TCLabs @ FinMMEval 2026: Systems for Tasks 1, 2 and 3

FinMMEval at CLEF 2026

Elvys Linhares Pontes*, Mohamed Benjannet

Trading Central Labs, Trading Central, Paris, France

Abstract

This paper presents the TCLabs participation to the FinMMEval Lab 2026, which focus on financial reasoning, multilingual financial question answering, and news-driven financial decision making. We participate in three tasks: (1) Financial Exam Question Answering, (2) Multilingual Financial Question Answering, and (3) Financial Decision Making. To address the diverse challenges posed by these tasks, we propose specialized multi-agent large language model (LLM) architectures designed for financial reasoning, multilingual evidence aggregation, retrieval-augmented analysis, and market forecasting. Our systems integrate domain-specialized expert agents, retrieval-augmented generation (RAG), adversarial debate mechanisms, and structured prompting strategies in order to improve reasoning robustness and factual consistency. Experimental results demonstrate that collaborative multi-agent reasoning can effectively support complex financial analysis tasks across multilingual and decision-oriented settings.

Keywords

Multi-agent analysis, Financial reasoning, Multilingual financial analysis, Financial decision making

1. Introduction

Financial reasoning represents one of the most challenging application domains for large language models (LLMs) because it requires the simultaneous integration of numerical analysis, domain-specific knowledge, regulatory interpretation, and evidence-based decision making. Unlike general-purpose reasoning tasks, financial applications often involve complex quantitative relationships, evolving regulations, multilingual information sources, and high-stakes decision environments where factual accuracy and interpretability are essential.

Recent advances in LLMs have significantly improved performance across a wide range of tasks, including text summarization [1] and question answering [2]. However, financial reasoning remains particularly difficult due to the need for precise numerical computation, contextual interpretation of financial documents, and the ability to reconcile heterogeneous sources of information. In practical financial environments, systems must often combine structured numerical evidence with unstructured textual information such as news articles, earnings reports, regulatory filings, and market commentary.

The FinMMEval Lab 2026 provides a comprehensive evaluation framework for studying these challenges [3]. The workshop covers three complementary dimensions of financial intelligence: financial examination question answering [4], multilingual financial question answering grounded in financial reports and news articles [5], and financial decision making for trading scenarios [6]. Together, these tasks evaluate both analytical reasoning and real-world financial decision support capabilities.

In this work, we propose specialized multi-agent systems tailored to the requirements of each task. Our approaches combine collaborative expert reasoning, retrieval-augmented generation (RAG), multilingual evidence integration, adversarial debate mechanisms, and structured prompting strategies. Rather than relying on a single monolithic model, our systems decompose complex financial reasoning into specialized sub-tasks handled by domain-focused agents. This design enables improved reasoning transparency, better factual grounding, and more robust decision making.

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

*Corresponding author.

elvys.linharespontes@tradingcentral.com (E. L. Pontes); mohamed.benjannet@tradingcentral.com (M. Benjannet)

0000-0002-9571-5193 (E. L. Pontes)

The remainder of this paper is organized as follows. Section 2 describes the three FinMMEval tasks. Section 3 reviews related work on financial LLMs, RAG, and multi-agent reasoning systems. Section 4 presents our proposed architectures and their official results in the competition. Finally, Section 5 concludes the paper with key insights and final remarks.

2. Task Description

2.1. Task 1: Financial Exam Q&A

Task 1 focuses on financial examination question answering [4]. Given a stand-alone multiple-choice question Q associated with four candidate answers $\{A_1, A_2, A_3, A_4\}$, the system must predict the correct answer A^*. Questions cover multiple financial domains, including accounting, valuation, ethics, corporate finance, economics, portfolio management, and financial regulation, and are available in Arabic, Chinese, English and Hindi [7, 8].

This task evaluates the ability of language models to perform professional-level financial reasoning under examination conditions. Many questions require multi-step analytical reasoning, numerical computations, interpretation of accounting standards, or identification of subtle regulatory distinctions. Consequently, strong performance requires not only factual knowledge but also robust logical reasoning and mathematical consistency.

2.2. Task 2: Multilingual Financial Q&A

Task 2 evaluates multilingual financial reasoning capabilities grounded in heterogeneous financial documents [5]. The input consists of financial reports R extracted from English 10-K and 10-Q filings, multilingual financial news articles, and a financial question Q [9]. The objective is to generate a concise answer A supported by relevant multilingual textual evidence.

Compared with traditional question answering tasks, this task introduces several additional challenges. First, systems must process multilingual information originating from diverse news sources and financial reports. Second, the task requires combining qualitative narratives from news articles with quantitative evidence extracted from structured financial filings. Finally, generated responses must remain concise while preserving factual accuracy and evidential support.

2.3. Task 3: Financial Decision Making

Task 3 is a daily news-driven financial forecasting and trading task. Systems receive market information and must generate a trading recommendation among three possible actions: BUY, HOLD, or SELL [6].

