Overview of FinMMEval 2026 Task 2: Multilingual Financial Short-Answer Question Answering

Zhuohan Xie[1,*], Xueqing Peng[2], Georgi Georgiev[3], Dimitar Dimitrov[3], Yuyang Dai[4], Rania Elbadry[1], Vanshikaa Jani[5], Lingfei Qian[2], Fan Zhang[6], Jimin Huang[2], Jiahui Geng[7], Yankai Chen[1,8], Ye Yuan[8,9], Haolun Wu[8,9], Yuxia Wang[4], Ivan Koychev[3], Veselin Stoyanov[1], Mingzi Song[10], Yu Chen[6], Xue Liu[1,8,9] and Preslav Nakov[1]

[1]*Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates*
[2]*The Fin AI, United States*
[3]*FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria*
[4]*INSAIT, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria*
[5]*University of Arizona, Tucson, United States*
[6]*The University of Tokyo, Tokyo, Japan*
[7]*Linköping University, Linköping, Sweden*
[8]*McGill University, Montreal, Canada*
[9]*Mila, Quebec AI Institute, Montreal, Canada*
[10]*Meiji Gakuin University, Tokyo, Japan*

Abstract

FinMMEval 2026 Task 2 evaluates short-answer financial question answering over multilingual evidence. Each final-test item pairs an English question with financial statements and news in English, Chinese, Japanese, Spanish, and Greek. Participating systems submit one concise answer per item in JSONL format. The final-test set contains 256 items, split evenly between easy and expert tiers; each tier contains four question templates instantiated over 32 company-report groups. Gold answers were withheld during submission, and systems were ranked by macro-averaged item-level ROUGE-1 F1 against organizer-held reference answers. The final leaderboard includes 12 ranked submissions. The strongest systems are closely clustered, with the top four separated by less than one percentage point in ROUGE-1 F1. The submitted system papers document retrieval-augmented generation, cross-lingual evidence handling, structured prompting, answer compression, and validation strategies.

Keywords

FinMMEval, financial question answering, multilingual evaluation, retrieval-augmented generation, ROUGE

1. Introduction

Financial question answering requires evidence selection, accounting and entity interpretation, and numerical or domain reasoning. In a short-answer setting, systems must also compress the selected evidence without losing the financial facts required by the question. Task 2 therefore evaluates concise answers by their lexical overlap with an organizer-held reference; this captures one dimension of answer quality rather than serving as a proxy for factual correctness [1]. It is part of the broader FinMMEval lab at CLEF 2026 [2], alongside companion tracks on multilingual multiple-choice QA in Task 1 [3] and live trading agents in Task 3 [4]. Task 2 builds on the PolyFiQA resources from MultiFinBen [5], a benchmark suite for multilingual and multimodal financial applications. Its hidden-test design requires systems to align financial statements with multilingual news and express the resulting evidence in an answer that can be evaluated against an organizer-held reference.

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)

2. Related Work

Task 2 is related to financial question answering, multilingual financial NLP, and automatic short-answer evaluation. FiQA introduced financial QA in an information-retrieval-oriented shared-task setting [6]. FinQA [7] and TAT-QA [8] focus on numerical and tabular reasoning over financial reports, while ConvFinQA adds a conversational setting in which systems must track evidence across turns [9]. FinanceBench studies open-book financial QA over public-company information [10]. FinChain [11] and RealFin [12] further emphasize verifiable reasoning steps and implicit financial assumptions.

Task 2 also relates to multilingual and document-grounded financial evaluation. MultiFinBen extends financial evaluation to multilingual and multimodal applications [5]. SAHM provides a complementary Arabic financial reasoning benchmark [13]. Finance-domain representation learning provides modeling context for financial reports, news, and terminology, from encoder-style FinBERT [14] to large-scale BloombergGPT [15]. Instruction and evaluation resources such as PIXIU [16] and FinGPT [17] broaden this toolset for financial LLM systems. Evidence selection is also central to intra-document evidence reranking for financial question answering in FinCARDS [18] and to localized disclosure reporting in FinReporting [19]. Herculean extends financial evaluation from static tasks to end-to-end agentic workflows [20]. Where many prior settings focus on a fixed document collection or a single question-answering style, Task 2 emphasizes multilingual financial information needs and concise answer generation under a shared hidden-test protocol.

