Overview of FinMMEval 2026 Task 3: Live Financial Decision-Making Agents

Zhuohan Xie[1,*], Lingfei Qian[2], Georgi Georgiev[3], Dimitar Dimitrov[3], Yuyang Dai[4], Rania Elbadry[1], Vanshikaa Jani[5], Xueqing Peng[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 3 evaluates financial decision-making agents in a live endpoint-based setting. Participating teams deployed public HTTP endpoints that received daily market payloads and returned BUY, HOLD, or SELL decisions for BTC and TSLA. Unlike static question-answering benchmarks, this task evaluates trading decisions produced by complete agent services under repeated daily calls; endpoint reachability and response validity are deployment constraints of the protocol. For each accepted endpoint, the official ranking averages the BTC and TSLA cumulative returns, each computed from a long/flat/short portfolio simulation with an initial balance of 100,000. Within the official scoring window, simulated returns differed sharply by asset: *Infraglyph* led the cross-asset mean CR, *Fin-Analyst* achieved the highest TSLA return, and *NLP-DE* achieved the highest BTC return. This divergence shows why a multi-asset evaluation should report both aggregate and asset-level results.

Keywords

FinMMEval, financial agents, live evaluation, trading agents, endpoint evaluation, cumulative return

1. Introduction

Task 3 evaluates whether deployed agents can convert price history, momentum features, news, and filing-derived context into daily positions scored against subsequent market prices under the official portfolio simulation. The protocol scores the admissible market action rather than the surface form of a generated explanation, thereby separating the target decision from other qualities of the output [1]. A live trading-agent task also introduces deployment constraints: systems must remain reachable and schema-compliant across repeated daily calls. It is part of the broader FinMMEval lab at CLEF 2026 [2], complementing the static Task 1 multiple-choice QA setting [3] and the Task 2 short-answer QA setting [4]. Task 3 uses a FinMMEval-specific HTTP endpoint protocol inspired by the live multi-market agent-evaluation setting introduced in When Agents Trade [5]. Participants submit endpoints rather than static files; each endpoint receives the same daily payload for a given asset and date and returns a recommended BUY, HOLD, or SELL action. The primary metric is Cumulative Return (CR), which connects daily decisions to their realized market consequences. By extending the lab from hidden

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)

static QA to live endpoint-based action selection, the task connects financial interpretation to simulated portfolio outcomes while exposing practical service constraints.

2. Related Work

Task 3 is related to finance-specific language models, financial agent evaluation, and market-decision benchmarks. Financial language models and instruction resources have shown the value of domain-specific pretraining, instruction data, and evaluation suites for financial NLP. FinBERT represents an early encoder-based direction [6], while BloombergGPT studies large-scale language modeling for finance [7]. PIXIU adapts instruction data and evaluation resources to broader financial tasks [8], and FinGPT provides an open framework for financial LLM applications [9].

Several financial reasoning benchmarks study complementary evidence-interpretation problems. RealFin [10] and SAHM [11] focus on financial reasoning under language- or region-specific settings. FinChain emphasizes verifiable chain-style reasoning in financial settings [12]. FinCARDS studies intra-document evidence reranking for financial question answering [13], while FinReporting examines localized disclosure-reporting workflows [14].

Recent shared tasks and benchmarks have further moved toward agentic financial scenarios. The FinNLP-AgentScen shared task studied finance-oriented LLM settings including classification, summarization, and single-stock trading [15]. Herculean studies broader agentic financial intelligence under finance-oriented decision settings [16]. When Agents Trade introduced a live multi-market trading arena for LLM agents [5]. FinMMEval Task 3 adapts this direction to a CLEF shared-task setting by requiring participants to maintain public endpoints while ranking systems by market return.

