UC-UCO-CICESE_UT3-Plenitas Team - Exploring in the PROFE 2025: Language Proficiency Evaluation

Yoan Martínez-López[1,2,*], Mayte Guerra Saborit[3], Yanaima Jauriga[3], Ireimis de las Mercedes Leguen de Varona[3], Julio Madera[3], Ansel Rodríguez-González[4], Carlos de Castro Lozano[1,2], Jose Miguel Ramírez Uceda[1,2] and José Carlos Arévalo Fernández[2]

[1]*Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*CICESE-UT3, Nayarit, México*

Abstract

PROFE2025 is a challenge that tests automated systems on authentic Spanish reading-comprehension exams used by the Instituto Cervantes for human learners, without providing task-specific training data. This setup emphasizes transfer learning and large generative models. Participants can choose from three subtasks: Multiple Choice, Matching, and Fill-in-the-Gap, each evaluated using accuracy metrics that align with human assessment. Teams completing all subtasks receive an exam-level score based on the proportion of exams passed (60% accuracy), enabling direct comparison to human performance. The competition utilizes the IC-UNED-RC-ES corpus and Hugging Face's transformer model ecosystem. In internal evaluations, three Qwen-3 variants were tested: Qwen 3.0, a "think" prompting version, and a DeepSeek-distilled Qwen 2.5 7B. Both base and distilled models achieved 32.22% overall accuracy, while the "think" variant scored only 10.19%, indicating room for improvement in reasoning-focused prompts. Among 40 submissions, only two teams completed all three subtasks. UC-CICESE_UT3-Plenitas placed second with 32.22%, 4.89%, and 10.19%, respectively, underscoring both potential and current limitations of LLMs in this domain. PROFE2025 highlights that large language models can engage meaningfully with complex, human-like exam conditions, but significant improvements are possible through specialized adaptation, fine-tuning, and advanced prompting strategies, positioning it as a key benchmark for progress in automated language understanding..

Keywords

LLMs, exam, accuracy, prompts

1. Introduction

Advanced language comprehension is critical for capturing semantic subtleties and supporting logical inference in natural language processing. A persistent challenge in this domain is the creation of new evaluation resources, which typically demands considerable human effort. To address this, existing human reading comprehension datasets—such as RACE [1]chave been widely reused, enabling the benchmarking of automated systems against human-level performance. Nonetheless, the majority of these datasets are predominantly available in English and often include extensive training data that closely resembles the test sets. This overlap may constrain the ability to generalize results, particularly when evaluating the reasoning capabilities of language models. PROFE 2025 is a shared task designed to evaluate the performance of automated systems in Spanish reading comprehension under conditions

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
†These authors contributed equally.
yoan.martinez@plenitas.com (Y. Martínez-López); mayte.guerra@reduc.edu.cu (M. G. Saborit); yanaima.jauriga@reduc.edu.cu (Y. Jauriga); ireimis.leguen@reduc.edu.cu (I. d. l. M. L. d. Varona); julio.madera@reduc.edu.cu (J. Madera); ansel@cicese.edu.mx (A. Rodríguez-González); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda); josecarlos@plenitas.com (J. C. A. Fernández)
0000-0002-1950-567X (Y. Martínez-López); 0000-0002-9556-5869 (M. G. Saborit); 0009-0000-6891-0068 (Y. Jauriga); 0000-0002-1886-7644 (I. d. l. M. L. d. Varona); 0000-0001-5551-690X (J. Madera); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda); XXXX-XXXX-XXXX-XXXX (J. C. A. Fernández)

comparable to those used for assessing human learners [2, 3, 4]. The challenge is based on real exams developed by the Instituto Cervantes to evaluate Spanish language proficiency at various levels. Unlike traditional NLP tasks, PROFE 2025 does not provide task-specific training data [5]. Instead, it emphasizes the use of transfer learning and large generative language models (LLMs), encouraging participants to adapt their systems to novel scenarios using limited prior examples.

The competition is structured into three subtasks: Multiple Choice, Matching, and Fill-in-the-Gap. Each subtask presents different linguistic and reasoning challenges, and is evaluated using standard accuracy metrics [5]. For teams completing all three subtasks, an additional exam-level score is calculated based on the proportion of exams passed (defined as achieving at least 60% accuracy), facilitating direct comparison between machine and human performance. In this paper, we present our participation in the PROFE 2025 Language Proficiency Evaluation, describing our methodological approach, model choices, and results across the different subtasks.

