SINAI at PROFE 2025: Testing the Reading Comprehension of GEMINI 2.0 Flash

María Victoria Cantero-Romero[1], Salud María Jiménez-Zafra[2]

[1]Department of Spanish Philology, SINAI, CEATIC, Universidad de Jaén, Spain
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, Spain

Abstract

This paper presents a zero-shot evaluation of a large language model applied to automatic reading comprehension in Spanish. The experiment was conducted as part of the PROFE 2025 shared task of IberLEF 2025 workshop. It focused on the first subtask, which consists of multiple choice questions based on DELE exam texts. The model received no training examples, and a task-specific prompt was carefully designed to simulate real test conditions. Specifically, the model tested was Gemini 2.0 Flash, which achieved an accuracy of 0.91, demonstrating strong performance on standardized reading comprehension tasks in Spanish. This result contributes to ongoing research on automatic reading comprehension by showing that large language models can successfully interpret real-world assessment items, provided that they are guided by well-structured and pedagogically motivated prompts.

Keywords

Automatic reading comprehension, DELE, multiple choice tests, Spanish, zero-shot, large language models

1. Introduction

Automatic reading comprehension is a key area of research within Natural Language Processing (NLP), concerned with evaluating the ability of language models to understand written texts [1]. The primary goal of this field is to develop systems capable of interpreting and reasoning on texts in a way that approximates human performance [2]. Previous studies have explored this task in various languages [3, 4], highlighting both methodological challenges and promising results.

The PROFE 2025 shared task [5], proposed as part of the IberLEF 2025 [6] evaluation campaign, seeks to assess the performance of transfer learning approaches and generative large language models (LLMs) in reading comprehension tasks, specifically in Spanish. For this purpose, a set of standardized exercises from the Diplomas de Español como Lengua Extranjera (DELE), developed by the Instituto Cervantes, has been employed.

According to the Diccionario de términos clave de ELE published by the Instituto Cervantes, reading comprehension is one of the core communicative skills. It entails not only the decoding of linguistic information, but also the activation of cognitive, perceptual, attitudinal, and sociocultural processes that contribute to textual interpretation [7].

Understanding to what extent LLMs can successfully engage with such multifaceted tasks is critical, as reading comprehension involves far more than syntactic or lexical processing alone. The outcomes of this study provide insights into the current capabilities and limitations of LLMs in Spanish, contributing to the broader discussion on multilingual comprehension and model generalization.

In this context, the SINAI research group contributed to the PROFE 2025 shared task addressing the first subtask, which targets multiple choice reading comprehension exercises in Spanish.

The remainder of this paper is organized as follows. Section 2 describes the PROFE 2025 shared task. Section 3 outlines the methodology and prompt design. Section 4 details the experimental setup and the model used. Section 5 presents the results and a discussion of the main findings. Finally, Section 6 provides the conclusions and outlines directions for future work.

IberLEF 2025, September 2025, Zaragoza, Spain
vcantero@ujaen.es (M. V. Cantero-Romero); sjzafra@ujaen.es (S. M. Jiménez-Zafra)
0009-0008-7052-7322 (M. V. Cantero-Romero); 0000-0003-3274-8825 (S. M. Jiménez-Zafra)

2. Task description

The PROFE shared task [5] is designed to evaluate the capacity of transfer learning approaches or generative LLMs to accurately solve reading comprehension exercises from the exams of the Diplomas de Español como Lengua Extranjera (DELE), issued by the Cervantes Institute. The organizers do not provide training data to evaluate the systems under the same conditions as humans, that is, the systems receive a set of exercises with their corresponding instructions, but no specific training material.

The task comprises three subtasks but, in the present study, we focus exclusively on the first one. A brief description of each subtask is provided below:

1. **Subtask 1: Multiple choice**. It involves reading a text, whose length varies according to the proficiency level, and answering a set of multiple choice questions. Each question presents several options, from which the model must select the most appropriate one.
2. **Subtask 2: Matching**. It requires establishing correspondences between two sets of texts. Items from one set must be matched with their corresponding elements in the other, based on semantic and pragmatic coherence. There is only one possible matching per text, but the first set can contain extra unnecessary texts.
3. **Subtask 3: Filling the gap**. It consists of completing a partially elided text by inserting appropriate fragments provided in a separate list.

