PFB at EVALITA 2026: Overview of the Prometeia Financial Benchmark

Alessandro P. Bardelli[2,†], Tolga Çekiç[3,†], Irem Demirtaş[3,†], Michele Filannino[2,†], Simona Scala[1,*,†], Andrea Galassi[4,†], Gianmarco Pappacoda[4,†] and Paolo Torroni[4,†]

[1]*Prometeia, Piazza Trento e Trieste 3, 40137, Bologna, Italy*
[2]*Prometeia, Via Brera, 18, 20121, Milan, Italy*
[3]*Prometeia, River Plaza, Kat 19 Büyükdere Caddesi Bahar Sokak No. 13, 34394, Levent | Istanbul | Turkey*
[4]*Università di Bologna, Dipartimento Informatica - Scienza e Ingegneria, Viale del Risorgimento 2, 40136, Bologna, Italy*

Abstract

The Prometeia Financial Benchmark (PFB) is the EVALITA 2026 shared task on finance questions across 3 languages: Italian, English, and Turkish, and 3 difficulty levels: easy, medium, and hard. The challenge is organized in two subtasks, one on Italian data and one on all three languages. For each subtask, we have received 2 submissions. Our main takeaways are that no significant performance differences stand out across languages and difficulty levels, and that PFB appears to be a challenging benchmark for models smaller than 3B, whereas 20B models already reach an overall accuracy of 90%.

Keywords

Large Language Models, Finance NLP, Multiple-choice QA, Multilingual benchmark, Prometeia Financial Benchmark

1. Introduction

Language Models (LMs) are increasingly adopted as general-purpose components for knowledge-intensive workflows, and finance is a natural target for their use [1, 2]. Finance-related language and reasoning, however, impose stricter requirements than many general-purpose benchmarks capture: terminology is specialized, concepts are tightly interdependent, and answers that appear plausible at the surface level may still be incorrect. Recent evidence highlights reliability limitations in finance-specific settings, including hallucination-related issues [3], motivating domain-targeted evaluation protocols.

This paper presents the *Prometeia Financial Benchmark* (PFB) shared task, organized as a track within EVALITA26 [4], to benchmark LMs on finance-domain multiple-choice question answering (MCQA). The task is based on a curated dataset of 1,500 questions released in aligned English, Italian, and Turkish versions, and is structured into an Italian-only and a multilingual subtrack. We describe the task definition, dataset construction, and evaluation protocol, and provide an overview of participant submissions and results.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author.
†These authors contributed equally.
alessandro.bardelli@prometeia.com (A. P. Bardelli); tolga.cekic@prometeia.com (T. Çekiç); irem.demirtas@prometeia.com (I. Demirtaş); michele.filannino@prometeia.com (M. Filannino); simona.scala@prometeia.com (S. Scala); a.galassi@unibo.it (A. Galassi); gianmarco.pappacoda@unibo.it (G. Pappacoda); p.torroni@unibo.it (P. Torroni)
0009-0009-7670-5668 (A. P. Bardelli); 0009-0008-2061-5893 (T. Çekiç); 0009-0007-5586-9773 (I. Demirtaş); 0000-0001-8208-2238 (M. Filannino); 0009-0003-5324-5466 (S. Scala); 0000-0001-9711-7042 (A. Galassi); 0009-0001-6609-4156 (G. Pappacoda); 0000-0002-9253-8638 (P. Torroni)

Dedicated to the memory of Annalina Caputo.

2. Motivation

Finance is a high-stakes domain in which language technology is expected to support decision-making, compliance-driven workflows, and knowledge access over complex documentation. Although modern LMs perform well on broad benchmarks, their fluency can mask brittle understanding and overconfident errors [5]. In finance, such failures are especially problematic: incorrect statements may propagate into reports, internal procedures, or user-facing advisory contexts, increasing operational and financial risk [3, 6]. A dedicated benchmark is therefore needed to quantify reliability on domain content rather than generic linguistic ability.

Many finance tasks require distinguishing between closely related concepts, interpreting definitions precisely, and selecting the only option that is fully consistent with the question context. MCQA provides a controlled evaluation setting: the decision space is explicit, distractors can be designed to be plausible within the domain, and scoring is unambiguous. This helps separate genuine domain comprehension from persuasive surface generation.

