TextSentinels @ FinMMEval 2026 Task 2: Chronological Stabilization of a BM25-Grounded Multilingual Financial Question Answering Pipeline

Durairaj Thenmozhi[1], Adithya Sivakumar[2,*], Aakash Balasubramanian[2], Amudhan A[2] and Amrit Gopinath[2]

[1]*Shiv Nadar University, Chennai, India*
[2]*Sri Sivasubramaniya Nadar College of Engineering, Chennai, India*

Abstract

The CLEF FinMMEval 2026 Task 2 (PolyFiQA) benchmark evaluates multilingual financial question answering over heterogeneous evidence sources, including SEC filings, financial statements, tabular disclosures, and multilingual news articles. Although recent retrieval-augmented generation (RAG) systems have demonstrated strong reasoning capabilities, we observed that benchmark performance was dominated by retrieval stability and lexical evidence preservation rather than increasingly sophisticated generation strategies. This paper documents the chronological engineering evolution of our system, beginning with a naïve prompting baseline and progressing through multiple iterations involving retrieval redesign, prompt refinement, multilingual evidence preservation, and infrastructure optimization. Throughout development, several seemingly promising techniques—including aggressive semantic retrieval, dense reranking, defensive prompting, and blind context truncation—consistently reduced ROUGE performance by disrupting financial table locality and multilingual evidence alignment. These observations motivated a lightweight production pipeline built around BM25 retrieval, chunk-level evidence preservation, deterministic prompting, and minimal post-processing. Local validation performance improved from approximately 0.04 to 0.28 ROUGE-1 F1 across successive system iterations, while the final benchmark submission achieved approximately 0.18 ROUGE-1 F1. Beyond reporting a practical system, this work highlights practical engineering lessons regarding multilingual retrieval stability, benchmark-oriented optimization, and the interaction between retrieval quality and overlap-based evaluation metrics.

Keywords

FinMMEval, PolyFiQA, Multilingual Financial Question Answering, BM25 Retrieval, ROUGE Optimization, System Iteration, Retrieval Stabilization

1. Introduction

Answering financial questions using multilingual enterprise documents is difficult for several interacting reasons. Regulatory filings are filled with structural noise, tabular disclosures are dense and format-inconsistent, and evidence frequently appears fragmented or partially translated across source documents. Compounding this, benchmark evaluation based on ROUGE overlap creates a strong bias towards lexical precision rather than free-form explanatory reasoning, meaning that a system's ability to *locate and preserve* the exact wording of evidence often matters more than its ability to reason about that evidence.

The CLEF FinMMEval 2026 Task 2 (PolyFiQA) challenge [1, 2, 5] requires systems to answer multilingual financial reasoning questions using evidence sources that include SEC filings, news reports, financial statements, and tabular disclosures. The inputs frequently mix multiple languages within the same document block, including English, Chinese, Japanese, Spanish, and Greek, and the expected output is a concise answer optimized for ROUGE overlap-based evaluation rather than discursive summarization. These properties make PolyFiQA a demanding testbed: a system must retrieve the correct evidence span, preserve its exact surface form, and avoid contaminating the answer with adjacent

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
thenmozhid@snuchennai.edu.in (D. Thenmozhi); adithya2410402@ssn.edu.in (A. Sivakumar); aakash2410407@ssn.edu.in (A. Balasubramanian); amudhan2410622@ssn.edu.in (A. A); amrit2410182@ssn.edu.in (A. Gopinath)


multilingual or numerically unrelated content.

Rather than presenting a single static architecture, this paper documents the engineering trajectory that produced our final submission. The development process was highly iterative and was repeatedly disrupted by retrieval failures, multilingual contamination, hallucination, and infrastructure throughput limitations. Across several weeks of experimentation, we observed a counterintuitive pattern: many retrieval and prompting strategies that appeared, on paper, to be more sophisticated actually degraded downstream performance. This paper reports both the successful design decisions and the negative results that shaped them, since we believe the latter are frequently omitted from shared-task system descriptions despite being of comparable practical value.

The central conclusion of our work is that, for this benchmark, the dominant obstacle was retrieval stabilization within a retrieval-augmented generation (RAG) pipeline [8] rather than reasoning maximization. Simpler retrieval mechanisms coupled with disciplined evidence preservation consistently outperformed aggressive ranking, filtering, and reranking approaches, a finding that runs against the general trend of increasingly elaborate retrieval architectures in the broader RAG literature.

