Fin_Agent at FinEval 2026: Task-Aware Methods for Multilingual Financial NLP

Mohotarema Rashid Tanzina Hossain
Xiaoying Song Lingzi Hong
University of North Texas
{MohotaremaRashid,tanzinahossain,XiaoyingSong}@my.unt.edu, Lingzi.Hong@unt.edu

Abstract

We present Fin_Agent, our submission to the FinEval 2026 shared task on multilingual financial language understanding. Our systems address Greek financial named-entity recognition, Japanese implicit communication recognition, and multilingual financial question answering. For Greek NER, we combine annotation-aware numeric rules with zero-shot Qwen2.5-32B classification, achieving 0.9521 public F1 and 0.8884 private F1. For Japanese implicit communication, an interpretable cue-based ordinal scorer achieves 98.0% test accuracy. For multilingual financial QA, BGE-M3 retrieval with Qwen3-32B generation achieves 0.3023 and 0.3419 ROUGE-1 F1 on the public and private splits. Key limitations include dataset-specific rules, handcrafted cues, and limited supervision. Implementation code is available https://github.com/Rashidnira/finagent-fineval-2026.

1 Introduction

Financial markets rely on heterogeneous information from corporate filings, financial statements, earnings announcements, news, and investor-facing communications (Rashid et al., 2026, 2025a,b; Zhao et al., 2026). Real-world financial analysis increasingly spans multiple languages and modalities, including reports, news, scanned filings, and meeting recordings (Peng et al., 2026). Because finance is a high-stakes domain, effective language processing requires both domain knowledge and contextual interpretation (Lin et al., 2026). These challenges are amplified in low-resource languages, where financial corpora, expert annotations, benchmarks, and pretrained models remain limited (Cao et al., 2026; Jørgensen et al., 2023; Peng et al., 2025). Although multilingual LLMs offer broad cross-lingual capabilities, their financial performance remains uneven across languages, particularly when specialized terminology and complex reasoning are involved (Peng et al., 2026).

Despite growing interest in applying AI and LLMs to finance (Wang et al., 2026), prior work has largely focused on tasks such as financial summarization, question answering, and stock-movement prediction, while multilingual and low-resource financial NLP remains comparatively underexplored (Cao et al., 2025). The FinEval 2026 Multilingual Financial Language Understanding shared task addresses this gap through three complementary problems: Subtask 1 focuses on Greek financial NER for extracting structured numeric and textual entities; Subtask 2 targets Japanese implicit communication recognition by modeling the degree of managerial commitment in investor-facing responses; and Subtask 3 evaluates multilingual financial question answering across SEC filings and contemporaneous news in English, Chinese, Japanese, Spanish, and Greek.

We develop task-specific systems tailored to the supervision and reasoning requirements of each subtask (Figure 1): a hybrid rule based LLM approach for Greek financial NER, an interpretable cue-based model for Japanese implicit communication, and a retrieval-guided generation pipeline for multilingual financial QA.

Our contributions are threefold: (1) we show that annotation-aware rules can effectively complement LLM-based classification under limited and asymmetric NER supervision; (2) we demonstrate that a lightweight, interpretable linguistic model can accurately capture ordinal commitment in Japanese financial communication without relying on LLM inference; and (3) we investigate retrieval-guided and question-level generation for multilingual financial QA and analyze their effects under controlled settings.

2 Related Work

2.1 Named-entity recognition in specialized domains

Financial NER extracts structured information from corporate filings, financial statements, and business documents, with recent work introducing

Proceedings of the 11th Workshop on Financial Technology and Natural Language Processing, pages 811–821
October 28, 2026 ©2026 Association for Computational Linguistics

domain-specific resources and evaluating LLMs for financial entity extraction (Loukas et al., 2022; Lu and Huo, 2025). However, financial NLP resources remain concentrated in high-resource languages, while domain-specific annotation is costly and scarce (Aguda et al., 2024; Uthayasooriyar et al., 2025), making non-English financial NER particularly challenging.

Like other low-resource languages, Greek NER suffers from limited annotated resources and morphological complexity (Lucarelli et al., 2007). Earlier approaches relied on rules and gazetteers, followed by statistical models and Greek-specific pretrained models such as GreekBERT (Koutsikakis et al., 2020; Michailidis et al., 2006). More recently, Peng et al. (2025) introduced Plutus and showed that Greek financial NER remains challenging even for modern multilingual and LLM-based models. To address this gap, Fin_Agent combines rule-based extraction for structured financial entities with zero-shot LLM classification for textual entities.

2.2 Implicit communication

Financial communication often conveys managerial intent indirectly through tone, uncertainty, and strategic wording, providing information beyond numerical disclosures (Bushee et al., 2018; Larcker and Zakolyukina, 2012; Li, 2008; Loughran and McDonald, 2011). However, financial NLP has largely focused on sentiment, tone, and deception rather than implicit managerial commitment (Kearney and Liu, 2014). Japanese financial NLP has mainly addressed corporate disclosures, financial benchmarks, and meeting understanding (Doi et al., 2020; Hirano, 2024; Zhu et al., 2025). We address this gap using interpretable linguistic cues for commitment, deferral, reaffirmation, and refusal.