Multilinguality is an additional practical driver. Financial institutions operate across markets, and key materials are often consumed in local languages, not only in English. Yet most available benchmarks remain predominantly English-centric, making it difficult to disentangle domain effects from language effects. Multilingual, aligned resources enable controlled cross-lingual comparisons and support analyses of robustness, transfer, and language-specific terminology handling [7, 8, 9, 10].

Finally, the shared-task format provides a transparent and reproducible comparison framework. It complements prior work on financial LLM evaluation and benchmarking [11, 12, 13] by fixing data, protocol, and metrics, while encouraging diverse modelling approaches and facilitating systematic error analysis.

3. Definition of the task

The shared task assesses a model's ability to *understand and reason over finance-domain content* in a controlled MCQA setting. Given a question and five candidate answers, systems must return the option that best answers the question. The task is open with respect to modelling choices: participants were free to submit systems based on different paradigms (e.g., encoder-only or decoder-only LMs, instruction-tuned LLMs, prompted systems) and to rely on either open or proprietary models, including both "small" and "large" language models.

For each test instance, participants submit a single answer choice in {A,B,C,D,E}. Systems may optionally provide short textual justifications. These explanations are not used for ranking, but they may support qualitative analyses (e.g., recurring failure modes, reasoning patterns, and domain-specific errors).

The shared task is organized into two subtracks. The first is an **Italian-only** track, which constitutes the primary focus within EVALITA. The second is a **multilingual** track covering Italian, English, and Turkish, designed to enable controlled comparisons across aligned instances. Participants may submit to one or both subtracks.

For this task, participating teams were required to submit their primary run along with information about the model architecture, inference strategy, and related implementation specifics. Additionally, each team was allowed to submit up to four secondary runs.

3.1. Dataset

The task is based on the *Prometeia Financial Benchmark* (PFB) [14], a curated collection of 1,500 finance-domain items released in three aligned languages (English, Italian, and Turkish). Questions were derived from heterogeneous finance-related sources (e.g., reports, papers, and regulatory texts) and curated with expert review to ensure domain relevance, internal consistency, and clarity.

Each instance is identified by a language-invariant `custom_id`, enabling one-to-one alignment across the three languages. Provenance is captured by a `category` label indicating the source family. Items

Table 1
Core fields in the PFB dataset release

Field	Description
`custom_id`	Unique identifier, shared across languages for alignment.
`question`	Question stem.
`choiceA–choiceE`	Five candidate answers.
`correct_answer`	Gold option in {A,B,C,D,E}.
`difficulty_level`	Categorical difficulty in {easy, medium, hard}.

follow a standard MCQA schema with the stem (`question`), answer options (`choiceA–choiceE`), and the gold label `correct_answer` in {A,B,C,D,E}. In addition, the dataset provides a categorical indicator, `difficulty_level` (easy, medium, hard).

Table 2
Distribution of correct choices across labels in PFB dataset

Label	A	B	C	D	E
Number	402	438	424	237	0
Percentage	26.8	29.1	28.3	15.8	0

Table 3
Distribution of difficulty in PFB dataset

Difficulty	Easy	Medium	Hard
Number	248	1155	98
Percentage	16.6	76.9	6.5

Dataset construction followed a two-stage process: a large pool of candidate questions was first produced from the selected sources: financial texts (financial books, and regulatory texts), academic papers and financial reports (Table 4). Then iteratively filtered and refined through expert review. Quality control relied on structured assessment signals recorded at the item level—including *soundness*, *ambiguity*, *factuality*, and *relevance*—complemented by free-text notes. These signals guided revisions of stems and distractors, removal of problematic items, and documentation of corner cases. The dataset was originally constructed with four answer options (A–D), each containing exactly one correct response. Later, a fifth option (E), labeled as "None of the above" (and its counterparts in Italian and Turkish datasets) was appended to all questions. This option was intentionally designed to be incorrect in every case and serves as a controlled distractor. Its inclusion enabled the evaluation of the models' susceptibility to uncertainty and their tendency to select generalized options.

