Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 487-497 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of PROFE at IberLEF 2025: Language Proficiency Evaluation

Resumen de PROFE en IberLEF 2025: Evaluación de la competencia lingüística

Álvaro Rodrigo[1], **Sergio Moreno-Álvarez**[1], **Alberto Pérez**[1], **Anselmo Peñas**[1], **Rodrigo Agerri**[2], **Javier Fruns-Jiménez**[3], **Inés Soria-Pastor**[3]
[1]NLP&IR group at UNED
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[3]Instituto Cervantes
{alvarory, smoreno, alberto.perez, anselmo}@lsi.uned.es
rodrigo.agerri@ehu.eus
{javier.fruns, isoria}@cervantes.es

Resumen: La comprensión del lenguaje exige captar los matices semánticos sutiles y las inferencias lógicas que sustentan el lenguaje, algo que va más allá del análisis de patrones superficiales del texto. La escasez de recursos de evaluación dado su alto coste de intervención humana para el etiquetado, suele limitar el progreso en este ámbito. Para subsanar esta carencia, la tarea propuesta, denominada PROFE 2025, reutiliza los exámenes de competencia en español desarrollados durante años por el Instituto Cervantes. Estos exámenes, concebidos originalmente para evaluar a estudiantes humanos, se transforman ahora en un banco de pruebas riguroso para medir el rendimiento de sistemas automáticos basados en modelos de lenguaje natural. En esta tarea, se invita a los participantes a desarrollar soluciones sustentadas en diversas estrategias, como el aprendizaje por transferencia o los modelos generativos, entre otras.
Palabras clave: Comprensión Lingüística, Aprendizaje por Transferencia, Evaluación Directa Zero-Shot, Modelos de Inteligencia Artificial aplicados al Lenguaje.

Abstract: Language comprehension requires capturing the subtle semantic nuances and logical inferences that underpin language, something that goes beyond the analysis of superficial text patterns. Given the high cost of human intervention for labeling, the scarcity of assessment resources often limits progress in this area. To address this shortcoming, the proposed task, PROFE 2025, reuses the Spanish proficiency tests developed over the years by the Instituto Cervantes. These exams, originally conceived to evaluate human learners, are now transformed into a rigorous testbed for measuring the performance of automatic systems based on natural language models. In this task, participants are invited to develop solutions based on various strategies, such as transfer learning or generative models, among others.
Keywords: Language Comprehension, Transfer Learning, Zero-Shot Evaluation, Large Language Models.

1 Introduction

Reading Comprehension (RC) has become a crucial benchmark for assessing the reasoning capabilities of Natural Language Processing (NLP) systems. RC evaluates various skills, from extracting concise insights hidden in large volumes of text to adapting across multiple domains, languages, and communication styles (Geva et al., 2021). With the advent of Large Language Models (LLMs), the significance of RC has increased even further for evaluating capabilities across different languages (Zhou et al., 2024). Deep language comprehension is essential for understanding semantic nuances, logical inference, entity recognition, and the relationships between document sentences (Jiang et al., 2021).

A major challenge in RC has been creating new resources, often requiring significant

ISSN 1135-5948 DOI 10.26342/2025-75-36

human intervention. This problem has led to the use of crowd workers, which sometimes results in collections with high lexical overlap, making it challenging to evaluate reasoning capabilities effectively. One solution has been to reuse human RC collections, such as RACE (Lai et al., 2017), allowing researchers to compare the performance of automated systems against human benchmarks. However, most of these resources have been created primarily in English and often contain a substantial amount of training data similar to the test set, which can limit the generalization of results regarding reasoning capabilities.

In this context, we have proposed the PROFE (Language Proficiency Evaluation) shared task at IberLEF 2025 (González-Barba, Chiruzzo, and Jiménez-Zafra, 2025). In this task, we propose to evaluate automatic systems under the same conditions in which human language understanding is assessed. To this end, we have used exams produced by Instituto Cervantes[1] to evaluate new language learners at different levels. Participant systems received a set of exercises with their corresponding instructions. Thus, PROFE proposes an RC task without specific training data, thereby pushing the use of Transfer Learning approaches or Large Generative Language Models. Only a few exams were provided as samples of the dataset.