This task evaluates the ability of LLMs to combine financial news understanding, technical analysis, forecasting reasoning, and decision-making capabilities under operational constraints. Since predictions must be generated within strict inference time limits, systems must balance reasoning quality with computational efficiency. The benchmark therefore reflects realistic financial deployment scenarios where rapid yet informed decisions are required.

3. Related Work

Research on LLMs for financial reasoning has expanded rapidly in recent years, particularly following the emergence of transformer-based architectures and instruction-tuned conversational models [10]. Financial natural language processing (FinNLP) applications increasingly rely on LLMs to address tasks such as sentiment analysis [11, 12], news summarization [13], and trading [14].

Early financial question-answering systems were primarily based on supervised classification models trained on domain-specific datasets. Although these approaches achieved strong performance on narrowly defined benchmarks, they often lacked interpretability and struggled with complex professional

reasoning tasks involving multi-step analysis or numerical computations. Professional certification examinations, in particular, require the integration of accounting standards, ethics, regulation, quantitative finance, and practical decision-making knowledge.

Recent transformer-based language models have demonstrated strong capabilities across several domains. Models such as GPT-4, Claude, and other instruction-tuned architectures have shown promising performance on legal, medical, and financial examinations.

RAG architectures introduced by Lewis et al. [15] improved factual grounding by combining neural language models with external document retrieval systems. In financial contexts, RAG approaches are particularly useful because accounting standards, ethical rules, and financial regulations evolve continuously. By grounding responses in authoritative knowledge bases, retrieval mechanisms help reduce hallucinations and improve factual consistency.

Recent advances in prompting techniques have further enhanced reasoning performance in large language models. Chain-of-thought prompting [16] demonstrated that explicitly decomposing reasoning steps substantially improves performance on mathematical and logical tasks. Similarly, zero-shot reasoning approaches [17] showed that reasoning-oriented prompts can induce structured analytical behavior even without task-specific demonstrations. ReAct [18] further combined reasoning and external tool usage within unified prompting frameworks.

In parallel, multi-agent reasoning systems have emerged as a promising direction for improving analytical robustness. Debate-based and collaborative reasoning frameworks enable multiple specialized agents to evaluate candidate solutions from complementary perspectives. Prior work on AI debate and multi-agent deliberation suggests that adversarial collaboration between agents can reduce reasoning errors and improve decision quality [19]. Such mechanisms are particularly relevant for financial applications, where decisions often require balancing competing interpretations of market information.

Several benchmarks have specifically targeted financial reasoning. FinQA [20], ConvFinQA [21], and related datasets highlighted the importance of numerical reasoning and table understanding in financial NLP systems. More recently, open-source financial language models such as FinGPT [22] enabled reproducible research and domain-specific adaptation for financial applications.

Tool-augmented reasoning frameworks have also become increasingly relevant in finance. Systems such as Gorilla [23] demonstrated how language models can interact with external APIs and computational tools, which is highly valuable for financial forecasting systems that rely on real-time market information and technical indicators.

Our work proposes the combination of domain-specialized expert agents, RAG, adversarial debate, iterative consensus, and mathematical verification within a unified framework tailored specifically for professional financial examinations, question-answering and financial decision making.

4. System Description

4.1. Task 1: Multi-Agent Financial Exam System

Professional financial certification examinations represent a challenge task for LLMs. Successfully solving these questions requires models to integrate conceptual understanding, numerical reasoning, regulatory interpretation, and practical financial expertise while avoiding logical inconsistencies and unsupported conclusions.

To address these challenges, we propose a collaborative multi-agent architecture composed of specialized financial expert agents. Instead of relying on a single reasoning process, the system decomposes financial analysis into multiple complementary expert perspectives in order to improve the quality of analysis and reduce hallucinations.

The reasoning pipeline begins with a *Financial Examiner Agent*. This agent performs an initial analysis of the input question by identifying the relevant financial concepts, formulas, regulatory dimensions, and reasoning strategies required to solve the problem. In addition, this agent reformulates the original question into a structured representation that facilitates downstream expert analysis.

To improve factual grounding and reduce hallucinations, the system subsequently activates a RAG Knowledge Agent. This agent retrieves supporting financial information from external resources, including financial references and educational materials such as Investopedia[1]. Retrieved evidence is incorporated into the reasoning context provided to the expert agents.

The enriched problem representation is then forwarded to a panel of specialized financial agents:

- *ValuationExpert* focuses on valuation methodologies, investment analysis, and asset pricing.
- *AccountingExpert* specializes in accounting standards, financial statements, and reporting interpretation.
- *EthicsRegulationExpert* analyzes ethical constraints, compliance issues, and regulatory frameworks.
- *CorporateFinanceExpert* evaluates capital structure, financing decisions, and corporate financial strategy.
- *MathematicalFinancialExpert* validates numerical computations, formulas, and quantitative consistency.

