AI_TLfanclub at FinMMEval 2026 Task 1: Distilled Qwen Models for Multilingual Financial Exam Question Answering

Answer-Surface Alignment for Multilingual Financial MCQ

Do Thanh Duc[1,*], Nguyen Nhat Minh[1] and Le Thanh Huong[1]

[1]Hanoi University of Science and Technology, Hanoi, Vietnam

Abstract

This paper describes the AI_TLfanclub system for Task 1 of the CLEF 2026 FinMMEval Lab, a multilingual financial multiple-choice question answering task. Our central technical claim is that Task 1 should be treated as an answer-surface alignment problem: the system should be optimized for the exact answer-label distribution used by the official metric, while reasoning traces serve mainly as supervision rather than as the final prediction interface. The submitted system instantiates this idea with schema normalization, calibrated answer-label scoring, teacher-rationale distillation paired with direct-answer twins, targeted fine-tuning, and validation-gated reward-guided ablations. We report two main findings. First, rationale distillation plus targeted fine-tuning is the strongest and most stable mainline in our internal studies, improving the base Qwen3.5-4B student from 63.87% to 83.94% on a 274-example checkpoint-selection validation set. Second, reinforcement learning is useful only when it preserves direct answer behavior and should not be assumed better than the best supervised checkpoint by default. On the official final-test leaderboards, AI_TLfanclub ranks 7th in English with 80.50% accuracy, 8th in Chinese with 70.50%, 8th in Arabic with 84.00%, and 8th in Hindi with 68.50%. The corresponding micro-average over the four final-test tracks is 75.88% (607/800). These results support a conservative conclusion: for this benchmark, answer-label scoring, rationale supervision, and targeted fine-tuning are more reliable than unvalidated late-stage alignment complexity.

Keywords

financial QA, multilingual MCQ, distillation, LoRA, calibration, FinMMEval

1. Introduction

FinMMEval 2026 evaluates financial AI systems across multilingual reasoning and decision-making tasks [1]. This paper studies Task 1, Financial Exam Question Answering, where each example is a stand-alone financial multiple-choice question and the system must output the correct option label. The task overview defines accuracy as the primary metric and reports final-test rankings separately for English, Chinese, Arabic, and Hindi [2].

Professional finance exams require a mixture of factual knowledge, regulatory awareness, numerical reasoning, and careful interpretation of distractors. This makes the task different from open-ended financial question answering: the final output is short, but the decision process must be stable under multilingual wording, domain-specific terminology, and option-order bias.

Our main design principle is therefore to optimize the answer-selection surface directly. Instead of relying only on long generated explanations and post-hoc parsing, the system scores answer labels as next-token candidates, calibrates these scores across option permutations, and uses teacher-generated rationales primarily as training supervision. Rather than claiming a universal new algorithm, this paper makes an evidence-based claim about this benchmark: when evaluation is exact answer-label accuracy, training, inference, and model selection should all be aligned to the final label surface.

We make four contributions.

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

*Corresponding author.

duc.dt242095m@sis.hust.edu.vn (D. T. Duc); minh.nm252184m@sis.hust.edu.vn (N. N. Minh); huong.lethanh@hust.edu.vn (L. T. Huong)

- We formulate Task 1 as an answer-surface alignment problem and show why direct label scoring and calibration are first-order design choices for multilingual financial MCQ.
- We study a paired rationale-distillation recipe in which each reasoning trace is coupled with a direct-answer twin, allowing the student to learn intermediate financial reasoning without drifting away from the official answer-label objective.
- We provide validation-gated evidence that targeted fine-tuning is a stronger mainline than simply taking the final RL checkpoint, and that GRPO-style alignment should be treated as a controlled extension rather than a guaranteed improvement.
- We report official leaderboard results together with language-specific analysis, showing stable English, Chinese, and Arabic behavior and a clear Hindi distribution-shift limitation.