We have proposed three subtasks: multiple-choice, matching, and fill-in-the-gap. While multiple-choice has been a classical evaluation of RC capabilities, the other two represent new challenges for automatic systems. The results show that current technologies can obtain outstanding results in this collection by combining different Large Language Models (LLMs) in ensembles or multi-agent systems.

2 Task Definition

PROFE proposes to evaluate the RC abilities of automatic systems under the same conditions under which humans are evaluated. With this idea in mind, PROFE proposes three subtasks. Teams can participate in any subtask or all of them, and each subtask contains several exercises of the same type. The subtasks included in this edition are:

- **Multiple choice**: Each exercise includes a text and a set of multiple-choice questions about the text, where only one answer is correct. Given a multiple-choice question, systems must select the correct answer among the candidates. Figure 1 shows an example of one exercise for this subtask.
- **Matching**: Each exercise contains two sets of texts. Systems must find, for each text in the second set, the text in the first set that best matches. There is only one possible match per text. Additionally, the first set may contain unnecessary texts. Figure 2 shows an example of one exercise for this subtask.
- **Fill in the gap**: Each exercise contains a text with several gaps corresponding to fragments that have been removed and presented disorderly as options. Systems must determine the correct position for each fragment. There is only one correct filling per gap, but there could be more candidates than gaps. Figure 3 shows an example of one exercise for this subtask.

While multiple-choice has been a standard format in other RC evaluations (Colelough, Bartels, and Demner-Fushman, 2025) or benchmarks (Rogers et al., 2020), RC evaluations based on matching and fill-in-the-gap are, to the best of our knowledge, new. In conjunction with the lack of specific training data, the variety of subtasks presents opportunities for research on how to approach them, including adapting different prompts when using generative models.

3 Dataset

In this task, we utilize an excerpt of *IC-UNED-RC-ES_v1* (Peñas et al., 2025), a Spanish corpus derived from authentic proficiency examinations administered by the Instituto Cervantes. Certified exam setters authored all items that cover the full Common European Framework of Reference for Languages (CEFR) spectrum. Concretely, the difficulty levels are A1, A1E, A2, A2B1E, B1, B1E, B2, C1 and C2. However, not all levels are included for the different subtasks.

The dataset is divided into a set of exams, which contains a set of exercises, each one of the type of the three subtasks proposed at PROFE: i) *multiple choice* containing all levels, ii) *matching* without the B1E level and, iii) *fill the gap* including B1, B2, C1 and C2 levels. For each exam, the dataset contains

[1] https://cervantes.org/

all the respective questions, ensuring that all content from the exam is included in the IC-UNED-RC-ES_v1 dataset. The excerpt used at PROFE contains a total of 1704 *multiple choice* questions, 818 *matching* questions, 108 *fill the gap* questions, and 138 complete exams. We show in Figures 1, 2, and 3 one example exercise for each subtask.

The created dataset is not homogeneous and uniform. Specifically, the number of questions and possible answers varies within a particular level, task, and exam. In this regard, for the *matching* task, the number of potential options to match could be higher or lower than the number of questions, which creates an environment with two specific cases: i) more questions than answers, and ii) more answers than questions. On the other hand, for the *fill the gap* subtask, the number of gaps is always lower than the possible options.

A dedicated team manually digitised each examination exercise by transcribing every item into a structured tabular database. Integrity fields based on checksums were incorporated to enable automated validation. Once the tabular data was complete, the records were exported to JSON for two main reasons: (i) to enable scripts to flag missing attributes, transcription errors, and checksum inconsistencies; and (ii) to distribute the corpus in a machine-readable format that is suitable for computational evaluation.

We intend not to distribute the gold standard to prevent overfitting in post-campaign experiments and data contamination in LLMs.

4 Evaluation

We have used traditional accuracy (proportion of correct answers) as the main evaluation measure. Participant systems receive evaluation scores from two different perspectives:

- At the question level, where correct answers are counted individually without grouping them.
- At the exam level, where scores for each exam are considered. Each exam contains several exercises of different types. An exam is deemed to be passed if an accuracy score (accounted as the proportion of correct answers) above 0.6[2] is reached. Then, the proportion of passed exams is given as a global score. This perspective is only applied to those teams participating in the three subtasks.

More in detail, the exact evaluation per subtask is as follows:

- Multiple choice subtask: We measure accuracy as the proportion of questions correctly answered
- Matching subtask: We measure accuracy as the proportion of correct texts matched.
- Fill in the gap subtask: We measure accuracy as the proportion of correctly filled gaps.