Each expert agent independently produces a structured analysis together with supporting justifications and confidence estimates. These intermediate outputs are subsequently evaluated by the *DebateCriticAgent*, which performs adversarial verification of the proposed solutions. The critic agent identifies contradictions, verifies formulas, detects unsupported assumptions, and requests refinements whenever inconsistencies are observed.

Finally, the *ConsensusJudgeAgent* aggregates the refined analyses using a weighted voting strategy based on evidence quality, confidence scores, and domain relevance. This final stage produces the system prediction while ensuring coherence across expert opinions. The prompts for these agents are provided in Annex A.

4.2. Official results

We evaluated three LLMs as the core reasoning engines driving the agents within our multi-agent architecture: GPT-4o-mini, Mistral-Small-3.2-24B-Instruct-2506, and Mistral-7B-Instruct-v0.3 [24]. Model performance was evaluated using answer accuracy on a subset of the "bharatgenai/BhashaBench-Finance" dataset [7]. In addition to the proposed multi-agent framework, we developed a baseline system by fine-tuning Mistral-7B-Instruct-v0.3 to directly generate answers without collaborative multi-agent reasoning or agent interaction mechanisms.

Among all evaluated configurations, the multi-agent architecture with GPT-4o-mini achieved the best accuracy and was therefore selected for the official submission.

Table 1 presents the official results released by the organizers for the Arabic, Chinese, English, and Hindi tracks. Our team, *TCLabs*, achieved strong performance on the Arabic and Hindi tracks, obtaining accuracies close to 90% and securing 3rd place in Hindi and 5th place in Arabic in the official leader rankings. Performance on the Chinese and English tracks was comparatively lower, ranging from 75.5% to 79%.

4.3. Task 2: Multi-Agent Multilingual Financial QA System

For Task 2, we propose a multilingual multi-agent architecture designed to integrate both qualitative and quantitative financial evidence originating from heterogeneous multilingual sources. The system is composed of four specialized agents that collaboratively analyze financial reports and multilingual news articles to generate relevant answers supported by news evidence.

The reasoning process begins with the *News Expert Agent*, which processes multilingual financial news articles related to the target company and associated question. Its objective is to identify the

[1] https://www.investopedia.com/

Table 1
Official results for Task 1 across the Arabic, Chinese, English, and Hindi languages. The reported values correspond to the accuracy (%) between the predicted and gold-standard answers. The best results are highlighted in bold and our results are in italic.

Team	Arabic	Chinese	English	Hindi
AI_TLfanclub	84	70.5	80.5	68.5
DS@GT - FinMMEval Task 1	76	64.5	69.5	73.5
FinEval	83.5	46.5	66.5	62.5
fosu ltw	94	92	93.5	**92**
NLP-DE	91.5	87	87.5	85.5
Paticipate	–	–	64.5	–
PooRi	**97.5**	**96.5**	**97.5**	**92**
pranshu rastogi	95	92	91	91.5
TCLabs	*89.5*	*75.5*	*79*	*90*
TextSentinels	86	76.5	81.5	–
UyoAngle	86.5	77	83	87
Vitt Manthan	–	26.5	29.5	31

most relevant qualitative evidence, extract key statements, summarize market narratives, and capture contextual information that may influence the interpretation of the input question.

In parallel, the *Report Expert Agent* focuses on structured financial information extracted from corporate filings such as 10-K and 10-Q reports. This agent identifies relevant numerical evidence from financial statements, including income statements, balance sheets, and cash flow statements related to the input question.

The analyses generated by the News Expert Agent and Report Expert Agent are subsequently processed by the *Financial Expert Agent*. This agent combines both analyses in order to evaluate consistency between market perception and financial reporting information to better answer the input question.

Finally, the consolidated analysis is passed to the *Answer Generator Agent*, which produces the final response while respecting formatting constraints and maximum word limits imposed by the task. The prompts for these agents are provided in Annex B.

4.3.1. Official results

Similar to Task 1, we evaluated GPT-4o-mini, Mistral-Small-3.2-24B-Instruct-2506, and Mistral-7B-Instruct-v0.3 for the core reasoning of our agents. Model performance was assessed using ROUGE-1 F1-score metric. Experiments were conducted on a subset of the "TheFinAI/PolyFiQA-Expert" dataset. The LLM GPT-4o-mini also achieved the best F1-score and was therefore selected for the official submission.

Table 2 presents the official results released by the organizers. Our model, *TCLabs*, achieved a F1-score of 0.250643 with a precision of 0.2321 and a recall of 0.32574. This performance placed the model in 9th position on the official leaderboard.

The imbalance between precision and recall contributed to the lower overall F1-score, suggesting that future improvements should focus on enhancing prediction accuracy while maintaining coverage of relevant information.

