Verifiable Financial Enterprise Question Answering via Inference-Time Grounding and Traceability

Anubha Kabra **Katie Jooyoung Kim** **Colin Zhiwei Kou**
Helene Sajer **Yimei Fan** **Gabriel Martinez Vidiri**
Bloomberg
{akabra16, jkim2425, ckou9, hsajer3, yfan258, gmartinezvi1} @bloomberg.net

Abstract

Financial enterprise AI systems deployed in high-stakes settings require responses that are verifiable, traceable, and auditable. We introduce a modular, model- and data-agnostic inference-time control framework, together with a deployment-aware evaluation strategy for verifiable financial enterprise question answering. Our method enforces faithfulness at inference time without retraining or changes to retrieval infrastructure. We deploy our method in a production financial enterprise assistant and evaluate it using a combination of intrinsic faithfulness metrics, baseline comparisons, and real-world user feedback. Our approach improves groundedness by 29% over baselines, reduces hallucinations to near-zero levels, and achieves near-perfect document-span traceability. Together, our results demonstrate that modular pipeline design combined with detailed, deployment-aware evaluation provides a practical and effective path toward verifiable financial enterprise QA systems.

Keywords: financial question answering, enterprise LLM systems, groundedness, span-level traceability, auditability, retrieval-augmented generation, faithful generation

1. Introduction

Institutions are increasingly deploying large language model (LLM) systems to support enterprise question answering across compliance, operations, product, and support functions (Huang et al., 2023; Gao et al., 2023). Prior work in financial NLP has explored domain-specific modeling and sentiment analysis for financial texts (Gao et al., 2023; Chen et al., 2024b; Huang et al., 2023), yet verifiable financial enterprise QA remains comparatively underexplored. In these high-stakes financial settings, responses must satisfy stricter standards than fluency or relevance alone; they must be verifiable, traceable to authoritative financial documents, and auditable under regulatory and operational scrutiny. Retrieval-augmented generation (RAG) pipelines, which pair LLMs with document retrievers to produce citation-backed responses, have shown strong performance on web-scale benchmarks (Kryscinski et al., 2020). However, despite their perceived interpretability, these systems often fall short of real-world verifiability and faithfulness requirements in financial enterprise QA. In regulated environments, users require not only fluent answers but guarantees that generated claims are factually supported, transparently verifiable, and robust to noisy or heterogeneous evidence sources (Chen et al., 2024b; Choubey et al., 2025).

Many failures in RAG systems arise at inference time rather than from retrieval or training deficiencies alone. During generation, LLMs must integrate evidence, synthesize claims, and assign citations under real-world constraints, yet existing pipelines offer limited support for auditing, monitoring, and intervention. As a result, deployed systems frequently produce unsupported claims, misattributed citations, or references that cannot be traced to concrete evidence spans, particularly when operating over long, unstructured enterprise documents (Joren et al., 2025; Choubey et al., 2025; Packowski et al., 2024). Such failures undermine interpretability, erode user trust, and hinder responsible deployment.

To study these challenges in practice, a month-long pilot deployment of an enterprise assistant built on a standard RAG architecture was conducted.

1.1. Pilot Deployment

The assistant used a state-of-the-art LLM[1] and retrieved the top-n documents from a heterogeneous corpus of internal sources, including wikis, policy manuals, knowledge base articles, and product documentation, using a customized retrieval backend optimized for low latency. Retrieved documents were incorporated into a task-specific prompt to generate natural language responses with inline citations.

The system was deployed for one month to 55 users across support, operations, compliance, and product roles, processing approximately 4,000 real-world queries. While users valued response fluency and relevance, the deployment revealed recurring failures such as incorrect citations, untraceable references, and unsupported claims.

Failure Analysis from Pilot Deployment

The pilot deployment revealed several recurring failure modes that limit the faithfulness of LLM-generated outputs in financial enterprise settings.

[1] Details withheld due to internal policies.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 39–48

Error Type	Description	Root Cause	Impact
Hallucinated Links	Nonexistent citations or URLs	Pattern-based generation	Erodes trust
Citation Drift	Cited passage doesn't support claim	Misaligned grounding	Reduces factual reliability
Limited Traceability	Hard to locate cited text	Buried content, weak anchors	Lowers transparency

Table 1: Key limitations of the pilot enterprise assistant deployment.

