiring reasoning over financial relationships for a sample intra-document multihop question.

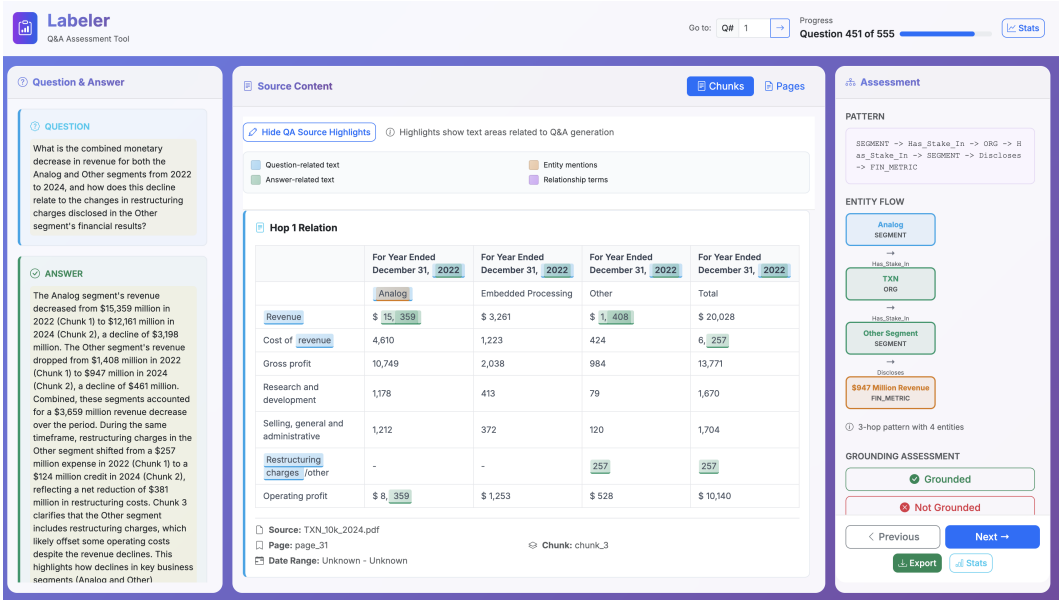

Figure 2: Additional view of the labeling tool showing source content, extracted triples, and reasoning patterns for connected context as in Figure 1. This illustrates the ongoing effort to build a cleaner, larger, and manually verified multi-hop financial QA dataset.