4.4. Task 3: Financial Decision-Making System

For Task 3, we propose a decision-making framework that integrates financial news, technical indicators, and LLM reasoning. The goal is to enable robust decision-making under real-world market conditions by combining heterogeneous sources of financial information.

Our approach brings together multiple complementary information streams, including organizer-provided financial news, Trading Central market news summaries, technical analysis indicators, and

Table 2
Official results for Task 2 on the English dataset. The reported values correspond to ROUGE-1 precision, recall, and F1-score. The best results are highlighted in bold and our results are in italic.

Team	F1-Score	Precision	Recall
AI_TLfanclub	0.303923	**0.364714**	0.30701
Calibrated Signals	**0.31182**	0.302498	0.376361
DS@GT FinMMEval	0.285309	0.257237	0.388839
Ethereum Team B CLEF	0.254618	0.308559	0.264243
HU_LLM_Fin	0.228886	0.258063	0.280321
IGT	0.307075	0.282108	**0.40444**
NLP-DE	0.204852	0.264342	0.190123
PooRi	0.297152	0.278093	0.379763
pranshu rastogi	0.309609	0.320841	0.35366
TCLabs	*0.250643*	*0.2321*	*0.32574*
TextSentinels	0.18734	0.30288	0.175564
The Lab Rats	0.273353	0.273697	0.34169

structured prompting strategies. By combining these elements, the system is able to capture both macroeconomic context and asset-specific signals.

The pipeline is organized into four sequential stages, each responsible for progressively refining and structuring the input data. The first stage, *Organizer Data Preprocessing*, focuses on processing the news articles provided by the challenge organizers. In this step, we isolate macroeconomic and geopolitical information in order to extract high-level market context that may influence asset behavior.

Building upon this, the second stage, *News Integration*, merges the organizer-provided news with Trading Central market summaries. This integration allows the system to combine broad contextual narratives with asset-specific financial analysis, resulting in a more complete and balanced representation of market conditions.

The third stage, *Technical Indicator Collection*, extracts and standardizes quantitative market signals. To further enrich the available information, we integrate a set of technical analysis indicators, including the daily and monthly price returns, the 50-day and 200-day simple moving averages, support and resistance levels, and the currently active chart pattern, when applicable. This enriched feature set provides a robust quantitative foundation that complements the news-based data.

Finally, in the *Decision-Making* stage, all processed information is consolidated into a structured prompt (see Annex C). This prompt includes clear task instructions, the integrated financial dataset, and explicit output format specifications, ensuring that the model receives a coherent and well-organized input, which in turn promotes consistent and reliable decision-making outputs. It is worth noting that, unlike the multi-agent pipeline used for Task 2 (see Annex B), we did not adopt a multi-agent architecture for this task, as the response time is constrained to 180 seconds; with a single decision step already requiring approximately 140 seconds, no margin remained for additional processing stages.

4.4.1. Training Configurations

To evaluate the impact of external information and strategies, we design three experimental configurations that progressively increase the level of information and model specialization.

In the *Baseline Configuration*, models are provided only with organizer-supplied data, without any external enrichment. This setting serves as a reference point for evaluating the intrinsic capability of each model under minimal information conditions.

In the *Augmented Configuration*, we extend the input by incorporating Trading Central news summaries and technical indicators in addition to the organizer-provided data. This configuration is designed to assess the benefit of enriched market context on model performance.

Finally, in the *Fine-Tuned Configuration*, models are further trained on the provided training dataset. This setting allows us to evaluate whether domain-specific adaptation can improve performance beyond

prompt-based strategies alone.

4.4.2. Evaluated Models

We evaluate three open-source LLMs that represent different architectural and reasoning paradigms: GPT-OSS-20B [25], Qwen2.5-72B-Instruct [26] and DeepSeek-R1-Distill-Llama-70B [27]. These models are selected to ensure diversity in reasoning capabilities while remaining compatible with the computational constraints of the FinMMEval challenge.

4.4.3. Experimental Evaluation

All experiments are conducted using the official FinMMEval Lab datasets and evaluation protocols. For Task 3, all systems are constrained by a strict inference time limit of three minutes per prediction, reflecting realistic deployment constraints in financial decision-making scenarios.

Table 3 summarizes the performance of all evaluated models across different configurations using precision, recall, F1-score, and accuracy. The best results were achieved by the DeepSeek-R1 model using external data of Trading Central.

The incorporation of news analysis and technical indicators from Trading Central enabled the models to better interpret market conditions and make more informed analysis, leading to consistent improvements across all metrics. In contrast, fine-tuned versions of LLMs exhibited a reduction in performance, likely due to constrained reasoning capacity introduced during fine-tuning process. Based on these results, we selected the DeepSeek-R1 model with Trading Central data as our official submission.

Table 3
Performance comparison across models and configurations (TC: Trading Central data, FT: fine-tuning). The best results are highlighted in bold.

