

VICOMTECH at PROFE 2025: LLM Size is not so Important

Alexander Platas^{1,†}, Aingeru Bellido^{1,†}, Cristian Parra^{1,†}, Elena Zotova^{1,*,†}, Pablo Turón^{1,†}
and Montse Cuadros^{1,†}

¹Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

Abstract

This paper presents Vicomtech’s participation in the PROFE 2025 shared task. Our team took part in all three proposed tasks, achieving the best results across the board by leveraging various Large Language Models (LLMs). The main strategies explored include experimenting with LLMs of different sizes, ensemble methods, and sequence-to-sequence approaches. The results demonstrate that while larger LLMs perform exceptionally well across all tasks, smaller models offer competitive performance with significantly lower computational cost, making them a more lightweight and affordable alternative.

Keywords

LLM-prompting, educational NLP, gap filling, multiple-choice question answering, NLP

1. Introduction

Understanding the deep meaning and logic behind natural language remains one of the core challenges in NLP. While numerous benchmarks aim to evaluate reading comprehension, they are often limited in language coverage—focusing predominantly on English—and exhibit high overlap between training and test data, which can overestimate a system’s reasoning abilities.

The PROFE 2025 shared task [1], organized as part of the IberLEF 2025 evaluation campaign [2], addresses these limitations by introducing a multilingual evaluation framework based on real Spanish proficiency exams developed by the Instituto Cervantes. Unlike conventional datasets, this benchmark provides no dedicated training material, requiring systems to generalize through zero-shot or few-shot In-Context Learning (ICL) and generative modelling.

The task comprises three distinct subtasks — multiple-choice, text matching, and gap filling — each inspired by actual exam formats and designed to evaluate different aspects of language understanding. To support model development, we created a custom internal dataset that mirrors these formats, enabling rigorous pre-evaluation and informed design of prompts. Given the absence of organizer-provided training data and the scarcity of similar open-source resources, we addressed this challenge by using machine translation to construct a training set for fine-tuning a Large Language Model (LLM).

This paper presents Vicomtech’s participation in all three subtasks of the PROFE 2025 shared task. Our submission explores a diverse range of strategies centred around large language models, combining models of different sizes, ensemble techniques, and sequence-to-sequence formulations tailored to each task. Rather than relying solely on the most powerful models, we systematically evaluate the trade-offs between model size, efficiency, and performance. Our findings show that while larger LLMs achieve the best results, smaller and more efficient models can still perform competitively, both significantly outperforming traditional sequence-to-sequence and Semantic Textual Similarity (STS) baselines.

The remainder of this paper is structured as follows: Section 2 surveys related work. Section 3 describes the task and dataset construction. Section 4 details our methodology. Section 5 reports

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

[†]These authors contributed equally.

✉ aplatas@vicomtech.org (A. Platas); abellido@vicomtech.org (A. Bellido); cdparra@vicomtech.org (C. Parra); ezotova@vicomtech.org (E. Zotova); pturon@vicomtech.org (P. Turón); mcuadros@vicomtech.org (M. Cuadros)

0009-0002-5501-017X (A. Platas); 0009-0008-4248-9533 (A. Bellido); 0009-0003-6139-0401 (C. Parra); 0000-0002-8350-1331 (E. Zotova); 0000-0002-5563-1120 (P. Turón); 0000-0002-3620-1053 (M. Cuadros)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

experimental results. Finally, Sections 6 and 7 provide a discussion of key findings and outline future directions.

2. Related Work

Reading comprehension tests involve reading a short text passage and answering a series of questions about that text. Automatic evaluation of these tests remains a challenging task. The vast majority of studies are conducted to evaluate English exams. For English, there are diverse Multiple-Choice QA datasets, such as RACE [3], and QuAIL [4].

For Spanish, the following question-answering datasets are available: SQuAD-es [5], a span-based dataset with explicitly stated answers, and Entrance Exams (EE) [6], a multiple-choice dataset requiring reasoning but limited by its small size. UNED-ACCESS 2024 [7] is a bilingual dataset which contains 1,003 questions from various subjects in the UNED Access Course for Over-25s, originally formulated in Spanish and professionally translated into English. ReCoRES dataset [8] extracted from actual university entrance examinations provided by Peruvian institutions that train students for entrance examinations, comprises 439 texts and 1,822 questions with 2-7 candidate answers each.

The methods in the recent studies are mostly based on the transformer architectures [9]. For instance, mT5-based models have been employed in a pipeline encompassing candidate answer extraction, answer-aware question generation, and distractor generation [10]. [11] have investigated the extent to which multilingual models can be trained in one language and applied to another for MCQ tasks. Findings indicate that both monolingual and multilingual models can be zero-shot transferred to different datasets and languages, maintaining performance levels. This approach is beneficial for languages with limited annotated data

The gap-filling task datasets are the following: the Cambridge Exams Publishing Open Cloze (CEPOC) [12], SCDE: Sentence Cloze Dataset with High Quality Distractors From Examinations [13]. The research on gap-filling tasks is mostly represented by a generation of gaps and distractors, and transformer-based methods such as encoder-decoder models [14, 15, 16, 17].