3. Dataset

3.1. Data Collection

The Task 3 live challenge focuses on BTC and TSLA. BTC represents a continuously traded crypto asset with rapid market movements, whereas TSLA represents a highly followed equity affected by company news, filings, earnings expectations, and broader market dynamics. The official reporting window uses BTC evaluations from 2026-05-11 through 2026-07-04 and TSLA evaluations from 2026-05-11 through 2026-06-26. Organizer-side payload-generation failures affected 17 calendar dates: 18–19 May, 31 May–11 June, and 24–26 June. Rows from those dates that entered official scoring were normalized to HOLD/flat, while non-trading TSLA dates were excluded according to the trading calendar. Validated BTC coverage extended through 4 July, whereas TSLA scoring ended on 26 June. Any open TSLA position was liquidated on the final scored date, and no further TSLA profit or loss was recorded.

For each active endpoint, asset, and evaluation date, the evaluation runner sent a JSON payload containing the date, current price information, recent price history, momentum features, news-derived information, and available filing context. All participants could use this shared payload directly or combine it with their own model architectures, memory modules, retrieval systems, forecasting components, rule-based signals, or external infrastructure. Decisions were generated and stored at the time of each daily call and ingested into the leaderboard database later; database update timestamps therefore record ingestion time rather than endpoint-call time. The raw archive preserves all collected endpoint rows, including TSLA rows from non-trading dates. Official scoring includes all 18 endpoints on every valid asset date and retains the agent and model identifiers, archived daily close price, and normalized action needed to reproduce the leaderboard simulation.

3.2. Data Annotations

Task 3 evaluation records contain decision labels produced by deployed participant agents rather than manually annotated static answers. For each asset-day request, an endpoint had to return valid JSON

containing one recommended action from the allowed set: BUY, HOLD, or SELL. BUY corresponds to a long position, SELL corresponds to a short position, and HOLD corresponds to a flat position.

3.3. Data Statistics

The frozen raw archive contains 1,813 collected rows, of which 1,584 enter official scoring after the validated asset windows and trading-day filters are applied. BTC contributes 990 scoring records across 55 dates, compared with 594 TSLA records across 33 dates, reflecting the shorter validated TSLA window (Table 1).

Table 1
Evaluation windows, endpoint coverage, and record counts for BTC and TSLA.

Metric	Value
BTC scoring window	2026-05-11 to 2026-07-04
TSLA scoring window	2026-05-11 to 2026-06-26
Active endpoints	18
Assets	2
Archived rows	1,813 (BTC: 990; TSLA: 823)
Scored rows	1,584 (BTC: 990; TSLA: 594)
Scoring dates	BTC: 55; TSLA: 33
Excluded TSLA non-trading rows	229

4. Evaluation Framework

4.1. Task Organization

The protocol sends one request to each active endpoint for each asset on each evaluation date. The action space is:

- BUY: take or maintain a long position;
- HOLD: stay flat for the asset on that day;
- SELL: take or maintain a short position.

The daily action replaces the previous position for that asset, allowing a system to move between long, flat, and short exposure as market evidence changes. Endpoint eligibility required public HTTP(S) reachability, valid HTTPS certificates where applicable, a response within 180 seconds, parseable JSON, and an allowed action. The official evaluation set retained one accepted endpoint per team, yielding 18 endpoints that the evaluation runner called daily within each asset's scoring window.

4.2. Evaluation Measures

The primary metric is Cumulative Return (CR) under the official daily long/flat/short portfolio simulation. Each asset starts with $B_0 = 100{,}000$. Executions incur a fee rate of 0.0006 (6 basis points) per trade leg and a slippage rate of 0.001 (10 basis points), implemented by increasing buy prices and decreasing sell prices by the slippage rate. An open position is marked to the current asset price before the new action is applied, and any position remaining on the final scored date is liquidated with the same costs. If B_T is the balance after final liquidation, the asset-level cumulative return is

$$\mathrm{CR} = \frac{B_T - B_0}{B_0},$$

where $B_0 = 100{,}000$ and B_T includes position returns, fees, and slippage over the validated asset window.