Model	Precision	Recall	F1-score	Accuracy
GPT-OSS-20B	48.0	48.0	48.0	48.7
Qwen2.5-72B-Instruct	49.0	48.0	48.5	48.7
DeepSeek-R1-Distill-Llama-70B	52.7	52.7	52.7	53.8
GPT-OSS-20B + TC data	51.0	50.5	50.75	51.7
Qwen2.5-72B + TC data	51.5	51.0	51.2	52.0
DeepSeek-R1 + TC data	**53.5**	**53.5**	**53.5**	**54.3**
GPT-OSS-20B + TC + FT	49.5	49.5	49.5	50.0
Qwen2.5-72B + TC + FT	50.5	51.5	51.0	51.3
DeepSeek-R1 + TC + FT	53.0	53.0	53.0	54.0

4.4.4. Official results

The competition officially commenced on May 11, 2026. As it was still ongoing at the time of writing, the results reported in Tables 4 and 5 represent a snapshot of the leaderboard recorded on July 3, 2026. Consequently, the rankings and performance metrics presented in this paper may differ from the final results released at the conclusion of the competition.

At the time of evaluation, the proposed TCAgent achieved cumulative returns of -1.27% for the TSLA trading task and -9.56% for the BTC trading task. Although these returns did not place the proposed approach among the top-performing submissions on the public leaderboard, they demonstrate a clear improvement over the corresponding Buy&Hold baseline strategies. Specifically, TCAgent outperformed the Buy&Hold benchmark by 13.55% on TSLA and 15.57% on BTC.

It is worth noting that an unexpected issue was encountered with the data source responsible for providing technical indicators for the BTC trading task. Because the problem arose after the official evaluation had begun, it could not be resolved without compromising model consistency throughout the

competition. As a result, technical indicators were available only for the TSLA task during the official evaluation. The absence of technical indicators for BTC not only reduced the amount of information available to the trading agent but also introduced a discrepancy between the data structure used during development and that used during evaluation. This limitation may have contributed to the lower performance observed for the BTC task.

Table 4
Official results for Task 3: Tesla Inc (TSLA) leaderboard. The best-performing agent is shown in bold, while our proposed TCAgent is shown in italics.

Agent	Return	Vs Buy&Hold	Sharpe / Win%
Fin_Analyst	**+13.51%**	**+28.33%**	**4.10 / 88.00%**
EdgeQuantAgent	+10.51%	+25.34%	3.34 / 67.00%
CalibratedSignals	+7.76%	+22.58%	9.64 / 67.00%
FINSAGE	+5.38%	+20.20%	1.82 / 80.00%
TrustMe	+3.44%	+18.26%	2.76 / 75.00%
PooRi Agent	+0.84%	+15.66%	8.13 / 100.00%
Enjamul Islam Khan	0.00%	+14.82%	0.00 / 0.00%
AAA	-0.78%	+14.05%	-0.21 / 50.00%
ai-tlfanclub-clef-task3	-0.88%	+13.95%	-0.73 / 33.00%
TCAgent	*-1.27%*	*+13.55%*	*-0.53 / 40.00%*
Buy&Hold	-14.82%	0.00%	–

Table 5
Official results for Task 3: Bitcoin (BTC) leaderboard. The best-performing agent is shown in bold, while our proposed TCAgent is shown in italics.

Agent	Return	Vs Buy&Hold	Sharpe / Win%
NLP-DE-ONE	**+7.00%**	**+32.14%**	**1.83 / 61.00%**
OhAzdaha	+2.38%	+27.51%	1.38 / 100.00%
FinSentinel-QwenTrader	+1.28%	+26.41%	0.46 / 47.00%
CLaC	+0.54%	+25.67%	0.28 / 33.00%
Enjamul Islam Khan	0.00%	+25.13%	0.00 / 0.00%
CalibratedSignals	-0.53%	+24.60%	-0.17 / 43.00%
PooRi Agent	-0.86%	+24.27%	-0.39 / 25.00%
ai-tlfanclub-clef-task3	-1.60%	+23.53%	-1.05 / 40.00%
TCAgent	*-9.56%*	*+15.57%*	*-4.38 / 20.00%*
Buy&Hold	-25.13%	0.00%	–

Overall, these results indicate that the proposed agent was able to substantially mitigate losses during the evaluation period, even under challenging market conditions. While the absolute returns remained negative, the consistent improvement relative to the passive Buy&Hold baseline suggests that the system effectively integrated information extracted from financial news, corporate reports, and, where available, technical indicators to support trading decisions.

5. Conclusion

This paper presented the TCLabs systems developed for the FinMMEval Lab 2026. Our approaches combine multi-agent reasoning, retrieval-augmented generation, multilingual evidence aggregation, structured prompting strategies, and market-oriented financial analysis.