3. Task Description and Datasets

The shared task focuses on the automatic resolution of official Spanish language exams designed by the Instituto Cervantes, targeting learners' levels from A1 to C2. The goal is to develop systems based on language models capable of accurately solving different types of exam exercises.

The task is divided into three subtasks, each corresponding to a specific exercise type commonly found in these exams:

- Subtask 1, Multiple-Choice: Select the correct answer among some options based on a given text.
- Subtask 2, Matching: Match textual fragments from two lists.
- Subtask 3, Gap Filling: Complete gaps from a text by correctly filling them with the missing fragments.

These subtasks aim to evaluate the capabilities of NLP models in understanding and processing language across a wide range of proficiency levels.

To evaluate our proposed approaches, we constructed a benchmark dataset by collecting multiple exercises for each subtask from publicly available online sources. These sources include official exams published by Instituto Cervantes [18], as well as freely accessible educational websites that provide similar Spanish language learning resources.

It is important to emphasize that the collected data was used strictly for evaluation purposes. Our goal was to obtain a reliable estimate of the performance of our models under realistic testing conditions. To ensure the evaluation was representative, we curated exercises for three subtasks and across a broad range of Spanish proficiency levels, from beginner (A1) to proficient (C2).

3.1. Multiple-choice

This subtask involves answering multiple-choice questions based on a reading passage in Spanish. The texts cover general-domain content, ranging from activity schedules at lower proficiency levels to narratives and articles at more advanced levels.

The number of questions per passage and the number of answer options per question may vary depending on the exam level. The overall goal is to evaluate different language models on this task, which requires a dataset with matching characteristics.

We initially used the Cambridge Multiple-Choice Questions Reading Dataset [19] for evaluation. This corpus contains a total of 120 reading comprehension texts in English, covering a wide range of proficiency levels. However, given that the model's performance can vary significantly across languages, we collected a new dataset consisting of 40 reading comprehension texts in Spanish, spanning various proficiency levels, as detailed in Table 1.

Table 1

Distribution of exercises by proficiency level in our evaluation dataset.

Spanish Multiple-Choice Questions Reading Dataset							
Proficiency level	A1	A2	B1	B2	C1	C2	Total
Number of exercises	6	10	13	6	4	1	40

These instances were semi-automatically gathered from preparatory exercises sourced from Instituto Cervantes website¹, Lingua website², Lingolia website³, Inmsol website⁴ and ProfeDeELE website⁵. Each text has multiple-choice questions. Table 2 provides statistics on the number of questions per text and the number of possible answers per question.

Table 2

Descriptive statistics of our multiple-choice evaluation dataset, including the number of texts, questions per text, and answer options per question.

Statistics	Total	Max.	Min.	Mean
Texts	40	-	-	-
Questions	221	10 per text	5 per text	5.53 per text
Possible answers	786	4 per question	2 per question	3.56 per question

3.2. Matching

The matching subtask is a challenge that involves identifying precise correspondences between two lists of text fragments. Each source item must be accurately paired with its target counterpart. While three illustrative examples are initially provided, the development of NLP based solutions requires access to a larger well structured dataset. We collected an evaluation dataset gathered from various online sources and aligned with the specifications required by the challenge. This dataset includes exercises spanning different proficiency levels, as shown in Table 3.

Table 3

Distribution of matching exercises by proficiency level in our evaluation dataset.

Spanish Matching Dataset							
Proficiency level	A1	A2	B1	B2	C1	C2	Total
Number of exercises	7	4	8	5	4	2	30

¹<https://exámenes.cervantes.es/>

²<https://lingua.com/es/espanol/lectura/>

³<https://espanol.lingolia.com/es/comprehension-lectora>

⁴<https://www.inmsol.com/>

⁵<https://www.profedeele.es/exámenes/>

The collected matching exercises consist of two parallel lists of text fragments and a general instruction that outlines the task. The number of items in the source and target sets may vary, and some exercises include distractor texts in the source set that do not correspond to any item in the target set. On average, 6 textual fragments in the source set and 7 items in the target set.

The dataset compiled includes several sources, including the Cervantes Virtual Center⁶, Obejetivo DELE (Diploma de Español como Lengua Extranjera)⁷, the Language Institute of the University of Seville⁸, and the Tía Tula Blog⁹. Part of the data was compiled from official exams used for international certification of Spanish proficiency within the DELE system. The other part contains exercises using an exam model tailored to the matching task. Two common examples of matching exercises are presented on our repository¹⁰.