The contributions of this work are threefold. First, we document the chronological engineering evolution of a multilingual financial QA system, including both successful and unsuccessful design choices, providing practical insights of a kind often omitted from shared-task reports that describe only the final architecture. Second, we demonstrate empirically that retrieval stability and lexical evidence preservation consistently outperform more complex retrieval and prompting strategies under ROUGE-based evaluation, a result with implications beyond this specific benchmark. Third, we analyze the interaction between multilingual evidence, retrieval locality, and benchmark-oriented prompting, showing how these factors jointly shaped the final system design and where they continue to limit performance.

2. Related Work

Multilingual financial question answering has recently emerged as a demanding extension of financial NLP, requiring systems to reason over heterogeneous evidence such as SEC filings, tabular disclosures, and news articles spanning multiple languages. The PolyFiQA task, introduced as part of CLEF FinMMEval [1, 5], formalizes this challenge by requiring concise, ROUGE-evaluated answers grounded in multilingual, multi-source evidence. Related benchmarking efforts such as MultiFinBen [2] similarly highlight the difficulty of multilingual and multimodal financial reasoning tasks, underscoring a persistent gap between general-purpose language model capability and the fine-grained evidence grounding that financial domains demand.

Retrieval-augmented generation [8] has become the standard paradigm for grounding language model outputs in external evidence, typically implemented through either sparse lexical retrieval such as BM25 [3] or dense semantic retrieval using sentence-embedding models [7]. Dense and reranking-based retrieval methods have shown strong results in open-domain question answering, but financial documents introduce distinct challenges: densely structured tables whose meaning depends on positional context, evidence fragments distributed across multiple languages within the same passage, and lexically strict evaluation metrics that reward exact overlap over semantic paraphrase. Under these conditions, retrieval sophistication does not straightforwardly translate into benchmark performance, since semantic reranking and filtering can just as easily discard the exact surface forms that ROUGE rewards.

This paper positions itself within that literature not as a proposal of a new retrieval algorithm, but as an engineering optimization study: an empirical account of how a benchmark-oriented multilingual financial QA pipeline was iteratively stabilized under real infrastructure and evaluation constraints, ultimately converging on the classical BM25 retrieval framework [3] rather than a more elaborate dense or hybrid alternative.

3. Methodology

3.1. Initial Baseline

Our initial approach used direct prompting over the raw context without any retrieval component: the entire context associated with a question was concatenated into a single prompt and passed to the inference model with a deliberately simple instruction to answer using the context and return a concise answer only. We experimented primarily with Llama-based models [4], specifically `llama-3.1-70b-versatile` and `llama-3.1-8b-instant`. Despite its apparent reasonableness, this baseline performed poorly under ROUGE evaluation [6], with local ROUGE-1 F1 scores remaining near 0.04.

The dominant failure mode in this baseline was hallucination: the model frequently synthesized financial explanations that were not supported by the context, an issue that was especially pronounced for statistical and numeric references. Many of these answers were financially plausible on their own terms, yet their overlap with reference answers was extremely low, revealing that ROUGE evaluation heavily favors exact figures, percentages, preserved year-over-year phrasing, and direct multilingual quote overlap. This observation shifted our optimization target away from improving reasoning quality and toward preserving lexical evidence.

In response, we introduced a defensive instruction directing the model to return an explicit "none" response whenever the answer was not stated in the context. Rather than resolving the hallucination problem, this change produced a new regression: the model became excessively conservative and returned "none" even when relevant evidence was present, which sharply reduced recall and caused a benchmark-level collapse. Local validation runs showed that excessive defensive prompting reduced ROUGE-1 F1 from 0.11 back down to approximately 0.06. This failure mode became especially severe in multilingual contexts, where evidence was fragmented across partial tables, translated snippets, and analyst commentary, and it was this baseline collapse that motivated the shift toward a dedicated retrieval stage described next.

3.2. Retrieval Evolution

Our next development phase focused on retrieval, under the initial assumption that more aggressive filtering of the context would improve answer quality. This assumption proved incorrect: early retrieval experiments involving embedding-similarity reranking, semantic filtering using multilingual sentence embeddings [7], top-k pruning, query expansion, and language-specific chunk filtering repeatedly introduced irrelevant fragments into the prompt, causing generation instability and numeric corruption rather than improved precision.

A related and equally damaging failure arose from context truncation. Initial versions of the pipeline truncated context using fixed token limits without regard for local document structure, which proved highly destructive for financial tables, quarterly statement rows, news evidence sections, and percentage explanations. Numeric fields critical to the correct answer frequently appeared near truncation boundaries, causing the model to lose denominator values, year-over-year comparison rows, unit descriptors, and currency indicators. One particularly damaging pattern occurred when truncation removed a table's header row while preserving its numeric rows, which led the model to hallucinate financial categories for values that had been detached from their labels. This specific failure mode was one of the major turning points in the pipeline's design, since it demonstrated that structural locality, not just topical relevance, needed to be preserved during retrieval.

