JFra-Team at PROFE 2025: A Multi-Agent Zero-Shot System for Spanish Language Proficiency Question Answering

Jesús M. Fraile-Hernández[1,*], Anselmo Peñas[1]

[1]UNED NLP & IR Group, Universidad Nacional de Educación a Distancia, Madrid, 28040, Spain

Abstract

This paper presents a multi-agent system designed to address the Multiple Choice subtask of the PROFE 2025 shared task, which involves selecting the correct answer to questions derived from Spanish language proficiency exams. The system operates in a zero-shot setting using the Gemini 2.5 Flash large language model. It comprises four specialised agent types working collaboratively to make informed and interpretable decisions. Evaluated on the official test set, the system achieved 92.90% accuracy, ranking among the top performers in the shared task.

Keywords

Multi-agent systems, Zero-shot learning, Large Language Models, Multiple-choice QA

1. Introduction

Question Answering (QA) has long been a central task in Natural Language Processing (NLP), aiming to develop systems capable of answering natural language questions based on contextual information. The advent of Large Language Models (LLMs), such as Gemini 2.5 [1] and GPT-4 [2], has dramatically advanced the state of the art, enabling zero-shot and few-shot capabilities that significantly reduce the need for task-specific data. These models have proven particularly valuable in multilingual and low-resource settings, where traditional supervised methods struggle.

Parallel to this, there is growing interest in multi-agent systems (MAS) for NLP, where multiple interacting agents with distinct roles collaborate to solve complex problems. When combined with LLMs, such systems can simulate diverse perspectives and provide a form of internal deliberation, leading to more robust and interpretable outputs [3].

The PROFE 2025 shared task [4], part of IberLEF 2025 [5], addresses the challenge of automatic Spanish language proficiency evaluation using real exam content from the Instituto Cervantes. It is structured into three subtasks: multiple-choice reading comprehension, matching, and gap-filling. In this work, we focus on Subtask 1, which requires selecting the correct answer from several options based on a short passage and question.

This paper presents a multi-agent framework built around zero-shot LLM to address subtask 1. Section 3 introduces the system architecture and agent roles. Section 4 presents the experimental results and comparative analysis. Section 5 discusses implications and limitations. Finally, Section 6 concludes the paper by summarising the main findings.

2. Subtask description

The Multiple Choice subtask within the PROFE 2025 framework is designed to evaluate the capabilities of automated systems in comprehending and interpreting Spanish texts. Each exercise presents a passage accompanied by a series of multiple-choice questions, each offering several answer options with only one correct choice. The primary objective for participating systems is to accurately identify

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

jfraile@lsi.uned.es (J. M. Fraile-Hernández); anselmo@lsi.uned.es (A. Peñas)

0009-0001-5474-4844 (J. M. Fraile-Hernández); 0000-0002-7867-0149 (A. Peñas)

the correct answer for each question, thereby demonstrating proficiency in language understanding and reasoning.

The dataset employed for this subtask is the IC-UNED-RC-ES corpus, derived from authentic examinations administered by the Instituto Cervantes. These examinations, meticulously crafted by language assessment experts, aim to measure Spanish language proficiency across various levels, ranging from A1 to C2 as per the Common European Framework of Reference for Languages (CEFR).

2.1. Dataset Description

The dataset developed for the Multiple Choice subtask comprises questions that span a wide range of proficiency levels, from A1 to C2. Each question is accompanied by a set of candidate answers, among which only one is correct. The core objective for participating systems is to select the correct response from the options provided, thereby demonstrating their capability to interpret and reason over linguistic input.

A distinctive characteristic of the dataset is the variable number of answer choices associated with each question. Specifically, the dataset includes questions with two, three, or four options, introducing an additional layer of complexity. The organisers have released a development set consisting of a very limited sample of ten questions, designed primarily for testing and debugging purposes. A detailed breakdown of the number of questions and corresponding answer options across each CEFR level, for both the development and testing sets, is presented in Table 1. This level distribution highlights not only the diversity of linguistic challenges posed by the task, but also the unbalanced nature of the dataset in terms of difficulty levels, which must be accounted for during model evaluation.

Level	Dev Questions	Dev Options	Test Questions	Test Options
A1	0	0	72	288
A1E	0	0	48	144
A2	0	0	488	1464
A2B1E	0	0	24	72
B1	10	30	488	1394
B1E	0	0	80	300
B2	0	0	306	852
C1	0	0	36	108
C2	0	0	162	486
Total	**10**	**30**	**1704**	**5108**

Table 1
Distribution of questions and answer options across CEFR levels in the development and testing sets

2.2. Evaluation

The primary evaluation metric employed for this subtask is *accuracy*, defined as the proportion of questions for which the system selects the correct answer among the provided alternatives. This metric is both intuitive and interpretable, facilitating benchmarking of system performance. Moreover, its use aligns with established practices in human language assessment, thereby enabling a direct and meaningful comparison between the results obtained by automatic systems and those achieved by human examinees under comparable conditions.