In particular, we observed the following caveats (See Table 1):

1. **Hallucinated links:** Generated citations or URLs that did not exist in the enterprise corpus.
2. **Citation drift:** Valid documents were cited but did not substantiate the associated claims.
3. **Limited traceability:** Even when citations were correct, verifying claims was difficult due to long or unstructured source documents.

To systematically analyze these failures, we characterize faithfulness along two dimensions. These are: (a) whether generated claims are factually supported by the cited source documents and (b) whether claims can be linked to specific, verifiable spans within those sources.

These observations motivate the following research questions:

RQ1: Can transparency into failure modes be incorporated into the pipeline design?

RQ2: Can we make LLMs more faithful by adding citation-level grounding and span-level traceability without retraining?

RQ3: Do improvements in LLM faithfulness translate to better downstream user satisfaction?

To address these research questions, we design and deploy `EvidenT`, a lightweight, post-hoc, model- and data-agnostic inference-time control system for extractive enterprise question answering. In light of **RQ1**, `EvidenT` adopts a modular pipeline design that enables step-wise inspection and evaluation of individual components, providing transparency into failure modes during system development. Addressing **RQ2**, the pipeline enforces citation-level groundedness and span-level traceability at inference time without requiring retraining or changes to retrieval infrastructure. This enforcement improves groundedness by 29%, achieves up to 99% span-level traceability, and reduces hallucinations to near-zero levels. With respect to **RQ3**, we introduce a detailed evaluation strategy that assesses faithfulness both prior to deployment and under real-world financial enterprise usage conditions, and observe that these targeted improvements are associated with improved downstream user satisfaction.

2. Related Work

2.1. Fine-tuning and Domain Adaptation

Fine-tuning LLMs has been widely explored to improve grounding in retrieval-augmented generation (RAG) systems (Huang et al., 2024; Penzkofer and Baumann, 2024; Zhang et al., 2024). However, such approaches require substantial domain-specific supervision, which is often impractical in financial enterprise settings. Moreover, fine-tuned models remain susceptible to hallucination or over-generalization under weak retrieval (Lee et al., 2025; Soudani et al., 2024), and primarily improve fluency rather than citation-level traceability (Ghosal et al., 2024). As a result, model-level optimization alone fails to guarantee transparent citation alignment or verifiable provenance (Ye et al., 2024; Huang et al., 2024), limiting applicability in high-stakes domains.

2.2. Basic RAG Systems

Despite their widespread use, RAG pipelines exhibit persistent failures in noisy and heterogeneous financial enterprise environments (Chen et al., 2024a). Retrieval remains a key bottleneck, as financial enterprise corpora are often fragmented, redundant, and inconsistently indexed (Sharma, 2025; Brown et al., 2025). Even when relevant documents are retrieved, *citation drift* – where cited passages do not support the generated claims – frequently occurs (Patel and Anand, 2024; Huang et al., 2024). Because retrieval and generation are loosely coupled, existing RAG systems lack explicit mechanisms to enforce span-level grounding, resulting in low citation precision and limited auditability.

2.3. LLM Faithfulness in real-world environments

While prior work has introduced citation-focused approaches and metrics to improve faithfulness in knowledge-grounded generation, for example in dialogue systems (Rashkin et al., 2021), more recent work has proposed citation-evaluation frameworks that assess citation quality in a more systematic way (Xu et al., 2025). However, much of this evaluation still works at a coarse level, often checking support only at the document or passage level. As a result, these approaches do not provide span-level source

attribution, which financial enterprise users need for fast and reliable verification. At the same time, widely used citation-generation benchmarks and setups (e.g., ALCE) are primarily grounded in public, research-style corpora and evaluation protocols (Gao et al., 2023), which can differ substantially from enterprise settings where collections are long, heterogeneous, and frequently unstructured (Anderson et al., 2024). Recent work on long-context settings further suggests that extracting and attributing evidence spans in long/unstructured inputs is itself challenging, reinforcing the need for span-level traceability beyond document-level citation correctness (Wright et al., 2025). As a result, current systems offer limited guarantees about where information comes from and do not fully support transparent, verifiable generation in real-world financial enterprise use.

3. Our Approach

Our approach consists of a model- and document-agnostic modular pipeline designed for robust citation generation in financial enterprise settings. An overview of how the pipeline processes queries is shown in Figure 1. The modular design aligns with RQ1, enabling transparent and incremental inspection of failure modes throughout pipeline development. This structure allows individual components to be analyzed and improved independently. The pipeline comprises the following components.