These repeated failures motivated a shift to sparse lexical retrieval using BM25Okapi [3] in place of the embedding-heavy approaches described above. BM25 produced markedly more stable results than semantic reranking because its scoring is grounded directly in term frequency and inverse document frequency within the original document, rather than in a learned embedding space that can place topically related but evidentially irrelevant multilingual passages close together. Because BM25 ranks whole chunks by lexical overlap with the query rather than by semantic similarity, it tends to preserve locally coherent evidence rather than aggressively mixing semantically related but structurally distant

chunks; we found this conservative behavior, rather than any additional ranking sophistication, to be the primary reason it outperformed dense alternatives on this benchmark. Accordingly, the retrieval pipeline was redesigned to prioritize chunk-locality preservation over semantic maximization, retrieving whole chunks via BM25 and passing them intact into the prompt rather than recombining fragments drawn from different parts of the document. Figure 1 summarizes the resulting end-to-end architecture.

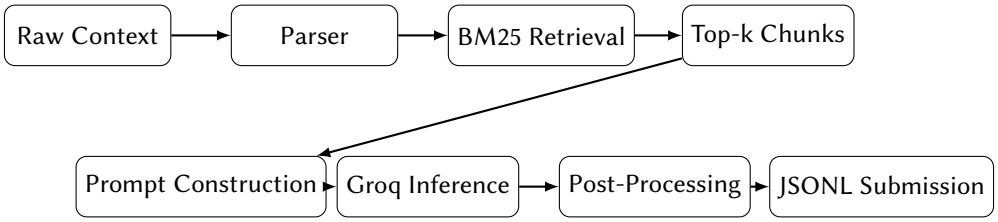

Figure 1: Final stabilized multilingual financial QA pipeline.

A further empirical discovery concerned the treatment of the *News Evidence* sections contained within each document. During early preprocessing, duplicated headlines, publisher metadata, repeated article fragments, and formatting artifacts were removed under the assumption that they constituted retrieval noise. Manual inspection of low-scoring predictions revealed that this cleaning strategy frequently eliminated the exact lexical evidence reproduced in the benchmark reference answers. Consequently, the preprocessing stage was simplified by retaining complete news evidence blocks within their original document chunks instead of filtering or rewriting them. These preserved blocks were indexed directly by BM25 alongside the surrounding financial context, allowing retrieval to operate over the original multilingual evidence without additional normalization. This modification substantially improved lexical overlap with the reference answers and increased local ROUGE-1 F1 from approximately 0.19 to 0.24, illustrating that benchmark-oriented retrieval favored evidence preservation over aggressive text cleaning.

3.3. Prompt Engineering

Our prompting strategy underwent several major iterations in parallel with the retrieval changes described above. Early prompts encouraged the model to explain its financial reasoning clearly, which produced unnecessarily verbose outputs with low lexical overlap against the concise reference answers. The final prompt was substantially narrower, instructing the model to answer using only the provided evidence, preserve exact figures and percentages, return a concise benchmark-style answer, and avoid explanation beyond the evidence; this reformulation substantially reduced hallucination rates relative to the earlier, more permissive instruction.

Deterministic decoding settings were adopted for the same reason: final inference used a temperature of 0.0, a top-p of 1.0, and a constrained output length. Under ROUGE evaluation [6], lexical consistency across runs mattered more than the expressive or semantic variation that non-deterministic decoding would otherwise introduce, so removing sampling variance was itself a stabilizing intervention rather than merely a convenience.

3.4. Infrastructure Constraints

We initially assumed that larger models would dominate benchmark performance, and qualitatively this was true in isolated inspection: the 70B model produced cleaner synthesis, stronger multilingual coherence, and better table interpretation than its 8B counterpart. However, this qualitative advantage did not survive contact with submission-scale inference. Under sustained, large-scale runs, the 70B configuration experienced API throttling, queue delays, unstable latency, interrupted inference batches, timeout cascades, and partial batch failures, all of which became highly problematic when generating predictions across the full benchmark rather than a small evaluation sample.

By contrast, `llama-3.1-8b-instant` delivered stable throughput, predictable latency, and reliable batch completion under the same conditions. Although the 8B model occasionally produced weaker synthesis quality and shorter evidence spans in isolated comparisons, its throughput stability was decisive for large-scale inference, since a qualitatively stronger model that cannot reliably complete a full benchmark run is of limited practical value. This tradeoff between per-example quality and submission-scale reliability constituted another major turning point in the system's design and ultimately determined the model used in the final pipeline.