3. Methodology

Given the limited availability of development data, we adopt a zero-shot approach leveraging a Large Language Model (LLM) without incorporating any additional fine-tuning or supervised adaptation. To

mitigate the challenges posed by low-resource evaluation and enhance answer reliability, we propose a multi-agent architecture specifically designed for the multiple-choice subtask.

The proposed system comprises four distinct groups of agents, each designed to fulfil a specialised role in the decision-making pipeline: (i) *Responder Agent*, which select the most appropriate answer from the available options; (ii) *Blind Responder Agent*, which generate a free-form answer without access to the predefined answer options. Their response is subsequently matched to one of the candidates using semantic similarity metrics, allowing us to evaluate alignment between unconstrained reasoning and constrained choices; (iii) *Evaluator Agents*, which receive the options not selected by the Responder Agent and independently assess whether these alternatives could be correct and (iv) a *Moderator Agent*, which acts as an adjudicator, resolving disagreements across agents and consolidating their opinions into a final, consensus-based answer.

This architecture is illustrated in Figure 1. All interactions within the system are conducted in Spanish, aligning with the linguistic nature of the task.

For all agents, we employ `Gemini-2.5-Flash` [1] as the underlying LLM, selected for its favourable trade-off between performance and efficiency in zero-shot multilingual reasoning tasks. Prompt engineering was conducted manually to ensure that each agent's role and expected behaviour were explicitly defined, enabling coherent and complementary interactions within the agent ensemble.

All implementation details, including the source code and the complete set of prompts used in the experiments, are publicly available in our GitHub repository.[1]

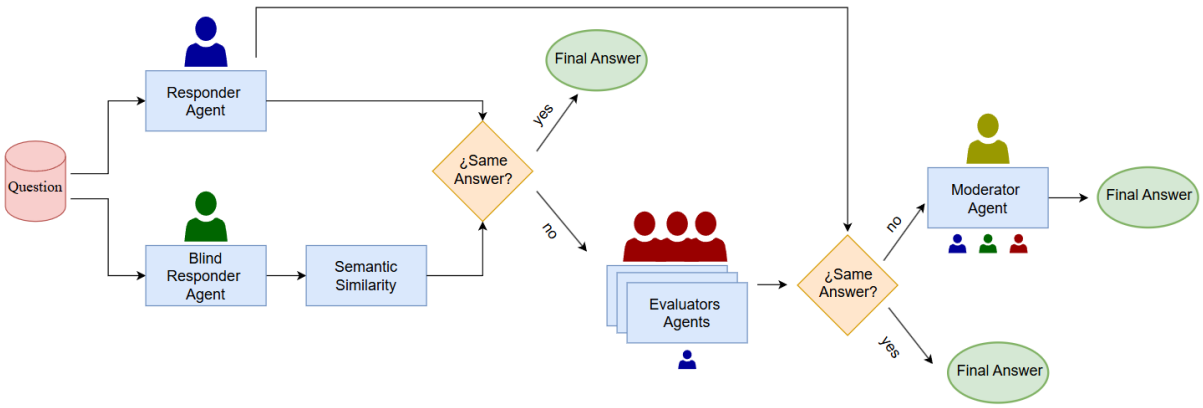

Figure 1: Overview of the multi-agent architecture employed in our system.

3.1. Responder Agent

The *Responder Agent* is tasked with selecting the most appropriate answer from the set of provided options, given a multiple-choice question and its associated context. This agent not only chooses an option but is also required to generate a textual justification supporting its decision. The rationale provides interpretability, allowing downstream agents and human evaluators to better understand the decision-making process.

This agent is based on the simulation of typical test-taking behaviour that reflects the same context and set of choices faced by human test takers.

3.2. Blind Responder Agent

The *Blind Responder Agent* addresses the same question as the Responder Agent but without access to the predefined answer choices. It generates an open-form answer based solely on the question and context. To reconcile this unconstrained response with the format of a multiple-choice task, a semantic

[1] https://github.com/JesusFraile/Profe2025-JFraTeam

similarity match is performed between the generated answer and each of the available options. This matching process is carried out using the `paraphrase-multilingual-MiniLM-L12-v2` [6] model, a lightweight and multilingual sentence embedding model known for its efficiency and robustness in semantic similarity tasks.