3.1. Document Retrieval Module

This module retrieves the top N relevant documents from heterogeneous financial enterprise data sources without relying on a centralized index. Instead, we retrieve documents independently from multiple distributed sources. Our solution is designed to be retrieval method-agnostic, enabling flexible integration with a variety of data sources.

3.2. Structured Passage Extraction Module

This module extracts candidate passages from retrieved documents using an LLM with structured output formatting. The prompt is designed to return verbatim spans from the document context. Each passage is represented in the following JSON structure:

```
{
  "passage_id": "<passage_id>",
  "url": "<url>",
  "content": "<passage_content>"
}
```

We utilize prior findings showing that LLMs achieve near-perfect performance on verbatim span extraction from long-context source documents (Hsieh et al., 2024) and structurally straightforward format transformations (Yang et al., 2026).

3.3. Source Alignment Filter Module

This module filters hallucinated passages based on n-gram overlap with their source documents. For each passage, we retrieve the full content of the source document using the provided `url` field. We then compute the n-gram overlap (typically $n = 5$) between the passage and the document content. Let p be an extracted passage and d its cited document. Define the n-gram overlap ratio as

$$\text{overlap}_n(p, d) = \frac{|G_n(p) \cap G_n(d)|}{|G_n(p)|}. \quad (1)$$

where $G_n(p)$ is the multiset of n-grams in p. We define the filtering action as follows:

- If **overlap** $>$ **threshold**, the passage is retained with the original `url`.
- If **0** $<$ **overlap** $\leq$ **threshold**, the passage is truncated to retain only the overlapping portion; the original `url` is preserved.
- If **overlap** $= 0$, we suspect citation drift. We iterate over all other retrieved documents to find one with overlap $>$ threshold. If found, the passage `url` is replaced with this.
- If none of these conditions apply, we drop the passage from the generated JSON.

To deliberately prioritize verifiability over textual abstraction, we employ an n-gram-based evaluation framework. We adopt lexical matching to accommodate the specialized finance domain, where documents are clause-driven and lexically precise. System identifiers, ticker symbols, and regulatory phrasing often carry specific operational meaning and cannot be freely paraphrased without altering intent (Kim et al., 2025; Li et al., 2025). These texts are structured and compliance-sensitive, differing substantially from open-domain text; the data is effectively out-of-distribution for semantic models. (Chen et al., 2024b; Anderson et al., 2024; Choi et al., 2025). Off-the-shelf semantic similarity models are trained on general-domain data and rely on subword tokenization, which can fragment rare financial identifiers and overlook clause-level distinctions that matter in compliance settings (Kudo and Richardson, 2018; Araci, 2019). Prior work shows that in such formulaic domains, lexical methods outperform semantic approaches (Choi et al., 2025; Thakur et al., 2021). In our early experiments, this approach showed superior performance compared to semantic metrics such as BARTScore, and also satisfied the strict latency requirements common in financial enterprise settings. In our setting,

the data distribution differs substantially from the training data of most semantic models. In practice, users can configure threshold values based on application requirements and desired alignment strictness. The framework remains extensible and can incorporate hybrid or semantic matching when domain-adapted semantic models are available.

3.4. Answer Generation Module:

A second LLM call generates the final answer using only the filtered and verified verbatim passages. Guided by a structured prompt, the LLM enforces strict grounding and formatting constraints to ensure maximal traceability while transforming the verbatim passages into a coherent, well-formed, and easily consumable answer without fragmentation. Each sentence in the final answer is cited using the associated URL fields, allowing the exact source location of the supporting content to be surfaced to the user. This design enhances user trust by enabling direct and transparent verification. In the user interface, we highlight the precise text spans corresponding to each citation in the source documents (see Figure 1). Clicking an inline citation takes the user directly to the exact location from which the referenced information originates.

3.5. Post-processing Module

The post-processing module programmatically refines the generation from the previous module to create a coherent final output to be presented to the users. The processing includes formatting paragraphs, removing repeated citation URLs, and normalizing company-specific terms to improve legibility.

3.6. Experimental Settings

We use two publicly available open-weight models of differing sizes – `LLaMA-3.1-8B` (M1) and `LLaMA-3.3-70B` (M2) – to demonstrate that our strategy is agnostic to model scale. We use open-weight models to adhere to data privacy constraints. The test set has approximately 500 financial enterprise QA data points, collected from our initial pilot study to closely reflect real user behavior. For maximal determinism, the temperature for each model call is set to 0.0.