We have used accuracy as the evaluation measure because there is only one correct option among candidates and because it is the measure applied to humans doing the same exams. Thus, we can compare the performance of automatic systems and humans under the same conditions.

5 Baselines

We have proposed three baselines, one per subtask. The *multiple-choice* baseline is conducted using the XLM-RoBERTa-Large transformer, fine-tuned on the RACE RC dataset. Fine-tuning is carried out for three epochs with an effective batch size of 12, a constant learning rate of $3e^{-5}$, and the AdamW optimizer. Evaluation metrics are obtained by reloading the best checkpoint produced under these hyperparameters and applying the identical configuration during inference on the task.

On the other hand, for the *matching* task, answer selection is carried out with the generative model Llama-3-8B[3]. This variant of the original architecture has been further tuned on carefully curated Spanish corpora. All experiments employ 4-bit asymmetric, non-uniform weight quantisation with double quantisation, while bfloat16 is retained for computation. A set of allowed tokens is defined to be selected for the model without sampling. However, the results may vary slightly depending on the prompts used.

[2]Instituto Cervantes requires having at least 60% of correct responses to pass an exam

[3]https://huggingface.co/Kukedlc/LLaMa-3-8b-Spanish-RAG-v2.1

Instrucciones: A continuación encontrará usted un texto y 10 preguntas sobre él. Marque la opción correcta en la Hoja de Respuestas Número 1.

Texto
ACTIVIDADES CULTURALES Y DE OCIO TEATRO 1. Carta de amor. La última obra de Fernando Rabal, se la ha dedicado a su madre y por eso está llena de sentimientos. El texto fue estrenado en enero en el teatro Reina Sofía, y se mueve entre la novela y la poesía. Entrada: 10 €. No hay entradas 2. Luces de Bohemia. Helena Pimenta, una de las directoras de teatro más importantes de España, dirige el drama "Luces de bohemia"en el teatro Cervantes de Teruel abierto al público desde 1479. La directora ha estado trabajando 2 años en esta obra. Entrada: de 6 a 9 €. 3. Pan con pan. Es un texto dirigido por Miguel Muñoz. La obra está llena de personajes en la que se recuerda a los más pobres de la sociedad. En esta obra hay humor, poesía y crítica social. Ha sido premio Max de teatro 2002. Entrada con descuento para estudiantes. 4. Defensa de Sancho Panza. El autor y director Fernando Fernán Gómez estrenó ayer esta obra en el teatro Infanta Isabel, en Madrid. En ella el protagonista tiene que defend-erse de algunos delitos. "La idea ni es original ni es mía", ha dicho el famoso director. Entrada: 15 €. MÚSICA 1. The Fairy Queen. Lindsay Kemp pensó en la película "El sueño de una noche de vera-no" para hacer esta obra. Es una mezcla de música, teatro y baile. Esta obra podremos verla las próximas Navidades. 2. La flauta mágica de Mozart, interpretada por la compañía El Teatro Negro de Praga. Durante el mes de octubre podemos ver y escuchar La flauta mágica en el teatro Maravillas de Jaén. La obra añade a la música el encanto de un espectáculo sin palabras. 3. Música religiosa en las catedrales. En las catedrales de Madrid, Plasencia y Salamanca se podrá escuchar este concierto el próximo mes de mayo, interpretado por La Orquesta de la Comunidad de Madrid dedicado a Juan Sebastián Bach. Está organizado por la Fundación Caja Madrid. 4. Broadway. Es el musical más famoso de todos los tiempos. Durante estos días se puede ver en El Teatro de la Ópera Nacional de Krasnodarsk en Rusia. Otro atractivo de esta obra será ver el espectáculo de luces y colores con los últimos avances tecnológicos. VIAJES 1. Hervás (Cáceres). Duración: del 4 al 7 de abril. Precio: a partir de 250 €. Organiza: Halcón Viajes. Programa de cuatro días con natación y paseos a caballo. El precio también incluye el alojamiento en un hotel de tres estrellas con pensión completa. Viaje ahora y pague en los próximos meses. 2. Doñana (Huelva). Duración: del 4 al 7 de abril. Precio: 200 €. Organiza: Gente Viajera. Cuatro días en contacto directo con la naturaleza. Se realizarán excursiones por la montaña. Alojamiento en hotel de tres estrellas, con pensión completa (desayuno, comida y cena). 3. París (Francia). Duración: del 4 al 7 de abril. Precio: 342 €, con vuelo desde Madrid y 350 € desde Barcelona. Preguntar precios desde otras ciudades. Organiza: Viajes Barcelona. Uno de los mejores precios de la temporada, ya que incluye el transporte desde el aeropuerto hasta el hotel, el alojamiento con desayuno, guía y seguro de viaje. 4. Milán, Florencia y Génova (Italia). Duración: del 2 al 7 de abril. Precio: 1.168 €, con vuelo desde Barcelona y 1.198 € desde Madrid. Organiza: Turismúsica. Además de los vuelos, alojamiento con desayuno en hoteles de cuatro estrellas y seguro de viaje, este programa ofrece visitas culturales guiadas por Milán, Florencia, Pisa y Génova. También incluye las entradas a tres conciertos. 5. Menorca (Islas Baleares). Duración: del 1 al 8 de abril. Precio: 301 €, con vuelo desde Madrid. Consultar precios desde otras ciudades. Organiza: Águila Viajes. Perfecto para familias que quieran unas vacaciones de sol y playa. Transporte, alojamiento en régimen de media pensión (desayuno, comida o cena) y seguro incluidos.

