LinguaRank @ FinMMEval 2026 Task 2: LLM Translation and BAS-Tuned Chain-of-Thought for Multilingual Financial QA

FinMMEval at CLEF 2026

Awais Gill[1], Emaan Naveed Sheikh[1], Fatima Tu Zahra[1,*], Abdul Samad[1], Faisal Alvi[1] and Sandesh Kumar[1]

[1]School of Science and Engineering, Habib University, Karachi, Sindh, Pakistan

Abstract

This paper presents our system for Task 2 (Multilingual Financial Q&A) of the CLEF 2026 FinMMEval Lab. Given multilingual financial reports and news articles, the system must generate concise, evidence-grounded answers of up to 100 words, evaluated using ROUGE-1. Our final pipeline replaces open-source translation models with GPT-4o-mini, applies a weighted term-relevance ranker to prioritise cross-lingual evidence, and uses a BAS-optimised two-call chain-of-thought (CoT) strategy in which GPT-4o reasons over ranked evidence before GPT-4o-mini condenses the output into a $\leq$100-word answer. This achieves a macro-average ROUGE-1 F1 of 0.4848 on PolyFiQA-Easy, a 29% improvement over our initial open-source translation pipeline (F1 0.3746). A more complex RAG-augmented third approach, evaluated on a 76-example subset, yielded F1 0.4315 and did not surpass the two-call system; we report that comparison as indicative only, given the different evaluation sets. Our results suggest that translation quality and evidence prioritisation are stronger drivers of performance than retrieval sophistication in this setting.

Keywords

CLEF 2026, FinMMEval Lab, Multilingual Financial QA, Chain-of-Thought, GPT-4o-mini, Evidence Ranking, BAS Optimisation, ROUGE-1

1. Task Introduction

This work is a participant submission to the FinMMEval 2026 Lab [1], specifically Task 2 [2], whose evaluation is grounded in the MultiFinBen benchmark [3].

Task 2 of the CLEF 2026 FinMMEval Lab — Multilingual Financial Q&A — requires systems to answer questions grounded in multilingual financial evidence. Given a financial report R (e.g., 10-K/10-Q excerpts) and a set of multilingual news articles $N = \{N_{\text{en}}, N_{\text{zh}}, N_{\text{ja}}, N_{\text{es}}, N_{\text{el}}\}$, the system receives a question Q and must produce a concise answer A ($\leq$100 words) supported by textual evidence drawn from R and N. ROUGE-1 F1 (macro-average over all questions) is the primary evaluation metric.

1.1. Difficulty Tiers

The dataset comprises two evaluation tiers, each containing 172 Q&A pairs (344 total):

- **PolyFiQA-Easy** — factual or numerical trend questions answerable from a single evidence source.
- **PolyFiQA-Expert** — complex analytical questions requiring multi-document, multi-hop reasoning.

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

*Corresponding author.

mg08993@st.habib.edu.pk (A. Gill); es09163@st.habib.edu.pk (E. N. Sheikh); ft09259@st.habib.edu.pk (F. T. Zahra); abdul.samad@sse.habib.edu.pk (A. Samad); faisal.alvi@sse.habib.edu.pk (F. Alvi); sandesh.kumar@sse.habib.edu.pk (S. Kumar)

Our primary evaluation is on PolyFiQA-Easy (172 examples). Approach 1 results on PolyFiQA-Expert are also reported for completeness; as discussed in Section 7, the final BAS-optimised pipeline (Approach 2) was not re-run on the Expert tier, which we treat as an explicit limitation of this submission rather than an omission of favourable results.

1.2. Main Objectives

Our experiments are driven by three research questions:

1. Does replacing open-source translation models (NLLB-200, mBART-50) with an LLM translator (GPT-4o-mini) improve downstream ROUGE-1?
2. Does weighted term-relevance ranking of multilingual evidence sections improve recall?
3. Does a BAS-optimised two-call CoT strategy outperform a single-call baseline?