This paper focuses on the submitted Task 1 system and its evaluation; implementation details are described only when they are necessary for reproducibility or methodological clarity.

2. Task and Submission Requirements

Task 1 asks participants to select the correct answer among four candidate options. The public FinMMEval description lists financial exam sources such as CFA-style English questions, CPA-style Chinese questions, Arabic financial and Shari'ah-compliant questions, and Hindi financial-domain questions, with additional multilingual resources available for training and transfer [3, 4, 5, 6, 7]. The official final-test leaderboard is released by language, and hidden labels and participant answers remain private.

The lab permits dataset expansion and retrieval for Task 1, but these components must be clearly disclosed in the system paper. Our system uses public training resources, internally generated teacher rationales, and local retrieval over the training-support pool. It does not use hidden development or final-test labels for training, and it does not use external web retrieval at inference time. The organizer-held development set is used for validation and the final-test set is used only for official reporting.

3. Related Work

Recent financial benchmarks show that finance-focused language modeling is not a single homogeneous problem. FinMMEval evaluates multilingual and multimodal financial competence across several tasks [1], while RealFin studies underspecified financial questions that require implicit reasoning and domain completion [3]. For Task 1 specifically, multilingual financial MCQ is shaped by heterogeneous source exams and uneven resource availability across languages such as Arabic and Hindi [2, 4, 5, 6, 7]. This makes language balance, terminology drift, and output formatting first-order modeling concerns.

Reasoning-oriented prompting is a natural baseline for difficult multiple-choice tasks. Chain-of-thought prompting and self-consistency often improve reasoning accuracy in large language models [8, 9], but exact-match MCQ evaluation still requires a stable mapping from long-form reasoning to the final option label. Few-shot calibration work likewise shows that language models can be biased by prompt priors and answer-position effects unless the prediction surface is corrected explicitly [10]. Our system builds on this insight by centering the answer-label surface throughout training, inference, and checkpoint selection.

The second foundation is compact model adaptation and alignment. Knowledge distillation transfers behavior from a stronger teacher to a smaller student [11], while LoRA and QLoRA make such adaptation feasible under tight accelerator budgets [12, 13]. Preference and reward optimization methods such as DPO and GRPO-style verifiable reward learning have shown strong results in reasoning-heavy settings [14, 15]. Our contribution is not a new alignment algorithm; it is an evidence-based recipe for deciding which of these tools improve multilingual financial MCQ and which ones regress the final answer-label objective.

Table 1
Data groups used by the Task 1 system. Official development and final-test sets are evaluation-only.

Data group	Role	Size
Official development	Organizer-held validation for English, Chinese, Arabic, and Hindi	338
Official final test	Final leaderboard evaluation, 200 examples per language	800
Broad baseline pool	Initial multilingual benchmark including English, Chinese, Spanish, Greek, and Arabic sources	1423
Balanced teacher-seed pool	Public and expanded examples used for teacher rationale generation	3007
Internal validation	Held-out checkpoint-selection set across six language tags	274
Strict distillation validation	Balanced validation for comparing distillation methods	252
Strict distillation train	Rationale and direct-answer training records after filtering and twinning	5400

4. System Overview

For a system paper, workflow descriptions are often more useful than source-code listings. Our system can be summarized as the following six-stage pipeline.

1. **Normalize questions and options.** Heterogeneous source formats are converted into a common record with a question, language tag, option labels, option texts, and a gold answer label when available.
2. **Build supervised and distillation data.** Public training examples are balanced by language; teacher-generated rationales and direct-answer examples are filtered and deduplicated.
3. **Train a compact student.** A Qwen-family student model [16] is adapted with LoRA/QLoRA [12, 13] using rationale supervision and direct-label supervision.
4. **Target weak cases.** Additional fine-tuning emphasizes languages and examples where the base student or early adapters are unstable.
5. **Calibrate inference.** At prediction time, the system scores answer labels directly, applies option-order calibration, and may combine validation-selected local few-shot retrieval and language-specific prompt variants.
6. **Select by validation.** Checkpoints are selected by exact-match accuracy and language-level stability, not by the final training step.

