AMSN at PFB: Overview of the Automated Multilingual Economics Question Answering Task

Ali Saleh Mohammadabad[1], Mehregan Nazarmohsenifakori[1]

[1]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Abstract

This report presents a comprehensive investigation into automated question answering for economics-based multiple choice questions across three languages (English, Italian, and Turkish). We explore three distinct methodological approaches: (1) fine-tuned transformer models, (2) prompt-engineered large language model agents using GPT-4o and GPT-5, and (3) a hybrid mistake detection and correction system. Our results demonstrate that while standalone transformer models achieve approximately 50% accuracy, GPT-4o with chain-of-thought prompting reaches 74% accuracy on English and Italian datasets. A hybrid system combining GPT-4o predictions with transformer-based error correction improves performance to approximately 78%. The newer GPT-5 model achieves superior performance without requiring mistake correction, ultimately serving as our final solution with test set accuracies of 89.11% (English), 91.21% (Italian), and 88.31% (Turkish).

Keywords

Question Answering, Economics, Large Language Models, Chain-of-Thought Prompting, Multilingual NLP, Transformer Models, Error Correction

1. Introduction

The task of automated question answering in specialized domains such as economics presents unique challenges, as it requires not only broad domain knowledge spanning microeconomics, macroeconomics, finance, and econometrics, but also complex multi-step reasoning and robust multilingual understanding. These requirements make economics QA a demanding benchmark for evaluating modern NLP systems.

This report documents our participation in the AMSN shared task at EVALITA 2026, where we systematically explored multiple approaches to address these challenges on a multilingual economics multiple-choice question dataset covering English, Italian, and Turkish. Our investigation progressed through three primary methodologies. We first examined fine-tuned transformer-based models trained directly on the task data, which served as an important baseline for understanding the limitations of supervised fine-tuning with limited domain-specific data. We then evaluated large language models (LLMs) accessed via API, specifically GPT-4o and GPT-5, using carefully designed chain-of-thought prompting strategies that leverage their extensive pre-trained economic knowledge. Finally, we developed a hybrid system that combines LLM predictions with a transformer-based mistake detection and correction module, aiming to capture the complementary strengths of both paradigms.

Through this progression, we demonstrate that advances in base model capability, as exemplified by GPT-5, can outperform both fine-tuned specialised models and hybrid correction approaches, achieving test accuracies of 89.11% (English), 91.21% (Italian), and 88.31% (Turkish).

2. System Description

2.1. Dataset Characteristics

The dataset provided for the AMSN shared task consists of multiple-choice economics questions spanning three languages: English (EN), Italian (IT), and Turkish (TR), each with 1,001 test questions. In addition to the test set, a validation split was provided for system development and hyperparameter

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

ali.saleh4@studio.unibo.it (A. S. Mohammadabad); mehrega.nazarmohseni@studio.unibo.it (M. Nazarmohsenifakori)

selection. Each question presents a stem followed by five answer options labelled A through E, where option E typically represents an exclusionary choice such as "None of the above." The questions cover a wide range of economics topics, including micro- and macroeconomics, finance, and quantitative methods, requiring models to combine domain-specific knowledge with logical reasoning.

2.2. Evaluation Metrics

Performance is measured using classification accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Questions}} \times 100\%$$

2.3. Approach 1: Transformer Model Training

2.3.1. Architecture and Implementation

We investigated transformer-based models for direct question answering, focusing primarily on multilingual architectures suitable for cross-lingual transfer learning.

Models Evaluated

- **mDeBERTa-v3** (Multilingual DeBERTa) [1] — Primary model
- Additional transformer variants for comparison

2.3.2. Results and Limitations

Table 1 presents the validation accuracy achieved by transformer models.

Table 1
Transformer Model Performance

Model	Validation Accuracy
mDeBERTa-v3	∼50%

The transformer models demonstrated several limitations:

1. **Complexity of Economic Reasoning:** The models struggled to capture the nuanced reasoning required for economics questions, which often involve multi-step logical deduction and domain-specific knowledge.
2. **Limited Training Data:** The available training data was insufficient for the model to learn robust patterns for all question types and economic concepts.
3. **Domain Knowledge Gap:** Fine-tuning alone proved inadequate for encoding the breadth of economic knowledge required.

2.4. Approach 2: Large Language Model Agents with Prompt Engineering

2.4.1. Motivation

Given the limitations of fine-tuned transformer models, we explored leveraging the pre-trained knowledge embedded in large language models through carefully designed prompting strategies. We evaluated two models accessed via the OpenAI API: GPT-4o [2] (snapshot `gpt-4o-2024-08-06`) and the newer GPT-5 (exact snapshot version to be specified by the authors; see Section 2.4.3).

