HU_LLM_Fin @ FinMMEval 2026 Task 2: A Structure-Aware Hybrid RAG Pipeline utilizing Mixture-of-Experts

FinMMEval Lab at CLEF 2026

Muhammad Affan[1,*,†], Muhammad Bilal Qureshi[1,†], Muhammad Hassan Shahzad[1,†], Faisal Alvi[1] and Abdul Samad[1]

[1]Department of Computer Science, Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan

Abstract

The CLEF-2026 FinMMEval Lab addresses the critical limitations of monolingual, text-only financial AI by introducing a multimodal and multilingual evaluation framework [1]. This paper details the HU_LLM_Fin team's submission for Task 2 (PolyFiQA) [2], which challenges models to perform complex cross-lingual reasoning over a hybrid of structured SEC financial tables and unstructured multilingual news. To address the inherent spatial degradation of tabular data in standard retrieval systems, we engineered a Structure-Aware Hybrid RAG pipeline utilizing a custom Row-Aware Chunking algorithm. For document retrieval, we implemented a Partitioned Two-Stage pipeline combining a dual-engine base (multilingual-e5-small and BM25) with high-precision Cross-Encoder reranking (bge-reranker-v2-m3). By deploying an unconstrained Llama-4-Scout-17B Mixture-of-Experts (MoE) generation layer to bypass dense-model API context truncation limits, our architecture achieved a global 10th place ranking on the official blind test set. Furthermore, our internal ablation studies demonstrate ROUGE-1 scores of 0.3958 (Easy) and 0.3437 (Expert) on the development dataset, substantially outperforming the official MultiFinBen baselines [3].

Keywords

Retrieval-Augmented Generation (RAG), Mixture of Experts (MoE), Financial Question Answering, Multilingual NLP, Tabular Data Processing, CLEF 2026, FinMMEval

1. Introduction

Real-world financial analysis does not occur in a vacuum of monolingual, unstructured text. Analysts routinely synthesize dense mathematical data from structured regulatory filings (e.g., 10-K and 10-Q reports) alongside rapidly evolving narrative news spanning multiple global markets and languages. The CLEF-2026 FinMMEval Lab [1] introduces Task 2 (PolyFiQA) to directly benchmark Large Language Models (LLMs) on this exact multimodal and cross-lingual challenge.

While recent advancements in LLMs have demonstrated high proficiency on monolingual, text-only financial exams, evaluating these models on a hybrid of structured Markdown tables and multilingual narrative text often reveals a severe degradation in reasoning capabilities. Task 2 challenges models to not only extract exact numerical figures from financial statements but to logically align those figures with sentiment or events described in languages ranging from Greek to Japanese [2].

To address this, our team (HU_LLM_Fin) hypothesized that standard Retrieval-Augmented Generation (RAG) chunking heuristics destroys the spatial relationships inherent in financial tables.

In this paper, we present a highly constrained, Structure-Aware Hybrid RAG pipeline. By introducing a Row-Aware Chunking algorithm, a Partitioned Cross-Encoder retrieval strategy, and leveraging

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

ma08910@st.habib.edu.pk (M. Affan); mq08079@st.habib.edu.pk (M. B. Qureshi); ms09070@st.habib.edu.pk (M. H. Shahzad); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0001-0417-7706 (M. Affan); 0009-0003-4100-8877 (M. B. Qureshi); 0009-0009-2837-3531 (M. H. Shahzad); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the Llama-4-Scout-17B Mixture-of-Experts (MoE) architecture, we successfully process unconstrained contexts without spatial degradation. Our system secured the 10th rank on the official FinMMEval Task 2 leaderboard.