4. Baselines

Building on the limitations in Section 2, we evaluate baselines that align with our goals of improving *faithfulness* without retraining or multi-stage orchestration.

Our design choices follow two principles. *(1) Scope alignment:* Our objective is to improve grounding and traceability in a model- and data-agnostic way; comparing with fine-tuned RAG systems (Asai et al., 2024; Lee et al., 2025) would conflate architectural complexity with our verification mechanism. *(2) Practical relevance:* Financial enterprise environments often preclude retraining or large-scale supervision due to privacy, fragmentation, and latency constraints, including strict limits on real-time LLM calls (Qian et al., 2025; Sun et al., 2024). Accordingly, we do not compare against verifier-based or self-reflective RAG systems, which rely on additional or iterative LLM calls and introduce latency, nondeterminism, and auditability challenges that are incompatible with our financial enterprise deployment constraints. We evaluate under realistic plug-and-play conditions. `EvidenT` remains complementary to advanced verifier-based systems and can be layered atop them for stronger grounding and traceability. We compare against two representative baselines:

Direct Prompting with Inline Citations: The LLM generates citations inline as part of its response, serving as a minimal citation-aware baseline without explicit verification (Singal et al., 2024; Lewis et al., 2020).

Ground-Every-Sentence: The LLM appends a citation after every sentence, ensuring each atomic statement is grounded in at least one retrieved document, enabling fine-grained evaluation of citation precision and coverage (Xia et al., 2025).

5. Evaluation

While the ultimate measure of an enterprise assistant's success lies in **user satisfaction**, such outcomes can only be reliably assessed post-deployment. During development, we therefore rely on *proxy metrics* that correlate with user trust and perceived reliability. Below, we present a detailed evaluation setup and results comparing our proposed approach with the two baselines introduced in Section 4. We report: (a) an *intrinsic evaluation*, (b) *comparative results* against baselines; and (c) *post-deployment metrics* based on real-world user feedback. Together, *(a)* and *(b)* form a practical proxy evaluation strategy for assessing groundedness, traceability, and trustworthiness during development, which can be applied to other modular pipelines like ours.

5.1. Intrinsic Evaluation (RQ1)

The pipeline is designed to support incremental evaluation based on observed failure modes, enabling targeted analysis of individual components rather than treating the system as a whole. This design allows us to isolate and address problematic

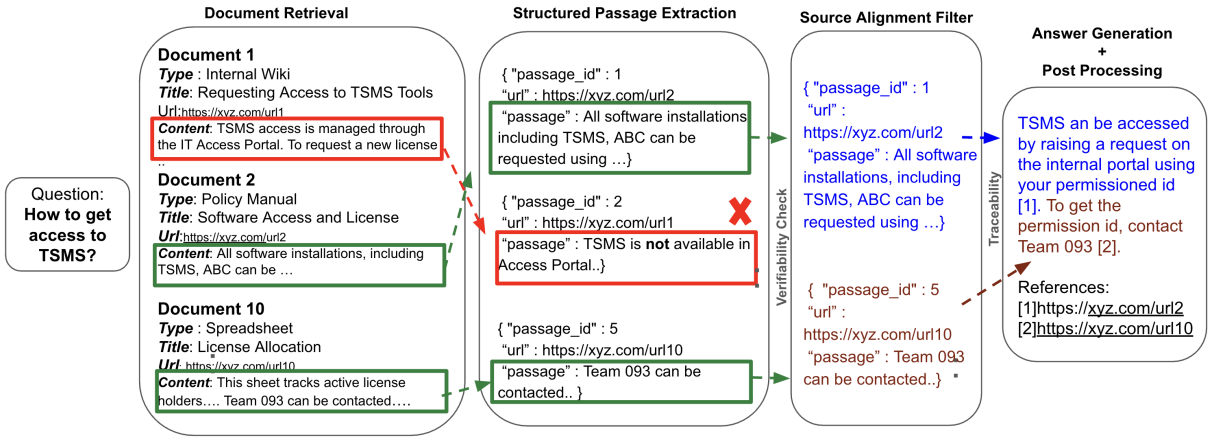

Figure 1: A step-by-step visual of our pipeline showing how a query is processed, from *document retrieval* to *answer generation* with example content that is entirely fictional and used only for illustration.

steps more effectively.

5.1.1. Evaluating Source Alignment Filter Module

The goal is to assess whether the passages generated in this module are factually aligned with the retrieved reference documents or not. To this end, we quantify two specific error categories: `URL` *drift* and *hallucinations*.