2.4.2. Prompting Strategies

We systematically evaluated five distinct prompting strategies on a sample of 50 validation questions. Table 2 summarizes the results.

Table 2
Prompting Strategy Comparison (50 validation questions)

Strategy	Score	Accuracy	Rank
Chain of Thought	45/50	**90.0%**	Best
Expert Reasoning	45/50	**90.0%**	Best
Careful Analysis	44/50	88.0%	
Baseline	43/50	86.0%	
Step-by-Step	43/50	86.0%	

2.4.3. Optimal Prompt Design

The chain-of-thought prompting strategy [3] emerged as the top performer. The complete prompt implementation consists of a system prompt that establishes the model's expertise and reasoning framework, combined with a user prompt containing the formatted question.

System Prompt. The system prompt establishes the model's role and reasoning methodology:

```
You are an expert economist with deep knowledge in microeconomics,
macroeconomics, finance, and econometrics.

For this multiple choice question:
1. First, carefully analyze each statement or option
2. Apply your economic knowledge to evaluate correctness
3. Consider edge cases and common misconceptions
4. After your analysis, provide your final answer

IMPORTANT: End your response with "ANSWER: X" where X is the letter (A, B,
C, D, or E).
```

Question Formatting. Each question is formatted as a user prompt with the question text followed by lettered options (A through E).

API Configuration. The prompts were executed with temperature 0.1 for consistent responses and a maximum of 1500 output tokens for detailed reasoning. We used the OpenAI Chat Completions API with model identifiers `gpt-4o-2024-08-06` (GPT-4o) and `gpt-5-YYYY-MM-DD` (GPT-5; replace with the exact snapshot version used).[1]

Answer Extraction. Responses were parsed using multiple extraction patterns including "ANSWER: X", "The correct answer is X", and fallback to the last letter A-E in the response, achieving near-perfect parsing accuracy.

2.4.4. Full Dataset Evaluation

Applying the chain-of-thought strategy to the complete dataset across all three languages yielded the results shown in Table 3.

2.4.5. Key Observations

Several important patterns emerged from the full-dataset evaluation. First, the chain-of-thought approach does more than improve accuracy: it exposes the model's step-by-step reasoning, providing interpretability and enabling downstream error analysis by inspecting the rationale behind each prediction. Second, the same prompting strategy generalised effectively across all three languages, achieving

[1] Readers wishing to reproduce results should note that OpenAI API model behaviour may vary across snapshot versions. The exact version strings used in our experiments should be substituted here by the authors.

Table 3
LLM Performance Across Languages (Validation Set)

Model	English (EN)	Italian (IT)	Turkish (TR)
GPT-4o	74%	74%	65%
GPT-5	**82%**	**81%**	**73%**

consistent performance on English and Italian. Third, a Turkish performance gap is evident, with accuracies 9–10 percentage points lower than English and Italian for both models, suggesting that coverage of Turkish economics terminology and concepts is comparatively weaker in the models' pre-training data. Finally, GPT-5 demonstrates substantial generational improvements over GPT-4o, with gains of approximately 8 percentage points across all three languages, underscoring the impact of advances in base model capability on downstream task performance.

2.5. Approach 3: Hybrid Mistake Detection and Correction

2.5.1. Rationale

While GPT-4o achieved respectable 70-75% accuracy, we investigated whether a specialized model trained to detect and correct GPT-4o's mistakes could push performance beyond this baseline.

2.5.2. System Architecture

The hybrid system, implemented in `Finance_Concepts_Transformer_Updated (1).ipynb`, consists of two components:

1. **Mistake Detector:** A transformer model trained to identify when the LLM's prediction is likely incorrect
2. **Answer Corrector:** A model that predicts the correct answer when mistakes are detected

The system uses the following input features:

- Question text
- All answer options
- GPT-4o's prediction
- **GPT-4o's reasoning** (extracted from chain-of-thought output)

2.5.3. Training Data

The hybrid components were trained using GPT-4o predictions on the validation set, for which ground truth labels were available. For the mistake detector, binary labels were derived by comparing GPT-4o's predicted answer against the ground truth: predictions matching the correct answer were labelled as correct, while mismatches were labelled as mistakes. For the answer corrector, the ground truth answer label for each question was used as the target class in a five-way multi-class classification setting.

2.5.4. Results

The hybrid system applied to GPT-4o predictions achieved approximately **77-79% accuracy**, representing a 3-5 percentage point improvement over standalone GPT-4o predictions on English and Italian datasets.