Failed or invalid actions and organizer-affected scoring rows are normalized to HOLD/flat, creating no directional exposure; non-trading TSLA raw dates are excluded by the trading calendar. The official ranking is the arithmetic mean of the reported BTC and TSLA cumulative returns. This gives the two assets equal weight despite their different validated horizons and does not normalize for duration or volatility.

4.3. HOLD Baseline

The HOLD baseline yields zero return; it is neither a participant entry nor a buy-and-hold benchmark. Participant systems are ranked only against one another using CR.

5. Results and Participant Approaches

5.1. Competition Results

Among the 18 accepted endpoints, *Infraglyph* led the cross-asset mean with 4.83%, followed by *Calibrated Signals* at 4.12% and *NLP-DE* at 4.03% (Table 2). The 0.09 percentage-point gap between second and third place conceals sharply different BTC and TSLA profiles, underscoring the need for asset-level comparison. Table 3 shows the asset-level divergence: *Fin-Analyst* achieved the highest TSLA return (13.33%), whereas *NLP-DE* achieved the highest BTC return (9.97%).

Table 2
Official Task 3 ranking by mean BTC and TSLA return after fees and slippage. Means are computed from unrounded asset-level returns; displayed values are rounded to two decimal places.

Rank	Team	Mean (%)	BTC (%)	TSLA (%)
1	Infraglyph	4.83	-0.68	10.34
2	Calibrated Signals	4.12	0.48	7.76
3	NLP-DE	4.03	9.97	-1.91
4	Fin-Analyst	3.94	-5.45	13.33
5	pjmathematician	1.45	2.05	0.84
6	The Lab Rats	0.56	4.26	-3.14
7	CLaC	0.37	3.50	-2.75
8	Enjamul Islam Khan	0.00	0.00	0.00
9	Amongus	-0.02	2.38	-2.42
10	TrustMe	-0.09	-3.62	3.44
11	Gentlemen	-0.15	-5.51	5.21
12	AI_TLfanclub	-1.24	-1.60	-0.88
13	AAA	-1.72	-2.50	-0.93
14	A	-5.03	-6.27	-3.79
15	TCLabs	-5.57	-9.70	-1.43
16	Club GAIA - ESCOM	-5.73	-4.40	-7.07
17	DS@GT ARC	-7.56	-12.55	-2.57
18	TextSentinels	-8.86	-11.51	-6.20

Table 3
Top Task 3 results ranked independently within BTC and TSLA. Closed positions count completed position episodes, including final forced liquidation.

Asset	Rank	Team	CR (%)	Closed
BTC	1	NLP-DE	9.97	18
BTC	2	The Lab Rats	4.26	17
BTC	3	CLaC	3.50	3
BTC	4	Amongus	2.38	2
BTC	5	pjmathematician	2.05	4
TSLA	1	Fin-Analyst	13.33	8
TSLA	2	Infraglyph	10.34	9
TSLA	3	Calibrated Signals	7.76	3
TSLA	4	Gentlemen	5.21	10
TSLA	5	TrustMe	3.44	4

5.2. Analysis of Results

The final ranking shows substantial asset specialization. *Infraglyph* led the cross-asset mean CR by combining a 10.34% TSLA return with a limited BTC loss, whereas *Fin-Analyst* and *NLP-DE* achieved the strongest single-asset returns on TSLA and BTC, respectively. Across the 18 systems, BTC and TSLA returns were weakly correlated (Pearson $r = 0.06$; Spearman $\rho = 0.09$), and only two systems achieved positive returns on both assets. The weak cross-asset correlation and limited number of jointly profitable systems justify reporting both within-asset ranks and the cross-asset mean. Seven systems achieved a positive cross-asset mean CR, one remained flat, and ten had a negative mean over the official window. Because the HOLD baseline is the only organizer baseline, positive CR does not establish outperformance of passive market exposure.