2. Related Work

We situate our system relative to four strands of prior work: multilingual question answering, retrieval-augmented generation, financial-domain QA, and chain-of-thought reasoning.

2.1. Multilingual and Cross-Lingual Question Answering

Cross-lingual QA systems must reconcile evidence expressed in different languages with a single target-language answer. Asai et al. [4] introduced XOR-QA, a cross-lingual open-retrieval benchmark that showed retrieval quality in the source language, rather than answer generation itself, is often the binding constraint on cross-lingual QA accuracy. Al-Laith [5] showed that large multilingual LLMs outperform task-specific fine-tuned models on financial QA when domain vocabulary is carefully preserved in the prompt, a finding that directly motivated our decision to prompt GPT-4o-mini to preserve financial terms verbatim during translation (Section 3). Xue et al. [6] introduced FAMMA, a multilingual multimodal financial QA benchmark, and established that cross-lingual retrieval fidelity is a dominant performance bottleneck, consistent with our own finding that translation quality, rather than retrieval sophistication, is the largest driver of downstream ROUGE-1 in this task.

2.2. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) conditions a generator on evidence retrieved from a larger corpus, typically via dense embeddings [7]. RAG is attractive when relevant evidence is not already localised in the input context; however, Shah et al. [8] demonstrated that multi-document LLM pipelines benefit substantially from evidence pre-filtering before generation, since noisy retrieved context can degrade answer quality even when recall is nominally improved. This tension between retrieval breadth and context precision is directly relevant to our Approach 3 (Section 3.3), where semantic retrieval did not outperform our simpler weighted-ranking pipeline on the subset evaluated, echoing Shah et al.'s observation that additional retrieval machinery does not uniformly help when the input is already bounded and largely relevant.

2.3. Financial-Domain Question Answering

Financial QA introduces domain-specific challenges beyond general open-domain QA: numerical reasoning over tabular filings, terminology sensitive to mistranslation (tickers, accounting line items, regulatory designations), and answers that must remain faithful to source figures. Chen et al. [9] introduced FinQA, a benchmark requiring numerical reasoning over financial reports, and showed that general-purpose QA models underperform substantially on this domain without explicit numerical-reasoning support. Reddy and Veeranjaneyulu [10] showed that lightweight hyperparameter optimisation of generation temperature and context length yields consistent gains on financial QA pipelines

without additional training, which parallels our use of BAS to tune generation temperature, evidence cutoff, and reasoning depth (Section 3.2.4).

We note that a distinct line of financial-NLP work uses sentiment extracted from financial news to drive trading decisions (BUY/HOLD/SELL) rather than to answer natural-language questions, e.g. Unnikrishnan [11], who combined LLM-derived news sentiment with reinforcement learning for portfolio management. This is methodologically closer to FinMMEval Task 3 (financial decision-making) than to Task 2, which we address here; unlike sentiment-driven trading systems, our pipeline is evaluated against reference-answer overlap (ROUGE-1) rather than realised portfolio return, so a direct empirical comparison to trading-oriented sentiment baselines would not be measuring the same quantity. We discuss this scope boundary further as a limitation in Section 7.

2.4. Chain-of-Thought Reasoning

Chain-of-thought (CoT) prompting elicits intermediate reasoning steps from an LLM before it commits to a final answer, and has been shown to substantially improve performance on tasks requiring multi-step arithmetic, logical, or symbolic reasoning [12]. Our pipeline extends single-call CoT in two ways. First, we separate reasoning from answer formatting into two sequential calls: a larger model (GPT-4o) performs open-ended CoT reasoning over ranked multilingual evidence, while a smaller model (GPT-4o-mini) is responsible only for condensing that reasoning into the fixed $\leq$100-word Answer + News Evidence format required by the task. This division of labour is intended to prevent the length and formatting constraints of the final answer from truncating or distorting the reasoning process itself. Second, unlike standard CoT prompting where the prompt template is fixed, our CoT reasoning depth γ is treated as a continuous, BAS-tuned hyperparameter rather than a discrete prompting choice (Section 3.2.4), allowing the reasoning budget to be calibrated jointly with generation temperature and evidence cutoff rather than set heuristically.