2. Methodology

2.1. Subtasks

The PROFE 2025 challenge consists of three subtasks, each corresponding to a specific type of exercise. Participating teams may choose to engage in any combination of these subtasks [2, 5]. Each subtask includes multiple exercises of the same format:

- Multiple Choice Subtask: Each exercise provides a short reading passage followed by a set of multiple-choice questions, each with a single correct answer. The system must identify the correct response among the given alternatives.
- Matching Subtask: Each exercise presents two sets of texts. The system must determine the most appropriate match for each item in the first set from the items in the second set. Each text in the first set has exactly one correct match, although the set may contain additional, unmatched texts to increase task complexity.
- Fill-in-the-Gap Subtask: Each exercise features a passage with several missing fragments, which are presented in a disordered list. The system must accurately determine the correct position of each fragment. There is only one correct placement per gap, although the number of candidate fragments may exceed the number of gaps.

These subtasks encourage the development of diverse modeling strategies, particularly when leveraging large generative models that require tailored prompting methods.

2.2. Evaluation

The primary evaluation metric used across all subtasks is accuracy, defined as the proportion of correctly answered items. Systems are evaluated from two perspectives:

- Question-level evaluation: Accuracy is calculated by individually scoring each question, regardless of the exam structure.
- Exam-level evaluation: This applies only to systems that participate in all three subtasks. Accuracy is aggregated across all exercises within an exam. An exam is considered passed if the system achieves an accuracy score greater than or equal to 60%. The final global score reflects the proportion of exams passed.

Detailed evaluation by subtask includes:

- Multiple Choice: Accuracy is computed as the proportion of correct answers.
- Matching: Accuracy reflects the proportion of correct text matches.
- Fill-in-the-Gap: Accuracy measures the proportion of correctly placed fragments.

Accuracy is chosen as the evaluation metric because each task has a single correct answer per item, mirroring the conditions under which human candidates are assessed in the same exams. This enables a direct comparison between system and human performance.

2.3. Dataset

The challenge uses a subset of the IC-UNED-RC-ES dataset, which comprises reading comprehension exercises from the official exams administered by the Instituto Cervantes to evaluate Spanish language learners across different proficiency levels [6, 7]. The dataset is jointly developed by the Instituto Cervantes (IC) and the Universidad Nacional de Educación a Distancia (UNED) and is distributed under the CC BY-NC-SA 4.0 license.

2.4. Hugging Face Platform

Hugging Face is a prominent platform and open-source community at the forefront of artificial intelligence (AI) and machine learning (ML). Frequently referred to as the "GitHub of Machine Learning," it provides a collaborative infrastructure that hosts a vast collection of pre-trained models, datasets, and interactive applications accessible via its Model Hub. Hugging Face is best known for its Transformers library, which offers a unified framework for state-of-the-art transformer-based models across multiple domains including natural language processing (NLP), computer vision, and audio analysis[8].

The platform supports thousands of pre-trained models based on the Transformer architecture, covering tasks such as text classification, summarization, translation, sentiment analysis, and more[9]. Furthermore, Hugging Face offers tools to deploy and test models directly within a web browser, facilitating experimentation and community-driven model development. Most of the models used in the PROFE 2025 competition were sourced from this hub, reflecting its central role in modern AI research and deployment.

2.5. Deep Learning and Large Language Models (LLMs)

Deep learning is a subfield of machine learning that utilizes artificial neural networks with multiple layers—referred to as "deep"—to model complex and abstract data patterns. These architectures are particularly effective for processing unstructured data such as text, images, and audio. Deep learning systems are capable of learning task-specific features from large volumes of data without manual rule definition [10, 11].

Large Language Models (LLMs) represent a significant advancement in deep learning. Trained on massive corpora of textual data, these models are designed to understand, generate, and manipulate human language. They are primarily built upon the Transformer architecture and excel in a wide array of NLP tasks, including text generation, question answering, summarization, machine translation, and code generation [12]. Examples of well-known LLMs include GPT (OpenAI) [13, 14], LLaMA (Meta)[15], Mistral (Mistral AI)[16, 17, 18], Claude (Anthropic)[19, 20], Qwen (Alibaba)[21, 22], DeepSeek(Deepseek AI)[23, 24] and Gemma (Google DeepMind)[25]. These models are typically trained using self-supervised learning techniques and operate on tokenized text sequences to predict subsequent tokens, thereby learning language patterns and contextual relationships.

2.5.1. Qwen-3 Model Family

Qwen-3 is a cutting-edge large language model developed using an advanced Transformer architecture. It incorporates multiple self-attention layers and expanded hidden dimensions, enabling it to capture complex linguistic structures, long-range dependencies, and refined reasoning capabilities. The model is pretrained on extensive text corpora—including web content, books, and code repositories—using objectives such as masked and causal language modeling[26, 27].

Post-training, Qwen-3 can be adapted for specialized domains via fine-tuning or employed in zero-shot/few-shot learning scenarios. At inference time, it processes input prompts through its transformer

layers, generating output by predicting tokens iteratively until reaching an end-of-sequence marker. Due to its size and capacity, Qwen-3 achieves high fluency and contextual accuracy across diverse applications. However, its deployment requires substantial computational resources. To address this, techniques such as prompt engineering and parameter-efficient fine-tuning (e.g., LoRA) are commonly used to adapt the model efficiently.