3. Dataset

3.1. Data Collection

The final-test set was constructed from organizer-held November versions of PolyFiQA-Easy and PolyFiQA-Expert. PolyFiQA was introduced as part of MultiFinBen [5]. Each tier contained 204 candidate items. We excluded 76 items whose source task identifier or full query matched the corresponding public PolyFiQA release, because those releases exposed both the questions and their reference answers. The remaining 128 items per tier formed a 256-question hidden test.

The questions and answer instructions are in English. The multilingual component lies in the evidence bundle: each item combines financial statements with financial news in English, Chinese, Japanese, Spanish, and Greek. Easy questions ask for factual or numerical findings from the statements and supporting news evidence, whereas expert questions ask for news-based synthesis supported by financial-statement evidence.

3.2. Data Annotations

Each final-test item pairs a unique task identifier and difficulty tier with an English question and its multilingual evidence context. Tier-specific answer templates distinguish easy responses, which provide an answer and news evidence, from expert responses, which provide an answer and supporting financial-statement evidence. Reference answers were retained from the source PolyFiQA rows after overlap removal and remained organizer-held. Each submission provided one non-empty answer per task identifier, with `None` reserved for cases in which no answer could be found.

3.3. Data Statistics

In each tier, 51 candidate source groups with four questions each produced 204 items; 19 groups already represented in the public release accounted for the 76 exclusions; and 32 retained groups yielded the remaining $204 - 76 = 128$ held-out items. Easy and Expert use the same 32 retained groups, providing four questions from each tier per group and 256 final-test items overall (Table 1).

Table 1
Held-out test construction by tier. Public-overlap items matched the corresponding public PolyFiQA release and were excluded.

Tier	Candidate items	Public overlap	Held-out items	Retained groups	Items/ group
Easy	204	76	128	32	4
Expert	204	76	128	32	4
Combined	408	152	256	32	8

4. Evaluation Framework

4.1. Task Organization

Under the hidden-test protocol, each submission was a JSONL file containing exactly one non-empty answer for each of the 256 official task identifiers. Reference answers and scores remained hidden until evaluation, and the final complete valid submission from each team entered the official leaderboard.

4.2. Evaluation Measures

The official metric is macro-averaged item-level ROUGE-1 F1 [21], computed between submitted answers and organizer-held reference answers. The scorer processes the complete string stored in each `answer` field, including section labels and evidence text. It applies Unicode NFKC normalization, lowercases the text, tokenizes words, numbers, CJK characters, Japanese kana, Arabic-script characters, and Greek/Latin ranges with a regular expression, and computes unigram multiset overlap. For each item, unigram precision P and recall R are computed against the reference answer, and item-level F1 is:

$$\text{ROUGE-1 F1} = \frac{2PR}{P+R}.$$

Because all reference answers in the final test are non-empty, an item receives an F1 score of 0 when $P + R = 0$. The leaderboard score averages the 256 item-level ROUGE-1 F1 values; mean item-level precision and recall are reported as diagnostic components of unigram overlap. The entire response contributes to ROUGE-1; the 100-word answer limit is a formatting guideline rather than a separate scoring term.

4.3. Development References and Baselines

Public PolyFiQA examples and the submission validator provided development references for answer formatting and end-to-end pipeline checks. The final leaderboard reports participant systems only and does not include an organizer model baseline as a ranked entry.

5. Results and Participant Approaches

5.1. Competition Results

The final Task 2 ranking contains 12 complete submissions spanning direct extraction, retrieval, cross-lingual processing, structured prompting, and answer validation. The top of the leaderboard is tightly clustered: *Calibrated Signals* ranked first with a macro-averaged ROUGE-1 F1 of 31.18%, and only 0.79 percentage points separated the top four systems (Table 2). The frozen leaderboard aggregates all 256 items; separate easy- and expert-tier scores were not released, so the official results do not support tier-specific performance comparisons.

Table 2
Official Task 2 final-test leaderboard with macro-averaged ROUGE-1 F1 and diagnostic precision and recall.