3.5. Final Pipeline

The final production system integrated the design decisions described above into a single architecture consisting of BM25 retrieval, chunk-level extraction, multilingual evidence preservation, financial table preservation, deterministic prompting, and lightweight post-processing. Each component corresponds directly to a failure mode identified during development: BM25 retrieval and chunk-level extraction address the semantic-contamination and truncation failures described in Section 3.2; multilingual and table evidence preservation address the news-evidence and table-fragmentation issues; deterministic prompting addresses the verbosity and hallucination issues described in Section 3.3; and the choice of inference backend addresses the throughput instability described in Section 3.4. The complete configuration of this final pipeline, including the specific components used at inference time, is reported in Section 4.

4. Experimental Setup

The system was evaluated on the CLEF FinMMEval 2026 Task 2 (PolyFiQA) dataset, which pairs multilingual financial questions with heterogeneous evidence documents drawn from SEC filings, financial statements, tabular disclosures, and multilingual news articles. Retrieval was performed using BM25Okapi [3] over parsed document chunks, with top-k chunk retrieval used to select the evidence passed into the prompt; chunk preservation logic ensured that financial tables and news evidence blocks retrieved by BM25 were passed through to the prompt intact rather than being split or re-filtered downstream, following the rationale described in Section 3.2. Generation was performed using `llama-3.1-8b-instant` served via the Groq API, selected over the larger `llama-3.1-70b-versatile` model for the throughput reasons discussed in Section 3.4, with deterministic decoding (temperature 0.0, top-p 1.0, constrained output length) as described in Section 3.3. Post-processing was limited to lightweight cleanup of model output rather than any aggressive reformatting, consistent with the evidence-preservation philosophy adopted throughout the pipeline. Predictions were submitted in JSONL format and evaluated using ROUGE-1 F1 [6], in line with the benchmark's official evaluation protocol. Table 1 summarizes this final configuration.

Before indexing, documents were parsed into contiguous retrieval units while preserving their original ordering. The parser avoided splitting financial tables and multilingual news evidence wherever possible so that numerical values, row headers, and contextual descriptions remained within the same retrieval unit. Retrieved chunks were concatenated directly without semantic reranking prior to prompt construction.

Figure 1 illustrates the complete inference workflow. Each question and its associated multilingual context were first parsed into retrieval units while preserving document ordering. The retrieval component ranked these chunks using BM25Okapi, after which the highest-ranked chunks were concatenated without additional semantic reranking. The resulting context was embedded directly into a deterministic prompt template and submitted to the Groq inference API. Generated responses underwent only lightweight post-processing to remove formatting artifacts before being written to the final JSONL submission file. No additional answer rewriting or semantic correction was performed after generation.

Table 1
Final Production Configuration

Component	Configuration
Model	llama-3.1-8b-instant
Inference Backend	Groq API
Retrieval Engine	BM25Okapi
Chunk Strategy	Contiguous document chunking
Chunk Preservation	Tables + News Evidence preserved
Decoding	Deterministic
Temperature	0.0
Post-processing	Lightweight cleanup only
Submission Format	JSONL
Evaluation Metric	ROUGE-1 F1

5. Results

5.1. Chronological ROUGE Progression

Table 2 reports the local validation trajectory across system iterations. Progress was gradual and, at several points, unstable: individual changes that seemed reasonable in isolation, such as the defensive "none" prompt or aggressive embedding-based reranking, sometimes reduced performance relative to the preceding iteration before subsequent corrections restored and then extended earlier gains. Overall, local ROUGE-1 F1 improved from 0.04 for the naïve direct-prompting baseline to 0.28 for the final stabilized pipeline, while the corresponding benchmark submission achieved approximately 0.18 ROUGE-1 F1.

Table 2
Chronological System Evolution

System Variant	ROUGE-1 F1
Naïve direct prompting	0.04
Defensive prompting baseline	0.06
Aggressive retrieval filtering	0.09
Embedding-heavy reranking	0.12
Basic BM25 retrieval	0.18
Chunk preservation pipeline	0.21
News Evidence preservation	0.24
Final stabilized pipeline	0.28

5.2. Precision vs Recall Tradeoff

The most persistent challenge throughout pipeline development was balancing retrieval precision against evidence recall. Overly permissive retrieval introduced multilingual contamination, irrelevant financial entities, and noisy generation, while overly restrictive retrieval caused missing percentages, dropped financial rows, and incomplete answer spans. In practice, the best-performing configuration was surprisingly conservative: it prioritized plain BM25 ranking, moderate chunk counts, minimal cleaning, and preserved evidence locality over any more aggressive filtering or reranking step, echoing the pattern observed throughout the retrieval evolution described in Section 3.2.