2.3 Multilingual Financial Question Answering

Financial question answering requires models to identify and combine numerical and narrative evidence from corporate disclosures. FinBen broadens evaluation across financial QA and retrieval-augmented tasks (Xie et al., 2024), while bilingual studies report performance differences between English and Spanish financial NLP (Zhang et al., 2024). However, these settings do not directly test analyst-style questions requiring joint reasoning over financial statements and contemporaneous multilingual news. MultiFinBen addresses this

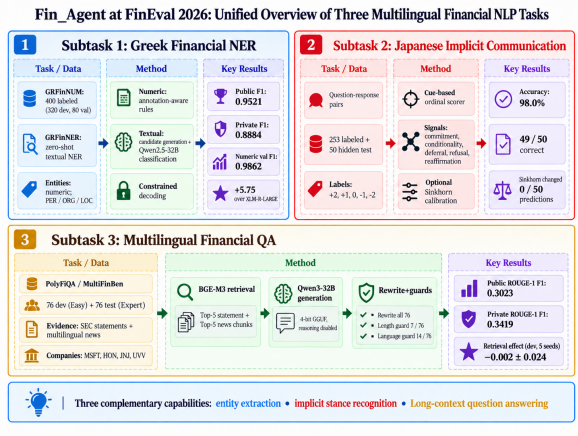

Figure 1: System architecture.

gap through PolyFiQA, which combines SEC statements, news in five languages, and analyst-style questions (Peng et al., 2026). Our system uses BGE-M3 retrieval, per-question Qwen3-32B generation, and controlled rewriting to produce concise, format-compliant answers.

3 Subtask 1: Greek Financial NER

3.1 Task and Data

Subtask 1 focuses on Greek financial NER using two datasets: GRFinNUM for numeric entities and GRFinNER for textual entities. GRFinNUM contains 400 labeled instances (320 development, 80 validation) covering money, percentages, temporal expressions, quantities, and other numeric entities, while GRFinNER targets persons, organizations, and locations and is evaluated in a zero-shot setting because no labeled training examples are provided.

3.2 System

We employ a hybrid architecture reflecting the asymmetric supervision available across GRFinNUM and GRFinNER.

Numeric entities are extracted using ordered regular expressions covering monetary amounts, percentages, dates, quantities, and other financial numeric expressions. Ambiguous numerals are classified using nearby lexical context, and span selection follows the annotation boundaries observed in the labeled data.

For textual NER, we adopt a two-stage candidate-classification approach. We first generate a high-recall set of candidate spans using capitalization patterns, corporate legal forms, gazetteers, person-name patterns, acronyms, and address-like expressions. Each candidate c_j is then classified by Qwen2.5-32B-Instruct-AWQ (Qwen Team, 2024)

Evaluation	System	F1	Rank
Public test	Hybrid system	0.9521	4
Private test	Hybrid system	0.8884	6
Validation	XLM-R-LARGE	0.9287	–
Validation	**Rules (ours)**	**0.9862**	–

Table 1: Subtask 1 official results and numeric-NER ablation.

as person, organization, location, or none, using the full passage x_i as context and the label set $\{\mathrm{PER}, \mathrm{ORG}, \mathrm{LOC}, \mathrm{NONE}\}$. Full rule details are in Appendix A. The predicted type is

$$\hat{t}_j = \arg\max_t p(t \mid x_i, c_j).$$

Because the model classifies only pre-generated spans rather than freely generating entities, all accepted predictions remain grounded in the source passage. A final constrained-decoding stage removes low-confidence, generic, and overlapping candidates.

3.3 Results and Analysis

Table 1 summarizes the main Subtask 1 results. Our full hybrid system ranked 4th on the public leaderboard with 0.9521 F1 and 6th on the private leaderboard with 0.8884 F1. Separately, on the numeric validation set, the proposed deterministic rules achieved 0.9862 F1, outperforming the XLM-R-LARGE baseline (Conneau et al., 2020) by 5.75 F1 points.

3.4 Limitations and Errors

For Subtask 1, Because gold labels for the textual GRFinNER data are unavailable, we assume the main errors involve numeric ambiguity, boundary or type confusion, and entity omission. Bare numbers could also be difficult to classify when local context is insufficient, while overlapping spans may lead to incorrect boundaries or ORG–LOC confusion. Some entities are likely to be missed as the candidate-generation stage fails to propose the correct span.

4 Subtask 2: Implicit Communication Recognition

4.1 Task and Data

Subtask 2 focuses on identifying the degree of commitment expressed in Japanese investor-relations responses. Given a question–response pair, the task is to determine whether the company's response expresses strong commitment, qualified commitment, neutrality or deferral, weak refusal, or strong refusal, corresponding to labels $+2$, $+1$, 0, -1, and -2, respectively. The released working data used by our system contain 253 labeled instances and 50 hidden test instances. Because the target label reflects the company's communicative stance, our system primarily analyzes the response, using the question only as supporting context.

Task	Public	Private	Metric
Subtask 2 Japanese ICR	100.0%	97.14%	Accuracy

Table 2: Official Subtask 2 results.

4.2 System

We employ a linguistically motivated scoring framework that maps each response to five pseudo-logit scores,

$$\mathbf{s}_i = [s_{+2}, s_{+1}, s_0, s_{-1}, s_{-2}],$$

using cues for explicit decisions, volition, conditionality, deferral, soft and hard refusal, reopening, and reaffirmation. Negation is treated contextually: expressions such as "there is no change" are interpreted as reaffirmation rather than refusal and are masked before hard-negation detection. The submitted system additionally applies Sinkhorn normalization (Sinkhorn and Knopp, 1967; Cuturi, 2013) under an assumed uniform class prior. The full cue inventory is in Appendix B.