The reason for introducing this agent is because of its ability to emulate an unbiased reasoning process, uninfluenced by distracting choices. This setup allows us to assess whether the correct answer can be naturally inferred from the question-context pair, and to compare this inference with the restricted scenario of multiple choice selection.

3.3. Evaluator Agents

For each multiple-choice question, the system instantiates a number of *Evaluator Agents* equal to the number of available answer options minus one. Each Evaluator Agent is assigned to independently assess one of the options not selected by the Responder Agent. The input provided to each Evaluator consists on the contextual passage, the question, and the specific answer option to be evaluated. Based on this information, the Evaluator determines whether the assigned option could plausibly be correct and generates a textual justification for its judgement.

This design promotes impartial evaluation by avoiding cross-contamination of reasoning between agents. By not exposing Evaluators to the choice of Responder Agent, each judgement remains independent. The inclusion of Evaluators helps to reinforce the critical reasoning capacity of the system, allowing it to reconsider discarded options and thus reducing the risk of overconfident or erroneous selections by the lead Response Agent.

3.4. Moderator Agent

The *Moderator Agent* acts as the final arbitrator, consolidating the outputs from all other agents to determine the definitive system answer. It reviews the Responder Agent's chosen answer and justification, the Blind Responder Agent's response, and the judgments and rationales from all Evaluator Agents. Based on this multi-source evidence, the Moderator selects the most appropriate answer, particularly in cases where disagreements or uncertainty arise among the other agents.

This design decision is inspired by the principle of deliberative consensus. The Moderator serves to synthesise divergent viewpoints into a coherent outcome, ensuring that the final response is not the result of a single heuristic or model decision.

3.5. System Pipeline

The multi-agent system follows a structured pipeline consisting of the following stages:

- **Initial Answer Generation.** The Responder and the Blind Responder Agent operate simultaneously. The Responder selects an answer from the available options and provides a justification. In parallel, the Blind Responder generates an open-form response without seeing the options. This free-form answer is subsequently aligned with the closest candidate using semantic similarity based on the `paraphrase-multilingual-MiniLM-L12-v2` model.
- **Agreement Check.** If both agents independently converge on the same answer (if the Blind Responder Agent's semantically closest option matches the Responder Agent's selected answer and the semantic similarity score exceeds a predefined threshold of 0.6) the system accepts this answer immediately. This mechanism leverages agreement between two independently reasoned responses to increase efficiency while minimising spurious matches due to chance or low semantic alignment.
- **Evaluator Stage.** If disagreement occurs, the system initiates the evaluation phase. A set of Evaluator Agents, one per unselected answer option, assess the acceptability of their assigned option. Each Evaluator receives only the context, the question, and the specific answer to evaluate,

System	Accuracy (%)	Gap to Best (%)
Responder Agent Only	**92.96**	**-2.58**
Blind Responder + Similarity Matching	55.87	-39.67
Full Multi-Agent System	92.90	-2.64
Ezotova	**95.54**	**0.00**
JFraTeam (Responder Agent)	92.96	-2.58
Ralucaginga	91.96	-3.58
Mvictoria	91.14	-4.40
Annamwinkler	32.51	-63.03
Ymlopez	32.22	-63.32

Table 2
Accuracy scores of our system configurations compared to the best submission from each team in the shared task leaderboard.

Resolution Stage	Percentage of Questions (%)
Agreement between Responder and Blind Responder	34.21
Evaluator Agents Activated	65.79
– Resolved Unanimously by Evaluators	49.18
– Escalated to Moderator Agent	16.61

Table 3
Distribution of questions by resolution stage in the multi-agent pipeline.

and returns a binary judgement along with a justification. If all Evaluators unanimously reject their assigned options, the original selection by the Responder Agent is confirmed.

- **Mediation.** In cases where at least one Evaluator gives a valid alternative as valid, the system delegates the decision to the Moderator Agent. This agent aggregates all justifications and outputs, resolves conflicts, and synthesises a reasoned consensus.
- **Finalisation.** The Moderator Agent produces the final answer accompanied by an integrated explanation derived from the system's internal deliberation. This ensures that the output is both well-justified and interpretable.

This modular and deliberative pipeline is designed to adapt its depth of reasoning based on the complexity of each input, combining fast agreement-based shortcuts with deeper analysis when required.

4. Results

To assess the performance of our proposed system, we conducted a evaluation on the official test set of the shared task. The analysis covers three configurations: (i) the standalone *Responder Agent*; (ii) the *Blind Responder Agent* coupled with the semantic similarity matching mechanism; and (iii) the full *Multi-Agent System*.

Table 2 summarises the results of our system in all three configurations. For comparative purposes, we also report the results from the shared task leaderboard, selecting only the best-performing submission from each participating team. Additionally, we include the relative accuracy gap between each configuration of our system and the highest-scoring model overall.