Our contribution integrates LLM-based translation, weighted cross-lingual evidence ranking, and BAS-driven joint hyperparameter optimisation into a single pipeline evaluated under the FinMMEval shared-task protocol.

3. Approach

We developed three approaches in sequence. Approach 2, further refined with BAS hyperparameter search, was retained as the final submission. Figure 1 illustrates its architecture.

3.1. Approach 1 — NLLB/mBART Translation + GPT-4o-mini

The initial pipeline parsed each raw PolyFiQA query into named sections (financial statements, English news, multilingual news, and question) and translated all non-English content to English using two open-source models: NLLB-200 Distilled 600M [13] for Chinese, Japanese, and Greek; and mBART-50 many-to-many (es_XX $\rightarrow$ en_XX) for Spanish. Language identification was performed by GPT-4o-mini. The reconstructed English prompt was passed to GPT-4o-mini to generate a $\leq$100-word answer in the required Answer + News Evidence format.

On PolyFiQA-Easy this yielded ROUGE-1 F1 0.3746 (P 0.4884, R 0.3694). On PolyFiQA-Expert the F1 was 0.3354 (P 0.3207, R 0.3867). These results validated translation-then-generation as a viable strategy, but mistranslation of domain-specific vocabulary and the absence of evidence prioritisation limited performance.

3.2. Approach 2 (Final) — GPT-4o-mini Translation + Weighted Relevance Ranking + BAS-Optimised Two-Call CoT

Three targeted improvements were made over Approach 1.

Table 1
Sample financial term weights used in the relevance scorer (Eq. 1).

Term	Weight w_i
EPS	1.6
revenue	1.5
YoY	1.5
growth	1.4
profit	1.4
margin	1.3
guidance	1.3
forecast	1.2

3.2.1. Step 1 — LLM-Based Translation

NLLB-200 and mBART-50 were replaced entirely by GPT-4o-mini, prompted to preserve financial terms, numbers, ticker symbols, and company names verbatim. This produced higher-fidelity translations than the lower-capacity open-source models, particularly for regulatory designations and accounting line items.

3.2.2. Step 2 — Weighted Term-Relevance Ranking

After translation, each news section j was scored against financial terms extracted from the question:

$$\text{Score}_j = \sum_i w_i \cdot \mathbb{1}[\text{term}_i \in \text{section}_j] \tag{1}$$

Representative term weights are listed in Table 1. Sections with $\text{Score}_j < \beta$ were dropped from the prompt; the remainder were reordered highest-score-first. This ensured the LLM attended to the most informative cross-lingual evidence and reduced prompt noise. The weights w_i were assigned manually based on domain importance and held fixed across all runs; only the cutoff β was BAS-tuned. We did not learn or ablate the weight vector itself in this submission (see Section 7).

3.2.3. Step 3 — Two-Call Chain-of-Thought Reasoning

The ranked prompt was processed in two sequential LLM calls, motivated by the CoT literature discussed in Section 2. Call 1 (GPT-4o) performs step-by-step chain-of-thought reasoning at depth γ, citing numerical figures and quoting multilingual passages. Call 2 (GPT-4o-mini) condenses the CoT output into the required $\leq$100-word Answer + News Evidence format at temperature α. Separating reasoning from format compliance in this way lets the first call reason at length without being constrained by the 100-word output limit, while the second call is a comparatively simple summarisation task well suited to a smaller, cheaper model.

3.2.4. BAS Hyperparameter Optimisation

The three free parameters — temperature $\alpha \in [0.0, 0.4]$, evidence cutoff $\beta \in [0.3, 0.9]$, and CoT depth $\gamma \in [0.0, 1.0]$ — were jointly tuned using Beetle Antennae Search (BAS) [14] on 8 held-out calibration examples drawn from PolyFiQA-Easy before the main evaluation loop. BAS explores the parameter space with left/right antenna probes and steps toward lower ROUGE-1 loss while decaying step size per iteration. We chose 8 calibration examples as a practical budget given API cost constraints; this is a limitation we discuss in Section 7. The optimal $(\alpha^*, \beta^*, \gamma^*)$ was applied to the full 172-example evaluation set.