4.3 Results and Analysis

For Subtask 2, the system achieved 15/15 accuracy on the public split and 34/35 on the private split, corresponding to a combined accuracy of 49/50 (98.0%) across both evaluation splits (Table 2).

For Japanese implicit communication, Sinkhorn calibration changes no test predictions showing that the linguistic scorer rather than batch-level prior matching determines the final decisions.

4.4 Limitations and Errors

For Japanese implicit communication, errors mainly occur between closely related commitment levels, such as qualified commitment vs. deferral and weak vs. strong refusal. The rule-based cues may also miss unseen or highly context-dependent expressions.

5 Subtask 3: Multilingual Financial Question Answering

5.1 Task and Data

Subtask 3 evaluates multilingual financial question answering using PolyFiQA from MultiFinBen (Peng et al., 2026). We use its two datasets: 76 quantitative questions from PolyFiQA-Easy for development and 76 analyst-style questions from PolyFiQA-Expert for evaluation. The evaluation set contains four questions for each of 19 financial-report groups covering MSFT, HON, JNJ, and UVV during 2020–2021. Both datasets use the same reports and evidence contexts but differ in their instructions and question styles. Each question includes SEC financial statements and related news in five languages. The required response contains a concise answer and supporting evidence and cannot exceed 100 words.

5.2 System

Because each context is long and combines several sources, we use retrieval to highlight relevant evidence without removing the original context. For each question, BGE-M3 retrieves the five most relevant financial-statement chunks and five most relevant news chunks (Chen et al., 2024). We insert these chunks as an evidence block before the final question. Qwen3-32B then generates one response per question using 4-bit GGUF inference with reasoning disabled (Yang et al., 2025). After generation, a rewrite pass targets 55–65 English words. If the response remains too long, a second pass targets 45 words. If the script check detects characters above U+2000, a language pass rewrites the answer in English with a 70-word target. Final checks flag potentially unsupported numbers and non-verbatim quotations without changing the response. Section 5.3 compares the public submission variants and reports a controlled development analysis.

We entered the model-based category and complied with its 70-billion-parameter limit (FinNLP-2026 Shared Task Organisers, 2026). Qwen3-32B and BGE-M3 contain approximately 33.3B parameters combined and run locally without model API calls; therefore, the OpenRouter requirement for API-based workflows does not apply. Four-bit quantization reduces memory use without changing the parameter count. Appendix C.1 provide the prompts and processing counts, and Appendix C.2 reports the parameter accounting.

Split	F1	Precision	Recall
Public (24)	0.3023	0.3040	0.3528
Private (52)	0.3419	0.3227	0.3899

Table 3: Official Subtask 3 results.

Configuration	Words	F1	P	R
Grouped, no retrieval	37.6	.2618	.344	.253
Per-question + BGE-M3†	91.8	.2987	.274	.386
Per-question + 55-word	61.2	.3022	.321	.329
Per-question + 72-word	71.8	.2944	.293	.342
Rewrite + normalization	69.4	**.3023**	.304	.353

Mean words use all 76 answers; other metrics use the 24 public instances. Final three rows are separate rewrite variants, not cumulative, and share no identical answers.
† Twenty-three answers exceeded 100 words. We retain this variant to show the performance before rewriting.

Table 4: Public-set submission variants.

5.3 Results and Analysis

Table 3 reports our official results. The final system scores higher on the private split than on the public split, with both precision and recall improving, especially recall. Table 4 records five public submission variants. The largest change occurs when we replace grouped generation without retrieval with question-level generation and retrieval. The three independent rewrite variants differ by less than 0.008 F1. Because they use separate scripts and share no identical answers, these results do not isolate the effects of answer length, language rewriting, or monetary normalization.

Because the largest change introduces retrieval and question-level generation together, the public results cannot isolate either effect. We therefore test all four combinations on PolyFiQA-Easy across five seeds while holding the context preprocessing, instruction format, model, decoding settings, and scorer constant. Mean F1 differs by less than 0.004 across the four configurations. The estimated effects of retrieval, generation granularity, and their interaction are smaller than their variation across seeds. We therefore find no stable advantage for either component under these conditions. Appendix C.3 reports the complete analysis.

Grouped answers average 33 words and question-level answers average 52 words, compared with a 48-word reference average. Despite this difference, the configurations achieve similar F1 scores. Thus, answer length alone does not explain performance in the controlled analysis.

5.4 Limitations and Errors

Our further error investigation finds six missing evidence labels and flag two numbers and five quotations for review. The two number flags and four quotation flags originate in generation; the length guard adds one quotation flag and all six missing labels (Appendix C.4). Manual inspection also finds one altered monetary scale, which the validator does not test. Although post-processing introduces some problems, it remains necessary because 23 raw answers exceed the 100-word limit. Hidden references prevent us from measuring the score effect of these cases. The controlled analysis uses only the development set and may not generalize to the evaluation questions. Repeated evaluation on the small public split may have influenced configuration selection. Finally, our validators flag but do not repair possible errors, and ROUGE-1 measures word overlap rather than factual accuracy.