- **% citation drift:** The percentage of cases where the generated `URL` does not match the reference `URL`, yet corresponds to a valid `URL` present in the retrieved documents.
- **% hallucination:** The percentage of generated passages that contain one or more hallucinations.

As shown in Table 2, we observe a considerable amount of citation drift. However, we are able to locate the correct citation within the retrieved documents that matches the passage and substitute accordingly. The larger model demonstrates stronger factual grounding, exhibiting significantly fewer hallucinated passages. Notably, when hallucinations do occur, the initial portion of the generated response is often accurate, but as the generation continues, it gradually diverges from the source material.

5.1.2. Human Evaluation of Generated Passages

After generating the source passages (*Structured Passage Extraction Module*) and processing them through the filtering (*Source Alignment Filter*), we conducted human evaluation to assess the relevance of the passages to the original queries. Evaluating relevance required *subject-matter expertise*, a capability that current LLMs do not inherently possess.

	M1	M2
% citation drift	24.19	13.5
% hallucinated	16.4	0.6

Table 2: N-gram filter metrics for models M1 and M2.

	M1	M2
% Relevant	42	59
% Partially relevant	58	87

Table 3: Human evaluation results of passage relevance for models M1 and M2.

Table 3 presents, for each query, the number of passages that evaluators judged as fully or partially relevant, which we then averaged across all queries. Two annotators reviewed 50 queries and corresponding passages, and we compute the final score by averaging their judgments.

The smaller model M1 returns a higher share of irrelevant passages, while the larger model M2 consistently produces more fully or partially relevant results, achieving up to 87% partially relevant passages. Both models can surface relevant passages from noisy, unstructured documents, indicating the pipeline effectively identifies meaningful content despite input noise. The second LLM call in the answer generation module further chooses what to surface from these passages, adding another layer of verification to the source documents.

5.2. Evaluating Faithfulness (RQ2)

5.2.1. Evaluating Groundedness

To evaluate citation groundedness in relation to RQ2, we compute the following metrics:

%Hallucination: A binary measure indicating whether an answer contains any hallucinated citations. In financial enterprise settings like ours, even

Method	%Groundedness		%Hallucination	
	M1	M2	M1	M2
Direct Prompting	52	73	31	18
Ground Every Sentence	47	71	54	23
EvidenT	**77**	**93**	**0**	**0**

Table 4: Comparison across groundedness and hallucination metrics.

Method	SemMax		SemRecall	
	M1	M2	M1	M2
Direct Prompting	3.280	4.100	-0.390	1.930
Ground Every Sentence	4.168	4.320	0.439	2.360
EvidenT	**5.360**	**5.990**	**4.390**	**5.450**

Table 5: Semantic similarity metrics across different approaches and model variants.

a single hallucinated citation is unacceptable; therefore, an answer receives a score of 0 if it contains any citation that does not correspond to a retrieved document, and a score of 1 otherwise. Let c' denote the set of citations generated in an answer and c denote the set of all retrieved URLs. Then:

$$\text{Hallucination} = \mathbb{1}\big[\,\mathrm{c}' \subseteq \mathrm{c}\,\big] \tag{2}$$

where $\mathbb{1}[\cdot]$ is equal to 1 if all generated citations exist within the retrieved documents, and 0 if any are hallucinated.

%Groundedness: A measure of whether the model's answer is substantively supported by authoritative source documents without introducing unsupported citations. We use human-annotated URLs that are sufficient to answer each test instance. Because financial enterprise QA operates in an open-world setting where multiple documents may independently support the same answer, the gold set is sufficient but not exhaustive.

An answer is considered grounded if (i) at least one generated citation overlaps with the expert-annotated sufficient set, and (ii) no generated citation is hallucinated (i.e., all citations correspond to retrieved documents). This definition preserves the open-world assumption while enforcing citation validity, yielding a stricter and deployment-aligned measure of faithfulness.

Let c' denote the set of citations generated in the model's output and c denote the set of gold citations. We define groundedness as:

$$\text{Groundedness} = \mathbb{1}\big[\,\mathrm{c}' \cap \mathrm{c} \neq \emptyset\,\big] \tag{3}$$

where $\mathbb{1}[\cdot]$ is 1 if there is any overlap between the gold-labelled and generated citations, and 0 otherwise.