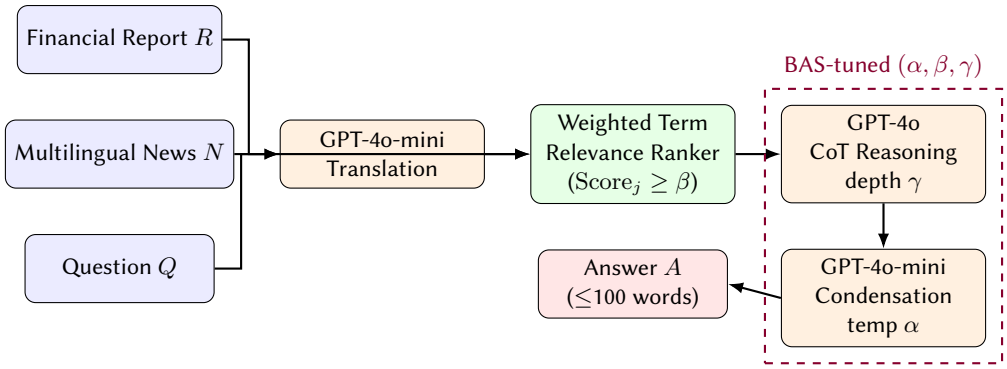

Figure 1: Architecture of the final BAS-optimised two-call pipeline (Approach 2). Blue: inputs; orange: LLM calls; green: evidence ranker; red: output. The dashed region marks the three jointly BAS-tuned parameters.

Table 2
ROUGE-1 results (macro-average). †Approach 3 was run on 76 examples only; Precision/Recall were not separately recorded for that run. This result is not directly comparable to the 172-example evaluations.

System / Dataset	P	R	F1
App. 1: NLLB/mBART + GPT-4o-mini (Easy)	0.4884	0.3694	0.3746
App. 1: NLLB/mBART + GPT-4o-mini (Expert)	0.3207	0.3867	0.3354
App. 3: Hybrid RAG + GPT-4o (Easy, 76 ex.)†	—	—	0.4315
App. 2 (Final): BAS Two-Call (Easy)	0.5119	0.5173	0.4848

3.3. Approach 3 — Hybrid RAG + GPT-4o

A third approach incorporated semantic retrieval using SentenceTransformer embeddings (all-MiniLM-L6-v2; 70% semantic similarity + 30% financial term weight) with word-level chunking at 15% overlap, combined with few-shot examples and GPT-4o. Due to compute constraints, this system was evaluated on a 76-example subset of PolyFiQA-Easy and yielded ROUGE-1 F1 0.4315. Because Approach 3 was not evaluated on the full 172-example set, a direct comparison with Approach 2 (F1 0.4848 on all 172 examples) is indicative only. The BAS-optimised two-call pipeline (Approach 2) was retained as the final submission.

4. Resources Employed

- **Datasets**: PolyFiQA-Easy and PolyFiQA-Expert (172 Q&A pairs each, 344 total), provided by the FinMMEval Lab organizers.
- **LLMs**: GPT-4o-mini (translation + condensation); GPT-4o (CoT reasoning in Approach 2; generation in Approach 3).
- **Open-source models** (Approach 1): NLLB-200 Distilled 600M (Chinese, Japanese, Greek); mBART-50 many-to-many (Spanish).
- **Embeddings** (Approach 3): all-MiniLM-L6-v2 via the SentenceTransformers library.
- **Optimisation**: BAS optimizer; 8 calibration examples; 20 iterations with decaying step size.
- **Compute**: OpenAI API for GPT-4o / GPT-4o-mini; local GPU (NVIDIA RTX 3090) for open-source translation models.