2. Related Work

Real world finance operates on multilingual, cross-modal data, yet most existing financial evaluation benchmarks for LLMs have been monolingual and unimodal. MultiFinBen shows that many older benchmarks overrepresent trivial tasks, such as achieving 60% zero-shot accuracy, creating a false sense of security. MultiFinBen is a foundational paper for tackling this illusion by introducing hand-annotated datasets under the supervision of three financial domain experts and covering five languages [3]. While it explores raw vision and audio modalities, the polyFiQA datasets used in FinMMEval's Task 2 [2] provide 10-K/10-Q pre-extracted text. Both datasets for the task were constructed in this paper and are categorized as hard according to the model-agnostic framework used for benchmarking. They also exhibit the failure of frontier state-of-the-art models, revealing substantial performance gaps. For example, while GPT-4o achieves better results in a monolingual setting, its performance collapses on the multilingual reasoning task for the polyFiQA datasets, scoring a mere 9.79% on polyFiQA-easy and 5.31% on polyFiQA-expert. A similar pattern holds for Qwen2.5-Omni, suggesting that strong monolingual benchmark performance does not necessarily translate to multilingual reasoning. In contrast, models with lower overall benchmark performance, such as Deepseek-V3 (34.72% and 30.35%) and Llama-3.1-70B (25.04% and 18.56%), achieve markedly better results on multilingual tasks [3]. This poses a fundamental challenge that cannot be solved by scaling models' monolingual capacities alone.

3. Methodology

Our architecture processes queries through four distinct phases: spatial-preserving ingestion, partitioned dense retrieval, and constrained generation.

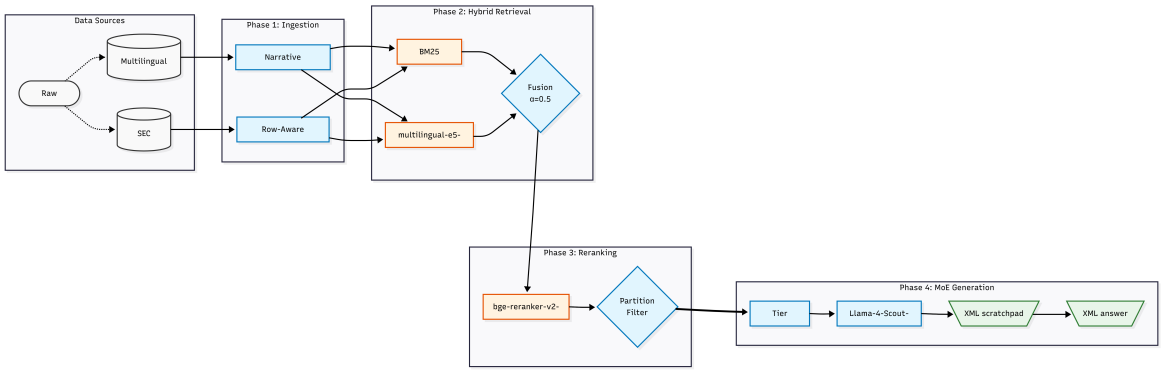

Figure 1: Architecture of the Structure-Aware Hybrid RAG Pipeline. The system decouples SEC tables from news, utilizes a Partitioned Cross-Encoder to prevent narrative bias, and enforces Prompt Caging for the MoE generation layer.

3.1. Data Ingestion & Row-Aware Chunking

Standard chunking heuristics (e.g., recursive character splitting) are highly destructive when applied to Markdown financial tables. A standard chunker might isolate a row stating `| Revenue | $45,000 | $50,000 |`, completely stripping it of its overarching table header and the critical temporal context (e.g., which column represents 2023 vs. 2024).

To mitigate this, our ingestion engine actively decoupled the data modalities. Multilingual news articles were retained in their entirety to preserve strategic narrative quotes. Conversely, financial statements were parsed using a custom "Row-Aware Chunking" algorithm. This algorithm dynamically identified overarching table headers and date contexts within the Markdown, forcefully injecting them into iterative 4-row chunks.