As shown in Table 4, the EvidenT approach yields a substantial improvement in factual groundedness while completely eliminating hallucinations. This improvement can be largely attributed to the *N-gram Filtering Module* (*Source Alignment Filter*), which proactively filters out hallucinated or citation-drift passages prior to generating the final response. In contrast, the baseline prompting strategies exhibit a notable degree of both hallucination and citation drift. Between the two model variants, M2 consistently outperforms M1 in both groundedness and hallucination resistance. Overall, EvidenT achieves a **29%** relative improvement in groundedness for M2, with zero instances of hallucination observed across the evaluated set.

5.2.2. Evaluating Traceability

In addition to the above, we use both semantic and lexical post-generation metrics to quantify how well an answer generated via EvidenT can be traced back to its source document at the span level pertaining to RQ2.

For all subsequent evaluations, we first extract factual statements by taking the text preceding each citation: for example, the blue and brown spans in Figure 1 illustrate two separate facts. For EvidenT, we additionally gather both the extracted facts and their corresponding source passages by matching cited URLs. These are then compared against the content of the cited documents, identified through the same cited URLs. Results are averaged across all queries.

Given a reference document D and a model-generated answer A, we quantify how traceable A is to D. We employ several methods to measure this. Let U and V denote the multisets of tokens from A and D, respectively. t denotes the token.

5.2.3. Word Overlap

Let $O = \sum_t \min(\text{count}_U(t), \text{count}_V(t))$ for each token t.

$$\text{AnsCov} = \frac{O}{|U|}, \qquad \text{DocFocus} = \frac{O}{|V|} \tag{4}$$

AnsCov measures how well the document covers the answer content, while DocFocus reflects how concentrated the document is on that answer.

5.2.4. n-gram Overlap

For $n \in \{2, 3, 5, 10\}$, let $\mathcal{G}_n(T)$ denote the multiset of n-grams in text T, and let $I_n = |\mathcal{G}_n(A) \cap \mathcal{G}_n(D)|$ be the number of overlapping n-grams between the answer A and document D. We define AnsCov@n, which measures how much of the answer's phrasing is covered by the document, and DocFocus@n, which captures how concentrated the document is on the answer content.

Method	Model	DocFocus	AnsCov	Focus@2	Cov@2	Focus@3	Cov@3	Focus@5	Cov@5	Focus@10	Cov@10
Direct Prompting	M1	0.053	0.733	0.031	0.457	0.025	0.377	0.021	0.331	0.014	0.292
	M2	0.047	0.887	0.032	0.644	0.026	0.528	0.020	0.436	0.012	0.324
Ground Every Sentence	M1	0.044	0.736	0.027	0.442	0.022	0.363	0.019	0.309	0.014	0.265
	M2	0.039	0.897	0.028	0.655	0.023	0.536	0.018	0.431	0.011	0.328
EvidenT	M1	**0.096**	**0.998**	**0.093**	**0.996**	**0.093**	**0.995**	**0.091**	**0.993**	**0.096**	**0.986**
	M2	**0.162**	**0.999**	**0.162**	**0.996**	**0.160**	**0.993**	**0.158**	**0.990**	**0.156**	**0.979**

Table 6: Answer Coverage (*AnsCov*) and Document Focus (*DocFocus*) metrics across methods and models. Higher *Coverage* indicates that a document captures more of the generated answer's content, while higher *Focus* reflects a greater proportion of the document being relevant to the answer.

$$\text{AnsCov@}n = \frac{I_n}{|\mathcal{G}_n(A)|}, \quad \text{DocFocus@}n = \frac{I_n}{|\mathcal{G}_n(D)|} \tag{5}$$

Larger n values emphasize near-verbatim phrasing and reduce tolerance for paraphrasing. Results are shown in Table 6. The high AnsCov values and low DocFocus values reflect the long and noisy nature of the source documents. EvidenT substantially improves coverage over both baselines. In particular, at $n = 10$, EvidenT achieves near-saturated scores (AnsCov: 0.999 vs. 0.897; Cov@10: 0.979 vs. 0.328), indicating that generated answers preserve long contiguous spans from the cited documents.

In compliance-sensitive financial settings, this degree of span-level alignment is operationally important: users must verify claims directly against authoritative clauses. The strong lexical coverage therefore reflects audit-grade traceability rather than incidental surface overlap.