The dataset used in the shared task, IC-UNED-RC-ES, is based on real Spanish language proficiency exams developed by the Instituto Cervantes. These exams, created by certified experts in language assessment, span CEFR levels from A1 to C2 and have been digitized for computational use. The complete dataset contains 282 exams and 855 individual exercises, amounting to a total of 6,146 evaluation points distributed across the three task types: multiple choice, matching, and filling the gap. For this edition of the task, the organizers have used approximately 50% of the available material, while the remaining data have been reserved for future evaluations.

Regarding the evaluation, the main evaluation metric used is accuracy, defined as the proportion of correctly answered items. Performance is assessed from two perspectives. First, at the item level, each correct response is counted individually. Second, at the exam level, a system is considered to have passed an exam if it achieves an accuracy above 60%. This exam-level evaluation is applied only to systems participating in the three subtasks. Accuracy was selected as the primary metric because each question has a single correct answer, and it is also the criterion used to evaluate human candidates in the original DELE exams, enabling direct comparison between human and model performance.

3. Methodology

As previously stated, this study focuses on the first subtask of the PROFE 2025 challenge: answering multiple choice reading comprehension questions based on texts from the DELE exams developed by the Instituto Cervantes.

To conduct our experiments, we designed a detailed prompt following established principles of prompt engineering [8]. The objective was to guide the language model to perform the task as reliably and accurately as possible. The design prioritized clarity, precision, and the inclusion of contextually relevant cues. Drawing on the recommendations of Morales-Chan (2023), the prompt included explicit task instructions and role assignment to optimize the model's alignment with the intended behavior.

The role assigned to the model was that of an expert in the Spanish language and its teaching as a foreign language, with the additional specification that it was acting as a DELE test developer. This role was selected to match the nature of the task: solving DELE-style comprehension exercises, which require not only linguistic understanding but also familiarity with the structure and intent of standardized language assessments.

In addition to role specification, the prompt incorporated constraints and guidance similar to those provided to human learners preparing for DELE exams [10]. This included contextualizing the task,

Prompt

Eres un experto en español y en su enseñanza como lengua extranjera. Vas a corregir un ejercicio de comprensión lectora. Este ejercicio es de nivel **NIVEL_DEL_EXAMEN**. Lee el siguiente texto y contesta a las preguntas.

Tienes que elegir una de las opciones. Debes comprender la idea principal, hay normalmente una pregunta sobre ello y también la información concreta que hay en el texto, NO sobreentiendas respuestas. Las respuestas verdaderas no siempre son las que más información tengan. En ocasiones debes prestar atención a toda la información y reflexionar, puesto que la respuesta no está en una frase, si no en el conjunto del texto, para ello fíjate en tiempos verbales, marcadores del discurso y pronombres. Recuerda que no siempre son las mismas palabras en el texto y en las preguntas. Algunas preguntas son sobre la intención del texto.

Cada pregunta tiene un identificador único (questionId) que debes respetar exactamente. Estos identificadores tienen este formato:
<nivel>_<fecha>_<ejercicio>_<pregunta>
Donde:
- <nivel> es el nivel del examen (por ejemplo, B2)
- <fecha> es la fecha del examen con año de cuatro cifras (por ejemplo, 2010-05-22)
- <ejercicio> es el ejercicio, por ejemplo E1
- <pregunta> es el número de la pregunta, como Q1, Q2, etc.

Ejemplo correcto de identificador:
B2_2010-05-22_E1_Q3

Debes usar exactamente los identificadores que se te muestran a continuación, sin modificarlos: **IDS_VÁLIDOS**

Devuelve las respuestas en formato JSON con estos identificadores como claves y una sola letra como valor (A, B, C...).

No inventes identificadores ni des explicaciones. No escribas 'ninguna de las anteriores'.
Ejemplo:
{
B2_2010-05-22_E1_Q3: A,
B2_2010-05-22_E1_Q4: B
}

TEXTO_EXAMEN

PREGUNTAS: **PREGUNTAS_EXAMEN**

Figure 1: Prompt used in the PROFE 2025 multiple choice subtask to evaluate the Gemini 2.0 Flash generative model.

emphasizing attention to textual details, and clarifying that the goal was to select the most appropriate option among several possible answers.