Table 4
Distribution of question categories in PFB dataset

Category	Financial Texts	Financial Papers	Financial Reports
Number	327	339	335
Percentage	32.6	33.9	33.5

The benchmark is distributed in three languages with instance-level alignment. The Italian and Turkish versions were generated via automatic translation (DeepL Enterprise) and subsequently manually post-edited to correct domain terminology, preserve the semantics of both stems and answer options, and avoid language-specific artifacts that could inadvertently cue the correct answer. For the

shared-task workflow, PFB is released with a public development split and a held-out test split used for final evaluation; all splits preserve the `custom_id` alignment across languages. The dataset has two splits: 500 questions have been published for examples and 1001 questions have been used for testing. The splits are the same across all three languages.

In Table 2, we show the distribution of the correct choices, in Table 3 the distribution of difficulty levels and in Table 4 the distribution of question categories.

3.2. Evaluation Measures

Systems are evaluated using *accuracy*, i.e., the fraction of instances for which the selected option matches the gold label to be chosen among the available options. Given a test set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{A, B, C, D, E\}$ is the correct option and $\hat{y}_i$ is the system prediction, the primary score is:

$$\mathrm{Acc}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = y_i\right]. \tag{1}$$

In addition to overall accuracy on the full test set, we also report accuracy on predefined *difficulty-based subsets* derived from the item difficulty labels (e.g., easy vs. hard), to support diagnostic comparisons across systems.

4. Results

4.1. Baselines

As baselines, we use a similarity-based approach based on an encoder model and two small-sized LLMs:

- **Similarity** (0.5B): we select the answer as the one most similar to the question. The similarity score is computed using a Sentence-BERT model [15]: *distiluse-base-multilingual-cased-v1*.[1]
- **Qwen** (1.7B) [16]: we use *Qwen3-1.7B*[2] (instruction tuned).
- **Llama** (3B) [17]: we use *Llama-3.2-3B-Instruct*.[3]

For the LLMs, our prompt contains the question, the list of possible answer, and their labels in the following format:

```
Question: [question]
A: [choiceA]
B: [choiceB]
C: [choiceC]
D: [choiceD]
E: [choiceE]
Answer:
```

All the baselines have been pre-trained, to some extent, in the three languages considered in this task: Italian, English, Turkish.

4.2. Systems Overview

We received two submissions. The **UNITOR** system by Borazio et al. [18] proposes an agentic architecture without any fine-tuning or domain adaptation that follows a three-stage orchestration pipeline. The first stage is semantic routing, used to distinguish between three different reasoning strategies:

[1] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1
[2] https://huggingface.co/Qwen/Qwen3-1.7B
[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

quantitative analysis, Boolean fact-checking, and knowledge-based research reasoning. Following routing, a specialized reasoning phase instantiates multiple parallel inference threads aimed at exploring parallel reasoning paths. Finally, an aggregation stage uses majority voting to select the most stable answer under sampling, following [19]. An iterative refinement step may fire if the consensus is weak. During refinement, the reasoning modules are re-invoked with a constrained version of the original prompt, in which low-support options are explicitly masked. The authors evaluate this architecture on three open-weight models: LLaMA 3.1 (8B) as a lightweight baseline, GPT-OSS-20B (21B) as the primary reference model, and DeepSeek v3.1 (671B) as a large upper bound. The top performer in the validation set used for the official submission is DeepSeek 3.1 for English, and GPT-OSS-20B for Italian and Turkish, achieving about 0.88 average accuracy across all languages.

The **AMSN** system by Mohammadabad and Nazarmohsenifakori [20] is the result of analysis of three different approaches: fine-tuning transformer-based models for question-answering, LLM prompting, and a hybrid approach. The first approach leverages mDeBERTa-v3. The second one is used to evaluate GPT-4o and GPT-5 with several prompting strategies. The hybrid approach uses a specialized trained model to detect and correct GPT-4o's mistakes. The top performer in the validation set used for the official submission is GPT-5, which achieves nearly 0.90 average accuracy across all languages.