3.3. Filling the gaps

The “filling the gaps” task involves a text where certain fragments have been removed, with candidates for these gaps presented in a disorderly fashion. Typically, the number of candidate fragments exceeds the number of gaps to be filled¹¹.

This type of exercise demands a high level of language comprehension and, as such, is typically included only in exams at the B1 proficiency level or higher. Students are generally required to engage in iterative reasoning over the available options before selecting the final answer.

To compare different systems designed for this task, a dataset of 20 instances was compiled, representing varied proficiency levels as detailed in Table 4. In the collected dataset, each exercise contains 7 to 8 gaps to be filled, and the number of fragments from which to choose is fixed at 6. Therefore, there are typically 1 to 2 additional fragments included as distractors.

Table 4

Evaluation dataset distribution for “filling the gaps” task.

Spanish Filling the Gaps Dataset						
Proficiency level	B1	B2	C1	C2	C1/C2	Total
Number of exercises	7	3	5	3	2	20

These instances were manually gathered from previous examinations and preparatory exercises sourced from the Instituto Cervantes website¹², Tía Tula Spanish School website⁹ and DELE Ahora Spanish learning website¹³. These sources were chosen due to their official alignment with the DELE exam format and their wide use in preparation contexts.

As this exercise format is specific to these exams, the number of collected instances is limited. As a result, fine-tuning an LLM with such a limited number of instances is challenging; however, the dataset has proven useful for selecting the most suitable ICL strategy (presented in Section 4.3).

4. Proposed Approach

In this section, the proposed approaches are presented. Table 5 lists the LLMs used for each task, including both open-source and commercial models of various sizes. Table 6 presents the embedding models, and Table 7 shows the STS models.

⁶<https://cvc.cervantes.es/>

⁷<https://objetivodele.com/>

⁸<https://institutoideiomas.us.es/>

⁹<https://blog.tiatula.com/2010/03/modelos-de-examen-dele.html>

¹⁰<https://github.com/Vicomtech/profe2025/tree/master/subtask2>

¹¹<https://github.com/Vicomtech/profe2025/tree/master/subtask3>

¹²<https://exámenes.cervantes.es/>

¹³<https://deleahora.com/>

Table 5

Used LLMs throughout subtasks. “–” indicates unknown data.

Model	Open Source	Size	Release Date	Subtask
Claude 3.7 Sonnet [20]	✗	–	Feb, 2025	1,2,3
Gemini 2.0 Flash Thinking [21]		–	Dec, 2024	3
Gemini 2.0 Flash Thinking [22]		–	Jan, 2025	2
Gemini 2.5 Flash [23]	✗	–	Apr, 2025	3
Gemini 2.5 Pro [24]		–	May, 2025	3
o4-mini [25]	✗	–	Apr, 2025	3
GPT-4o [26]		–	May, 2024	2
QwQ [27]	✓	32B	Mar, 2025	1
Qwen 2.5 [28]	✓	72B	Dec, 2024	1
		14B		1,2
		7B		1
Qwen3 [29]	✓	30B-A3B	Apr, 2025	3
		32B		3
DeepSeek R1 [30]		685B		1,3
DeepSeek R1 Distill Qwen [30]	✓	14B	Jan, 2025	1,3
DeepSeek R1 Distill Llama [30]		70B		2
Llama 3.3 [31]	✓	70B	Dec, 2024	1,2,3
Gemma 3 [32]	✓	12B	Mar, 2025	1,2
		4B		1
		27B		3
Mistral Large 2 [33]		123B	Nov, 2024	1,3
Mistral NeMo [34]	✓	12B	Jul, 2024	1
Ministral [35]		8B	Oct, 2024	1,2
Phi 4 [36]	✓	14B	Dec, 2024	1,2,3
		3B		1

Table 6

Used embedding models throughout subtasks.

Model	Open Source	Dimension	Subtask
paraphrase-multilingual-MiniLM-L12-v2 [37]	✓	384	3
paraphrase-multilingual-mpnet-base-v2 [37]	✓	768	2
text-embedding-ada-002 [38]	✗	1536	2
bge-m3 [39]	✓	1024	2
text-embedding-3-small [40]	✗	1536	2

Table 7

Used encoder models throughout subtasks.

Model	Open Source	Dimension	Subtask
deberta-base-long-nli [41]	✓	1024	1
deberta-v3-base-tasksource-nli [41]	✓	768	1
A2T_RoBERTa_SMFA_ACE [42]	✓	1024	1
longformer-base-4096-bne-es-nli [43]	✓	768	1

4.1. Multiple-choice

For multiple-choice task, we explored several approaches ranging from zero-shot In-Context Learning (ICL) with both commercial and open-source LLMs to more traditional language models for STS. Additionally, we conducted fine-tuning experiments with LLMs and implemented ensemble methods using LLMs and encoder models.

