

Overview of FinMMEval 2026 Task 1: Multilingual Financial Multiple-Choice Question Answering

Zhuohan Xie^{1,*}, Yuyang Dai², Rania Elbadry¹, Vanshikaa Jani³, Georgi Georgiev⁴,
Dimitar Dimitrov⁴, Fan Zhang⁶, Xueqing Peng⁵, Lingfei Qian⁵, Jimin Huang⁵, Jiahui Geng⁷,
Yankai Chen^{1,8}, Ye Yuan^{8,9}, Haolun Wu^{8,9}, Yuxia Wang², Ivan Koychev⁴, Veselin Stoyanov¹,
Mingzi Song¹⁰, Yu Chen⁶, Xue Liu^{1,8,9} and Preslav Nakov¹

¹Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

²INSAIT, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

³University of Arizona, Tucson, United States

⁴FMI, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

⁵The Fin AI, United States

⁶The University of Tokyo, Tokyo, Japan

⁷Linköping University, Linköping, Sweden

⁸McGill University, Montreal, Canada

⁹Mila, Quebec AI Institute, Montreal, Canada

¹⁰Meiji Gakuin University, Tokyo, Japan

Abstract

FinMMEval 2026 Task 1 evaluates multilingual financial multiple-choice question answering in English, Chinese, Arabic, and Hindi. The task tests whether systems can select the correct answer to finance questions involving domain terminology, numerical interpretation, and conceptual financial reasoning across languages and scripts. The final-test set contains 800 questions, with 200 questions per language; gold answers were withheld during submission, and each language was ranked independently by accuracy. The final leaderboards contain 13 English, 11 Chinese, 11 Arabic, and 10 Hindi ranked submissions. Top accuracies range from 92.0% in Hindi to 97.5% in English and Arabic, with the same leading teams appearing near the top across all four languages. The documented systems used retrieval augmentation, direct answer-option scoring, language-specific prompting, selective self-consistency, confidence checks, and LLM-based review stages.

Keywords

FinMMEval, financial QA, multilingual evaluation, multiple-choice QA, financial reasoning

1. Introduction

Financial question answering requires systems to combine domain knowledge with numerical interpretation, reporting conventions, and implicit assumptions. Depending on the question, a system may need to locate evidence in a company report, perform a calculation, or distinguish between closely related financial concepts. Multilingual evaluation adds variation in scripts, terminology, and source conventions to these reasoning demands. Task 1 isolates answer correctness from surface-form variation by requiring systems to return a single option label for each question. This design evaluates the target capability separately from other properties of generated output [1]. For each of the four languages—English, Chinese, Arabic, and Hindi—systems answer 200 final-test questions and submit one option for every item. Providing all systems with the same candidate answers eliminates variation in free-form wording while preserving the need for financial knowledge, numerical calculation, and multilingual terminology. The task is part of the broader FinMMEval lab at CLEF 2026 [2], which also includes short-answer financial QA in Task 2 [3] and live trading-agent evaluation in Task 3 [4].

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

*Corresponding author.

✉ zhuohan.xie@mbzuai.ac.ae (Z. Xie); Preslav.Nakov@mbzuai.ac.ae (P. Nakov)



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

2. Related Work

FinMMEval Task 1 is connected to financial QA, finance-oriented language modeling, and multilingual financial evaluation. FiQA established a retrieval-oriented shared-task setting for financial opinion mining and QA [5], and FinanceBench later framed open-book QA over public-company information as a finance-specific evaluation problem [6]. FinQA [7] and TAT-QA [8] focus on numerical and tabular reasoning over financial reports. ConvFinQA extends this line to conversational question answering, where systems must track financial evidence across turns [9]. FinChain emphasizes verifiable reasoning chains in financial tasks [10], while RealFin targets implicit assumptions in realistic financial reasoning settings [11]. Evidence selection and report construction are also important in finance-oriented QA pipelines. FinCARDS studies intra-document evidence reranking for financial question answering [12], while FinReporting examines localized financial disclosure reporting [13]. Herculean extends this broader evaluation direction toward agentic financial intelligence [14]. In parallel, finance-specific encoders such as FinBERT [15] and large language models such as BloombergGPT [16] provide modeling foundations for financial NLP. Instruction-tuned resources such as PIXIU [17] and FinGPT [18] further broaden the reusable toolset for shared-task systems. Task 1 also follows the multilingual direction represented by MultiFinBen [19]. SAHM studies Arabic financial-reasoning settings [20], and BhashaBench provides Hindi/English India-centric domain-evaluation context [21].