We also evaluated whether the mistake detection system could improve GPT-5 predictions. However, the correction mechanism did not yield improvements, suggesting that GPT-5's reasoning capabilities already capture the patterns the mistake detector was designed to identify.

3. Results

3.1. Validation Performance

Table 4 presents a comprehensive comparison of all approaches on the validation set.

Table 4
Performance Comparison Across Approaches (Validation Set)

Approach	EN	IT	TR	Avg
Transformer (mDeBERTa-v3)	50%	50%	50%	50.0%
GPT-4o (Chain-of-Thought)	74%	74%	65%	71.0%
GPT-4o + Mistake Correction	78%	77%	—	77.5%
GPT-5 (Chain-of-Thought)	**82%**	**81%**	**73%**	**78.7%**
GPT-5 + Correction (no improvement)	82%	81%	73%	78.7%

3.2. Final Test Set Results

After finalizing our approach with GPT-5, we evaluated the model on the held-out test set of 1,001 questions per language. Table 5 presents the final test accuracies.

Table 5
Final Test Set Performance (1,001 questions per language)

Language	Correct	Incorrect	Accuracy	Improvement
English (EN)	892	109	**89.11%**	+7.11 pp
Italian (IT)	913	88	**91.21%**	+10.21 pp
Turkish (TR)	884	117	**88.31%**	+15.31 pp
Average	**896.3**	**104.7**	**89.54%**	**+10.84 pp**

The test set results demonstrate exceptional performance, with GPT-5 achieving nearly 90% average accuracy across all three languages. Notably, the test performance significantly exceeds validation performance, with improvements ranging from 7.11 to 15.31 percentage points.

3.3. Answer Distribution Analysis

Table 6 presents the answer distribution across all three languages.

Table 6
Predicted Answer Distribution by Language (1,001 questions each)

Language	A	B	C	D	E	Total
English	26.9%	27.6%	28.1%	16.2%	1.3%	100%
Italian	27.5%	28.2%	27.1%	14.9%	2.4%	100%
Turkish	27.3%	26.2%	29.7%	15.8%	1.1%	100%
Average	**27.2%**	**27.3%**	**28.3%**	**15.6%**	**1.6%**	**100%**

Key observations from the distribution analysis:

- Options A, B, and C account for approximately 82-83% of all predictions across languages
- Option D represents only 15-16% of predictions
- Option E ("None of the above") appears in only 1-2% of predictions
- The remarkable similarity in distribution patterns across languages indicates robust multilingual performance

4. Discussion

4.1. Comparative Analysis

Our investigation reveals a clear performance hierarchy:

1. **GPT-5 Test Performance (89.54% avg):** Achieves exceptional performance on the held-out test set, significantly exceeding validation performance (78.7%).
2. **GPT-5 Validation (78.7% avg):** Strong performance during development, selected as final solution.
3. **GPT-4o + Mistake Correction (77.5% avg):** Demonstrates that targeted error correction can improve baseline LLM performance.
4. **GPT-4o with Chain-of-Thought (71.0% avg):** Leverages pre-trained knowledge effectively, outperforming fine-tuned models by 21 percentage points.
5. **Fine-tuned Transformers (50%):** Limited by training data and domain knowledge constraints.

4.2. Key Success Factors

The results across our three approaches point to several factors that drove performance. The most decisive factor was the breadth of pre-trained knowledge embedded in large language models: GPT-4o and GPT-5 have been exposed to extensive economics literature during pre-training, providing a foundation for domain reasoning that supervised fine-tuning of smaller models on limited task data simply cannot replicate.

Chain-of-thought prompting contributed not only to accuracy but also to reasoning transparency, making the system's decision process interpretable and facilitating downstream error analysis. For the GPT-4o baseline specifically, the hybrid correction system proved beneficial by exploiting complementary strengths: the LLM contributes broad knowledge and reasoning, while the specialist transformer contributes targeted pattern recognition for detecting and correcting systematic errors. Finally, the results illustrate the impact of model generation: GPT-5's improvements over GPT-4o were large enough to make the separate correction mechanism redundant, suggesting that architectural and training advances in frontier models can subsume the role of specialised post-processing.

4.3. Analysis of Results

The results reveal several noteworthy patterns. For GPT-4o, the mistake detector successfully identifies systematic errors in the model's predictions, and the subsequent correction step yields measurable accuracy gains of 3–5 percentage points on English and Italian. This confirms that GPT-4o makes identifiable, recurring mistakes that a specialised classifier can learn to detect.

In contrast, the same correction mechanism applied to GPT-5 predictions produced no improvement. This outcome indicates that GPT-5's stronger reasoning capabilities already address the failure modes that the mistake detector was designed to capture, making the additional correction layer redundant. Consequently, GPT-5 with chain-of-thought prompting was selected as the final submission system, without any post-processing.