#	**Pregunta**	**(A)**	**(B)**	**(C)**
1	En la obra «Carta de amor» el autor ha dedicado la obra a...	los sentimientos.	a su madre.	a la reina.
2	¿Para cuál de estas obras de teatro ya no se pueden comprar entradas?...	«Carta de amor».	«Luces de bohemia».	«Pan con pan».
3	¿En qué viaje están incluidas *todas* las comidas?...	Doñana.	París.	Menorca.

Figure 1: Corpus of a question within a specific exam for the *multiple choice* task.

Finally, the *fill the gap* baseline utilizes the Mixtral-8×7B[4] generative model, which is based on the main GPTQ branch configured for 4-bit quantization. All other hyperparameters follow their default settings, and the computation data type is set to bfloat16. Inference employs a maximum sequence length of 8192 tokens with no group-size constraint,

[4]https://huggingface.co/TheBloke/Mixtral-8x7B-Instruct-v0.1-GPTQ

Instrucciones: Usted va a leer seis enunciados que resumen los textos de algunos signos del horóscopo para la próxima semana y diez textos que corresponden a distintos signos. Relacione los enunciados (1–6) con los textos del horóscopo (A–J). **HAY TRES TEXTOS QUE NO DEBE RELACIONAR.** Marque las opciones elegidas en la Hoja de Respuestas.

Textos del horóscopo

A *Aries* – Esta semana debe cuidarse: cansancio, dolor de cabeza, molestias digestivas. Relájese y duerma bien. Mejor día: domingo. Nº suerte: 8.

B *Tauro* – Ánimo bajo que se refleja en su físico. El ejercicio en compañía le sentará genial. Mejor día: sábado. Nº suerte: 7.

C *Géminis* – Salud, trabajo y familia irán bien; evite discusiones de pareja. Mejor día: jueves. Nº suerte: 5.

D *Cáncer* – Semana llena de energía, sobre todo en el amor: alguien especial podría aparecer. Mejor día: viernes. Nº suerte: 2.

E *Leo* – La suerte le ronda; pruebe la lotería o un viaje de negocios para mejorar su carrera. Mejor día: lunes. Nº suerte: 0.

F *Virgo* – Dedíquese a su ocio: libro pendiente, buena música, llamar a amigos, cocinar su plato favorito. Mejor día: viernes. Nº suerte: 1.

G *Piscis* – Estrés laboral; estudie las ofertas que tiene y mantenga el optimismo. Mejor día: domingo. Nº suerte: 3.

H *Escorpio* – Deberá tomar una gran decisión; escuche a sus padres y personas cercanas. Mejor día: martes. Nº suerte: 1.

I *Sagitario* – Recibirá dinero extra: buen momento para ese viaje o reformar la casa. Mejor día: miércoles. Nº suerte: 4.

J *Acuario* – Éxito profesional, pero controle las compras y las tarjetas de crédito. Mejor día: sábado. Nº suerte: 3.