2.5.2. DeepSeek-Qwen-7B

The model referred to as DeepSeek-Qwen-7B results from a distillation process that combines capabilities from the Qwen and DeepSeek model families. It is based on Qwen 2.5 7B—a model developed by Alibaba Cloud featuring 7 billion parameters—and enhanced using knowledge distilled from DeepSeek-R1, a series of reasoning-focused models developed by DeepSeek AI. DeepSeek-R1 models are noted for their advanced training strategies, including large-scale reinforcement learning, to enhance logical reasoning[28].

Model distillation in this context transfers reasoning competencies from the more complex DeepSeek-R1 model to the smaller, more efficient Qwen 2.5 7B. This process enables the resulting DeepSeek-Qwen-7B model to achieve improved performance on reasoning tasks while maintaining manageable computational requirements. It represents a promising balance between reasoning capability and operational efficiency, particularly suitable for deployment in resource-constrained environments[29].

3. Results

In this research, the DeepSeek-Qwen2.5 and Qwen3.0 models were used to solve the tasks presented in the proposed challenge. The DeepSeek-Qwen2.5 model was configured using Hugging Face's text-generation pipeline. It was loaded along with its corresponding tokenizer via AutoTokenizer and AutoModelForCausalLM, with automatic hardware assignment enabled through device_map="auto" to utilize the GPU when available. The pipeline was initialized with max_new_tokens=10 to constrain output length and a temperature of 0.6 to balance coherence and diversity in generation. This setup was well-suited for short, focused responses in educational tasks requiring controlled variability, such as reading comprehension or automated assessments. The Qwen3.0 model was configured similarly using Hugging Face's text generation pipeline and its official tokenizer. Given its smaller size, the model did not require a GPU and ran efficiently on CPU using the same device_map="auto" setting. The configuration also included max_new_tokens=10 and a temperature of 0.6, which allowed for concise, moderately diverse outputs. This setup proved effective for simpler educational tasks or as a lightweight component in automated systems, particularly in environments with constrained computational resources.

3.1. Internal Comparison of Model Configurations

The performance analysis of the three submitted models highlights a distinct disparity attributable to their respective configurations. Both the DeepSeek-Qwen2.5 and Qwen3.0 models achieved identical accuracy scores of 32.22%, suggesting that the integration of DeepSeek enhancements into the base Qwen architecture did not yield a measurable improvement for the given task. In contrast, the Qwen3.0-Think variant obtained a significantly lower score of 10.19%. This outcome indicates that the inclusion of reasoning-driven prompt strategies—designed to simulate more deliberative processing—may have introduced computational or cognitive overhead without enhancing prediction accuracy. These findings support the robustness of the default Qwen3.0 configuration, while also underscoring the need for further optimization and validation of advanced prompting techniques before they can be considered advantageous in practical settings.

Final Results

A total of 40 submissions were received during the competition, with multiple teams showcasing a variety of approaches and levels of performance. Among these, two teams distinguished themselves

by successfully completing all three subtasks—Multiple Choice, Matching, and Fill-in-the-Gap—thus qualifying for comprehensive evaluation under the exam-level scoring framework.

The "ezotova" team achieved outstanding results across all subtasks, with an accuracy of 89.67% in the Multiple Choice task, 84.31% in Matching, and 92.59% in Fill-in-the-Gap. These scores positioned the team in first place overall, reflecting highly effective system performance and generalization capabilities.

The UC-CICESE_UT3-Plenitas team (identifier: ymlopez, our team) also completed all three subtasks and secured second place overall among the fully evaluated participants. Despite facing several technical challenges, the team attained 32.22% accuracy in Multiple Choice, 4.89% in Matching, and 10.19% in Fill-in-the-Gap. These results, while modest compared to the leading team, underscore the team's commitment to addressing the full scope of the competition and provide a strong foundation for further system improvement.

A summary of results for selected participants is presented in Table 1, which illustrates the comparative accuracy across subtasks and highlights the diversity in model capabilities and task performance.

Table 1
Performance scores of selected teams across the three PROFE 2025 subtasks. Bold values highlight the UC-CICESE_UT3-Plenitas team results.