4.3. Experimental Results

Table 5
Results on the two subtasks

	Subtask 1	**Subtask 2**
AMSN [20]	0.91	0.90
UNITOR [18]	0.88	0.88
Llama [17]	0.37	0.37
Qwen [16]	0.33	0.29
Similarity [15]	0.21	0.20

Table 6
Breakdown of the results according to languages and difficulty of the instances (all, hard, medium, and easy)

Language	**IT**				**EN**				**TR**				**ALL**			
Difficulty	**A**	**H**	**M**	**E**	**A**	**H**	**M**	**E**	**A**	**H**	**M**	**E**	**A**	**H**	**M**	**E**
AMSN [20]	**0.91**	**0.92**	**0.92**	0.86	**0.89**	0.88	**0.90**	0.84	**0.88**	**0.88**	**0.89**	0.88	**0.90**	**0.89**	**0.90**	0.86
UNITOR [18]	0.88	0.84	0.88	**0.89**	**0.89**	**0.91**	**0.90**	**0.88**	**0.88**	**0.88**	0.88	**0.89**	0.88	0.88	0.89	**0.89**
Llama [17]	0.37	0.41	0.36	0.38	0.43	0.48	0.42	0.46	0.31	0.33	0.31	0.29	0.37	0.41	0.36	0.38
Qwen [16]	0.33	0.32	0.32	0.36	0.26	0.25	0.26	0.29	0.29	0.34	0.30	0.25	0.29	0.30	0.29	0.30
Similarity [15]	0.21	0.18	0.20	0.19	0.21	0.20	0.21	0.21	0.18	0.17	0.19	0.16	0.20	0.18	0.20	0.19

Table 7
Breakdown of the results according to languages and category: Financial Texts (**T**), Financial Papers (**P**), Financial Reports (**R**)

Language	**IT**			**EN**			**TR**			**ALL**		
Difficulty	**T**	**P**	**R**	**T**	**P**	**R**	**T**	**P**	**R**	**T**	**P**	**R**
AMSN [20]	0.89	0.93	0.92	0.88	0.92	0.87	0.88	0.90	0.86	0.88	**0.92**	0.89
UNITOR [18]	0.88	0.86	0.90	0.88	0.88	0.92	0.87	0.86	0.91	0.88	0.87	**0.91**
Llama [17]	0.40	0.43	0.27	0.45	0.48	0.37	0.32	0.35	0.25	0.39	0.42	0.29
Qwen [16]	0.33	0.36	0.29	0.24	0.34	0.22	0.31	0.37	0.19	0.29	0.36	0.23
Similarity [15]	0.20	0.19	0.23	0.20	0.22	0.22	0.17	0.22	0.15	0.19	0.21	0.20

AMSN is the best-performing model on both subtasks, with an accuracy of 0.91 and 0.90, as shown in Table 5. UNITOR performs slightly worse, with a score of 0.88 in both tasks. The baselines do not reach a satisfactory accuracy, with the best one, Llama, obtaining a score of 0.37. The Similarity approach yields the worst result, with an accuracy of about 0.20, similar to random choice.

Table 6 shows detailed results. The performance of the participant systems is comparable across languages, and slightly different across difficulty levels. For instance, on Italian data, almost counterintuitively, AMSN performs better on hard questions (0.92) than on easy questions (0.86), whereas UNITOR performs better on the easy questions (0.89) than on the hard ones (0.84). On Turkish data, they both obtain a similar accuracy score across difficulty levels. Similar patterns are also observed in the baselines.

Analyzing the distribution of the answers across the 5 labels, we observe that models from participants do not exhibit any specific bias towards any label and the distribution of their answers follows the distribution of the gold standard.

Table 7 shows results across language and categories. The performance is comparable across systems, however we observe a slight tendency of questions from financial texts (T) to be harder for participants' systems and baselines.

In general, the difference between the results of the participants is small across all subtasks, languages and categories. The size of the underlying models is likely to have a significant impact on the performance, given the gap between approaches that use models above 20B and the baselines, based on models with 3B parameters or less. A remarkable result is that 20B models are competitive against much larger models.

