Explainable & Faithful RAG for Financial QA: Citation-Disciplined Answers with Span-Level Verification

Anonymous ACL submission

Abstract

We propose a retrieval-augmented conversational assistant for finance that generates short, citation-disciplined answers grounded in SEC filings and FOMC texts. Our system enforces *attribution* via inline references and employs a *span-level verification* layer using natural language inference (NLI) to reduce hallucinations. We will evaluate answer correctness, faithfulness to evidence, attribution precision/recall, and refusal behavior when evidence is insufficient. We outline a realistic plan, datasets, baselines, ablations, and a midterm milestone.

1 Introduction & Motivation

Financial QA assistants must be *factual, attributable, and conservative* when evidence is weak. Retrieval-Augmented Generation (RAG; Lewis et al., 2020) improves factuality but often fails to (i) attach precise citations to atomic claims and (ii) decline when support is missing. In high-stakes domains (filings, earnings calls, central bank statements), these gaps erode trust. We target a modeling-first solution (no heavy serving work): (1) disciplined citation formatting; (2) automatic claim–evidence linking; (3) NLI-based support checks; (4) refusal or revision when contradictions/insufficiency are detected.

2 Relevant Literature & Key Takeaways

Retrieval-augmented generation. Lewis et al. (2020) demonstrated end-to-end RAG, while Izacard and Grave (2021a) showed that fusion-in-decoder improves evidence integration. We will adopt a modern dense retriever + reranker (e.g., Izacard and Grave, 2021b; Xiao et al., 2023; Nogueira and Lin, 2020) to raise recall and evidence quality.

Faithfulness, attribution, and hallucinations. Rashkin et al. (2023) formalize *Attributable to Identified Sources*, directly relevant to citation discipline. Manakul et al. (2023) and Min et al. (2023) present automatic faithfulness checks; we will operationalize faithfulness via sentence-level NLI on claim–evidence pairs (e.g., He et al., 2021). These works guide our metric design and refusal policy.

Financial QA datasets. Chen et al. (2021) (FinQA) and Zhu et al. (2021) (TAT-QA) capture numeric reasoning over reports/tables. We will start with text spans (sec. 4.1) and add light numeric grounding as a stretch. For multi-hop reasoning and evaluation ideas, we borrow structure from Yang et al. (2018).

What we adopt. From RAG/FiD: strong retrieval and compact evidence packing. From attribution work: explicit source-linked claims and conservative refusal. From faithfulness metrics: span-level NLI to quantify support and contradictions.

3 Planned Contribution & Innovation

Our novelty lies in **tying answer generation to verifiable spans**:

1. **Citation discipline.** Answers must include inline refs [1][2] that map to an evidence list; we constrain decoding to *only* use retrieved text.

2. **Span-level verification.** We decompose answers into atomic claims (simple clause splitter), align each to candidate sentences from cited passages, and apply NLI (*entail/contradict/neutral*). Unsupported claims trigger revision or refusal.

3. **Refusal policy.** If no sentence entails a claim, the system returns “Insufficient evidence” and suggests what evidence would be needed (e.g., guidance section, period, or table).

4. **(Stretch) Numeric grounding.** Regex/unit-aware copy of numbers from cited spans to

curb numeric hallucinations.

4 Project Plan

4.1 Data & Acquisition

Primary corpora: (i) **SEC filings** (10-K/10-Q) for 5–8 companies (150–250 docs total), parsed to text and chunked (300–500 tokens, 20–40% overlap). (ii) **FOMC statements/minutes** (3–5 years).
QA sets: (1) Curate 100–150 short factoid questions with gold answers and doc IDs; (2) import a text-only subset of **FinQA** and **TAT-QA** for external validation (no tables needed initially).