Finally, the prompt concluded with instructions for formatting the output clearly and consistently, ensuring that the model's responses could be evaluated automatically and without ambiguity.

To arrive at the final prompt configuration, a series of tests were conducted using texts extracted from DELE preparation books in order to evaluate prompt effectiveness prior to the official task submission. Furthermore, texts were grouped and evaluated by CEFR level, as the model was found to perform more accurately.

The full prompt used is presented in Figure 1.

4. Experimental set up

This study was conducted using a zero-shot [11] approach, in which the language model was not provided with any examples prior to answering the questions. This setting was selected to evaluate the model's intrinsic ability to perform reading comprehension based solely on the knowledge acquired during pretraining.

The objective was to assess the generalization capacity of the model when faced with unseen domain-specific tasks, specifically standardized reading comprehension exercises from the DELE exams, without relying on fine-tuning or in-context learning.

For the experiments, we used Gemini 2.0 Flash [12], developed by Google, a state-of-the-art large language model designed for high-speed inference and optimized performance across a variety of NLP tasks. No additional training or adaptation was applied prior to the evaluation. Gemini 2.0 Flash was accessed via the official *google.genai* Python SDK. All experiments were executed in Google Colab, using Python 3.11 in a standard notebook environment without GPU acceleration.

The experiments were performed by CEFR levels, generating one response for each exam of the corresponding level. For this, the dataset, provided in JSON format, was filtered to select only the exams of a specific CEFR level (A1, A2, B1, B1, B2, C1 or C2), and each exam was processed as a complete unit including the exam text and all its associated multiple-choice questions. Thus, the model Gemini 2.0 Flash independently evaluated each exam with a dynamically generated prompt (Figure 1). In addition, to ensure robustness, a retry mechanism was implemented that handled possible API speed limits or server errors (*RESOURCE_EXHAUSTED*), introducing controlled delays between attempts. The parameters dynamically included in the prompt were:

- NIVEL_DEL_EXAMEN: Level corresponding to the exam to be evaluated (A1, A2, B1, B1, B2, C1 or C2).
- IDS_VÁLIDOS: Explicit control over the inclusion of only valid question IDs.
- TEXTO_EXAMEN: Text of the exam to be solved.
- PREGUNTAS_EXAMEN: Exam questions for which an answer is to be generated.

5. Results and discussion

The results obtained in this study are highly promising and suggest that large language models, even in zero-shot configurations, possess substantial capabilities in performing reading comprehension tasks in Spanish. Specifically, the model achieved an accuracy of 0.91, significantly outperforming the provided task baseline of 0.64. This represents a 27 point improvement, underscoring the effectiveness of the prompt design.

Figure 2 presents a bar chart that compares the scores obtained by the different participating teams. As shown, our team, SINAI, achieved a result very close to that of the top-performing team, highlighting the competitiveness of our approach.

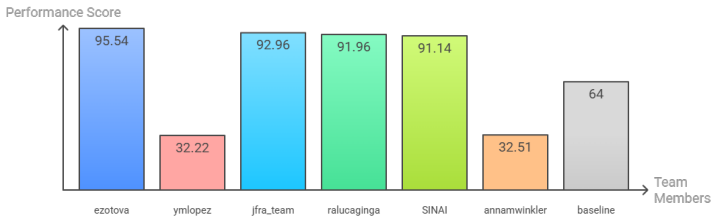

Figure 2: Results of the multiple choice subtask. Performance Score represents the accuracy measure.

This level of performance, achieved without prior examples or task specific adjustments, suggests that current language models not only internalize general linguistic patterns but also adopt task oriented reasoning strategies when guided by well-structured instructions.

Nevertheless, it is necessary to assess whether this effectiveness remains consistent across different item types and proficiency levels. Although the overall accuracy is high, it would be important to determine whether the model struggles with specific cognitive challenges, such as inference, figurative language, or discourse coherence—phenomena that are often difficult even for human learners.

In addition, during the experimental phase, it was observed that the prompt design had a direct impact on the model's performance. In particular, prompts that provided clearer, more specific instructions and reflected the communicative context of the exam tended to yield more accurate and coherent responses. This observation suggests that, even in zero-shot settings, the way a task is presented can serve as a trigger for activating the model's relevant capabilities. It is not merely a matter of issuing an instruction, but of doing so in a manner aligned with the internal logic of the test and the type of processing it requires.