6 Conclusion

We present task-aware systems for three complementary multilingual financial NLP problems. Our results on Greek NER and Japanese implicit communication show that specialized and interpretable methods can remain highly effective despite the increasing use of general-purpose LLMs. In particular, annotation-aware rules substantially outperform the XLM-R-LARGE baseline for numeric NER, while the Japanese linguistic scorer achieves 98.0% test accuracy independently of Sinkhorn calibration. For Subtask 3, our long-context QA pipeline combines retrieval and generation, but the controlled development analysis finds no stable advantage for either retrieval or generation strategy.

References

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145. ELRA and ICCL.

Brian J. Bushee, Ian D. Gow, and Daniel J. Taylor. 2018. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121.

Yupeng Cao, Zhi Chen, Prashant Kumar, Qingyun Pei, Yangyang Yu, Haohang Li, Fabrizio Dimino, Lorenzo Ausiello, K. P. Subbalakshmi, and Papa Momar Ndiaye. 2025. RiskLabs: Predicting financial risk using large language model based on multimodal and multisources data. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 901–909.

Yupeng Cao, Haohang Li, Weijin Liu, Wenbo Cao, Anke Xu, Lingfei Qian, Xueqing Peng, Minxue Tang, Zhiyuan Yao, Jimin Huang, K. P. Subbalakshmi, Zining Zhu, Jordan W. Suchow, and Yangyang Yu. 2026. FinTrace: Holistic trajectory-level evaluation of LLM tool calling for long-horizon financial tasks. *arXiv preprint arXiv:2604.10015*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2292–2300.

Nobushige Doi, Yusuke Oda, and Toshiaki Nakazawa. 2020. TDDC: Timely disclosure documents corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3719–3726. European Language Resources Association.

FinNLP-2026 Shared Task Organisers. 2026. The FinNLP-2026 shared task on multilingual financial language understanding. `https://www.finnlp2026sharedtask.info/`. Retrieved September 4, 2026.

Masanori Hirano. 2024. Construction of a Japanese financial benchmark for large language models. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 1–9. Association for Computational Linguistics.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909. Association for Computational Linguistics.

Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020)*, pages 110–117.

David F. Larcker and Anastasia A. Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540.

Feng Li. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2):221–247. Special issue: Economic Consequences of Alternative Accounting Standards and Regulation.

Shengyuan Lin, Kaiwen He, Jaisal Patel, Qinchuan Zhang, Chris Ding, James Tang, Keyi Wang, Yupeng Cao, Yan Wang, Kairong Xiao, Vincent Caldeira, Matt White, and Xiao-Yang Liu Yanglet. 2026. Evaluation and benchmarking suite for financial large language models and agents. *arXiv preprint arXiv:2602.19073*.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431. Association for Computational Linguistics.

Yi-Te Lu and Yintong Huo. 2025. Financial named entity recognition: How far can LLM go? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 164–168. Association for Computational Linguistics.

Georgios Lucarelli, Xenofon Vasilakos, and Ion Androutsopoulos. 2007. Named entity recognition in Greek texts with an ensemble of SVMs and active learning. *International Journal on Artificial Intelligence Tools*, 16(6):1015–1045.

Ionas Michailidis, Konstantinos Diamantaras, Spiros Vasileiadis, and Yannick Frère. 2006. Greek named entity recognition using support vector machines, maximum entropy and onetime. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource Greek finance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30176–30202, Suzhou, China. Association for Computational Linguistics.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and 28 others. 2026. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16934–16963, San Diego, California, United States. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mohotarema Rashid, Lingzi Hong, Zunhua Ding, and K. S. M. Tozammel Hossain. 2026. Fin-analyst at FinMMEval 2026 task 3: A live hybrid trading agent with LLM specialists and rule-based signals. *arXiv preprint arXiv:2607.12233*.

Mohotarema Rashid, Lingzi Hong, and Sarah Ryan. 2025a. Investigating fake financial news: A systematic review of trends, themes, LLM-based detection methods, and theories. *Aslib Journal of Information Management*. Ahead-of-print.

Mohotarema Rashid, Lingzi Hong, Sarah Ryan, Mahfuja Malik, and Jodi Philbrick. 2025b. Modeling the predictors of fake financial news sharing on social media using behavioral reasoning theory: Evidence from retail investors of USA. In *Proceedings of the Association for Information Science and Technology*, volume 62, pages 533–546.

Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Benno Uthayasooriyar, Antoine Ly, Franck Vermet, and Caio Corro. 2025. Training LayoutLM from scratch for efficient named-entity recognition in the insurance domain. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 101–110. Association for Computational Linguistics.

Yan Wang, Xuguang Ai, Jaisal Patel, Xueqing Peng, Fengran Mo, Yupeng Cao, Haohang Li, Mingyu Cao, Lingfei Qian, and Víctor Gutiérrez-Basulto. 2026. AUDITFLOW: Executable symbolic environments for structured financial reporting verification. *arXiv preprint arXiv:2606.03031*.