5.3. Participant Approaches

Table 4 summarizes components reported in 13 Task 3 Working Notes papers. Blank cells mean that a component was not reported. News or filing context appears throughout the documented systems, but teams differ in how they transform it into actions: some combine sentiment with technical or forecasting signals, while others emphasize memory, multi-agent decomposition, reinforcement learning, or rule-based gating.

Table 4
Signals and architectural components reported by the 13 documented Task 3 systems.

Team	Signals				Architecture			
	News/filings	Sentiment	Technical	Forecasting	Memory/retrieval	Multi-agent	Training	Rules/gating
TrustMe [17]	✓	✓			✓			✓
DS@GT ARC [18]	✓	✓	✓	✓		✓		✓
Infraglyph [19]	✓				✓			✓
TextSentinels [20]	✓	✓	✓	✓	✓			✓
Calibrated Signals [21]	✓	✓	✓			✓		✓
CLaC [22]	✓	✓	✓				✓	
Enjamul Islam Khan [23]	✓	✓	✓	✓		✓		
Amongus [24]	✓		✓	✓	✓	✓		✓
The Lab Rats [25]	✓	✓	✓					✓
Club GAIA - ESCOM [26]	✓	✓	✓	✓			✓	
pjmathematician [27]	✓	✓	✓		✓			✓
Fin-Analyst [28]	✓	✓	✓			✓		✓
TCLabs [29]	✓		✓					

***TrustMe* [17]**
(LLM trading, FinBERT sentiment, short-term memory, reflection, turbulence guard)
TrustMe reports a momentum-following LLM trading agent for Task 3. The system combines an LLM news analyzer, local FinBERT sentiment, a 10-day memory component, and reflection over recent decisions. Its action space is constrained by prompting and deterministic safeguards, including a turbulence guard that can override the usual BUY or HOLD decision and issue SELL.

***DS@GT ARC* [18]**
(tool use, agentic LLM trading, sentiment-aware news processing, random-forest forecasting)
DS@GT ARC studies the role of tools in agentic LLM-based trading systems. The reported *DS@GT StockTron* system combines specialist analyst signals, sentiment- and section-aware news processing, an LLM portfolio manager, structured decision storage, and a random-forest forecasting tool.

***Infraglyph* [19]**
(tiered memory, ChromaDB retrieval, temporal filtering, asset-specific personas)
Infraglyph describes *EdgeQuant Agent* as using a tiered cognitive memory architecture. The system stores news, filing context, and reflections in memory layers, retrieves them through embedding-based search, and applies temporal filtering to avoid look-ahead information. Asset-specific personas and conviction rules determine how the retrieved evidence is converted into an action.

***TextSentinels* [20]**
(FinBERT, XGBoost, recursive sentiment memory, technical indicators)
TextSentinels treats Task 3 as a multi-asset financial decision-making problem and uses machine-learning and transformer components. The system combines FinBERT sentiment features, technical indicators, recursive sentiment memory, and XGBoost-based decision logic before producing the required action label.

Calibrated Signals **[21]**
(multi-agent reasoning, quantitative indicators, LLM synthesis, trading intelligence)
Calibrated Signals combines live market enrichment, deterministic quantitative indicators, LLM-based news and filing analysis, and a multi-agent synthesis stage for BUY, HOLD, or SELL decisions.

CLaC **[22]**
(reinforcement learning, sentiment augmentation, alpha reward)
CLaC reports a sentiment-augmented deep reinforcement-learning approach for active trading. Its state representation combines financial sentiment with EMA, RSI, MACD, Bollinger Bands, and volume-derived indicators, while an alpha-reward objective guides the trading policy.

Enjamul Islam Khan **[23]**
(adaptive multi-agent framework, sentiment analysis, filing analysis, forecasting)
The *Enjamul Islam Khan* system combines specialized components for sentiment, technical analysis, event reasoning, SEC filing analysis, financial semantics, and multi-horizon forecasting before selecting a final action.