Team	Model	Multiple Choice (%)	Matching (%)	Fill-in-the-Gap (%)
ezotova		89.67	84.31	92.59
UC-CICESE_UT3-Plenitas	**Qwen3.0**	**32.22**	**4.89**	**10.19**
jfra_team responder		92.96	–	–
ralucaginga		91.96	–	–
mvictoria		91.14	–	–
annamwinkler		32.51	–	–
djanr2		43.77	–	–

4. Conclusions

The results obtained in the PROFE 2025 competition reaffirm the effectiveness of Large Language Models (LLMs) in language proficiency evaluation tasks, despite variations in performance across different subtasks. Our team secured second place among those who completed all three subtasks, highlighting the competitive potential of LLM-based approaches even in challenging areas such as matching and fill-in-the-gap exercises. These findings emphasize the need for continued refinement of LLMs through task-specific optimization, including advanced prompting strategies and fine-tuning techniques. Overall, PROFE 2025 provided a valuable benchmark for assessing the capabilities of automated systems under conditions comparable to those used for evaluating human performance.

Declaration on Generative AI

During the preparation of this work, the author(s) used GP4o in order to: Grammar and spelling check. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: Large-scale reading comprehension dataset from examinations, arXiv preprint arXiv:1704.04683 (2017).

[2] 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.
[3] K. C. Lion, D. A. Thompson, J. D. Cowden, E. Michel, S. A. Rafton, R. F. Hamdy, E. F. Killough, J. Fernandez, B. E. Ebel, Impact of language proficiency testing on provider use of spanish for clinical care, Pediatrics 130 (2012) e80–e87.
[4] A. D. Sobel, J. M. Ramirez, D. F. Walsh, S. F. Defroda, A. I. Cruz Jr, Evaluation of spanish language proficiency and resources available in academic pediatric orthopaedic centers, Journal of Pediatric Orthopaedics 40 (2020) 310–313.
[5] 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).
[6] M. D. Cervantes-Kelly, Translation and interpretation as a means to improve bilingual high school students' English and Spanish academic language proficiency, The University of Arizona, 2010.
[7] S. M. Alvarez, Evaluating the role of the spanish department in the education of us latin@ students: Un testimonio, Journal of Latinos and Education 12 (2013) 131–151.
[8] S. M. Jain, Hugging face, in: Introduction to transformers for NLP: With the hugging face library and models to solve problems, Springer, 2022, pp. 51–67.
[9] J. Jones, W. Jiang, N. Synovic, G. Thiruvathukal, J. Davis, What do we know about hugging face? a systematic literature review and quantitative validation of qualitative claims, in: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2024, pp. 13–24.
[10] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.
[11] N. Rusk, Deep learning, Nature Methods 13 (2016) 35–35.
[12] M. Johnsen, Large language models (LLMs), Maria Johnsen, 2024.
[13] M. Zhang, J. Li, A commentary of gpt-3 in mit technology review 2021, Fundamental Research 1 (2021) 831–833.
[14] K. I. Roumeliotis, N. D. Tselikas, Chatgpt and open-ai models: A preliminary review, Future Internet 15 (2023) 192.
[15] O. Analytica, Meta llama leak raises risk of ai-linked harms, Emerald Expert Briefings (2023).
[16] O. Aydin, E. Karaarslan, F. S. Erenay, N. Bacanin, Generative ai in academic writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma, arXiv preprint arXiv:2503.04765 (2025).
[17] F. Hamzah, N. Sulaiman, Multimodal integration in large language models: A case study with mistral llm (2024).
[18] A. Mistral, Mixtral of experts, Fecha de Publicación 11 (2023).
[19] A. Priyanshu, Y. Maurya, Z. Hong, Ai governance and accountability: An analysis of anthropic's claude, arXiv preprint arXiv:2407.01557 (2024).
[20] A. J. Adetayo, M. O. Aborisade, B. A. Sanni, Microsoft copilot and anthropic claude ai in education and library service, Library Hi Tech News (2024).
[21] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).
[22] A. Cloud, Qwen 2.5 (2024).
[23] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al., Deepseek-v3 technical report, arXiv preprint arXiv:2412.19437 (2024).
[24] L. Xiong, H. Wang, X. Chen, L. Sheng, Y. Xiong, J. Liu, Y. Xiao, H. Chen, Q.-L. Han, Y. Tang, Deepseek: Paradigm shifts and technical evolution in large ai models, IEEE/CAA Journal of Automatica Sinica 12 (2025) 841–858.
[25] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[26] C. Sun, Y. Li, D. Wu, B. Boulet, Onioneval: An unified evaluation of fact-conflicting hallucination for small-large language models, arXiv preprint arXiv:2501.12975 (2025).

[27] R. O. Popov, N. V. Karpenko, V. V. Gerasimov, Overview of small language models in practice, in: CEUR Workshop Proceedings, 2025, pp. 164–182.
[28] J. Wang, F. Meng, J. Zhou, Deep reasoning translation via reinforcement learning, arXiv preprint arXiv:2504.10187 (2025).
[29] S. Papicchio, S. Rossi, L. Cagliero, P. Papotti, Think2sql: Reinforce llm reasoning capabilities for text2sql, arXiv preprint arXiv:2504.15077 (2025).

5. Online Resources

The results are available via:

- Prelimanary Results,
- Codabench.