This design is deliberately conservative. The final submission relies on the components that improved held-out answer accuracy most consistently: direct answer-label scoring, rationale distillation, and targeted fine-tuning. Reinforcement learning and preference-pair construction are evaluated as ablations and future extensions rather than assumed to be beneficial by default.

5. Data Construction and Integrity

Table 1 summarizes the main data groups used in our workflow. The official development and final-test sets are reported separately from training and diagnostic validation data. Because Task 1 permits dataset expansion, we also use synthetic and teacher-annotated examples, but we audit them to avoid overlap with public evaluation files.

We apply three safeguards. First, option formats are normalized before training so that the model learns the task rather than dataset-specific layout artifacts. Second, distillation examples are deduplicated by both identifiers and question-option fingerprints. Third, final predictions are generated from hidden-test questions only, without feedback from hidden labels.

5.1. Generated supervision and filtering

The main purpose of generated supervision is not to add arbitrary volume, but to expose the compact student to financial reasoning patterns that are difficult to learn from answer labels alone. Teacher traces are therefore filtered before training. We prefer examples with a single unambiguous final answer, a reasoning path that supports that answer, and no conflict between the rationale and the final label. Examples with malformed options, missing final-answer markers, or inconsistent labels are removed or converted into direct-answer records only when the gold label is trusted.

For Chinese and Hindi, where the accessible supervised pool is smaller or more distributionally variable, synthetic and auxiliary examples are used more cautiously. They are treated as training aids rather than evaluation evidence. This distinction is important for transparent system reporting: official performance must be attributed to the final held-out leaderboards, while generated data should be described as part of the system design and not as a substitute for official evaluation.

In particular, Hindi-heavy auxiliary pools are not used at raw scale. We filter them toward MCQ-compatible records and rebalance them against the other languages so that multilingual training is not dominated by a single auxiliary source.

5.2. Leakage control

Because Task 1 allows participants to reorganize public training data and to use dataset expansion, leakage control becomes a central experimental concern. We use two complementary checks. Identifier-based filtering removes exact matches when stable identifiers are available. Fingerprint-based filtering hashes normalized question and option text so that near-duplicate public-holdout examples are excluded from distillation packs. The strict distillation audit reported zero identifier overlaps and zero question-option fingerprint overlaps with the public evaluation files. This makes the internal ablations more credible, although only the organizer's final-test leaderboard should be treated as the official result.

6. Methods

6.1. Answer-label scoring

For a question prompt x and valid answer labels $L = \{A, B, C, D\}$, the base prediction rule is

$$\hat{y} = \arg\max_{l \in L} \log p_\theta(l \mid x). \tag{1}$$

This directly matches the official evaluation surface. It also reduces failures caused by multilingual free-form generation, such as producing a correct explanation but omitting a clean final answer label.

To reduce position bias, the inference pipeline can apply a content-free calibration term and cyclic option permutations. The calibrated score is

$$s(l, x) = \log p_\theta(l \mid x) - \log p_\theta(l \mid x_0), \tag{2}$$

where x_0 is a short prompt with no question content. This follows the spirit of contextual calibration for prompt-based language modeling [10]. Scores are then averaged across option orderings so that each candidate appears in multiple positions before the final vote. In the main submission path, direct label scoring is preferred over free-form generation because it avoids answer parsing failures entirely.

6.2. Prompting and local retrieval

The system uses language-aware prompts. Direct-answer prompts are used for label scoring, while reasoning prompts are used during distillation and selected inference arms. A local BM25 retriever [17] selects few-shot examples from the training-support pool. This is not web retrieval: the retriever only accesses local, disclosed examples available before evaluation. Retrieval and few-shot prompting are retained only when they improve held-out direct accuracy; otherwise the system falls back to the

plain answer-label scorer. For Hindi and Arabic, translation-aware or bilingual prompts are considered because domain terminology can be unstable under direct prompting.