Unlike most free-form financial QA tasks, Task 1 constrains systems to select from predefined options. This design removes ambiguity in answer normalization and enables exact accuracy scoring while retaining the need for financial knowledge, multilingual terminology, and calculation-oriented reasoning.

3. Dataset

3.1. Data Collection

The final-test set contains 800 questions, with 200 questions in each of English, Chinese, Arabic, and Hindi. Each language split consists of normalized financial multiple-choice items with a question, a fixed answer-option set, and a single gold option label. The selected final-test items and their gold labels were withheld from the FinMMEval development leaderboards, so the portal did not reveal official answers before submission. The English and Chinese splits were drawn from CFA-style and CPA-style financial exam questions released with RealFin [11]. The Arabic split was drawn from the Arabic accounting and finance resources used in SAHM [20]. The Hindi split was drawn from the Hindi financial multiple-choice resources in BhashaBench [21].

The four-language design evaluates transfer beyond English-dominant financial QA benchmarks while covering languages with different scripts and resource profiles. Chinese and Arabic require systems to handle non-Latin scripts and domain terminology; Hindi adds an Indic-language financial QA setting.

3.2. Data Annotations

Each released Task 1 item contains a question, a finite set of answer options, and a unique question identifier. Each item has one gold answer label used for official scoring. The released data follow a shared multiple-choice schema across the four languages, while final-test labels remained hidden until the official results were computed. Most final-test items have four answer options; a subset retains two, three, or five options from the source resources instead of being forced into a uniform four-option format.

3.3. Data Statistics

The final test contains 200 questions in every language, while the public-development sets range from 70 to 100 items. Four-choice questions dominate all four splits, ranging from 79.0% in Chinese to 100%

in Hindi; Arabic is the only split with two-choice items, while English and Chinese retain smaller three- and five-choice subsets (Table 1).

Table 1

Dataset sources, split sizes, and answer-option distributions by language.

Language	Source	Dev	Test	2 opt.	3 opt.	4 opt.	5 opt.
English	CFA-style finance exams [11]	70	200	0	29	159	12
Chinese	CPA-style finance exams [11]	71	200	0	30	158	12
Arabic	Arabic finance exams [20]	97	200	20	0	180	0
Hindi	Hindi financial MCQs [21]	100	200	0	0	200	0

4. Evaluation Framework

4.1. Task Organization

Task 1 uses a public development release and a hidden final-test release under the same multiple-choice submission schema. The development leaderboards provided validation scores and baselines, while each final-test submission contained one JSON or JSONL prediction file per language. Gold answers, scores, and ranks remained hidden until the official results were released. For a submission to be ranked, its prediction file had to cover the full language split with one valid option label per item; rationales and confidence scores were not evaluated.

4.2. Evaluation Measures

The official metric is accuracy:

$$\text{Accuracy} = \frac{\# \text{ correct answers}}{\# \text{ test questions}}.$$

Separate language leaderboards preserve language-specific performance and allow teams to participate in a subset of languages. Only complete, valid 200-item submissions are ranked. Tied systems retain the published display order but share the same accuracy. Because the four language splits contain different questions and option-count distributions, their scores are reported independently and should not be interpreted as a controlled comparison of language difficulty.

4.3. Baselines

The public-development release included four interpretable baselines: a fixed-seed random selector, an Always-A rule, a round-robin option selector, and a zero-shot Qwen2.5-0.5B-Instruct run. The random baseline sampled uniformly from the valid answer options shown for each item using a fixed seed. The Always-A baseline selected option A whenever it was available, and the round-robin baseline cycled through valid option labels in item order. The Qwen2.5-0.5B-Instruct baseline was run zero-shot without task-specific fine-tuning. Under uniform random selection, the expected accuracies implied by the option distributions are 25.9% for English, 26.0% for Chinese, 27.5% for Arabic, and 25.0% for Hindi. Together, the baselines span chance, simple heuristics, and lightweight zero-shot modeling.