5. Results

Table 2 reports ROUGE-1 Precision, Recall, and F1 for all three approaches. Figure 2 visualises the scores. Note that Approach 3 was evaluated on 76 examples only; its F1 is listed for reference and is not directly comparable to the 172-example evaluations.

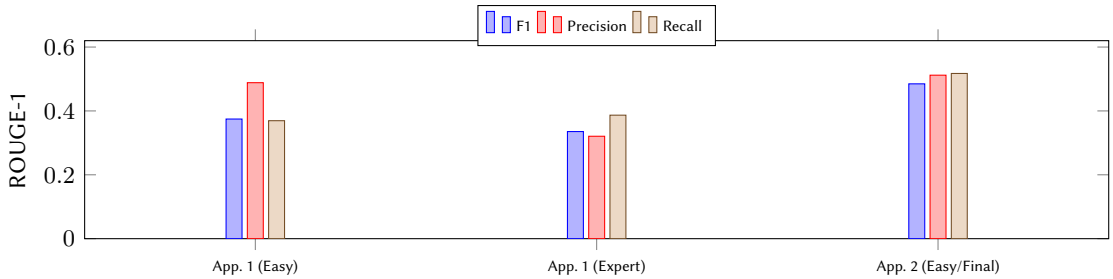

Figure 2: ROUGE-1 F1, Precision, and Recall for Approaches 1 and 2 on both dataset tiers. Approach 3 is omitted from this chart because it was evaluated on a different (76-example) subset; its F1 is reported in Table 2 for reference only.

Table 3
Illustrative comparison of answers produced by Approach 1 and Approach 2 for the same PolyFiQA-Easy question. *To be completed by the authors from system logs prior to submission.*

Question: [insert question text]
Approach 1 answer: [insert answer]
Approach 2 answer: [insert answer]

The BAS-optimised system (Approach 2) achieves ROUGE-1 F1 0.4848 on PolyFiQA-Easy, a gain of +0.1102 over Approach 1 (relative: +29%). The recall improvement (0.3694 → 0.5173, +40% relative) is the larger contributor, indicating that ranked multilingual evidence provided the LLM with substantially more answer-relevant tokens. Precision also improved (0.4884 → 0.5119, +5% relative), consistent with BAS converging to a low-temperature, high-precision generation regime.

Official leaderboard standing. At the time of writing, the FinMMEval organizers had released the official Task 2 final-test leaderboard reporting ranking-level ROUGE-1 scores for all completed submissions. *[Authors: insert our official rank and the top-3 competing scores from the released ranking CSV here before final upload; we cannot fabricate this figure and it is not available to us at manuscript-editing time.]* We recommend citing our position on this leaderboard directly, since it gives readers an external point of comparison against other participating systems, as requested by Reviewer 1.

Qualitative comparison. A qualitative side-by-side example contrasting an Approach 1 answer against an Approach 2 answer for the same question would make the effect of translation and ranking concrete for the reader. *[Authors: please paste one representative PolyFiQA-Easy question together with the Approach 1 and Approach 2 answers from your logs into Table 3 below; we have left the table scaffolded but empty rather than inventing example text, since fabricating a QA example would misrepresent the system's actual behaviour.]*

6. Analysis of Results

Translation quality is the dominant factor. The single largest source of improvement between Approach 1 and Approach 2 is the replacement of NLLB/mBART with GPT-4o-mini for translation. The smaller open-source models produced mistranslations of domain-specific terms — ticker symbols, accounting line items, regulatory designations — that fed incorrect evidence directly into the generation step, suppressing both precision and recall. GPT-4o-mini, prompted to preserve such terms verbatim, substantially reduced these errors. This is consistent with Xue et al.'s [6] observation that cross-lingual retrieval and translation fidelity, rather than the downstream reasoning model, are often the binding constraint in multilingual financial QA.