6.3. Rationale distillation

Teacher rationales are used to teach the student intermediate financial reasoning without requiring a large model at final-test inference. Each accepted rationale example is paired with a direct-answer twin. The rationale example trains the model to explain and conclude; the direct twin trains the one-token answer surface used by the official metric. The supervised loss is masked to assistant tokens:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t \in A} \log p_\theta(y_t \mid x, y_{<t}), \tag{3}$$

where A denotes assistant-token positions. This pairing is important because a model may learn to generate plausible explanations without becoming well calibrated on the exact answer-letter distribution. Final-test inference therefore does not depend on generating a rationale; rationales are used mainly as offline supervision.

6.4. Distillation variants

We evaluated three distillation strategies.

- **Rationale SFT**: sequence-level imitation of teacher rationales and direct labels.
- **Option-level KL**: supervised training plus KL divergence between teacher and student distributions over the valid answer labels.
- **Full-logits KD**: supervised training plus full-vocabulary KL divergence over assistant-token positions.

The option-level KL objective is

$$\mathcal{L} = \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{KD}} T^2 \mathrm{KL}\left(q_T(\cdot \mid x) \,||\, p_{\theta,T}(\cdot \mid x)\right), \tag{4}$$

where q_T is the teacher distribution over answer labels, $p_{\theta,T}$ is the student distribution, and T is the distillation temperature. Although online KL methods are attractive, our ablations show that the simpler rationale-SFT path is more reliable for direct answer accuracy.

6.5. Targeted fine-tuning and reward alignment

After distillation, targeted fine-tuning focuses on weak-language and error-prone examples. This stage is short and checkpointed frequently. The best checkpoint is selected by held-out direct accuracy and language-level stability.

We also evaluate reward-guided alignment. The GRPO-style reward follows the idea of verifiable reward optimization used in mathematical-reasoning systems [15], but is adapted to MCQ answer verification. Positive reward is assigned for the correct final label and clean answer format; penalties are assigned for invalid labels, conflicting answers, excessive length, and fallback-only parses. Preference-pair construction is related to DPO-style alignment [14]. In our current evidence, these reinforcement-learning components are useful research directions, but they are not treated as automatically superior to the best supervised checkpoint.

Longer reward runs are therefore early-stopped and checkpoint-selected rather than assumed to help monotonically. This matters in practice: several reward-optimized branches improve mid-run and then regress on the direct answer-label metric.

7. Experimental Setup

All official numbers in this paper come from organizer-provided development and final-test leaderboards. Development results are used as validation evidence; final-test results are the official leaderboard scores. The final-test leaderboard ranks each language separately, so the micro-average over four languages is reported only as an aggregate descriptive statistic.

For internal experiments, we use an initial broad benchmark, a 274-example checkpoint-selection validation set, and a 252-example strict distillation validation set. The strict distillation audit reports zero identifier overlap and zero question-option fingerprint overlap with public evaluation files. Training uses parameter-efficient adapters, 4-bit NF4 quantization where needed, and offline Kaggle-compatible model/data packaging.

7.1. Selection protocol and evidence hierarchy

We keep four evidence tiers separate. The official development leaderboard is used for external validation before final submission. The official final-test leaderboard is used only for official reporting. A 274-example labeled internal set is used for checkpoint selection in targeted fine-tuning and reward-alignment studies. A balanced 252-example labeled split is used to compare distillation variants under equal language coverage. Small CoT and public-reasoning subsets are diagnostic only and are never treated as substitutes for direct answer accuracy.

This separation matters because different artifacts measure different behaviors. Direct accuracy, reasoning-format compliance, response length, and modal-consensus diagnostics do not always move together. The paper therefore reports official answer accuracy first, then uses internal labeled studies only to justify design choices.