5. Results and Participant Approaches

5.1. Competition Results

Across the language-specific top-five results in Table 2, leading accuracies ranged from 92.0% in Hindi to 97.5% in English and Arabic. *pjmathematician* ranked first in English, Chinese, and Arabic, while

fosu ltw and *pjmathematician* shared the highest Hindi accuracy. The same three teams, *pjmathematician*, *fosu ltw*, and *Pranshu Rastogi*, formed the leading group in every language, although their order and score margins varied. English attracted the most valid ranked submissions (13), followed by Chinese and Arabic (11 each) and Hindi (10).

Table 2

Top-five final-test results by language.

Language	Team	Order	Correct	Acc. (%)
English	pjmathematician	1	195/200	97.5
English	fosu ltw	2	187/200	93.5
English	Pranshu Rastogi	3	182/200	91.0
English	NLP-DE	4	175/200	87.5
English	UyoAngle	5	166/200	83.0
Chinese	pjmathematician	1	193/200	96.5
Chinese	fosu ltw	2	184/200	92.0
Chinese	Pranshu Rastogi	3	184/200	92.0
Chinese	NLP-DE	4	174/200	87.0
Chinese	UyoAngle	5	154/200	77.0
Arabic	pjmathematician	1	195/200	97.5
Arabic	Pranshu Rastogi	2	190/200	95.0
Arabic	fosu ltw	3	188/200	94.0
Arabic	NLP-DE	4	183/200	91.5
Arabic	TCLabs	5	179/200	89.5
Hindi	fosu ltw	1	184/200	92.0
Hindi	pjmathematician	2	184/200	92.0
Hindi	Pranshu Rastogi	3	183/200	91.5
Hindi	TCLabs	4	180/200	90.0
Hindi	UyoAngle	5	174/200	87.0

5.2. Analysis of Results

The four leaderboards show similar high-end performance but different competitive profiles. English, Chinese, and Arabic all exceeded 96% at the top, whereas Hindi topped out at 92.0%; however, Hindi also had the narrowest top-five spread. Chinese showed the widest spread, at 19.5 percentage points between first and fifth place, followed by English, Arabic, and Hindi. These differences describe the submitted systems on four non-parallel question sets; they do not isolate language effects from differences in content or team participation.

5.3. Participant Approaches

Table 3 summarizes components reported in eight Task 1 Working Notes papers. Checkmarks indicate components explicitly reported by the authors; blank cells mean that the component was not reported. Output validation was common, while retrieval, language routing, self-consistency, and ensembling were used more selectively.

Table 3

Components reported by the eight documented Task 1 systems.

Team	Backbone		Evidence	Prompting		Post-processing	
	Gemini GPT	Qwen	Search/retrieval	Language routing Few-shot	Self-consistency	Validation	Ensemble
pjmathematician [22]	✓		✓			✓	
fosu ltw [23]		✓				✓	
Pranshu Rastogi [24]	✓			✓	✓	✓	
TCLabs [25]		✓	✓			✓	✓
UyoAngle [26]		✓					
AI_TLfanclub [27]			✓			✓	
TextSentinels [28]		✓	✓				✓
DS@GT [29]		✓	✓	✓			

pjmathematician* [22](Gemini, Google Search grounding, retrieval augmentation, multiple-choice financial QA)*

The system uses Gemini with Google Search grounding to provide external context before selecting the final answer option. Strict answer parsing and lightweight post-processing preserve the exact option-label format required by the evaluation protocol.

fosu ltw* [23](GPT-5.5, Doubao Expert Mode, expert review, confidence tiers)*

The *fosu ltw* workflow uses GPT-5.5 to generate initial predictions and Doubao Expert Mode to review financial formulas, accounting standards, policy timelines, and confidence tiers. The reviewed outputs are converted into single answer labels, while confidence tiers support internal consistency checks.