In addition to performance metrics, we conducted a analysis of the decision flow within the multi-agent system. This aimed to quantify how frequently each component was activated during inference on the test set. Specifically, we measured the proportion of questions resolved directly by agreement between the Responder and Blind Responder Agents, those that required evaluation by the Evaluator Agents, and those that ultimately triggered the Moderator Agent due to evaluator disagreement. The results are summarised in Table 3.

5. Discussion

The results shown in Table 2 and Table 3 provide a detailed evaluation of the multi-agent architecture's performance and internal dynamics. Notably, both the *Responder Agent Only* and the *Full Multi-Agent System* achieved accuracies of 92.96% and 92.90%, respectively. These scores position them within 2.58 and 2.64 percentage points of the top-performing system on the leaderboard. The Responder Agent would rank second overall, while the full multi-agent system would take third place. This demonstrates the system's robustness and its capacity to approximate top-tier performance through collaborative decision-making.

Conversely, the *Blind Responder + Similarity Matching* configuration attained a markedly lower accuracy of 55.87%. The semantic similarity module employed in this configuration used the lightweight `paraphrase-multilingual-MiniLM-L12-v2` model. While computationally efficient, this lightweight approach struggled to deliver reliable selections. It would be worthwhile in future work to assess the impact of integrating a more powerful semantic similarity model.

Further insights into the system's internal behaviour were obtained by analysing the contribution frequency of each agent in the decision pipeline (Table 3). Direct agreement between the Responder and Blind Responder Agents resolved 34.21% of the questions, enabling swift decision-making using the most efficient configuration. The remaining 65.79% of cases required additional processing: 49.18% were resolved unanimously by the Evaluator Agents, while 16.61% necessitated final arbitration by the Moderator Agent. These results validate the design of the system's hierarchical resolution mechanism, in which more computationally intensive processes are activated only when necessary.

In terms of computational cost, the experiments involved a total of approximately 4.333 million processed tokens and 6,373 model requests across the three tasks. These operations correspond to a total of 1,704 questions. On average, this corresponds to approximately 2,543 tokens and 3.74 requests per question.

Despite the strong performance of the full multi-agent system, it is important to note that the *Responder Agent Only*, operating in a traditional zero-shot setting, still achieved the highest accuracy within our framework and matched the best-performing single submission on the leaderboard. This suggests that while the multi-agent pipeline is effective in resolving ambiguous or conflicting cases, it does not necessarily surpass the capabilities of a well-designed standalone agent when provided with sufficient context.

To better understand this discrepancy, further analysis is needed. In particular, it would have been valuable to examine how the accuracy of each system varies across different levels of question difficulty or linguistic complexity. Identifying patterns in the types of errors made by the full system versus the Responder Agent could offer insights into the strengths and limitations of multi-agent deliberation, and guide the development of more adaptive and context-aware coordination strategies.

6. Conclusion

This paper has introduced a multi-agent architecture designed to solve the task of multiple-choice question answering in Spanish, using zero-shot prompting of a large language model without any task-specific fine-tuning. Our approach leverages the collaborative reasoning of four distinct agent types each with clearly defined roles and contributions to collective decision-making.

Through comprehensive evaluation on the official test set of the PROFE 2025 shared task, we have demonstrated that the system achieves competitive performance, closely matching the best-performing submissions. In particular, the Responder Agent alone reached an accuracy of 92.96%, outperforming most leaderboard entries, while the full multi-agent system offered a comparable performance with added interpretability and robustness mechanisms.

Acknowledgments

This work was supported by DeepInfo (PID2021-127777OB-C22) project.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, Gemini: A Family of Highly Capable Multimodal Models, 2025. URL: http://arxiv.org/abs/2312.11805. doi:10.48550/arXiv.2312.11805, arXiv:2312.11805 [cs].

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774, arXiv:2303.08774 [cs].

[3] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, H. D. Nguyen, Multi-Agent Collaboration Mechanisms: A Survey of LLMs, 2025. URL: http://arxiv.org/abs/2501.06322. doi:10.48550/arXiv.2501.06322, arXiv:2501.06322 [cs].

[4] Á. Rodrigo, S. Moreno-Álvarez, A. P. García-Plaza, A. Peñas, R. Agerri, J. Fruns-Jiménez, I. Soria-Pastor, Overview of PROFE at IberLEF 2025: Language Proficiency Evaluation, Procesamiento del Lenguaje Natural 75 (2025).

[5] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[6] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL: http://arxiv.org/abs/1908.10084. doi:10.48550/arXiv.1908.10084, arXiv:1908.10084 [cs].