Amongus **[24]**
(Agentic Mk1, Qwen2.5, deterministic indicators, TimesFM forecasts, HOLD fallback)
Amongus reports *Agentic Mk1*, a structured multi-agent trading system built around Qwen2.5-7B-Instruct and TimesFM-style forecasts. The system combines forecasting, deterministic technical indicators, state handling, and agentic reasoning to support daily BTC and TSLA decisions. Its endpoint returns HOLD when it cannot produce an admissible directional decision.

The Lab Rats **[25]**
(FinSentinel v2, deterministic trading, online threshold calibration, cross-asset regime gating)
The Lab Rats describe *FinSentinel v2*, a deterministic trading endpoint with online threshold calibration and cross-asset regime gating. The endpoint combines news sentiment, price-derived technical signals, and momentum labels, with rolling threshold calibration and recent-accuracy-based signal weighting. The final live endpoint emphasizes repeatable inference without calling an LLM at decision time.

Club GAIA - ESCOM **[26]**
(GRPO fine-tuning, quantized LLMs, multimodal financial decision making)
Club GAIA - ESCOM explores compact GRPO fine-tuning of quantized LLMs for multimodal financial decision making. The system adapts efficient language models for Task 3's action-selection setting by combining news, price sequences, scalar features, sentiment, and forecast summaries into structured prompts.

pjmathematician **[27]**
(multi-signal ensemble, technical indicators, search grounding, endpoint decision logic)
The *pjmathematician* Task 3 system combines technical indicators, filing-based fundamentals, news sentiment, and search-grounded context before applying an LLM-based arbitration step. The arbitration step aggregates these signals before issuing the endpoint response.

Fin-Analyst **[28]**
(aggregator LLM, multi-specialist architecture, rule-based BTC voting)
Fin-Analyst uses an aggregator LLM in its live trading agent. For TSLA, the system uses a multi-specialist architecture that processes disclosures, fundamentals, analyst forecasts, social media sentiment, technical indicators, and daily news. For BTC, it uses a rule-based vote over price trends, the Crypto Fear and Greed Index, and momentum before returning the final action.

***TCLabs* [29]**
(news preprocessing, technical indicators, structured prompts, action selection)
The *TCLabs* Task 3 system combines the daily task inputs with Trading Central-style news preprocessing, technical indicators, and structured LLM prompts before converting the evidence into a trading action.

Across the 13 documented systems, no single architecture dominates: systems combine financial text interpretation with different mixtures of retrieval, technical indicators, memory, reinforcement learning, fine-tuning, forecasting, and deterministic control. Taken together, the diversity of high-ranking designs and the weak correlation between BTC and TSLA returns do not identify a universally superior model family; they instead motivate asset-specific ablations in future evaluations.

6. Conclusion and Future Work

Task 3 introduced a live, endpoint-based evaluation in which deployed financial agents produced daily BUY, HOLD, or SELL decisions for BTC and TSLA. It attracted 18 accepted endpoints, and 13 working-notes papers describe a diverse set of LLM, retrieval, forecasting, reinforcement-learning, rule-based, and multi-agent trading methods. The final ranking shows substantial asset specialization: *Infraglyph* led the cross-asset mean CR, *Fin-Analyst* achieved the strongest TSLA asset-level result, and *NLP-DE* achieved the strongest BTC asset-level result.

The first edition highlighted endpoint reachability, schema compliance, and latency as deployment constraints that do not arise in a standard file-upload benchmark; it also showed the importance of organizer-side input validation for a live evaluation. Future live tasks should define default-action rules, input-quality checks, and the final validation cutoff before evaluation begins. Finally, CR measured on BTC and TSLA over a single evaluation window should not be interpreted as evidence of real-world profitability; it is a controlled shared-task metric for comparing systems under the same inputs, dates, and scoring policy.

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, maintaining endpoints, 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, 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.