7.2. Implementation details

The validated submission path uses Qwen-family causal language models with parameter-efficient adapters and 4-bit NF4 quantization when memory is constrained. For Qwen3.5 inference we disable internal thinking mode in the chat template, because otherwise the model may emit long hidden reasoning before the answer label, which hurts latency and complicates answer extraction. For the CFA and CPA sources, data are loaded from parquet snapshots rather than the default dataset loader because of schema incompatibilities in the public dataset cards. Final predictions are produced with deterministic answer-label scoring under fixed prompts and valid option labels.

7.3. Reproducibility and submission generation

The system is designed for an offline Kaggle-style environment. Model weights, adapters, generated data, and dependency wheels are mounted or packaged before execution. This avoids implicit changes caused by downloading new model revisions during evaluation. The answer-label scorer is deterministic given a fixed model checkpoint and prompt variant. Sampling is used for rationale generation and some self-consistency arms, but final submission generation relies on validation-selected settings rather than unrestricted sampling.

The final JSON predictions contain only the required answer labels. Confidence scores are not required for Task 1. For each language, the submission file is produced separately and checked for full coverage before upload. The official final-test leaderboard reports 100% coverage for AI_TLfanclub in all four languages, which indicates that every hidden-test example received a valid answer label.

7.4. What the system does not rely on

Several negative statements are worth making explicit. The submitted Task 1 system does not use hidden final-test labels, does not fine-tune on organizer-held development leaderboard items, and does not query external web search at inference time. Retrieval, when enabled, is local retrieval over the

Table 2
Official Task 1 development and final-test results for AI_TLfanclub. The final-test tracks are ranked separately by language.

Language	Development	Final test	Final rank
English	57/70 (81.43%)	161/200 (80.50%)	7
Chinese	50/71 (70.42%)	141/200 (70.50%)	8
Arabic	84/97 (86.60%)	168/200 (84.00%)	8
Hindi	84/100 (84.00%)	137/200 (68.50%)	8
Micro-average	275/338 (81.36%)	607/800 (75.88%)	–

Table 3
Shift from official development to final test, measured in percentage points.

Language	Final − development
English	-0.93
Chinese	$+0.08$
Arabic	-2.60
Hindi	-15.50

disclosed support pool. The reinforcement-learning components are not presented as a guaranteed final improvement; they are retained as ablations because validation evidence shows that long reward-optimization runs can regress direct accuracy.

8. Results

8.1. Official development and final-test results

Table 2 reports the system's official development and final-test performance. The final-test ranks are language-specific leaderboard ranks from the organizer's released CSV files.

English, Chinese, and Arabic show stable development-to-final behavior. Hindi is the main outlier: it drops from 84.00% on development to 68.50% on final test. This suggests that Hindi requires either better domain coverage or stronger distribution-shift controls, rather than simply more global training.

Table 3 shows that the main generalization problem is language-specific rather than global. The system transfers cleanly from development to final test in English, Chinese, and Arabic, but not in Hindi. This is consistent with the broader data audits: the Hindi auxiliary pool is both larger and more heterogeneous than the other languages, so naive scaling can improve apparent validation coverage while still leaving a benchmark-specific mismatch at test time.

8.2. Baseline and diagnostic ablations

The earliest compact baseline reached 43.9% accuracy over a 1423-example multilingual benchmark. This baseline was useful for diagnosing language imbalance: English and Arabic Business were relatively strong, while Chinese and Greek were weak. The later distillation pipeline was developed to address this imbalance.

Table 4 gives the most relevant checkpoint-selection ablation on a 274-example internal validation set. These are not official leaderboard scores, but they explain the final system choice. Rationale distillation plus targeted fine-tuning is the strongest supervised mainline, improving the base Qwen3.5-4B student from 63.87% to 83.94%. A GRPO checkpoint preserved this accuracy on the rationale branch, but did not improve it.

Table 5 compares distillation variants on the strict 252-example validation split. Rationale SFT is the best direct-answer method, while full-logits KD is competitive and stronger on the reasoning-subset

Table 4
Diagnostic internal validation results used for checkpoint selection.