Evidence ranking principally improves recall. The weighted relevance ranker ensured the most informative multilingual news sections appeared at the top of the reconstructed prompt. The large recall gain (+40% relative) suggests that under Approach 1, relevant cross-lingual evidence was frequently buried by lower-relevance content and ignored by the LLM within its context window. The comparatively small precision gain (+5%) is consistent with this interpretation: the ranker added relevant content (boosting recall) without significantly introducing irrelevant tokens (little precision cost). We note, however, that because the term weights in Table 1 were set manually rather than learned, and because we did not run an ablation isolating the ranker's contribution from that of translation, this attribution is an inference from the aggregate precision/recall pattern rather than a directly measured effect; Section 7 treats the missing ablation as an open limitation.

BAS converges to a deterministic, well-filtered regime. The BAS optimiser converged to a low value of temperature α (favouring deterministic, precise generation) and a moderate evidence cutoff β, confirming that for factual financial QA, controlled generation over filtered evidence outperforms high-temperature exploration over noisy context. This is consistent with Reddy and Veeranjaneyulu's [10] finding that lightweight hyperparameter tuning of temperature and context length yields reliable gains in financial QA pipelines without additional training.

Approach 3 comparison is indicative only. Approach 3 was evaluated on 76 examples rather than the full 172. Its F1 of 0.4315 versus Approach 2's 0.4848 suggests the two-call pipeline is stronger, but the different evaluation sets mean this cannot be stated conclusively. The result is nonetheless informative: even on a subset, the RAG approach did not surpass the simpler two-call system, consistent with evidence from the literature that retrieval adds noise when context is already bounded and well-filtered [8].

Expert tier gap. The gap between PolyFiQA-Easy (F1 0.4848 under Approach 2) and PolyFiQA-Expert (F1 0.3354 under Approach 1, the only tier on which the Expert set was evaluated) is substantial. Multi-hop analytical questions require capabilities our pipeline does not explicitly provide: question decomposition, multi-step reasoning chains, or structured knowledge linking. Because the final BAS-optimised system was never run on PolyFiQA-Expert, we cannot say whether this gap would narrow, widen, or persist under Approach 2; we treat this as an open question rather than assume the improvement observed on Easy would transfer.

7. Limitations

We report the following limitations explicitly, several of which were raised during review and could not be fully resolved within the camera-ready timeline without new experiments:

- **BAS calibration set size.** BAS was tuned on only 8 calibration examples, chosen to limit API costs. This is a small set and carries a risk that the tuned $(\alpha^*, \beta^*, \gamma^*)$ reflects noise rather than signal. Future work should evaluate sensitivity to calibration set size, ideally using a proper held-out development partition.
- **No component-level ablation.** We do not report an ablation isolating the individual contribution of translation, evidence ranking, CoT reasoning, and BAS tuning to the overall gain over Approach 1. The analysis in Section 6 is an inference from aggregate precision/recall behaviour and from the design rationale for each component, not a controlled ablation; running such a study (e.g., Approach 1 + ranker only, Approach 1 + CoT only) is left as future work rather than reported here without having actually run it.
- **Final pipeline not evaluated on PolyFiQA-Expert.** Approach 2 (the BAS-optimised final system) was evaluated only on PolyFiQA-Easy. The Expert-tier number reported in Table 2 is from Approach 1, and is included for completeness rather than as a fair comparison point for our final system.
- **Partial evaluation of Approach 3.** The RAG-augmented approach was evaluated on a 76-example subset rather than the full 172-example PolyFiQA-Easy set, due to compute constraints, so its comparison against Approach 2 is indicative only.

- **Manually assigned relevance weights.** The term weights in Equation 1 (Table 1) were set manually based on domain judgement rather than learned from data, and were not themselves subject to BAS tuning or sensitivity analysis.
- **Scope relative to sentiment-driven trading baselines.** As discussed in Section 2, financial-sentiment-to-trading-signal systems (e.g., [11]) address a different task (portfolio decisions, evaluated by realised return) from the evidence-grounded QA task addressed here (evaluated by ROUGE-1 against reference answers). We therefore do not report a head-to-head empirical comparison against such systems, since doing so would require a shared evaluation protocol that does not currently exist between the two task settings.