The official results demonstrate the effectiveness of collaborative expert reasoning for professional financial analysis and decision-making. In particular, our multi-agent architecture achieved strong performance on the professional financial examination benchmark, reaching accuracy scores of 89.5% for Arabic and 90% for Hindi.

Declaration on Generative AI

During the preparation of this work, the authors used **ChatGPT** and **Grammarly** in order to: **Grammar and spelling check**, **Paraphrase** and **reword**. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. L. Pontes, C.-E. González-Gallardo, M. Benjannet, C. Qu, A. Doucet, L3iTC at the FinLLM challenge task: Quantization for financial text classification & summarization, 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. 141–145. URL: https://aclanthology.org/2024.finnlp-2.14/.

[2] X. Wang, J. Chi, Z. Tai, T. S. T. Kwok, H. He, Z. Li, Y. Hua, M. Li, P. Lu, S. Wang, Y. Wu, H. Jerry, J. Tian, F. Mo, Y. Cui, L. Zhou, Finsage: A multi-aspect rag system for financial filings question answering, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 6144–6152. URL: https://doi.org/10.1145/3746252.3761587. doi:10.1145/3746252.3761587.

[3] Z. Xie, Y. Dai, R. Elbadry, V. Jani, X. Peng, L. Qian, G. Georgiev, D. Dimitrov, F. Zhang, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of FinMMEval 2026: Multilingual and multimodal financial evaluation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[4] Z. Xie, Y. Dai, R. Elbadry, V. Jani, G. Georgiev, D. Dimitrov, F. Zhang, X. Peng, L. Qian, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 1: Multilingual financial multiple-choice question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[5] Z. Xie, X. Peng, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, L. Qian, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 2: Financial question answering and summarization, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[6] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, X. Peng, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 3: Financial decision making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

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

[8] 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, 2026. URL: https://arxiv.org/abs/2604.19098. doi:10.48550/arXiv.2604.19098. arXiv:2604.19098.

[9] X. Peng, L. Qian, Y. Wang, R. Xiang, Y. He, Y. Ren, M. Jiang, V. J. Zhang, Y. Guo, J. Zhao, H. He, Y. Han, Y. Feng, Y. Jiang, Y. Cao, H. Li, Y. Yu, X. Wang, P. Gao, S. Lin, K. Wang, S. Yang, Y. Zhao, Z. Liu, P. Lu, J. Huang, S. Wang, T. Papadopoulos, P. Giannouris, E. Soufleri, N. Chen, Z. Deng, H. Fu, Y. Zhao, M. Lin, M. Qiu, K. E. Smith, A. Cohan, X.-Y. Liu, J. Huang, G. Xiong, A. Lopez-Lira, X. Chen, J. Tsujii, J.-Y. Nie, S. Ananiadou, Q. Xie, MultiFinBen: Benchmarking large language models for

multilingual and multimodal financial application, 2025. URL: https://arxiv.org/abs/2506.14028. doi:10.48550/arXiv.2506.14028. arXiv:2506.14028.

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

[11] E. L. Pontes, M. Benjannet, Contextual sentence analysis for the sentiment prediction on financial data, in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4570–4577. doi:10.1109/BigData52589.2021.9672027.

[12] E. Linhares Pontes, C.-E. González-Gallardo, G. Bordea, J. G Moreno, M. Ben Jannet, Y. Zhao, A. Doucet, Backtesting sentiment signals for trading: Evaluating the viability of alpha generation from sentiment analysis, in: F. Bechet, A.-G. Chifu, K. Pinel-sauvagnat, B. Favre, E. Maes, D. Nurbakova (Eds.), Actes de la session industrielle de CORIA-TALN 2025, ATALA \\& ARIA, Marseille, France, 2025, pp. 17–32. URL: https://aclanthology.org/2025.jeptalnrecital-industrielle.2/.

[13] N. Reche, E. L. Pontes, J.-M. Torres-Moreno, Financial news summarization: Can extractive methods still offer a true alternative to llms?, in: L. Martínez-Villaseñor, R. A. Vázquez, G. Ochoa-Ruiz (Eds.), Advances in Soft Computing, Springer Nature Switzerland, Cham, 2026, pp. 169–180.

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

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[18] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: International Conference on Learning Representations (ICLR), 2023.

[19] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

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

[21] 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, Proceedings of EMNLP 2022 (2022).

[22] Y. Liang, Y. Liu, N. Wang, H. Yang, B. Zhang, C. D. Wang, Fingpt: enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched llms, AAAI 2025 Workshop GoodData (2025).

[23] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, Gorilla: Large language model connected with massive apis, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 126544–126565. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/e4c61f578ff07830f5c37378dd3ecb0d-Paper-Conference.pdf. doi:10.52202/079017-4020.

[24] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[25] OpenAI, gpt-oss-120b gpt-oss-20b model card, 2025. URL: https://arxiv.org/abs/2508.10925. arXiv:2508.10925.