5.2.5. Semantic Matching

For semantic scoring, we create overlapping windows (256 tokens, stride 50) for both A and D. We use the off-the-shelf sentence cross-encoder to compute similarity scores between each answer $\{a_i\}$ and each document window $\{d_j\}$. The model jointly encodes each text pair and outputs a scalar relevance score, which we use without normalization.

$$\text{SemMax} = \max_{i,j} s(a_i, d_j), \tag{6}$$

$$\text{SemRecall} = \frac{1}{|\{a_i\}|} \sum_i \max_j s(a_i, d_j) \tag{7}$$

where $s(a_i, d_j)$ denotes the raw relevance score produced by the cross-encoder. SemMax captures the strongest localized semantic match between the answer and the document, while SemRecall reflects the overall semantic coverage of A by D.

Table 5 shows that SemMax scores are consistently higher than SemRecall across all methods and model sizes. This indicates that generated answers generally contain at least some segments that align well semantically with the source documents. However, when semantic alignment is averaged across all answer segments (i.e., SemRecall), weaker methods, particularly for M1, exhibit substantially lower coverage, suggesting fragmented or inconsistent grounding. Across both model sizes, EvidenT achieves the strongest performance on both metrics. For M2, EvidenT improves SemMax by approximately 39% relative to the strongest baseline and yields more than a 100% improvement in SemRecall. Similar trends hold for M1, where gains are even more pronounced for SemRecall. These results indicate that EvidenT not only produces answers with stronger localized semantic matches, but also maintains significantly more consistent semantic alignment with the source documents overall.

5.2.6. LLM-as-a-Judge Entailment

We employ an LLM-as-a-judge by explicitly reformulating the evaluation as a fine-grained entailment task. Instead of relying on a holistic or impressionistic judgment over an entire answer, we decompose the output into individual factual statements in A. For each statement, the model assigns a ternary label of *Yes*, *Partial*, or *No*, depending on whether the cited source document fully supports, partially supports, or does not support the claim. This formulation sharply constrains the role of the LLM and eliminates much of the ambiguity inherent in broad LLM-as-a-judge setups. In this setting, the LLM effectively acts as a semantic entailment model grounded in explicit evidence, rather than as an unconstrained evaluator. We use a different off-the-shelf LLM for this assessment and report results using only the larger model (M2), given its clear advantage over M1. As shown in Table 7, EvidenT achieves the highest entailment accuracy, with substantially more fully supported (Yes) statements and the fewest unsupported (No) ones, demonstrating stronger factual alignment with cited sources than all baselines.

5.3. Post-Deployment Evaluation (RQ3)

EvidenT explicitly targets improvements in faithfulness; we therefore evaluate whether gains along this dimension are associated with changes in

Method	% Yes	% Partial	% No
Direct Prompting	59.41	**27.58**	8.22
Ground Every Sentence	56.35	27.25	11.27
EvidenT	**74.38**	23.14	**1.65**

Table 7: Entailment metrics from LLM-as-a-Judge.

Metric	Pilot	EvidenT
Helpfulness	64%	92%
Relevance	75%	89%

Table 8: Post-deployment user metrics.

user satisfaction following deployment. We analyze user feedback collected during live usage of the financial enterprise assistant before and after introducing EvidenT, focusing on two metrics: *helpfulness*, measured via mandatory thumbs-up / thumbs-down feedback, and *relevance*, collected as a follow-up signal for helpful responses.

As shown in Table 8, deployment of EvidenT is followed by substantial improvements across both metrics. Helpfulness increases by 28%, while relevance improves by 19%. In addition, we observe a post-deployment user retention rate of 90%, indicating sustained engagement with the system.

While these metrics reflect overall user experience and may be influenced by multiple system-level factors, the targeted nature of the pipeline changes suggests that improvements in faithfulness contribute meaningfully to the observed gains. We do not claim causality; rather, the results indicate a strong association between increased faithfulness and improved user satisfaction in a real-world financial enterprise setting.

6. Conclusion

We presented a modular inference-time pipeline designed to improve faithfulness in financial enterprise extractive question answering systems. Beyond improving verifiability, the modular structure of the pipeline enables fine-grained inspection and debugging of individual components, facilitating the identification and mitigation of failure modes during system development and deployment (**RQ1**). In parallel, we introduced a comprehensive evaluation strategy that combines intrinsic metrics, comparative baselines, and post-deployment user feedback to assess groundedness, traceability, and downstream impact under real-world constraints. Our results demonstrate that citation-level groundedness and span-level traceability can substantially improve factual faithfulness without requiring model retraining (**RQ2**), and that these improvements translate into measurable gains in downstream user satisfaction (**RQ3**). Together, our findings demonstrate that modular pipeline design, coupled with detailed and stage-aware evaluation, provides a practical and effective strategy for building trustworthy LLM systems in industry settings.