These findings contribute to the field of automatic reading comprehension (ARC) by demonstrating that large language models can effectively solve structured comprehension tasks in Spanish, even without prior task-specific training. The use of authentic test materials and a zero-shot setup highlights the potential of these models for evaluating linguistic and interpretive competence in realistic educational scenarios.

6. Conclusions and future work

This study has shown that large language models, even in a zero-shot setting, can achieve strong performance in Spanish reading comprehension tasks, as demonstrated by the accuracy of 0.91, which significantly exceeds the task baseline. The success of the proposed approach can be attributed, in part, to the design of the prompt, which closely mirrored the instructions typically given to students preparing for DELE exams.

Beyond the strong results obtained, the findings highlight the growing potential of language models as tools in educational contexts. Their ability to interpret structured tasks and generate reliable responses opens the door to applications such as autonomous exam preparation, exercise generation, and even formative assessment in foreign language teaching.

These findings contribute to ongoing research on automatic reading comprehension, offering evidence that large language models can perform effectively on real world assessment tasks in Spanish, even in zero-shot conditions.

For future work, it would be valuable to conduct a more detailed evaluation of the model's performance across different CEFR levels and item types. Additionally, it would be of interest to apply the proposed approach to other reading comprehension tasks included in the competition, such as subtask 2 (text matching) and subtask 3 (filling the gap), in which we did not participate in this edition. This would help assess the model's ability to handle a broader range of evaluation formats.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and Project FedDAP (PID2020-116118GA-I00) and Project Trust-ReDaS (PID2020-119478GB-I00) supported by MICINN/AEI/10.13039/501100011033. The research work conducted by Salud María Jiménez-Zafra has been supported by the grant RYC2023-044481-I, funded by MICIU/AEI/10.13039/501100011033 and by ESF+.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4-turbo in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Zeng, S. Li, Q. Li, J. Hu, J. Hu, A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets, Applied Sciences 10 (2020) 7640. doi:10.3390/app10217640.

[2] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: Large-scale reading comprehension dataset from examinations, 2017. URL: https://arxiv.org/abs/1704.04683. arXiv:1704.04683, arXiv preprint.

[3] Álvaro Rodrigo, A. Peñas, Y. Miyao, E. Hovy, N. Kando, Overview of clef qa entrance exams task 2015, in: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum, volume 1391, CEUR-WS.org, Toulouse, France, 2015, pp. 171–182. URL: https://ceur-ws.org/Vol-1391/171-CR.pdf.

[4] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco: A human generated machine reading comprehension dataset, 2016. URL: https://arxiv.org/abs/1611.09268. arXiv:1611.09268, arXiv preprint.

[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] 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.

[7] Centro Virtual Cervantes, Comprensión lectora, https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/comprensionlectora.htm, s.f. Consultado el 30 de mayo de 2025.

[8] Prompt Engineering Guide, Prompt sin entrenamiento previo (zero-shot), https://www.promptingguide.ai/es/techniques/zeroshot, 2024. Consultado el 31 de mayo de 2025.

[9] M. A. Morales-Chan, Explorando el potencial de chat gpt: Una clasificación de prompts efectivos para la enseñanza, Tesario Virtual (2023). URL: http://biblioteca.galileo.edu/tesario/handle/123456789/1348, consultado el 31 de mayo de 2025.

[10] E. Puertas Moya, ¡Dale al DELE! A1: nuevos modelos de examen, enClave-ELE, Madrid, 2020. URL: https://enclave-ele.net/product/dale-al-dele-a1-nuevos-modelos-de-examen/, incluye acceso al libro digital en Blinklearning y audios descargables.

[11] D. Bergmann, ¿qué es el aprendizaje zero-shot?, https://www.ibm.com/es-es/think/topics/zero-shot-learning, 2024. Consultado el 31 de mayo de 2025.

[12] Google DeepMind, Gemini 2.0 flash: Lightweight multimodal model for fast reasoning, https://deepmind.google/technologies/gemini, 2024. Consultado el 31 de mayo de 2025.