Rank	Team	F1 (%)	Prec. (%)	Rec. (%)
1	Calibrated Signals	31.18	30.25	37.64
2	Pranshu Rastogi	30.96	32.08	35.37
3	IGT	30.71	28.21	40.44
4	AI_TLfanClub	30.39	36.47	30.70
5	pjmathematician	29.72	27.81	37.98
6	DS@GT	28.53	25.72	38.88
7	The Lab Rats	27.34	27.37	34.17
8	Ethereum Team B CLEF	25.46	30.86	26.42
9	TCLabs	25.06	23.21	32.57
10	HU_LLM_Fin	22.89	25.81	28.03
11	NLP-DE	20.49	26.43	19.01
12	TextSentinels	18.73	30.29	17.56

5.2. Analysis of Results

The diagnostic precision and recall values reveal different lexical-overlap profiles. *IGT* recorded the highest recall (40.44%), whereas *AI_TLfanClub* recorded the highest precision (36.47%). Because the official score is macro-averaged ROUGE-1 F1, different balances between unigram precision and recall can produce similar F1 values; neither diagnostic column directly measures factual correctness. Because the release includes no organizer baseline, tier-specific results, or paired uncertainty analysis, the leaderboard supports comparisons among submitted systems but does not establish absolute task difficulty or statistically reliable differences among the tightly clustered leaders.

5.3. Participant Approaches

Table 3 summarizes components reported in ten Task 2 Working Notes papers. Blank cells mean that a component was not reported. Structured prompting appears throughout the documented systems, while evidence handling varies among direct extraction, retrieval, and reranking.

Table 3
Components reported by the ten documented Task 2 systems.

	Evidence				Reasoning		Output		
Team	**Extraction**	**Retrieval**	**Reranking**	**Tools**	**Agent graph**	**Structured prompting**	**Translation/routing**	**Compression**	**Validation/MBR**
IGT [22]	✓					✓	✓		
Calibrated Signals [23]						✓		✓	✓
TextSentinels [24]		✓				✓			
AI_TLfanClub [25]		✓				✓			✓
DS@GT [26]		✓	✓	✓	✓	✓	✓		✓
The Lab Rats [27]			✓			✓	✓	✓	
Pranshu Rastogi [28]	✓					✓	✓		
HU_LLM_Fin [29]		✓	✓			✓	✓		
TCLabs [30]					✓	✓			
pjmathematician [31]						✓		✓	✓

***IGT* [22]**
(question-type prompting, targeted extraction, multilingual financial QA)
The system handles structured numerical questions with direct keyword and regular-expression extraction over filing text; for synthesis-oriented questions, it applies rule-based selection of multilingual passages before concise generation. Its official implementation uses Claude Sonnet 4 through AWS Bedrock with type-specific prompts for the easy and expert question families.

***Calibrated Signals* [23]**
(inference-only prompting, deterministic decoding, schema enforcement)
Calibrated Signals reports an inference-only system for cross-lingual financial question answering. The system uses a fixed deployable prompt, deterministic decoding, schema enforcement, and one constrained retry when the required answer format is violated.

***TextSentinels* [24]**
(BM25 retrieval, evidence preservation, multilingual QA)
TextSentinels builds a BM25-grounded method for multilingual financial QA. Its final pipeline preserves chunk-local evidence, limits overly aggressive retrieval, and constrains generation with the selected passages.

***AI_TLfanClub* [25]**
(FiSCO-RAG, cross-lingual evidence recall, BM25 retrieval, context packing)
AI_TLfanClub proposes *FiSCO-RAG*, a retrieval-augmented approach designed for cross-lingual evidence recall in financial QA. The system combines layout-aware chunking, CJK-aware BM25 retrieval, partition-aware context packing, schema-conditioned prompting, verifier filtering, and IDF-regularized minimum Bayes risk answer selection.

***DS@GT* [26]**
(FinNexus, LangGraph orchestration, Weaviate retrieval, reranking, tool execution)
DS@GT introduces *FinNexus*, an agentic retrieval-augmented reranking system for multilingual financial QA. The system uses LangGraph orchestration with Weaviate retrieval, reranking, and Python tool execution for numerical analysis. Retrieved passages are evaluated by a judging agent before tool-assisted reasoning and short-answer generation.

***The Lab Rats* [27]**
(GPT-4o-mini translation, evidence ranking, GPT-4o reasoning, answer compression)
The Lab Rats use GPT-4o-mini for translation and evidence ranking, GPT-4o for chain-of-thought reasoning, and GPT-4o-mini again for answer compression. The system addresses cross-lingual inputs by normalizing the question context before reasoning and generation. The final response is produced through a second call that compresses the reasoning output into the required short-answer format.