#	**Opciones**
1	Esta semana va a tener problemas de salud.
2	Es un buen momento para cambiar de trabajo.
3	Es bueno para usted hacer un poco de deporte.
4	Debería escuchar a su familia con atención.
5	Le recomiendan controlar sus gastos.
6	Puede conocer a una persona muy especial.

Figure 2: Corpus of a question within a specific exam for the *matching* task.

thereby lowering VRAM requirements. In this regard, the model is set to provide sampling outputs with temperature set to 0.5, top_k to 40, and top_p to 0.95. Similarly, results could vary based on prompt selection.

6 Participants and Results

We organized the submissions using the Codabench platform. While 19 teams filled the registration form at the website[5], 14 teams registered at the Codabench competition[6], and 8 teams submitted their runs. We allowed up to 5 runs per team in each subtask. We show in Table 1 the information about the participant groups and the reference to their reports.

All participants emphasized in their reports the fact that they did not receive task-specific training data, which led to the use of zero-shot approaches or training with similar data.

The multiple-choice subtask was the most popular, with 24 runs, while we received 15 runs for the matching subtask and 11 for the fill-in-the-gap subtask. On the other hand, we received only 9 submissions from 2 different teams for all the subtasks, to apply the evaluation at the exam level. We describe the results and the participants' approaches in the following sections.

[5] https://sites.google.com/view/profe2025

[6] https://www.codabench.org/competitions/5521/

Instrucciones: Lea el siguiente texto, del que se han extraído seis fragmentos. A continuación lea los ocho fragmentos propuestos (A–H) y decida en qué lugar del texto (19–24) hay que colocar cada uno de ellos. **HAY DOS FRAGMENTOS QUE NO TIENE QUE ELEGIR.** Marque las opciones elegidas en la Hoja de respuestas.

Texto

Tamales panameños El tamal es probablemente el producto más apreciado, y sin duda el más popular, de la cocina panameña. Sin embargo, debido a la dificultad de su elaboración, a los panameños nos da bastante miedo preparar en casa nuestros propios tamales. Para que cualquier persona se atreva a preparar este riquísimo plato tradicional, vamos a aconsejarles sobre los ingredientes que deben utilizarse en el auténtico tamal panameño. **(19)** De todos modos, el resultado final dependerá también de la calidad de los productos que utilice y de su mano en la cocina.

Comencemos por el componente fundamental: el grano de maíz. **(20)** Los tamales preparados a base de maíz nuevo se elaboran con los granos de la mazorca recién cosechada. Su preparación es más sencilla, ya que la mazorca se ralla y los granos se deshacen para formar la masa del tamal. En cambio, el maíz viejo es el grano seco, y es necesario cocinarlo hasta que esté tierno, por lo que prepararlo toma más tiempo. **(21)**

Otro componente importante es el relleno: los tamales más típicos en nuestra mesa se preparan con cerdo o gallina. Probablemente lo mejor sea con gallina de patio, pero requiere una cocción larga para lograr una carne más suave. **(22)**

Además de la carne del relleno, hay otros ingredientes que dan a los tamales panameños unas características especiales. **(23)** También en otros países se usan productos propios de cada región. Un ejemplo lo vemos en el sur de México, donde llevan chile y se envuelven con hojas de parra.

Por último, quiero recordarles algo muy importante: hay que tener en cuenta que, una vez cocidos, los tamales se deben conservar en el frío. **(24)** En caso contrario, el maíz se comienza a descomponer y produce un sabor agrio, desagradable al paladar y nocivo para nuestra salud.

Fragmentos propuestos

A En el interior del país se prefiere el tamal de maíz nuevo, mientras que en la capital es habitual el de maíz viejo.

B La mezcla de cebollas, ají criollo y tomates cocida lentamente es lo que distingue nuestros tamales panameños de los del resto de América.

C Estamos seguros de que, si elige los ingredientes adecuados y sigue las instrucciones de cualquier receta tradicional, obtendrá un tamal excelente.

D No obstante, el esfuerzo merece la pena porque el sabor que da es mucho más intenso que el del pollo.

E Por ejemplo, este ingrediente es la mejor opción para que el tamal panameño quede sabroso.

F Por ello, si no los vamos a consumir en los tres días siguientes, es necesario congelarlos.