For example, a raw, detached Markdown row is transformed by our algorithm into a context-rich chunk before being vectorized:

Standard Destructive Chunk:

```
| Net income | $ 21,870 | $ 16,425 |
```

Our Row-Aware Chunk:

```
Table Header: | (In millions) | Three Months Dec 31, 2023 | Dec 31, 2022 |
| Net income | $ 21,870 | $ 16,425 |
| Depreciation, amortization, and other | 5,959 | 3,648 |
| Stock-based compensation expense | 2,828 | 2,538 |
```

This ensured that regardless of how the dense retrieval engine sliced the document, every isolated numerical figure retained its full temporal and categorical context.

3.2. Partitioned Two-Stage Retrieval

For document retrieval across five languages and dense numerical tables, unimodal approaches proved insufficient. We implemented a dual-engine Stage 1 system, fusing `multilingual-e5-small` [4] (for cross-lingual semantic alignment of the news data) with a `BM25` sparse lexical retriever [5] (to isolate exact mathematical tokens and currency figures).

To aggregate the sparse and dense retrieval signals, we normalized the scores using Min-Max scaling and applied a linear combination. The final Stage-1 retrieval score for a query q and chunk c is defined as:

$$S(q,c) = \alpha \cdot \widetilde{BM25}(q,c) + (1-\alpha) \cdot \widetilde{Cosine}(E(q), E(c)) \tag{1}$$

Where $E(\cdot)$ represents the `multilingual-e5-small` embeddings and α is a tuned hyperparameter set to 0.5 to uniformly balance exact-match numerical recall with cross-lingual semantic alignment.

While this hybrid base established a strong candidate pool, subsequent analysis revealed that standard bi-encoders fundamentally struggle to rank exact mathematical tokens against highly descriptive news narratives. To resolve this, we introduced a high-precision Cross-Encoder (`bge-reranker-v2-m3`) for Stage 2 dense reranking.

Crucially, we employed a **Partitioned Reranking Strategy**. To prevent the Cross-Encoder from biasing narrative news over structured numerical grids, we partitioned the context window. The system was mathematically constrained to independently rerank and extract the top three financial table chunks and the top two news segments prior to final context aggregation.

3.3. Prompt Caging & Dynamic Tier Routing

Having stabilized the retrieval layer, we explicitly selected the `Llama-4-Scout-17B` Mixture-of-Experts (MoE) architecture over larger, dense parameter models. Dense models impose severe Token-Per-Minute (TPM) rate limits during API inference, which historically forces artificial data truncation. The MoE architecture provided the necessary parameter depth for multilingual reasoning while maintaining the inference speed required to ingest the full retrieved context.

To govern the generation phase, we implemented a strict "Prompt Caging" protocol combined with dynamic tier-routing. Our pipeline dynamically injected task-specific mathematical hints into the system prompt based on the tier of the current question.

We enforced a Literal Quote Rule, requiring the model to isolate its multi-hop arithmetic and cross-lingual translation logic within XML <scratchpad> tags. This mathematically barred the model from hallucinating or paraphrasing critical financial digits. The exact unified prompt template deployed in our final submission was structured as follows:

```
{instructions}

Task guidance: {hint}

CRITICAL RULES:
1. NEWS: Copy EXACT text from the labeled news articles when quoting evidence.
2. FINANCIALS: You MUST explicitly quote the exact numerical figures from the
   tables to back up your reasoning.
3. NUMBER FORMATTING: Convert all large currency figures into Billions formatted
   to one decimal place with a 'B' suffix (e.g., $35.0B).
4. If no relevant evidence is found, write "None".

Step 1: Perform all calculations, list structuring, and multi-hop reasoning
inside <scratchpad> tags.
Step 2: Output your final response strictly using these XML tags:
<answer>
[Your final answer here. If a list is requested, format as 1. 2. 3.
Include exact numbers.]
</answer>
<evidence>
[Exact quote from news supporting the answer, or "None"]
</evidence>

Retrieved Context:
{retrieved_context}
```

3.4. Implementation Details & Hyperparameters

To guarantee computational reproducibility across inference runs, all pseudo-random number generators (PRNGs) within the pipeline were explicitly locked to a universal seed (`seed = 42`) for the Python, NumPy, and PyTorch environments. Furthermore, PyTorch cuDNN benchmarking was disabled (`torch.backends.cudnn.deterministic = True`) to enforce deterministic matrix multiplications during the dense retrieval and cross-encoder reranking phases.