***Pranshu Rastogi* [28]**
(Gemini, prompt engineering, tier-aware prompts, cross-lingual report/news QA)
Pranshu Rastogi uses prompt-engineered Gemini models for Task 2. The system treats cross-lingual financial report and news QA as a prompt-design problem. It uses separate prompts for easy and expert questions and targeted numerical-extraction passes where needed, without adding a retrieval component.

***HU_LLM_Fin* [29]**
(hybrid RAG, row-aware chunking, partitioned reranking, mixture-of-experts)
HU_LLM_Fin reports a structure-aware hybrid RAG system using mixture-of-experts generation. The system preserves table structure through row-aware chunking, combines dense and lexical retrieval, applies partitioned cross-encoder reranking, and generates answers with the mixture-of-experts model Llama 4 Scout using tier-specific prompts.

***TCLabs* [30]**
(multilingual QA, expert agents, structured prompts, answer generation)
TCLabs describes a Task 2 multilingual QA system with separate news, report, and financial-reasoning experts before final answer generation. An answer-generation agent consolidates the news and report analyses into a concise response.

***pjmathematician* [31]**
(Gemini, structured prompting, structural validation, answer compression)
The *pjmathematician* Task 2 system uses Gemini to produce concise, evidence-grounded answers from financial filings and multilingual news inputs. Its validation step checks output structure and answer length, aiming to preserve the core evidence while satisfying the short-answer format.

Across the documented systems, recurring components included sparse or dense retrieval, evidence reranking, translation, tier-specific prompting, schema validation, and answer compression. The tightly clustered leaders use prompt-only, extraction-oriented, and retrieval-augmented approaches, but results for complete systems do not isolate the components responsible for the small score differences.

6. Conclusion and Future Work

FinMMEval 2026 Task 2 evaluated systems on a 256-item hidden test of financial short-answer QA grounded in multilingual evidence and derived from PolyFiQA. The official leaderboard, based on macro-averaged ROUGE-1 F1, contains 12 ranked submissions. The documented systems include both retrieval-augmented and prompt-only approaches, with different balances between precision and recall under lexical-overlap scoring. Task 2 is a compact hidden-test setting for open-ended financial QA beyond multiple-choice evaluation. Future editions should report performance by question type, evidence language, and difficulty tier, and should complement ROUGE with semantic or human-audited factuality checks to better distinguish wording effects from financial understanding.

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. Zarrieß (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, 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, X. Liu, P. Nakov, Overview of FinMMEval 2026 Task 1: Multilingual Financial Multiple-Choice 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] 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.

[6] 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.

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[22] Y. Chiu, IGT @ FinMMEval 2026 Task 2: Question-type prompting with targeted extraction for multilingual financial QA, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[23] G. Jain, H. Hitesh, S. Sundriyal, S. Bhardwaj, H. Behl, P. Gautam, S. Zimmerman, Calibrated Signals @ FinMMEval 2026 Task 2: Inference-only strategies for cross-lingual financial question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[24] D. Thenmozhi, A. Sivakumar, A. Balasubramanian, A. A, A. Gopinath, TextSentinels @ FinMMEval Task 2: Chronological stabilization of a BM25-grounded multilingual financial question answering pipeline, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[25] D. T. Duc, N. N. Minh, L. T. Huong, AI_TLfanClub at FinMMEval 2026 Task 2: FiSCO-RAG for cross-lingual evidence recall in financial question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[26] D. Liu, S. Li, S. Tian, DS@GT at FinMMEval 2026 Task 2: FinNexus-agentic retrieval-augmented reranking orchestration for multilingual financial question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[27] A. Gill, E. N. Sheikh, F. T. Zahra, A. Samad, F. Alvi, S. Kumar, The Lab Rats @ FinMMEval 2026 Task 2: LLM translation and BAS-tuned chain-of-thought for multilingual financial QA, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[28] 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.
[29] M. Affan, M. B. Qureshi, M. H. Shahzad, F. Alvi, A. Samad, HU_LLM_Fin @ FinMMEval 2026 Task 2: A structure-aware hybrid RAG pipeline utilizing mixture-of-experts, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[30] 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.
[31] 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.