[26] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[27] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

A. Prompts for the Task 1

This annex provides the complete system prompts used by the multi-agent architecture in Task 1. The prompts define the role, responsibilities, and expected reasoning process of each agent, ensuring consistent behavior throughout the evaluation.

Table 6 summarizes the prompts of the supporting agents responsible for question analysis, knowledge retrieval, quality assurance, and final decision making. These include the Financial Examiner Agent, which analyzes the question and identifies potential traps; the RAG Knowledge Agent, which retrieves relevant supporting information; the Expert Agent template; the Debate Critic Agent, which evaluates the consistency and factual correctness of the experts' reasoning; and the Consensus Judge Agent, which synthesizes all available evidence to produce the final answer.

Table 7 presents the domain and expertise to the domain-specific expert panel, including the Valuation, Accounting, Ethics & Regulation, Corporate Finance, and Mathematical Finance experts.

B. Prompts for the Task 2

This annex presents the system prompts used in the Multilingual Financial Q&A task. The workflow is organized into four sequential agents. First, the News Expert Agent extracts relevant insights from multilingual news sources. Second, the Report Expert Agent retrieves and structures quantitative evidence from financial statements. Third, the Financial Expert Agent synthesizes both qualitative and quantitative inputs to produce an integrated analysis. Finally, the Answer Generator Agent produces a concise, grounded final response supported by evidence from both news and financial reports.

Table 8 summarizes the prompts and responsibilities of each agent in this pipeline.

C. Prompt for the Task 3

This annex presents the prompt used for the Financial Decision Making task. Unlike the multilingual Q&A pipeline described in Annex B, this task relies on a single prompt that combines news, sentiment, momentum, and technical indicators into one query, asking the model to predict whether the asset price will close higher or lower on the following day. Table 9 summarizes the role and prompt used in this task.

Table 6
System prompts and behavioral rules assigned to the supporting agents for Task 1.

Agent	Role	Prompt
Financial Examiner Agent	Analyzes the exam question before expert reasoning begins.	Analyze the following financial exam question: {QUESTION}. Identify: (1) the primary domain and subtopic; (2) any traps, wording tricks, or distractors (e.g., most likely, except, double negatives); (3) the concepts, formulas, and standards required to answer correctly; (4) a clear reformulation of what the question is actually asking; and (5) which expert agents should focus on the question.
RAG Knowledge Agent	Retrieves external knowledge to support expert reasoning.	Retrieve relevant reference material and similar financial questions from trusted knowledge sources and the web. Return the most relevant passages to support subsequent expert analysis.
Expert Agent Template	Base prompt shared by all domain experts.	You are a world-class {DOMAIN} expert and senior financial exam coach. Your expertise covers {EXPERTISE}. You are one of several independent expert agents in a multi-agent reasoning system. Your objective is to perform a rigorous analysis of a multiple-choice financial exam question. Methodology: (1) reason step by step; (2) evaluate every answer choice as a candidate solution; (3) eliminate incorrect options with precise justification; (4) verify consistency with established concepts, formulas, and standards; (5) report a confidence score between 0 and 100. Cite relevant formulas, accounting standards (IFRS, US GAAP), CFA Standards, or other authoritative frameworks whenever applicable. If the question falls outside your primary expertise, provide your best analysis while appropriately lowering your confidence.
Debate Critic Agent	Reviews expert outputs for quality, consistency, and hallucinations.	Critically evaluate the experts' reasoning. Identify: (1) logical fallacies or unsupported conclusions; (2) factual errors or hallucinations (e.g., incorrect formulas or standards); (3) disagreements between experts and whether they can be resolved; (4) overconfident claims lacking evidence; and (5) failures to account for exam traps identified by the Financial Examiner Agent. Set revision_needed = true if experts substantially disagree, factual hallucinations are detected, retrieved evidence contradicts the proposed answer, or confidence scores appear unjustifiably high.
Consensus Judge Agent	Produces the final authoritative answer.	Combine the analyses from all experts, the Debate Critic, and the retrieved knowledge to determine the final answer. Weighting methodology: (1) consider the raw vote count; (2) weight votes by confidence while penalizing unjustified overconfidence; (3) strengthen answers supported by retrieved evidence; (4) discount reasoning flagged as hallucinated or weak; and (5) resolve ties by prioritizing the most relevant domain expert (e.g., Accounting Expert for IFRS questions). Return the final answer together with a concise justification and overall confidence score.

Table 7
Roles and system prompts of the expert-panel agents employed in Task 1.