[5] L. Qian, X. Peng, H. Smith, Y. Han, Y. He, H. Li, Y. Cao, Y. Yu, G. Xiong, P. Lu, Y. Wang, V. J. Zhang, H. He, A. Lopez-Lira, J. Huang, J. Nie, S. Ananiadou, When agents trade: Live multi-market trading arena for LLM agents, in: H. Hacid, Y. Maarek, F. Bonchi, I. Guy, E. Yilmaz (Eds.), Proceedings of the ACM Web Conference 2026, Association for Computing Machinery, Dubai, United Arab Emirates, 2026, pp. 7833–7844. URL: https://doi.org/10.1145/3774904.3792821. doi:10.1145/3774904.3792821.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Q. Xie, J. Huang, D. Li, Z. Chen, R. Xiang, M. Xiao, Y. Yu, V. Somasundaram, K. Yang, C. Yuan, Z. Luo, Z. Liu, Y. He, Y. Jiang, H. Li, D. Feng, X.-Y. Liu, B. Wang, H. Wang, Y. Lai, J. Suchow, A. Lopez-Lira, M. Peng, S. Ananiadou, FinNLP-AgentScen-2024 shared task: Financial challenges in large language models - FinLLMs, in: C.-C. Chen, T. Ishigaki, H. Takamura, A. Murai, S. Nishino, H.-H. Huang, H.-H. Chen (Eds.), Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning, Jeju, South Korea, 2024, pp. 119–126. URL: https://aclanthology.org/2024.finnlp-2.11/.

[16] 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.

[17] S. Kundu, A. Kaneshiro, M. Imai, Working notes of TrustMe team in CLEF 2026 FinMMEval track 3, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[18] M. Heil, A. Pramov, B. Gong, DS@GT ARC at FinMMEval 2026: On the role of tools in agentic LLM-based trading systems, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[19] U. Kava, H. Patel, T. Mandl, S. Modha, Infraglyph @ FinMMEval 2026 Task 3: EdgeQuant Agent – a tiered cognitive memory architecture, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[20] D. Thenmozhi, A. A, A. Gopinath, A. Balasubramanian, A. Sivakumar, TextSentinels @ FinMMEval 2026 Task 3: Multi-asset financial decision making using machine learning and financial transformers, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[21] G. Jain, H. Hitesh, S. Bhardwaj, H. Behl, S. Sundriyal, P. Gautam, S. Zimmerman, Calibrated Signals @ FinMMEval 2026 Task 3: Where quant meets reasoning: A multi-agent framework for trading intelligence, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[22] A. Neagu, E. Khan, L. Kosseim, CLaC @ FinMMEval 2026 Task 3: Sentiment-augmented deep reinforcement learning for active trading – an alpha-reward approach, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[23] E. I. Khan, A. Kuila, Adaptive multi-agent financial intelligence framework for financial decision making under dynamic market conditions, in: CLEF 2026 Working Notes, CEUR Workshop

Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[24] P. Panyasombat, T. Denkavin, N. Kertkeidkachorn, Amongus @ FinMMEval 2026 Task 3: Agentic Mk1 with structured multi-agent trading and TimesFM forecasts, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[25] A. Gill, E. N. Sheikh, F. T. Zahra, A. Samad, F. Alvi, S. Kumar, The Lab Rats @ FinMMEval 2026 Task 3: Online threshold calibration and cross-asset regime gating for live financial decision making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[26] J. B. Cruz, J. B. P. Cuatianquiz, R. E. M. Roldán, D. R. Ramos, Club GAIA - ESCOM @ FinMMEval 2026 Task 3: Multimodal financial decision making via compact GRPO fine-tuning of quantized LLMs, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[27] 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.
[28] M. Rashid, L. Hong, J. Ding, K. S. M. T. Hossain, Fin-Analyst at FinMMEval 2026 Task 3: A large language model (LLM) based live trading agent, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[29] 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.