Pranshu Rastogi* [24](Gemini, multilingual prompting, few-shot examples, self-consistency)*

Pranshu Rastogi uses prompt-engineered Gemini models for Task 1. The system uses language-specific prompt construction, few-shot examples, confidence-based decision logic, and selective self-consistency. The prompting strategy is adapted to the language and question type rather than applied uniformly across all four splits.

TCLabs* [25](GPT-4o-mini, multi-agent reasoning, retrieval augmentation, structured prompts)*

TCLabs uses GPT-4o-mini within a multi-agent architecture containing specialized financial-reasoning, retrieval, and verification roles. Structured prompts and debate-style checks screen unsupported answer choices before the final option is returned.

UyoAngle* [26](GPT-5.2 zero-shot, Qwen2.5 ablations, tokenizer analysis, confidence routing)*

The official *UyoAngle* submission uses GPT-5.2 in a zero-shot setting with structured single-label output in all four languages. The paper separately analyzes Qwen2.5-14B-Instruct prompting variants for tokenizer sensitivity, confidence-margin routing, definition retrieval, and parse-failure fallback; these experiments are not components of the official submission.

***AI_TLfanclub* [27]**

(Qwen, distillation, answer-label scoring, calibration)

AI_TLfanclub reports a Task 1 system based on distilled Qwen models for multilingual financial exam-style QA. The system emphasizes direct answer-label scoring, output calibration, and component selection based on development-set performance.

***TextSentinels* [28]**

(Qwen3-32B, GPT-OSS-20B, question-type routing, FAISS retrieval)

The final system uses Qwen3-32B as both the question-type router and a factual/conceptual scorer, with GPT-OSS-20B as a reasoning-oriented scorer. It assigns route-dependent FAISS retrieval depth and combines option-wise scores from the two models before returning the selected label.

***DS@GT* [29]**

(Qwen3, Qwen2.5, language routing, exemplar retrieval, direct option scoring)

DS@GT routes each language to either Qwen3-14B or Qwen2.5-14B and retrieves solved exemplars from language-specific indices. Its Retrieval-Augmented Direct Scoring procedure compares next-token probabilities for the available option labels instead of generating an unconstrained answer. Weighted reciprocal-rank fusion provides a fallback when native-language exemplars are insufficient.

The documented systems pursue the same constrained objective through different combinations of direct option scoring, language-specific prompting or routing, exemplar retrieval, and output control. Across these papers, model families vary, but several systems separate evidence acquisition or exemplar retrieval from constrained option selection and parsing.

6. Conclusion and Future Work

FinMMEval 2026 Task 1 was a shared-task track with a fixed evaluation protocol for financial multiple-choice QA in English, Chinese, Arabic, and Hindi. The task used hidden gold answers, a fixed answer-label schema, and language-specific accuracy leaderboards. The top systems achieved high accuracy, but participation, score spread, and top performance differed by language.

Future iterations should report accuracy by question type and financial topic alongside participation by language, helping distinguish language effects from differences in content and the set of participating systems. A unified multilingual score could complement, rather than replace, language-specific leaderboards if future editions use parallel or otherwise normalized cross-language test designs and attract sufficient four-language participation.

Acknowledgments

We thank the CLEF 2026 organizers for hosting the lab and supporting the working-notes process. We also thank all participating teams for submitting systems and providing feedback on the submission portals and evaluation workflow. We are grateful to Georgi Georgiev for supporting the FinMMEval awards, which helped recognize strong participant submissions across the tasks. The work of Dimitar Dimitrov and Ivan Koychev is partially supported by the project UNITE BG16RFPR002-1.014-0004 funded by PRIDST and also by the EU NextGenerationEU project, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No. BG-RRP-2.004-0008.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI GPT-5.5 for grammar checks, wording revisions, style improvements, and LaTeX formatting assistance.