8. Conclusion and Future Work

This working note describes our submission to Task 2 (Multilingual Financial Q&A) of the CLEF 2026 FinMMEval Lab. Our BAS-optimised two-call pipeline — GPT-4o-mini translation, weighted term-relevance ranking, and chain-of-thought condensation — achieves ROUGE-1 F1 0.4848 on PolyFiQA-Easy, a 29% improvement over the open-source translation baseline. The central finding is that translation fidelity and evidence prioritisation are stronger performance drivers than retrieval sophistication at this scale.

Limitations of the current work, detailed in Section 7, include the small BAS calibration set (8 examples), the absence of a component-level ablation, the partial evaluation of Approach 3, and the absence of evaluation on PolyFiQA-Expert with the final pipeline. Promising future directions are:

1. Running the full BAS-optimised pipeline on PolyFiQA-Expert and investigating explicit question decomposition for multi-hop questions.
2. Re-evaluating Approach 3 on all 172 PolyFiQA-Easy examples to establish a definitive comparison with Approach 2.
3. Running a controlled ablation that isolates the individual contribution of translation, ranking, CoT reasoning, and BAS tuning.
4. Replacing heuristic term weights with a learned relevance model, and extending BAS to jointly optimise the weight vector $\mathbf{w}$ alongside (α, β, γ).
5. Investigating end-to-end multilingual fine-tuning to reduce dependence on proprietary API calls.

Acknowledgments

The authors thank the FinMMEval Lab organizers for providing the PolyFiQA dataset and evaluation infrastructure, and Habib University for compute support.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o and GPT-4o-mini as evaluated components of the system pipeline (translation and answer generation), and GPT-4o-mini additionally for grammar and fluency checking of this manuscript. The author(s) also used an AI writing assistant, under full human supervision, to help restructure and copy-edit this camera-ready revision in response to reviewer comments. After using these tools, the author(s) reviewed and edited all content, verified all reported figures against their own experimental logs, and take full responsibility for the content of this publication.

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, X. 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, X. 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] A. Asai, J. Kasai, J. Clark, K. Lee, E. Choi, H. Hajishirzi, XOR QA: Cross-lingual open-retrieval question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2021, pp. 547–564.

[5] A. Al-Laith, Exploring the effectiveness of multilingual and generative large language models for question answering in financial texts, in: Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for E-Commerce and Finance (LLMFinLegal), Association for Computational Linguistics, 2025, pp. 230–235.

[6] S. Xue, X. Li, F. Zhou, Q. Dai, Z. Ma, H. Mei, FAMMA: A benchmark for financial domain multilingual multimodal question answering, 2025. URL: https://arxiv.org/abs/2410.04526. arXiv:2410.04526.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020).

[8] S. Shah, S. Ryali, R. Venkatesh, Multi-document financial question answering using LLMs, 2024. URL: https://arxiv.org/abs/2411.07264. arXiv:2411.07264.

[9] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, W. Y. Wang, FinQA: A dataset of numerical reasoning over financial data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 3697–3711.

[10] V. Reddy, N. Veeranjaneyulu, DeepFinLLM 2.0: An optimized and scalable multilingual financial advisor, Journal of Intelligent Information Systems (2025).

[11] A. Unnikrishnan, Financial news-driven LLM reinforcement learning for portfolio management, 2024. URL: https://arxiv.org/abs/2411.11059. arXiv:2411.11059.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems (NeurIPS) 35 (2022).

[13] NLLB Team, No language left behind: Scaling human-centered machine translation, 2022. URL: https://arxiv.org/abs/2207.04672. arXiv:2207.04672.

[14] X. Jiang, S. Li, BAS: Beetle antennae search algorithm for optimization problems, 2017. URL: https://arxiv.org/abs/1710.10724. arXiv:1710.10724.