Agent	Domain	Expertise
Valuation Expert	Valuation	DCF, WACC, CAPM, enterprise value multiples (EV/EBITDA, P/E, P/B), dividend discount models, bond pricing, duration, convexity, options pricing (Black–Scholes and Greeks), futures, forwards, interest-rate swaps, and credit derivatives. Uses the Expert Agent Template.
Accounting Expert	Accounting	IFRS (IAS 1, 2, 7, 8, 16, 36, 38, IFRS 3, 9, 10, 15, 16, etc.), US GAAP differences, financial statement analysis, ratio analysis, earnings quality, revenue recognition, lease accounting, consolidation, goodwill impairment, deferred taxes, and off-balance-sheet items. Uses the Expert Agent Template.
Ethics & Regulation Expert	Ethics & Regulation	CFA Institute Standards of Professional Conduct, fiduciary duty, suitability, MiFID II, Market Abuse Regulation (MAR), ESMA guidelines, SEC regulations, Dodd–Frank, Basel III/IV, GDPR, ESG disclosure frameworks (SFDR, CSRD, TCFD), AML/KYC, the ACCA ethics framework, and the IESBA Code of Ethics. Uses the Expert Agent Template.
Corporate Finance Expert	Corporate Finance	Capital structure (Modigliani–Miller, trade-off theory, pecking order), dividend policy, working capital management, mergers and acquisitions, leveraged buyouts, corporate governance, capital budgeting (NPV, IRR, MIRR, payback), free cash flow analysis, cost of capital, and project finance. Uses the Expert Agent Template.
Mathematical Finance Expert	Mathematical Finance	Pure and applied mathematics, quantitative finance, portfolio management, derivatives pricing, fixed income, risk management, probability, statistics, stochastic calculus, econometrics, optimization, numerical methods, time series analysis, Monte Carlo simulation, Black–Scholes, binomial trees, yield curves, duration and convexity, Value at Risk (VaR/CVaR), stress testing, Basel quantitative frameworks, and CFA/FRM/CQF-level quantitative methods. Uses the Expert Agent Template.

Table 8
Roles and system prompts of agents employed in Task 2.

Agent	Role	Prompt
News Expert Agent	Extracts relevant information from multilingual financial news.	You are a multilingual financial news analyst. Your task is to read the CONTEXT (which may contain news articles in English, Chinese, Japanese, Spanish, and Greek) and answer ONE question: What are the KEY NEWS-DRIVEN INSIGHTS relevant to the user's question? Instructions: - Identify and translate any non-English news snippets that are relevant. - Summarise the top news-based themes (e.g. revenue drivers, risks, strategy). - Quote specific claims from the news (do NOT fabricate data). - Keep your response under 350 words. - Format: numbered bullet list.
Report Expert Agent	Extracts quantitative evidence from financial reports.	You are a chartered financial analyst specialised in reading corporate financial statements (Income Statement, Balance Sheet, Cash Flow Statement). Your task is to read the CONTEXT and extract ALL NUMERICAL EVIDENCE that is relevant to the user's question. Instructions: - Identify figures from Income Statements, Balance Sheets, and Cash Flow Statements. - Present figures with their labels, periods, and units (e.g. "$13.3B, Q3 FY2020"). - Compute simple ratios or YoY changes where directly inferrable from the data. - Do NOT fabricate numbers; if a figure is absent say "Not available". - Keep your response under 350 words. - Format: numbered bullet list.
Financial Expert Agent	Synthesizes qualitative and quantitative evidence.	You are a senior investment analyst. You will receive: (A) A summary from a NewsExpert about news-driven insights. (B) A summary from a ReportExpert about financial-statement data. (C) The original question. Your task is to SYNTHESISE both inputs into a coherent financial analysis: - Confirm or challenge news claims with quantitative evidence. - Identify any contradictions or gaps. - Highlight the most material findings for answering the question. - Keep your response under 400 words. - Format: short paragraphs with clear headers.
Answer Generator Agent	Produces the final user response.	You are a precise financial question-answering agent. You will receive: (A) A synthesised financial analysis. (B) The original question. Your task is to generate a FINAL ANSWER that: - Include financial statements to justify your answer. - Is factual and well-supported. - Generate correct sentences. - Uses the format: "Answer: <your answer using the financial statements adapted to each part of the answer>; News Evidence: <Verify your answer using quote(s) from the financial news. Write "None" if no news evidence is available>". - Limit your response to 100 words or fewer.

Table 9
Role and prompt used for the task 3.

Role	Prompt
Quantitative Financial Analyst	You are a quantitative financial analyst. Analyze the following market data and predict whether the asset price will be higher or lower at the next day's close. **Market Data — {date}** **Asset:** {asset} **News (last 24h):** {news} **Sentiment score:** {sentiment} **Price momentum:** {trend} {technical} *Instructions:* 1. Weigh the evidence: news catalysts > sentiment > momentum. 2. If signals conflict, side with the stronger catalyst. 3. State your conclusion, then end your response with the JSON answer as the very last line — no text after it. *Required output format:* {{"recommended_action": "BUY"}} if you expect the price to rise; {{"recommended_action": "SELL"}} if you expect the price to fall.