We emphasize that the contribution of this work lies not in proposing algorithmic novelty, but in demonstrating that carefully designed inference-time control mechanisms can deliver verifiable, auditable behavior under strict latency and operational constraints. In high-risk financial enterprise environments, where retraining, multi-stage verification loops, and nondeterministic generation may be impractical, such engineering-oriented interventions provide a scalable and deployment-aligned path toward trustworthy LLM systems.

7. Ethical Considerations and Limitations

Our study is limited to text-based financial enterprise extractive question answering and focuses on groundedness and traceability as the primary optimization objectives. Accordingly, we do not evaluate free-form reasoning, creative generation, or standard public benchmarks, as these settings do not reflect financial enterprise operational and trust constraints. Due to data privacy requirements, the underlying financial enterprise datasets cannot be publicly released. We rely on lexical and semantic traceability metrics as proxies for downstream trust; controlled causal analyses, such as A/B testing, are out of scope for this work, although post-deployment user feedback provides an initial signal of real-world impact. Additionally, our experiments are restricted to open-weight models due to deployment and privacy constraints, and we do not consider multimodal inputs or non-textual evidence. Extending the pipeline and evaluation framework to these settings remains future work.

8. Bibliographical References

Eric Anderson, Jonathan Fritz, Austin Lee, Bohou Li, Mark Lindblad, Henry Lindeman, Alex Meyer, Parth Parmar, Tanvi Ranade, Mehul A Shah, et al. 2024. The design of an LLM-powered unstructured analytics system. *arXiv preprint arXiv:2409.00847*.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint arXiv:2508.06401*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhiyu Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Ruth Petzold, and William Yang Wang. 2024b. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*. Survey Certification.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. 2025. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the ICLR 2025 Workshop on Advances in Financial AI*.

Prafulla Kumar Choubey, Xiangyu Peng, Shilpa Bhagavath, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. 2025. Benchmarking deep search over heterogeneous enterprise data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Industry Track)*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. Understanding finetuning for factual knowledge extraction. In *International Conference on Learning Representations*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. 2025. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems. *arXiv preprint arXiv:2503.15191*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, pages 66–71.

Zhan Peng Lee, Andre Lin, and Calvin Tan. 2025. Finetune-rag: Fine-tuning language models to resist hallucination in retrieval-augmented generation. *arXiv preprint arXiv:2505.10792*.

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*.

Xiaochen Li, Domenico Bianculli, and Lionel Brian. 2025. Tracing content requirements in financial documents using multi-granularity text analysis. *Requirements Engineering*, 30(1):109 – 132.

Sarah Packowski, Inge Halilovic, Jenifer Schlotfeldt, and Trish Smith. 2024. Optimizing and evaluating enterprise retrieval-augmented generation (rag): A content design perspective. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pages 162–167.

Maya Patel and Aditi Anand. 2024. Factuality or fiction? benchmarking modern llms on ambiguous qa with citations. *arXiv preprint arXiv:2412.18051*.

Vinzent Penzkofer and Timo Baumann. 2024. Evaluating and fine-tuning retrieval-augmented language models to generate text with accurate citations. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. 2025. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *SIGIR-AP*, pages 47–54.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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)*, Online. Association for Computational Linguistics.

Chaitanya Sharma. 2025. Retrieval-Augmented Generation: A comprehensive survey of architectures, enhancements, and robustness. *arXiv preprint arXiv:2506.00054*.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Fact Extraction and VERification Workshop (FEVER)*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine-tuning vs. retrieval-augmented generation for less popular knowledge. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (Asia-Pacific)*.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the NeurIPS 2021 Track on Datasets and Benchmarks*.

Dustin Wright, Zain Muhammad Mujahid, Lu Wang, Isabelle Augenstein, and David Jurgens. 2025. Unstructured evidence attribution for long context query focused summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025. Ground Every Sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics*.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. 2025. Citeeval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32759–32778.

Jialin Yang, Dongfu Jiang, Tony He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. 2026. StructEval: Benchmarking LLMs' capabilities to generate structural outputs. *Transactions on Machine Learning Research*.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language models to domain-specific rag. In *Conference on Language Modeling*.