References

- [1] Z. Xie, T. Cohn, J. H. Lau, The next chapter: A study of large language models in storytelling, in: C. M. Keet, H.-Y. Lee, S. Zarri   (Eds.), *Proceedings of the 16th International Natural Language Generation Conference*, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 323–351. URL: <https://aclanthology.org/2023.inlg-main.23/>. doi:10.18653/v1/2023.inlg-main.23.
- [2] 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, Jena, Germany, 2026.
- [3] Z. Xie, X. Peng, G. Georgiev, D. Dimitrov, Y. Dai, R. Elbadry, V. Jani, L. Qian, 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 Task 2: Multilingual Financial Short-Answer Question Answering, in: *CLEF 2026 Working Notes*, *CEUR Workshop Proceedings*, CEUR-WS.org, Jena, Germany, 2026.
- [4] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, Y. Dai, R. Elbadry, V. Jani, X. Peng, 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 Task 3: Live Financial Decision-Making Agents, in: *CLEF 2026 Working Notes*, *CEUR Workshop Proceedings*, CEUR-WS.org, Jena, Germany, 2026.
- [5] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, A. Balahur, WWW’18 open challenge: Financial opinion mining and question answering, in: P. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), *Companion Proceedings of The Web Conference 2018*, Association for Computing Machinery, Lyon, France, 2018, pp. 1941–1942. URL: <https://doi.org/10.1145/3184558.3192301>. doi:10.1145/3184558.3192301.
- [6] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, B. Vidgen, FinanceBench: A new benchmark for financial question answering, 2023. URL: <https://arxiv.org/abs/2311.11944>. doi:10.48550/arXiv.2311.11944. arXiv:2311.11944.
- [7] 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: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3697–3711. URL: <https://aclanthology.org/2021.emnlp-main.300/>. doi:10.18653/v1/2021.emnlp-main.300.
- [8] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, T.-S. Chua, TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, 2021, pp. 3277–3287. URL: <https://aclanthology.org/2021.acl-long.254/>. doi:10.18653/v1/2021.acl-long.254.
- [9] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, W. Y. Wang, ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6279–6292. URL: <https://aclanthology.org/2022.emnlp-main.421/>. doi:10.18653/v1/2022.emnlp-main.421.
- [10] Z. Xie, D. Orel, R. Thareja, D. Sahnan, H. Madmoun, F. Zhang, D. Banerjee, G. N. Georgiev, X. Peng, L. Qian, J. Huang, J. Su, A. Singh, R. Xing, R. Elbadry, C. Xu, H. Li, F. Koto, I. Koychev, T. Chakraborty, Y. Wang, S. Lahlou, V. Stoyanov, S. Ananiadou, P. Nakov, FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*