G Este puede ser nuevo o viejo, y el sabor del tamal será completamente diferente en cada caso.

H Por lo tanto, debemos comprar el que esté más fresco para asegurarnos de que el relleno esté jugoso.

Figure 3: Corpus of a question within a specific exam for the *fill the gap* task.

6.1 Multiple-Choice Subtask

Most of the teams participated in the multiple-choice subtask. All the teams, except the URJC and djanr2 participated. The results for this subtask are presented in Table 2. We can see that several participants achieved results above the proposed baseline, with some systems even attaining over 90% accuracy. In fact, the best system, from the *Vicomtech* team, obtains a score of 95,54. These results show that this subtask was quite feasible for current technologies.

Regarding the approaches of the participants, most of them relied on LLMs like Gemini (JFra-Team and SINAI) or Qwen (UCO-CICESE_UT3-Plenitas). The best-performing systems went beyond by combining several LLMs or refining the prompts to exploit these systems better. Vicomtech performed several experiments during the development period and selected the most promising ones for each subtask. For the multiple-choice task, they experimented with LLMs of different sizes, and they also fine-tuned some models on two multiple-choice QA datasets: Belebele (Bandarkar et al., 2024) and RetrievalQA [7]. Besides, they employed two ensemble strategies: one based on majority voting and the other on a Random Forest classi-

[7] https://github.com/wln20/Retrieval_QA

Team	Report
UC-UCO-CICESE_UT3-Plenitas	(Martínez-López et al., 2025)
SINAI	(Cantero-Romero and Jiménez-Zafra, 2025)
Vicomtech	(Platas et al., 2025)
JFra-Team	(Fraile-Hernzández and Peñas, 2025)
NILUCM	(Winkler and Díaz, 2025)
URJC	(Rodríguez-García et al., 2025)
djanr2	-
The Vikings	-

Table 1: Participants and references to their reports.

fier.

The JFra-Team created a system using four specialized agent types that work collaboratively to make informed and interpretable decisions in a zero-shot setting using Gemini 2.5 Flash. While one agent, the responder, selected the most plausible option, another agent, the blind responder, extracted a candidate answer from the text. An evaluator agent assesses whether the not-selected alternatives could be correct, and finally, a moderator resolves disagreements among agents and selects the final answer of the system.

The SINAI team proposed a zero-shot system based on a task-specific prompt using Gemini 2.0 Flash. The prompt included explicit task instructions and role assignment to optimize the results. The role assigned to the model was that of an expert in the Spanish language, with the additional specification that it was acting as a developer of the Diplomas of Spanish as a Foreign Language (DELE) tests. Their high results show the importance of using an appropriate prompt for the task.

The NILUCM team relied on BERT-based models already fine-tuned on the SQuAD dataset (Rajpurkar et al., 2016). Then, they extracted answers from the text and calculated token overlap scores between each model's answer and each candidate, returning the candidate with the highest score.

Finally, the UC-UCO-CICESE_UT3-Plenitas team uses a DeepSeek-Qwen2.5, a Qwen3.0, and a Qwen3.0-Think model, which obtain the same results. As for the Vikings team, we did not receive any information about their approach.

6.2 Matching Subtask

Matching was the second subtask with more participants, and two of them, the URJC and the djanr2 teams, only participated in it. We show results for the matching subtask at Table 3. At this subtask, we can also see how the best-performing system, from the Vicomtech team, obtains an accuracy close to perfect performance (95,91). As for the other participants, there is still room for improvement.

The Vicomtech team tested the use of several LLMs using zero-shot in-context learning. Besides, they also tested an approach based on measuring the semantic similarity of embedding representations for both sets of texts.

The URJC team fine-tuned three Spanish BERT-based models on a synthetic collection created from the SQAC dataset (y Jordi Armengol-Estapé y Marc Pàmies y Joan Llop-Palao y Joaquin Silveira-Ocampo y Casimiro Pio Carrino y Carme Armentano-Oller y Carlos Rodriguez-Penagos y Aitor Gonzalez-Agirre y Marta Villegas, 2022). Then, they apply a voting technique to determine the best match for each question.

The UC-UCO-CICESE_UT3-Plenitas team uses a DeepSeek-Qwen2.5, a Qwen3.0 and a Qwen3.0-Think model, which obtain the same results. As for the djanr2 team, we did not receive any information about their approach.