The most striking pattern in the results is the gap between validation and test performance, discussed in detail in Section 4.4. Rather than reflecting overfitting, this improvement is consistent across all three languages and all metrics, suggesting genuine differences in question characteristics between the two splits. Given GPT-5's superior standalone performance, it was the natural choice for the final submission.

4.4. Validation-to-Test Performance Gap

A notable feature of our results is the consistent improvement from validation to test performance, ranging from $+7.11$ percentage points (English) to $+15.31$ percentage points (Turkish). While the

magnitude of this gain is striking, we do not believe it reflects data leakage or overfitting, for the following reasons.

First, GPT-5 was not fine-tuned on the validation set in any way; the only information derived from validation data was the selection of the chain-of-thought prompting strategy, which was evaluated on a small 50-question sample. The model itself was called via API in a zero-shot inference mode on both splits. Second, the improvement is consistent across all three languages, including Turkish, which argues against a language-specific anomaly.

The most plausible explanation is a distributional difference between the validation and test splits. In particular, the test questions may draw more heavily on canonical economic theory and textbook-level definitions that are well represented in GPT-5's pre-training corpus, whereas the validation questions may include a higher proportion of edge cases, ambiguous phrasing, or niche topics. A secondary contributing factor could be question phrasing: test questions may be more unambiguously worded, making chain-of-thought reasoning more reliable.

We acknowledge that a full investigation of this gap would require access to ground-truth labels for the validation set alongside a systematic analysis of question-level difficulty, topic distribution, and linguistic complexity across the two splits. We leave this as an important direction for future work, and encourage the task organisers to share validation gold labels to enable the research community to conduct such analyses.

5. Conclusion

This work demonstrates a systematic exploration of approaches to multilingual economics question answering, progressing from fine-tuned transformers (50%) to GPT-4o (71%) to GPT-4o with mistake correction (77.5%) to GPT-5 validation (78.7%), and ultimately achieving exceptional test performance (89.54% average).

Three key findings emerge from this work. First, advances in base model capability can subsume the benefits of specialised post-processing: GPT-5 outperformed GPT-4o with hybrid correction, demonstrating that investing in stronger foundation models can be more productive than engineering correction layers on top of weaker ones. Second, transformer-based mistake correction is effective but model-dependent: the correction system improved GPT-4o by approximately 6.5 percentage points on average but yielded no gain for GPT-5, confirming that such systems are most valuable when the underlying model has identifiable, systematic weaknesses. Third, GPT-5 with chain-of-thought prompting achieved strong test accuracies of 89.11% on English, 91.21% on Italian, and 88.31% on Turkish (average 89.54%).

The exceptional cross-lingual performance (88-91%) validates the robustness of our GPT-5 approach across all three languages. Italian achieved the highest accuracy (91.21%), while Turkish showed the most significant improvement from validation to test (+15.31 pp), effectively closing the performance gap.

Future work should focus on understanding the validation-to-test performance gap, investigating whether similar performance can be achieved with cost-effective open-source alternatives, and exploring the remaining 10% error cases to identify systematic failure patterns.

6. Declaration on Generative AI

In the preparation of this work, the authors utilized the following generative AI tools:

Research Subject (Models Under Investigation):

- **GPT-4o and GPT-5** (OpenAI's large language models) were the primary subject of investigation for multilingual economics question answering. These models were evaluated through their APIs to generate predictions on economics multiple-choice questions across English, Italian, and Turkish languages using chain-of-thought prompting strategies.

- **mDeBERTa-v3** transformer model was employed for baseline comparisons and as part of the hybrid mistake detection and correction system implemented in `Finance_Concepts_Transformer_Updated (1).ipynb`.

Writing Assistance:

- **Claude (Anthropic)** was used to assist in drafting portions of this manuscript, including structuring sections, formatting LaTeX code, refining language clarity, and organizing the presentation of results and methodologies. The authors reviewed, edited, and verified all AI-generated content for accuracy and appropriateness.

The experimental design, hypothesis formulation, data analysis, interpretation of results, and scientific conclusions were conducted entirely by the human authors. All code implementation, data processing, model evaluation, and statistical analysis were performed by the authors without AI assistance.

Acknowledgments

We thank the organizers of the FinanceNLP shared task for providing the multilingual economics question answering dataset and evaluation framework.

References

[1] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: International Conference on Learning Representations (ICLR), 2023.

[2] OpenAI, Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems (NeurIPS), volume 35, 2022, pp. 24824–24837.