Qianqian Xie, Weiqiang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yuzhi Dai, Duanyu Feng, Yijing Xu, Hao Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zhe Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 14 others. 2024. FinBen: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems*, volume 37.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiqiang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, Jimin Huang, and Qianqian Xie. 2024. Dólares or dollars? unraveling the bilingual prowess of financial LLMs between Spanish and English. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6236–6246.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Hanqi Jiang, Yi Pan, Junhao Chen, Yifan Zhou, Wei Ruan, Zeyu Zhang, Ruitong Sun, Gengchen Mai, Ninghao Liu, and Tianming Liu. 2026. Revolutionizing finance with LLMs: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

Jie Zhu, Junhui Li, Yalong Wen, Xiandong Li, Lifan Guo, and Feng Chen. 2025. MFinMeeting: A multilingual, multi-sector, and multi-task financial meeting understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 244–266. Association for Computational Linguistics.

A Subtask 1: Methods in Detail

We entered Subtask 1 as a model-based solution. The only neural model, Qwen2.5-32B-Instruct-AWQ, has 32B total parameters and is therefore below the 70B limit. It was served locally with vLLM 0.6.6.post1. Numeric NER uses no model; textual NER uses grounded candidate classification.

A.1 Numeric extraction

The extractor operates on character offsets and returns ordered `(start,end,surface,type)` records. A character-level claim array prevents a lower-priority rule from assigning an overlapping span. Accepted surfaces are never normalized, and repeated surface–type pairs are retained because the evaluation treats them as multisets.

The base numeral expression matches digits followed by zero or more period- or comma-delimited digit groups. The exact expression and all ordered rules are available in the accompanying code repository. Rules are applied in this order: (i) percentages, including ranges; (ii) three-part dates, decomposed into day, month, and year; (iii) law, article, paragraph, registry, and accounting-standard identifiers; (iv) named-month and partial dates, durations, and bare years; and (v) contextual classification of remaining numerals. Thus, for example, *31.12.2020* yields three temporal mentions, while a monetary phrase normally yields the bare numeral rather than its currency or magnitude word. This follows the released annotation convention: all labeled money, temporal, and other-numeric spans are single-token, as are 98% of percentages.

For a remaining numeral at offsets $[a, b)$, the default context is the 45 characters before a and 40 characters after b. Immediate monetary and unit tests use narrower windows of 14 characters before and 22 after. Currency symbols and amount expressions assign $\mathrm{XPHMATA}$; count nouns, share terminology, parentheses, and physical units assign $\Pi\mathrm{O\Sigma OTHTE\Sigma}$; temporal expressions assign $\mathrm{XPONIKA}$; and an attached percent sign assigns $\Pi\mathrm{O\Sigma O\Sigma TA}$. Before fallback classification, the extractor rejects numbers embedded in Latin names, street addresses, or hyphenated expressions such as *COVID-19*. Decimal-like unmatched numbers default to money when document-level money evidence is present; otherwise the fallback is quantity.

On the fixed 320/80 split of the 400 labeled numeric examples, the submitted rules obtained the following exact entity-level multiset scores.

Split	Precision	Recall	Micro-F1
Training (320)	0.9877	0.9683	0.9779
Validation (80)	0.9882	0.9843	0.9862

Table 5: Subtask 1 numeric-rule performance.

A.2 Textual candidate classification

Because no labeled GRFinNER examples were supplied, textual NER is zero-shot. The tokenizer preserves dotted abbreviations, Unicode words, numbers, and punctuation. Candidate sources comprise maximal capitalized runs; their prefixes, suffixes, and contiguous sub-runs; spans ending in Greek or international legal forms; gazetteer matches; person-shaped bigrams; title-anchored names; parenthesized acronyms; lowercase geo-

graphic heads; and street-number extensions. Exhaustive sub-runs are capped at six tokens, whereas maximal and legal-form spans are capped at sixteen. Before inference, we discard candidates shorter than two characters, without letters, longer than twelve tokens, or beginning with lowercase text.

We built a dictionary of Greek person, organization, and location names from eight existing NER datasets. : `wikiann_el`, `elNER4`, `elNER18`, `openner_core_elNER_ell`, `greek_legal_ner`, `polyglot_ner_el`, `mapa_el`, and `finerweb_ell`. The released code provides the gazetteer-construction procedure and records the contributing resources.

For each candidate, the model receives the complete passage and candidate span and predicts one label: person, organization, location, or none. We use deterministic, single-token decoding and normalize the returned label log-probabilities. Outputs that do not contain a valid label are treated as none.

For candidate c, we define entity confidence as $p_{\text{ent}}(c) = 1 - p(D \mid c)$. Candidates with $p_{\text{ent}}(c) \geq 0.90$ are decoded greedily into non-overlapping spans, prioritizing longer spans and then higher confidence. Each retained candidate is assigned the highest-probability entity class among person, organization, and location.

B Subtask 2: Methods in Detail

Our method of Subtask 2 is rule based. It maps the Japanese response to scores in label order $[+2, +1, 0, -1, -2]$. The patterns below give the exact submitted cue inventory in compact form.

B.1 Cue inventory

REAFFIRM spans are removed before counting hard negation. This prevents 変更はありません (“there is no change”) from being treated as a refusal while retaining a separate genuine refusal in the same response. Bare ありません is deliberately absent from NEG_HARD because it often negates a fact rather than the proposed action.