6.3 Fill-the-gap Subtask

At the fill-the-gap subtask, we only received submissions from three participants. We think this is due to the nature of the task, which differs from other everyday tasks faced in Natural Language Processing.

We show the results for this subtask in Table 4. Again, we can see extraordinary re-

Team	Run	Task 1: Multiple Choice
Vicomtech	ezotova 1	95.54
JFra-Team	responder	92.96
JFra-Team	multi_agent	92.9
The Vikings	submission4_G2	91.96
SINAI	zero_shot_multiple_choice	91.14
Vicomtech	ezotova 3	90.9
The Vikings	submission5_G3	89.79
SINAI	zero_shot_multiple_choice_1	89.67
Vicomtech	ezotova 4	89.67
Vicomtech	ezotova 2	87.32
SINAI	few_shot_multiple_choice_2	78.11
baseline	baseline	64.40
The Vikings	submission3_0S	63.38
JFra-Team	blind_responder	55.87
Vicomtech	ezotova 5	55.34
The Vikings	TB	35.09
NILUCM	roberta_large	32.51
NILUCM	bert_large	32.28
NILUCM	roberta_base	32.51
UC-UCO-CICESE_UT3-Plenitas	qwen3.0	32.22
UC-UCO-CICESE_UT3-Plenitas	qwen-think	32.22
UC-UCO-CICESE_UT3-Plenitas	deepseek-qwen	32.22

Table 2: Results for Task 1: Multiple Choice.

Team	Run	Task 2: Matching
Vicomtech	ezotova 2	95.91
Vicomtech	ezotova 1	89.39
Vicomtech	ezotova 4	84.31
Vicomtech	ezotova 5	71.93
Vicomtech	ezotova 3	57.35
djanr2	djanr2 ID 285662	43.77
djanr2	djanr2 ID 285677	43.28
baseline	baseline	40.05
URJC	URJC_TEAM	34.6
djanr2	djanr2 ID 285670	22.25
djanr2	djanr2 ID 285674	21.52
djanr2	djanr2 ID 285672	19.19
UC-UCO-CICESE_UT3-Plenitas	deepseek-qwen	4.89
UC-UCO-CICESE_UT3-Plenitas	qwen-think	4.89
UC-UCO-CICESE_UT3-Plenitas	qwen3.0	4.89

Table 3: Results for Task 2: Matching.

sults, with a best score of 93,52, obtained by two participants (Vicomtech and the Vikings teams). Unfortunately, we did not receive any report from the Vikings team, so we cannot describe their approach.

Vicomtech tested several approaches for this task, including in-context learning, Retrieval Augmented Generation (RAG), agentic RAG, semantic search, and an ensemble of several models. Again, the UC-UCO-CICESE_UT3-Plenitas team uses a DeepSeek-Qwen2.5, a Qwen3.0, and a Qwen3.0-Think model, which obtain the same results.

7 Exam Level

Finally, we include results at the exam level in Table 5. Evaluation at the exam level requires participating in the three subtasks, which were only done by two participants: Vicomtech and UC-UCO-CICESE_UT3-Plenitas. We also include results from our baselines. Scores in this table refer to the percentage of exams passed (with an accuracy of at least 60%) by each system.

As shown in the Table, only submissions from Vicomtech can pass more than half of the exams. Their first four runs passed almost all the exams, with the best system passing 98,55% of the exams.

8 Conclusions and Future Directions

We have described in this paper the main findings of the PROFE@IberLEF 2025 evaluation, which aimed to assess the Spanish language proficiency of automatic systems. We have proposed three subtasks that reuse exams from the Instituto Cervantes examinations: multiple-choice, matching, and fill-in-the-gap. While the first subtask has been previously employed for evaluating Reading Comprehension (RC), the other two are new, to the best of our knowledge, for this purpose. While the multiple-choice substask has attracted most participants, we expect to increase participation in the other two subtasks in the next edition.

Participants mostly employ Large Language Models (LLMs) with different configurations. The most promising results were obtained by combining several LLMs using ensembles or multi-agent architectures, paying special attention to creating adequate prompts tailored to the specific subtask. Other approaches fine-tuned BERT-based models on other RC collections, such as SQuAD or SQAC, or synthetic datasets derived from those collections.