6. Discussion

6.1. Failure Analysis

Three recurring failure patterns remained even in the final stabilized pipeline. First, cross-language numeric drift occurred when nearby multilingual chunks discussed different reporting periods, occasionally causing the model to mix quarterly values, yearly percentages, and currency references; this was difficult to eliminate completely given the fragmented multilingual structure of the source documents. Second, partial table fragmentation persisted for SEC tables that exceeded chunk boundaries despite the chunk-preservation logic described in Section 3.2, in which cases the model occasionally generated incomplete financial summaries. Third, multilingual quote compression occurred when the model paraphrased quotes into shorter English renderings; although semantically reasonable, this behavior reduced lexical overlap and therefore ROUGE score, illustrating once more the tension between semantically adequate generation and overlap-based evaluation.

6.2. Engineering Lessons

Taken together, these results support several practical lessons. BM25 outperformed dense and reranking-based retrieval on this benchmark primarily because its lexical scoring preserves locally coherent evidence rather than mixing topically related but structurally distant multilingual chunks, which matters more than ranking sophistication when evaluation rewards exact overlap. Dense retrieval and semantic reranking underperformed for the same reason in reverse: embedding-space similarity does not guarantee structural or positional coherence, and in this benchmark that mismatch manifested as numeric corruption and generation instability rather than improved relevance. Multilingual evidence preservation, and specifically the decision not to strip duplicated news fragments during preprocessing, mattered because reference answers frequently reused the exact wording of that evidence, so any cleaning step that discarded it directly capped attainable ROUGE performance regardless of downstream retrieval or generation quality. More broadly, ROUGE itself shaped the system's design at nearly every stage: because the metric rewards exact lexical overlap rather than semantic adequacy, prompts, retrieval, and post-processing were all tuned toward evidence preservation rather than explanatory quality. Finally, larger models were not uniformly better: the 70B model's qualitative advantages in synthesis and multilingual coherence were outweighed at submission scale by its throughput instability, indicating that infrastructure reliability is itself a first-class engineering constraint for benchmark participation rather than a secondary concern behind model capability.

7. Conclusion and Future Work

This paper presented the chronological engineering evolution of a multilingual financial question answering system developed for CLEF FinMMEval 2026 Task 2. Rather than converging on an increasingly sophisticated retrieval or reasoning architecture, our experiments repeatedly demonstrated that retrieval stability, evidence-locality preservation, and inference throughput reliability were the dominant factors determining benchmark performance under ROUGE-oriented multilingual financial QA evaluation. Several design choices that appeared promising in isolation, including aggressive semantic retrieval, defensive hallucination-suppression prompting, and larger-model inference, produced negative results once evaluated at submission scale, and it was the resolution of these specific failures, rather than any single architectural innovation, that produced the final system.

The final benchmark submission achieved approximately 0.18 ROUGE-1 F1, compared with 0.04 for the initial naïve baseline and 0.28 for the best local validation configuration; the gap between local validation and benchmark performance is itself consistent with the cross-language drift and table-fragmentation failures discussed in Section 5.1. Future work will focus on structure-aware table retrieval that preserves header-value associations across chunk boundaries, multilingual financial entity alignment to reduce cross-language numeric drift, retrieval-aware chunk stitching for tables that exceed

single-chunk boundaries, and hybrid sparse-dense retrieval strategies designed specifically to avoid the contamination effects that caused dense retrieval to underperform in the present system.

Acknowledgments

We thank the organizers of FinMMEval 2026 for providing the multilingual benchmark datasets, evaluation infrastructure, and experimental framework used throughout this work.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 5 for iterative drafting assistance, structural refinement, and formatting support. All experimental decisions, engineering analysis, evaluations, and technical conclusions were independently verified and reviewed by the authors. The authors take full responsibility for the contents of this manuscript.

References

[1] Xie, Z. et al.: Overview of FinMMEval 2026: Multilingual and Multimodal Financial Evaluation. CLEF 2026.
[2] Peng, X. et al.: MultiFinBen: Benchmarking Large Language Models for Multilingual and Multimodal Financial Application. arXiv:2506.14028 (2025)
[3] Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval (2009)
[4] Meta AI: Llama 3.1 Technical Report. Meta Research (2025)
[5] Xie, Z. et al.: Overview of the FinMMEval 2026 Task 2: Financial Question Answering and Summarization. CLEF 2026 Working Notes.
[6] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. ACL Workshop on Text Summarization Branches Out (2004)
[7] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP (2019)
[8] Lewis, P. et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS (2020)