B.2 Scoring function

Let c_f be the number of matches from family f and $m_k(c) = \min(c, k)$. Scores begin at zero. Define

$$d = 2.0m_2(c_{\text{DECIDED}}) + 0.8m_2(c_{\text{DONE}}), \quad (1)$$

$$u = 0.7m_2(c_{\text{NUM}}) + 0.7m_2(c_{\text{DATE}}). \quad (2)$$

If $d > 0$, add $d + u$ to s_{+2} and $0.5d$ to s_{+1}; under hard negation, instead add $d + u$ to s_{-2} and $0.4d$ to s_{-1}. Hard negation adds $2.2m_2(c_{\text{NEG_HARD}})$ to s_{-2} and $0.9m_2(c_{\text{NEG_HARD}})$ to s_{-1}. Soft negation adds $2.0m_2(c_{\text{NEG_SOFT}})$ to s_{-1} and $0.7m_2(c_{\text{NEG_SOFT}})$ to s_{-2}. Negation with REOPEN adds 1.8 to s_{-1} and subtracts 1.4 from s_{-2}; without REOPEN it adds 0.9 to s_{-2}.

VOLIT and COND add

$$w\big[1.5m_2(c_{\text{VOLIT}}) + 0.9m_2(c_{\text{COND}})\big] \quad (3)$$

to s_{+1}, where $w = 0.5$ under hard negation and 1 otherwise. NOPROMISE adds 1.0 to s_{+1} and subtracts 1.5 from s_{+2}. VOLIT or COND without d subtracts 1.0 from s_{+2}. REAFFIRM without hard negation adds 1.2 to s_{+2} and 0.6 to s_0. NEUTRAL adds $1.6m_3(c_{\text{NEUTRAL}})$ to s_0; a response with no negation, VOLIT, or decision adds another 1.5 to s_0, whereas any negation subtracts 1.2. No polarity reversal is applied when the question proposes a negative action. NumPy `argmax` breaks ties in label order $[+2, +1, 0, -1, -2]$.

B.3 Split, calibration, and results

The training set contains 253 labeled examples, distributed as $+2$: 29, $+1$: 139, 0: 72, -1: 11, and -2: 2. We construct a fixed 26-example development split by sampling six examples from each class, except -2, for which both available examples are used.

The pipeline applies 80 log-space Sinkhorn iterations with a uniform target prior, $\boldsymbol{\pi} = (0.2, 0.2, 0.2, 0.2, 0.2)$, temperature 1.0, and strength 1.0. Final labels are selected using row-wise argmax. Sinkhorn changes none of the 50 evaluation predictions, indicating that the final predictions are determined by the rule-based scorer. As Table 7 shows, the likely test-set improvement reflects shorter, clearer test responses with more explicit decision and refusal cues, rather than a benefit from Sinkhorn calibration. Since the evaluation labels and their distribution are unavailable, we cannot conclusively determine which dataset characteristics produced the large development–evaluation difference.

C Subtask 3: Methods in Detail

C.1 Prompts and Processing Counts

The submitted pipeline uses one generation stage followed by an initial rewrite and two conditional guards. Every model call uses Qwen3-32B with 4-bit GGUF inference and reasoning disabled. We

Family	Submitted Japanese cues (regular-expression alternatives)
DECIDED	決議し/いた/済, 決定し/いた/済, 承認し/いた/済, 締結し/済/いた, 確定して/しま/済, 正式に決, 可決, 合意に達し, 署名し, 決まりました, 取締役会で決議/承認/確認/決定
DONE	完了します/しました/いたします, 開始します, 実施します, 移行します, 適用します, 充当します, 返済します, 継続します, 実行します, 導入します, 払い戻します, 上場します, 着工します
NEG_HARD	いいえ, 予定/計画/こと/事はありません・ございません, する考えはありません, 意思/方針はありません, しません, いたしません, 行いません, 実施しません, 提供しない, 検討していません, 中止, 終了します, 撤回, 除外して, 白紙, 社内規程, お答えできません, 回答を差し控え, 開示する考えはない, ノーコメント
NEG_SOFT	難しい, 困難と, 見送る/見送り, 時期尚早, 優先順位は低い, 考えていません, 踏み切る考えは, 慎重に, 控えます, 差し控え
REOPEN	ただし/但し, 一方で, 改めて検討/判断/見直し, 再検討, 見直します, 余地は/がある, 状況が変われば, 環境が変わる, 整った段階, 数年後, 将来的には検討, 選択肢の一つ, 高まれば, 柔軟に
VOLIT	たいと考え, たいと思う/です, 努めて, 目指し/目指す, 注力して, 進めてまいり, 拡大したい, 引き上げたい, 検討してまいり, 検討を進め, 前向き, 想定している, 判断します, 排除しておらず
COND	れば, 次第, 整えば, 得られれば, できれば, 想定どおり, 条件が, 確認できれば, 進捗次第, 市況次第, 場合には, 可能性があります, 応じて
NEUTRAL	検証して, 精査して, 分析して, 策定過程/策定中, 評価している, 確認している, 段階です, 据え置き, 次回の説明, 発表時に説明, お示しします, 公表します, 双方, 総合的に勘案/判断, 継続的に評価, 検討している段階
NOPROMISE	確約はできない, 約束は/できない, 保証は/できない, 決まっておりません, 確定しておりません, 決定した事実はありません, 決定しておりません
REAFFIRM	変更は/もありません・ございません, 変更はない/無い, 変更せず, 変更しません, 計画どおり継続, 従来どおり変更なし
NUM/DATE	数字followed by 億円, 百万円, 万株, 株, 円, %, 拠点, 店, 名, 件, 年, 月, 日; 20XX年, 月日, 今期末, 来年度/下期, 本日, 来月, 次期, 期限まで

Table 6: Compact rendering of the exact Subtask 2 regex inventory. Slash-separated forms denote alternatives; the released source preserves the precise grouping and suffixes.