5. Conclusions

The *Prometeia Financial Benchmark* (PFB) shared task was designed to benchmark language models on finance-domain multiple-choice question answering (MCQA) on three languages. We received two full submissions from teams who explored enconder- and decoder-based systems of varying complexity, reaching an accuracy of about 90%. A comparison with the performance obtained by our baselines highlights that PFB is challenging for models smaller than 3B, and that 20B models may be adequate for the task. As future work, we want to investigate more thoroughly the relationship between the models' efficiency and their accuracy. Moreover, we would like to explore the possible benefit of exploiting the difficulty label, which is a feature that the participants did not have, but may be used in ensemble approaches.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT 5.2 in order to: Paraphrase and reword, and Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. Dong, F. Wu, K. Zhang, Y. Dai, S. Zhang, W. Ye, S. Chen, Z.-Q. Cheng, Large language model agents in finance: A survey bridging research, practice, and real-world deployment, in: Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 17889–17907.

[2] A. T. Khan, S. Li, X. Cao, Bridging finance and AI: a comprehensive survey of large language models in financial system, Digital Finance 7 (2025) 679–701.

[3] H. Kang, X.-Y. Liu, Deficiency of large language models in finance: An empirical examination of hallucination, arXiv preprint arXiv:2311.15548 (2023). URL: https://arxiv.org/abs/2311.15548.

[4] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, EVALITA 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for

italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[5] Y. Guo, Z. Xu, Y. Yang, Is ChatGPT a financial expert? evaluating language models on financial natural language processing, arXiv preprint arXiv:2310.12664 (2023). URL: https://arxiv.org/abs/2310.12664.

[6] P. Winder, C. Hildebrand, J. Hartmann, Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors, PloS one 20 (2025) e0325459.

[7] R. Jørgensen, O. Brandt, M. Hartmann, X. Dai, C. Igel, D. Elliott, Multifin: A dataset for multilingual financial NLP, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 894–909.

[8] X. Peng, L. Qian, Y. Wang, R. Xiang, Y. He, Y. Ren, M. Jiang, J. Zhao, H. He, Y. Han, et al., MultiFinBen: A multilingual, multimodal, and difficulty-aware benchmark for financial LLM evaluation, arXiv preprint arXiv:2506.14028 (2025).

[9] S. Xue, X. Li, F. Zhou, Q. Dai, Z. Chu, H. Mei, Famma: A benchmark for financial domain multilingual multimodal question answering, arXiv preprint arXiv:2410.04526 (2024).

[10] X. Zhang, R. Xiang, C. Yuan, D. Feng, W. Han, A. Lopez-Lira, X.-Y. Liu, M. Qiu, S. Ananiadou, M. Peng, et al., 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, 2024, pp. 6236–6246.

[11] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, et al., Finben: A holistic financial benchmark for large language models, Advances in Neural Information Processing Systems 37 (2024) 95716–95743.

[12] G. Matlin, M. Okamoto, H. Pardawala, Y. Yang, S. Chava, Financial language model evaluation (FLaME), in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 22633–22679.

[13] N. Tatarinov, S. Sukhani, A. Shah, S. Chava, Language modeling for the future of finance: A quantitative survey into metrics, tasks, and data opportunities, arXiv preprint arXiv:2504.07274 (2025).

[14] Prometeia, Prometeia financial benchmark: Benchmarking language models in the financial domain, Prometeia Insights, 2025. URL: https://www.prometeia.com/it/about-us/insights/article/prometeia-financial-benchmark-benchmarking-language-models-in-the-financial-domain-26892170.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 3980–3990.

[16] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[17] A. Grattafiori, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[18] F. Borazio, S. A. Mousavian Anaraki, S. I. Rai, D. Croce, R. Basili, UniTor at PFB: Constrained agentic prompting and self-consistency for financial Multi-Choice QA, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[19] K. Tian, E. Mitchell, H. Yao, C. D. Manning, C. Finn, Fine-tuning language models for factuality, in: ICLR, OpenReview.net, 2024.

[20] A. S. Mohammadabad, M. Nazarmohsenifakori, Multilingual economics multiple choice question answering: A comparative study of transformer models, large language models, and hybrid correction systems, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.