- (Volume 1: Long Papers), Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 14529–14553. URL: <https://aclanthology.org/2026.acl-long.662/>. doi:10.18653/v1/2026.acl-long.662.
- [11] Y. Dai, Y. Lin, Z. Xie, Y. Wang, RealFin: How well do LLMs reason about finance when users leave things unsaid?, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), Findings of the Association for Computational Linguistics: ACL 2026, Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 25050–25080. URL: <https://aclanthology.org/2026.findings-acl.1255/>. doi:10.18653/v1/2026.findings-acl.1255.
 - [12] Y. Zhou, F. Zhang, Y. Chen, H. Zhang, P. Nakov, Z. Xie, FinCARDS: Card-based analyst reranking for financial document question answering, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), Findings of the Association for Computational Linguistics: ACL 2026, Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 24836–24852. URL: <https://aclanthology.org/2026.findings-acl.1244/>. doi:10.18653/v1/2026.findings-acl.1244.
 - [13] F. Zhang, M. Song, R. Elbadry, Y. Chen, S. Wang, Y. Zhou, X. Zheng, Y. He, Y. Dai, G. N. Georgiev, A. Gull, M. U. Safder, F. Wu, L. Meng, F. Ji, J. Zhao, X. Peng, J. Huang, Y. Chen, X. Liu, P. Nakov, Z. Xie, FinReporting: An agentic workflow for localized reporting of cross-jurisdiction financial disclosure, in: G. Durrett, P. Jian (Eds.), Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 728–735. URL: <https://aclanthology.org/2026.acl-demo.71/>. doi:10.18653/v1/2026.acl-demo.71.
 - [14] X. Peng, Z. Xie, Y. Cao, H. Li, L. Qian, Y. Wang, V. J. Zhang, H. He, X. Ai, L. Ma, R. Xiang, Y. He, Y. Han, S. Wang, Y. Guo, M. Jiang, Y. Zhao, Y. Dong, X. Wang, Y. Chen, Y. Yuan, Q. Zhang, F. Lyu, H. Wu, Y. Yang, Z. Zhao, Y. Dai, F. Zhang, R. Elbadry, A. Gull, M. U. Safder, N. Chen, F. Zhu, T. Cai, Z. Wang, P. Giannouris, Y. Jiang, Z. Liu, M. Kabir, Y. Wang, Y. Zheng, Y. Yu, W. Liu, W. Cao, A. Xu, P. Lu, J. Huang, M. Lin, P. Tiwari, Y. Zhao, V. Gutiérrez-Basulto, X.-Y. Liu, K. E. Smith, J. Pei, A. Cohan, J. Huang, Y. Tang, A. Lopez-Lira, X. Chen, X. Liu, J. Tsujii, J.-Y. Nie, S. Ananiadou, Herculean: An agentic benchmark for financial intelligence, 2026. URL: <https://arxiv.org/abs/2605.14355>. doi:10.48550/arXiv.2605.14355. arXiv:2605.14355.
 - [15] Z. Liu, D. Huang, K. Huang, Z. Li, J. Zhao, FinBERT: A pre-trained financial language representation model for financial text mining, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4513–4519. URL: <https://doi.org/10.24963/ijcai.2020/622>. doi:10.24963/ijcai.2020/622, special Track on AI in FinTech.
 - [16] S. Wu, O. Irsoy, S. Lu, V. Dabrovolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, BloombergGPT: A large language model for finance, 2023. URL: <https://arxiv.org/abs/2303.17564>. doi:10.48550/arXiv.2303.17564. arXiv:2303.17564.
 - [17] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, J. Huang, PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 33469–33484. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/6a386d703b50f1cf1f61ab02a15967bb-Paper-Datasets_and_Benchmarks.pdf.
 - [18] H. Yang, X.-Y. Liu, C. D. Wang, FinGPT: Open-source financial large language models, 2023. URL: <https://arxiv.org/abs/2306.06031>. doi:10.48550/arXiv.2306.06031. arXiv:2306.06031.
 - [19] 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, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, San Diego,

- California, United States, 2026, pp. 16934–16963. URL: <https://aclanthology.org/2026.acl-long.770/>. doi:10.18653/v1/2026.acl-long.770.
- [20] R. Elbadry, S. Ahmad, A. Heakl, D. Bouch, M. Ahsan, M. AlMahri, M. E. 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, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 34509–34536. URL: <https://aclanthology.org/2026.acl-long.1593/>. doi:10.18653/v1/2026.acl-long.1593.
 - [21] 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>. doi:10.48550/arXiv.2510.25409. arXiv:2510.25409.
 - [22] P. Vachharajani, pjmathematician @ FinMMEval 2026: Systems for Tasks 1, 2 and 3, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [23] T. Liang, K. Lin, Y. Qi, Z. Han, fosu ltw @ FinMMEval 2026 Task 1: A traceable autoresearch workflow for multilingual financial multiple-choice answer review, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [24] P. Rastogi, Pranshu Rastogi @ FinMMEval 2026: Systems for Tasks 1 and 2 – prompt-engineered Gemini for multilingual and cross-lingual financial reports QA, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [25] E. L. Pontes, M. Benjannet, TCLabs @ FinMMEval 2026: Systems for Tasks 1, 2 and 3, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [26] L. Ouyang, UyoAngle @ FinMMEval 2026 Task 1: Tokenizer- and confidence-aware prompting for multilingual financial multiple-choice QA, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [27] D. T. Duc, N. N. Minh, L. T. Huong, AI_TLfanclub @ FinMMEval 2026 Task 1: Distilled Qwen models for multilingual financial exam question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [28] D. Thenmozhi, A. Gopinath, A. Sivakumar, A. A. A. Balasubramanian, TextSentinels @ FinMMEval 2026 Task 1: A multilingual routed retrieval-augmented system for financial multiple choice questions, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
 - [29] J. Ayela, K. Sahni, Language-routed RAG and direct option scoring for multilingual financial QA: DS@GT at FinMMEval, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.