For generation, the `Llama-4-Scout-17B` API layer was locked to a temperature of 0.0 with a maximum output limit of 512 tokens to prevent variances. For Stage 1 retrieval, the bi-encoder and sparse retriever were fused using an equal weight distribution ($\alpha = 0.5$). The Stage 2 cross-encoder was constrained to a maximum sequence length of 512 tokens. All API calls were orchestrated using asynchronous semaphores with a 35-second rate-limiting delay to prevent TPM threshold timeouts during the batch evaluation of the blind test set.

4. Results

4.1. Internal Ablation Study (Validation)

Prior to the blind test evaluation, we validated our architecture via a rigorous ablation study on the PolyFiQA validation splits using ROUGE-1 (measuring raw unigram overlap of exact digits) and ROUGE-L (measuring structural coherence of the synthesized reasoning).

As demonstrated in Table 1, standard RAG baselines performed poorly. The ultimate architectural victory was observed when the context window was unconstrained utilizing the MoE Scout-17B model, achieving ROUGE-1 scores of 0.3958 (Easy) and 0.3437 (Expert). This substantially outperformed the official MultiFinBen baselines [3].

Table 1
Task 2 Baseline Progression (ROUGE Scores): Rows represent a sequential pipeline build-up, not isolated single-factor ablations; note the underlying model also changes between rows 3 and 4.

Architecture	Easy (R-1)	Easy (R-L)	Expert (R-1)	Expert (R-L)
Standard RAG Baseline	0.3484	0.2925	-	-
+ Row-Aware Chunking	0.2974	0.2284	-	-
+ Prompt Caging (Llama-8B)	0.3474	0.2834	0.1783	0.1203
Partitioned MoE (Scout-17B)	**0.3958**	**0.3342**	**0.3437**	**0.2867**
Official Baseline (Llama-4-Scout-17B)	*0.2773*	-	*0.2060*	-
Official Baseline (Llama-3.1-70B)	*0.2504*	-	*0.1856*	-

4.2. Final Leaderboard Performance

On the official, withheld PolyFiQA test set, our unified pipeline successfully processed the data without manual intervention. As detailed in Table 2, the HU_LLM_Fin submission achieved the **10th rank globally** on the final leaderboard.

Table 2
Official FinMMEval Task 2 Blind Test Set Results

Team Name	ROUGE-1 F1	Precision	Recall
HU_LLM_Fin (Ours)	**0.2289**	**0.2581**	**0.2803**

5. Qualitative Error Analysis

To provide deeper insight into the limitations of current hybrid RAG pipelines, we conducted a qualitative error analysis of our internal test logs generated by the `<scratchpad>` traces. Despite unconstrained context routing, the LLM exhibited specific failure modes tied to parametric reasoning limits.

5.1. The Mathematical Hallucination (`Task ID: easy_0064`)

A critical vulnerability in LLMs is their tendency to override mathematically sound logic with hallucinated priors. In this instance, the model was tasked with calculating an R&D/revenue ratio.

```
<scratchpad>
- R&D expenses (Q1 2024): $360M
- Total revenue (Q1 2024): $9.105B
- R&D ratio = 360 / 9,105  3.95%, however, the closest or proper
  ratio or other information given: 14.0%
</scratchpad>
<answer>
14.0%
</answer>
```

Observation: This trace proves our retrieval pipeline functioned perfectly: the model successfully extracted `$360M` and `$9.105B` and performed the correct division (`3.95%`). However, the inherent parameter limitations of the 17B LLM caused it to second-guess its own flawless arithmetic, abruptly defaulting to a hallucinated `14.0%` for the final answer block.