Decision rule	Accuracy	Macro-recall
Raw rule-score argmax	0.3846	0.4000
Sinkhorn, strength 1.0	0.2692	0.3000

Table 7: Subtask 2 results on the fixed 26-item development split.

use temperature 0.7, top-p 0.8, top-k 20, a 4,096-token output limit, and a 32,768-token context window. BGE-M3 retrieves the five highest-ranked financial-statement chunks and five highest-ranked news chunks for each question. The headings below report each stage's trigger and processing count.

Evidence block (all 76). We insert the retrieved chunks into the organisers' query immediately before its final Question: line without replacing the original context.

> Highlighted Evidence (automatically retrieved, verbatim excerpts from the context above; check these first, but the full context above remains authoritative): [Financial statement excerpts] … [News excerpts]

Rewrite (all 76). Applied to every generated answer.

> You rewrite financial-QA answers to a strict house style. Rewrite the answer below. Rules:
> — Target 55–65 words total, NEVER more than 75.
> — ENGLISH ONLY. If the answer contains any Chinese, Japanese, Greek, or Spanish text (including quotes), replace it with a faithful English rendering of the same information. No non-Latin characters may remain.
> — Keep the exact labeled structure of the original (e.g. "Answer: …" and the "… Evidence: …" line with its exact label).
> — Write monetary amounts compactly: $27.2B, $416M, $3.5M — but ONLY when the magnitude is clear from the original wording (e.g. "US$14,014 million" → "$14.0B"). If the scale of a raw number is not explicit, keep it unchanged. Never alter what a number means. Keep percentages as printed (e.g. 18.6%).
> — Keep every distinct fact. Reuse the question's key terms in the wording.
> — Remove filler, hedging, repetition, and meta-commentary.
> — Output ONLY the rewritten answer text, nothing else.

Length guard (7 of 76). Applied only when the rewritten answer still exceeded 100 words.

> Rewrite the answer below to AT MOST 45 words total. Keep the exact labeled structure (Answer: / … Evidence: lines), keep every figure exactly as written, keep at most ONE short verbatim quote (drop the rest; never alter or translate inside quotation marks). Output ONLY the rewritten answer.

Language guard (14 of 76). Applied when the script detector found characters above U+2000 after the rewrite and length-guard stages. This check detects Chinese and Japanese characters and some typographic punctuation, but not Greek, Cyrillic, or Spanish written in the Latin script.

> The answer below contains non-English text (Chinese, Japanese, Greek, or Spanish). Rewrite it in ENGLISH ONLY, at most 70 words: translate the non-English parts faithfully into English prose (no quotation marks around translations), keep the labeled structure (Answer: / … Evidence: lines), keep all figures exactly. Not a single non-Latin character may remain. Output ONLY the rewritten answer.

Of the 76 generated answers, 23 exceeded the 100-word limit before rewriting, with the longest containing 153 words. Seven still exceeded the limit after the initial rewrite and therefore received the length guard.

C.2 Parameter Accounting

The shared task defines two participation categories. The published participation requirements

state that "Participants may submit either a model-based solution or an API-based agent/workflow solution." For model-based solutions, "The submitted model must not exceed 70 billion total parameters (70B)." For API-based agent and workflow solutions, "All model API calls must be made through OpenRouter," and "Each model used in the workflow must not exceed 70 billion total parameters (70B)." For Mixture-of-Experts models, "Eligibility is determined by the model's total parameter count, not the number of parameters activated during inference" ((FinNLP-2026 Shared Task Organisers, 2026)). We entered the model-based category, used the released weights without further training, and made no model API calls. The OpenRouter requirement is scoped to the API-based category and does not apply to our submission. Table 8 includes every learned component. We conservatively sum the generator and retriever rather than counting the generation model alone.

Component	Role	Parameters
Qwen3-32B	Generation/rewriting	32,762,123,264
BGE-M3	Passage retrieval	567,754,752
Deterministic code	Chunking/validation	0
Total		33,329,878,016

Table 8: Learned components and total parameter count for Subtask 3.

We serve Qwen3-32B using 4-bit GGUF weights. Quantization reduces storage precision without removing parameters. Both models use open weights and run on local hardware without contacting a hosted inference service. BGE-M3 generates no text; it ranks source passages that are inserted into the generation prompt. Qwen3-32B performs all answer generation and rewriting, while the chunker and validation checks use no learned parameters. The parameter counts come from the released model artefacts. The Qwen3-32B count is taken from the `general.parameter_count` field of the served model file, and the BGE-M3 count is calculated from the tensor shapes in its released weight file.

C.3 Controlled Development Analysis

Table 4 compares variants that differ in more than one respect at a time. To separate the two design choices, we built a grid of four configurations and ran each on the 76 development instances with five seeds.

Label	Generation	Retrieval
G	Grouped (4 questions/call)	None
G+R	Grouped	Per-question evidence
Q	One call per question	None
Q+R	One call per question	Evidence before question

Table 9: Controlled-analysis configurations.