Variant	Correct/total	Accuracy
Base Qwen3.5-4B	175/274	63.87%
Hybrid option-KL KD	193/274	70.44%
Hybrid KD + targeted FT	205/274	74.82%
Hybrid KD + targeted FT + GRPO	206/274	75.18%
Rationale SFT initialization	227/274	82.85%
Rationale SFT + targeted FT	230/274	83.94%
Rationale SFT + targeted FT + GRPO	230/274	83.94%

Table 5
Strict distillation-suite comparison on a balanced 252-example validation split.

Method	Direct acc.	Arabic acc.	CoT subset
Base student	61.90%	–	–
Rationale SFT	88.10%	80.95%	22.22%
Full-logits KD	86.11%	76.19%	50.00%
Option-level KL	67.06%	57.14%	27.78%

diagnostic. Option-level KL is less stable in this setting.

9. Discussion

The official final results and internal ablations lead to five main observations.

First, answer-label calibration is essential. The official metric rewards only the selected option, so long-form reasoning is useful only when it improves the final label. Direct-label examples and logit scoring align the model with this surface.

Second, reasoning quality and direct-answer quality can decouple. In our internal studies, some branches improve strict reasoning-format diagnostics or CoT cleanliness without improving the exact answer label. Cleaner rationales are therefore not enough by themselves; the model must preserve a sharp decision boundary over $A/B/C/D$.

Third, rationale distillation is more reliable than more complex online KD in our current runs. The likely reason is that teacher rationales provide structured financial reasoning while direct twins preserve the answer-label objective. This pairing gives the student process supervision without making final inference dependent on long generated explanations.

Fourth, checkpoint selection matters more than late-stage optimization complexity. Later training steps and final RL checkpoints are not automatically better. Several runs peaked earlier and then regressed, especially when reward groups became flat or when the model overfit stylistic answer markers. This is why the paper treats reward alignment as validation-gated rather than inherently superior.

Fifth, Hindi remains the largest robustness gap. The development score suggests that the system can answer Hindi financial MCQs under some conditions, but the final-test drop indicates a distribution shift. Future work should prioritize higher-quality Hindi finance data, stricter synthetic filtering, and Hindi-specific calibration rather than simply adding more mixed-language training.

10. Threats to Validity and Scope

Three scope conditions matter for interpreting this paper. First, the strongest external evidence is the official final-test leaderboard, whereas most design decisions are justified by two internal labeled validation sets of 274 and 252 examples. These are appropriate for checkpoint selection and ablation,

but they are not substitutes for repeated hidden-test evaluation, multiple random seeds, or confidence intervals.

Second, this paper deliberately distinguishes validated components from exploratory ones. The supported mainline is the one that improved held-out direct accuracy most consistently: calibrated answer-label scoring, rationale distillation, and targeted fine-tuning. Longer reward-optimization and preference-based branches remain research extensions unless they beat the supervised checkpoint under the same validation protocol.

Third, the paper claims competitiveness only where the evidence supports it. English, Chinese, and Arabic are stable from development to final test, but Hindi is not. The correct interpretation is therefore a competitive multilingual Task 1 system with a clear remaining robustness gap, not a solved multilingual financial QA model. Closely spaced leaderboard ranks should also not be over-interpreted without confidence intervals or repeated hidden-test measurements.

11. Future Work

Future work will focus on three directions. The first is data quality: we plan to expand verified Hindi and Chinese finance examples, filter synthetic examples more aggressively, and cap any single-language auxiliary source before multilingual training. The second is stronger but still validation-gated alignment: denser answer-surface distillation, short RL stages, and preference-pair mining should be tested only when they beat the supervised mainline under identical held-out selection. The third is robustness and evaluation: per-language calibration, retrieval weights, seed stability, and confidence-aware analysis should be tuned on strictly held-out validation sets and audited for hidden-test leakage risk.