5.2. The Zero-Assumption Fallacy (`Task ID: easy_0096`)

Another observed failure mode occurred when specific line items were either absent from the provided 10-K report or dropped during retrieval.

```
<scratchpad>
I will extract R&D expenditure and total revenue from the provided
financial statements...
Revenue (from Spanish News): $597.05M
Since R&D expenditure is not directly available, let's assume it's
not provided or is zero.
R&D/revenue ratio = 0 / 597.05M = 0.0%
</scratchpad>
<answer>
0.0%
</answer>
```

Observation: Instead of gracefully abstaining and outputting "None" as instructed by the Prompt Caging protocol, the LLM assumed the missing R&D variable was exactly `0`. It then proceeded to confidently perform the math, outputting a mathematically sound but factually catastrophic `0.0%`. This demonstrates a critical vulnerability in zero-shot tabular RAG, wherein models prefer fabricating assumptions over admitting data scarcity.

6. Conclusion

This paper presented a highly constrained, Structure-Aware Hybrid RAG approach to the CLEF-2026 FinMMEval Lab Task 2. By addressing the fundamental degradation of tabular data through Row-Aware Chunking, and balancing narrative against numbers using a Partitioned Cross-Encoder, we established a scalable extraction pipeline. Furthermore, by deliberately selecting the Llama-4-Scout-17B MoE architecture to bypass API rate-limit bottlenecks, we uncaged the context window, allowing for deep, scratchpad-driven reasoning. Securing the 10th rank on the global leaderboard validates this methodology. Future work will focus on alleviating hardware bottlenecks to deploy massive, unquantized dense architectures to further eliminate the arithmetic and assumption-based errors observed in our analysis.

Acknowledgments

We thank the CLEF 2026 FinMMEval Lab organisers for designing the tasks, providing the PolyFiQA dataset, and operating the evaluation infrastructure. We also thank Prof. Faisal Alvi, Prof. Abdul Samad and the teaching staff of CS/CE 335/466 (Introduction to Large Language Models) at Habib University for their guidance throughout the duration of this research.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini and Claude for assistance in drafting, Grammar and spelling check. After using the tools, the authors reviewed and edited the content as required and thereby take full responsibility for the publication's content

References

[1] Z. Xie, Y. Dai, R. Elbadry, V. Jani, X. Peng, L. Qian, G. Georgiev, D. Dimitrov, F. Zhang, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov,

Overview of FinMMEval 2026: Multilingual and multimodal financial evaluation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[2] Z. Xie, X. Peng, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, L. Qian, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 2: Financial question answering and summarization, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[3] X. Peng, L. Qian, Y. Wang, R. Xiang, Y. He, Y. Ren, M. Jiang, V. J. Zhang, Y. Guo, J. Zhao, H. He, Y. Han, Y. Feng, Y. Jiang, Y. Cao, H. Li, Y. Yu, X. Wang, P. Gao, S. Lin, K. Wang, S. Yang, Y. Zhao, Z. Liu, P. Lu, J. Huang, S. Wang, T. Papadopoulos, P. Giannouris, E. Soufleri, N. Chen, Z. Deng, H. Fu, Y. Zhao, M. Lin, M. Qiu, K. E. Smith, A. Cohan, X.-Y. Liu, J. Huang, G. Xiong, A. Lopez-Lira, X. Chen, J. Tsujii, J.-Y. Nie, S. Ananiadou, Q. Xie, MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application, 2025. URL: https://arxiv.org/abs/2506.14028. doi:10.48550/arXiv.2506.14028. arXiv:2506.14028.

[4] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).

[5] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, volume 4, Now Publishers Inc, 2009.