The cells differ only in retrieval and generation strategy. The generation strategy includes question grouping, the number of model calls, and output allocation. The 4,096-token output limit never binds: across the 190 grouped calls the longest completion was 694 tokens and every call ended normally. We built all four prompt sets from a single source. Each retrieval prompt records where its evidence block was inserted, so removing that span recovers the original query exactly, and the same span is what we re-insert for the retrieval cells. Every cell uses the same raw context, the same model and quantization, the same decoding settings, the same output limit, and the same scoring code. We apply no rewriting or normalization. One difference cannot be removed: answering four questions in one call requires an instruction stating how to separate the answers. That instruction is identical in G and G+R, so it cannot affect the retrieval comparison. We ran this analysis after the evaluation phase closed, with a 40,960-token context window and the Vulkan backend of our inference server on an H100 GPU, sampling at temperature 0.7 with seeds 1 to 5. This environment differs from the one that produced our submission, so the grid is a controlled analysis rather than a reproduction of the submitted runs. Using one environment for all four cells supports internally consistent comparisons.

Configuration	Default F1	Multiscript F1
G	0.4547 ± 0.0151	0.4396 ± 0.0191
G+R	0.4525 ± 0.0103	0.4391 ± 0.0097
Q	0.4564 ± 0.0054	0.4383 ± 0.0087
Q+R	0.4559 ± 0.0080	0.4406 ± 0.0092

Table 10: Development ROUGE-1 F1, mean and standard deviation over five seeds.

We compute each effect within a seed and then summarise the five differences, which accounts for seed-level variation shared across the four configurations.

Contrast	Default	Multiscript
G+R−G	-0.0022 ± 0.0239	-0.0005 ± 0.0269
Q+R−Q	-0.0005 ± 0.0072	$+0.0023 \pm 0.0141$
Q−G	$+0.0017 \pm 0.0195$	-0.0013 ± 0.0231
(Q+R)−(G+R)	$+0.0034 \pm 0.0113$	$+0.0015 \pm 0.0110$
Interaction	$+0.0017 \pm 0.0250$	$+0.0028 \pm 0.0273$

Table 11: Within-seed paired effects on development ROUGE-1 F1 (mean ± s.d. over five seeds).

Every estimate is smaller than the variation across the five runs, and the four configurations fall within 0.004 F1 of one another. With five seeds these figures describe the spread we observed; they are not a significance test. A small estimate means an effect we could not resolve at this sample size, not the absence of one.

The largest descriptive differences appear for R&D and cash-flow questions, but the patterns vary with retrieval. Each question type contains only 19 instances, so we draw no general conclusions.

We release the prompt templates and scripts required to generate the four configurations and reproduce the controlled analysis. A consistency audit checks that the cells differ only as described, that every run produced 76 answers, that all generations used the pinned model at the recorded seeds and window, and that replaying a cell from cache reproduces its answers exactly.

Configuration	Mean words	Ratio to reference
Reference	47.9	1.00
G	33.3	0.70
G+R	35.3	0.74
Q	51.6	1.08
Q+R	55.8	1.16

Table 12: Mean answer length in the controlled development analysis.

Configuration	Revenue	Balance sheet	Cash flow	R&D
G	.422	.460	.312	.625
G+R	.420	.464	.309	.617
Q	.424	.483	.340	.578
Q+R	.442	.482	.308	.591

Table 13: Development F1 by question type, default scorer, averaged over five seeds. Each type contains 19 instances.

C.4 Sources of Remaining Validator Flags

Our pipeline applies three post-processing stages after generation. To find out whether the flags raised on our submitted answers come from the model or from our own processing, we reproduced each stage from the cached responses and applied the same reference-free validator after each one. The final stage reproduces the submitted file byte for byte, so the trace describes the system we submitted.

The three classes of validator flags arise at different stages. Untraceable-number flags originate in generation, and the rewrite clears three of the five affected rows without adding any. Four quotation flags originate in generation, while the length guard adds one. The length guard introduces all six missing evidence labels, dropping the label from six answers while compressing them below the word limit. That last result is a failure to follow an explicit instruction rather than a missing one. The length-guard prompt in C.1 requires the labeled structure to be kept, and the model still dropped the label in six of the seven answers the stage touched. We conclude that a required output element needs a deterministic check after generation, and that a validator which only flags such cases, as ours does, will not prevent them.

Flagged rows	Raw	+Rewr.	+Len.	+Lang.
Untraceable numbers	5	2	2	2
Non-verbatim quotes	4	4	5	5
Missing evidence label	0	0	6	6
Rows with any flag	19	8	11	11

Table 14: Validator flags after each stage. Values are numbers of rows. The final row reports the union of all flagged rows, including checks not listed separately, chiefly answers longer than 110 words at raw generation. Because one answer may carry multiple flags, this value need not equal the sum of the three displayed classes. The label check tests for an exact string and is reliable; the number and quotation checks are heuristic and may flag derived figures or normalised wording.

Stage	Mean words	Max words	Over 100
Raw generation	91.8	153	23
+ Rewrite	73.7	126	7
+ Length guard	69.4	99	0
+ Language guard	69.4	99	0

Table 15: Answer length after each stage of the submitted post-processing pipeline. The two compression variants in Table 4 were produced by separate scripts and are not sequential stages of this pipeline.

After the language guard, 10 answers still triggered the code-point check because they contained curly quotation marks or dashes, not Chinese or Japanese text.