4.2 Models

Retrieval: Dense embeddings (e.g., `bge-base/large`) with FAISS; optional cross-encoder reranker (`bge-reranker` or MonoT5).
Generator: Open 7B–9B instruct model (Llama-3.1-8B, Mistral-7B, or Gemma-2-9B-it).
Verifier: NLI model (e.g., DeBERTa-v3-Large-MNLI) for claim–span entailment.
(Optional) PEFT: LoRA for citation obedience and terse style if needed.

4.3 Method

Pipeline: retrieve top-k $\rightarrow$ rerank top-m $\rightarrow$ pack evidence list $\rightarrow$ generate concise answer with inline refs $\rightarrow$ split into claims $\rightarrow$ select candidate evidence sentence per claim (BM25-over-sentences or highest sim) $\rightarrow$ NLI verify $\rightarrow$ (if unsupported) revise or refuse.

4.4 Evaluation

Automatic metrics:

- **Answer accuracy:** EM/F1 against gold.
- **Faithfulness:** % claims *entailed* by cited spans (NLI).
- **Attribution P/R:** precision/recall of citation $\leftrightarrow$ claim alignment (does cited span actually support the tagged claim?).
- **Hallucination rate:** % answers with any unsupported claim.
- **Refusal quality:** precision of "Insufficient evidence" (no entailed spans exist) and false refusal rate.

Human eval (50–100 items): helpfulness, specificity, and citation adequacy on 3-point scales.

4.5 Baselines & Ablations

Baselines: B0: No-RAG LLM; B1: Vanilla RAG (dense only); B2: RAG+Reranker.
Our method: RAG+Reranker+Citation discipline+Span-level NLI verification (with refusal).
Ablations: chunk size/overlap; top-k/m; prompt variants (with/without refusal rule); sentence- vs. paragraph-level evidence packing; verifier on/off; numeric-copy on/off.

4.6 Milestone (Halfway Checkpoint)

By the midpoint:

- Ingested $\geq$150 filings and 3–5 years of FOMC texts; FAISS index built.
- B0/B1/B2 implemented; first pass EM/F1 and retrieval recall.
- Citation formatting functional on dev set; preliminary faithfulness script working on 50 items.

4.7 Feasibility & Risks

Data availability. SEC and FOMC texts are publicly available; collection is straightforward. We will scope to a few tickers to keep indexing/eval tractable.

Evaluation realism. We will author 100–150 QA pairs tied to specific documents and validate with a subset from FinQA/TAT-QA; metrics follow prior work, and NLI-based checks provide scalable faithfulness estimates.

Baselines & ablations. Clearly defined (above) and feasible to run on a single GPU; no serving infrastructure required.

5 Deliverables

Reproducible code (`train/eval/analyze` scripts), a concise demo notebook (question $\rightarrow$ evidence $\rightarrow$ answer with citations $\rightarrow$ faithfulness report), and a final report with tables/plots showing accuracy, faithfulness, hallucinations, and ablation deltas.

What We Hope to Learn

(1) How much of the hallucination problem in financial QA can be mitigated by disciplined evidence packing + span-level verification; (2) which retrieval/rerank configurations most affect faithfulness; (3) when refusal improves overall utility.

References

Wenhu Chen, Yufei Xie, Hongmin Zhang, Muhao Chen, and William Yang Wang. 2021. Finqa: A dataset for numerical reasoning over financial data. In *EMNLP*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Gautier Izacard and Edouard Grave. 2021a. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Gautier Izacard and Edouard Grave. 2021b. Unsupervised dense information retrieval with Contriever. In *NeurIPS*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladislav 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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sewon Min and 1 others. 2023. Factscore: Fine-grained evaluation of factual precision in long-form text generation. In *EMNLP*.

Rodrigo Nogueira and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of EMNLP*.

Hannah Rashkin and 1 others. 2023. Measuring attribution in natural language generation. In *ACL*.

Binhao Xiao, Yidong Li, and 1 others. 2023. Bge embeddings: A lightweight general text embedding model. *arXiv preprint arXiv:2307.02888*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Fengbin Zhu, Chenyan Xiong, Zhiyuan Yu, and Jiawei Han. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *ACL*.