12. Conclusion

AI_TLfanclub submitted a Task 1 system built around answer-surface alignment for multilingual financial MCQ: distilled Qwen models, direct answer-label scoring, rationale supervision, targeted fine-tuning, and calibrated inference. On the official FinMMEval 2026 final-test leaderboards, the system achieved 80.50% accuracy in English, 70.50% in Chinese, 84.00% in Arabic, and 68.50% in Hindi, corresponding to ranks 7, 8, 8, and 8 respectively. The main empirical lesson is that rationale distillation plus targeted fine-tuning is the strongest validated mainline in the current evidence, whereas reinforcement learning should remain validation-gated rather than assumed superior. The strongest remaining challenge is Hindi robustness under final-test distribution shift.

Acknowledgements

The authors would like to express their gratitude to Associate Professor Dr. Le Thanh Huong (Hanoi University of Science and Technology) for her invaluable guidance and support. Furthermore, we thank the CLEF 2026 FinMMEval organizers for preparing the shared task, evaluation resources, and official leaderboards.

Declaration on Generative AI

During the preparation of this work, the author used large language model tools for text preparation assistance, language editing, and LaTeX formatting. The author reviewed and edited the content and takes 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, Y. Dai, R. Elbadry, V. Jani, G. Georgiev, D. Dimitrov, F. Zhang, X. Peng, L. Qian, 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 the FinMMEval 2026 task 1: Multilingual financial multiple-choice question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[3] Y. Dai, Y. Lin, Z. Xie, Y. Wang, RealFin: How well do LLMs reason about finance when users leave things unsaid?, 2026. URL: https://arxiv.org/abs/2602.07096. doi:10.48550/arXiv.2602.07096. arXiv:2602.07096.

[4] V. Devane, M. Nauman, B. Patel, A. M. Wakchoure, Y. Sant, S. Pawar, V. Thakur, A. Godse, S. Patra, N. Maurya, S. Racha, N. K. Singh, A. Nagpal, P. Sawarkar, K. V. Pundalik, R. Saluja, G. Ramakrishnan, BhashaBench V1: A comprehensive benchmark for the quadrant of indic domains, 2025. URL: https://arxiv.org/abs/2510.25409. arXiv:2510.25409.

[5] R. Elbadry, S. Ahmad, A. Heakl, D. Bouch, M. Ahsan, M. AlMahri, M. Elsaid Khalil, Y. Wang, S. Lahlou, S. Ananiadou, V. Stoyanov, J. Huang, X. Peng, P. Nakov, Z. Xie, SAHM: A benchmark for arabic financial and shari'ah-compliant reasoning, 2026. URL: https://arxiv.org/abs/2604.19098. doi:10.48550/arXiv.2604.19098. arXiv:2604.19098.

[6] 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.

[7] L. Qian, X. Peng, Y. Wang, V. J. Zhang, H. He, H. Smith, Y. Han, Y. He, H. Li, Y. Cao, Y. Yu, A. Lopez-Lira, P. Lu, J.-Y. Nie, G. Xiong, J. Huang, S. Ananiadou, When agents trade: Live multi-market trading benchmark for LLM agents, 2025. URL: https://arxiv.org/abs/2510.11695. doi:10.48550/arXiv.2510.11695. arXiv:2510.11695.

[8] 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, in: Advances in Neural Information Processing Systems, 2022.

[9] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations, 2023.

[10] T. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, M. Gardner, Calibrate before use: Improving few-shot performance of language models, in: International Conference on Machine Learning, 2021.

[11] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[13] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems, 2023.

[14] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference opti-

mization: Your language model is secretly a reward model, in: Advances in Neural Information Processing Systems, 2023.

[15] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, in: arXiv preprint arXiv:2402.03300, 2024.

[16] Qwen Team, Qwen3 Technical Report, 2025. arXiv:2505.09388.

[17] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, in: Foundations and Trends in Information Retrieval, volume 3, 2009, pp. 333–389.