The best results were quite outstanding and close to a perfect performance, demonstrating that current technologies can effectively assess human examinations for RC at different levels of proficiency.

Acknowledgments

This work has been partially supported by the DeepInfo (PID2021-127777OB-C22) and the DeepKnowledge (PID2021-127777OB-C21) projects funded by MCIN/AEI/10.13039/501100011033 and by FEDER.

References

Bandarkar, L., D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August. Association for Computational Linguistics.

Cantero-Romero, M. V. and S. M. Jiménez-Zafra. 2025. SINAI at PROFE 2025: Testing the reading comprehension of GEMINI 2.0 Flash. 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.*

Colelough, B., D. Bartels, and D. Demner-Fushman. 2025. Overview of the cliniqlink 2025 shared task on medical question-answering.

Fraile-Hernzández, J. M. and A. Peñas. 2025. JFra-Team at PROFE 2025: A Multi-Agent Zero-Shot System for Spanish Language Proficiency Question Answering. 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.*

Team	Run	Task 3: Fill in the gap
Vicomtech	ezotova 2	93.52
The Vikings	ralucaginga submission4_G2	93.52
Vicomtech	ezotova 4	92.59
Vicomtech	ezotova 3	89.81
Vicomtech	ezotova 5	89.81
Vicomtech	ezotova 1	88.89
The Vikings	ralucaginga submission5_G3	83.33
baseline	baseline	37.00
UC-UCO-CICESE_UT3-Plenitas	deepseek-qwen	10.19
UC-UCO-CICESE_UT3-Plenitas	qwen-think	10.19
UC-UCO-CICESE_UT3-Plenitas	qwen3.0	10.19

Table 4: Results for Task 3: Filling.

Team	Run	Exam Level
Vicomtech	ezotova 4	98.55
Vicomtech	ezotova 2	98.55
Vicomtech	ezotova 1	97.83
Vicomtech	ezotova 3	94.2
Vicomtech	ezotova 5	55.07
baseline	baseline	44.90
UC-UCO-CICESE_UT3-Plenitas	ymlopez submissiondeepseek-qwen.zip	0.0
UC-UCO-CICESE_UT3-Plenitas	ymlopez submissionqwen-think.zip	0.0
UC-UCO-CICESE_UT3-Plenitas	ymlopez submission.zip	0.0
UC-UCO-CICESE_UT3-Plenitas	ymlopez_submissionqwen3.0.zip	0.0

Table 5: Results for Exam Level.

Geva, M., D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. 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.*

Jiang, Z., J. Araki, H. Ding, and G. Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Martínez-López, Y., M. G. Saborit, Y. Jauriga, I. de las Mercedes Leguen de Varona, J. Madera, A. Rodríguez-González, C. de Castro Lozano, J. M. R. Uceda, and J. C. A. Fernández. 2025. UC-UCO-CICESE_UT3-Plenitas Team - Exploring in the PROFE2025: Language Proficiency Evaluation. 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.

Peñas, A., A. Rodrigo, J. Fruns-Jiménez, I. Soria-Pastor, and J. Reyes-Montesinos. 2025. Ic-uned-rc-es: Spanish reading comprehension dataset developed by instituto cervantes and uned. Zenodo. doi: https://doi.org/10.5281/zenodo.15790139.

Platas, A., A. Bellido, C. Parra, E. Zotova1, P. Turón, and M. Cuadros. 2025. VICOMTECH at PROFE2025: LLMsize is not so important. 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.*

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Rodríguez-García, M. Á., R. Cabido, M. Maes-Bermejo, and S. Montalvo. 2025. URJC-Team at PROFE2025@ IberLEF: Deep Language Comprehension Using Transformers Voting Architecture. 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.*

Rogers, A., O. Kovaleva, M. Downey, and A. Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Winkler, A.-M. and A. Díaz. 2025. NIL-UCM at PROFE 2025: Adapting QA Models to MultipleChoice Tasks. 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.*

y Jordi Armengol-Estapé y Marc Pàmies y Joan Llop-Palao y Joaquin Silveira-Ocampo y Casimiro Pio Carrino y Carme Armentano-Oller y Carlos Rodriguez-Penagos y Aitor Gonzalez-Agirre y Marta Villegas, A. G.-F. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Zhou, C., Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, et al. 2024. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65.