4.1.1. Zero-shot ICL using LLMs

For the zero-shot experiments with LLMs, we selected models demonstrating the highest performance according to the current state-of-the-art. Table 5 presents a list of the models evaluated.

We also explored the use of few-shot prompting; however, this approach did not yield significant improvements. This is likely due to the constraint that the model must return only a single letter corresponding to the correct answer, as we issued one prompt per question rather than prompting the model to answer all questions in a given exercise at once.

Furthermore, considering that smaller models (ranging from 8B to 14B) did not perform significantly worse than larger ones, and that their errors occurred on different questions, we implemented an ensemble of LLMs using Gemma 3 (12B), Phi 4 (14B), Qwen 2.5 (14B), and Ministral (8B), aiming to outperform larger models such as DeepSeek R1 (685B). Two ensemble strategies were employed:

1. **Majority voting:** selecting the most frequent answer among the models. In case of a tie, priority was given to the model with the best zero-shot performance.
2. **Random Forest classifier:** training a Random Forest on the predictions of the four models using the training set. The classifier learns to weight each model’s output and selects the most likely correct answer based on learned patterns.

4.1.2. Finetuning LLM

After evaluating the LLMs in the zero-shot setting, we selected the model with the best size-performance trade-off for fine-tuning. We also conducted an exhaustive search for multiple-choice QA datasets in Spanish with contextual information and identified only two:

1. **Belebele** [44]: A human-annotated multiple-choice reading comprehension dataset spanning 122 language variants. In this case, we used the Spanish subset, which contains 900 questions. Each question has four multiple-choice answers and is linked to a short passage.
2. **RetrievalQA** [45]: An automatically generated dataset that contains 196 document-question pairs, where each document is a short text about the history, culture, or other information of a country or region.

Despite the limited data, we performed an initial fine-tuning using LoRA¹⁴. Due to the scarcity of datasets with these specific characteristics, we translated a portion of the ReAding Comprehension dataset from Examinations (RACE) [3] dataset. This dataset is a machine reading comprehension dataset consisting of 27,933 passages and 97,867 questions from English exams. It is divided into middle and high school level questions. We selected the 2,500 questions with the longest contexts from the high school subset, since model performance showed more errors on C1–C2 level exams.

We performed machine translation using two different models:

1. **Itzuli:** A Neural Machine Translation (NMT) system accessible via API upon request¹⁵. The NMT approach has demonstrated its robustness for Basque-Spanish translation [46], but the Itzuli system supports English-Spanish translation as well.
2. **Gemma3** (27B) [32]: A family of lightweight, state-of-the-art open LLM from Google, built from the same research and technology used to create the Gemini models.

We used Itzuli for the reading passages and Gemma3 for the questions and answer choices. This decision was based on the observation that specialized translation models such as Itzuli are more accurate for full-sentence translation from English to Spanish, but often introduce gender, number and

¹⁴<https://github.com/Vicomtech/profe2025/tree/master/subtask1#fine-tuning-hyperparameters>

¹⁵<https://itzuli.vicomtech.org/api/>

verb conjugation errors when translating incomplete sentences. In many cases, the question consists of an incomplete sentence, with the missing part provided among the answer choices. It is important to emphasize the necessity for the answers to be grammatically compatible with the question, as otherwise the model may disregard the correct option if it lacks linguistic coherence. An illustrative example is shown in Table 8.

Table 8

Translation difference of incomplete sentences using a NMT model vs a LLM

Original	Question: According to the article, in the future, 3-D printing technology will probably... A) change the way people make products B) be applied as widely in our daily lives as computers C) forbid many countries to make purchases of weapons D) take the place of normal printers and save lots of energy
NMT	Question: Según el artículo, en el futuro, la tecnología de impresión 3D probablemente... A) cambiar la forma en que la gente hace productos B) se aplique tan ampliamente en nuestra vida diaria como las computadoras C) prohibir a muchos países la compra de armas D) tomar el lugar de las impresoras normales y ahorrar mucha energía
LLM	Question: Según el artículo, en el futuro, la tecnología de impresión 3D probablemente... A) cambiará la forma en que las personas fabrican productos B) se aplicará tan ampliamente en nuestra vida diaria como las computadoras C) prohibirá a muchos países comprar armas D) reemplazará a las impresoras normales y ahorrará mucha energía

4.1.3. Semantic Similarity

This method expects that pretrained STS models are used to measure the semantic closeness between pairs of text and answer options. These models generate embeddings for both the input text and each candidate option. The semantic similarity between each pair is then computed using cosine similarity.

In order to make the result more robust, we calculate the ensemble score with four models, using majority voting and giving the x2 coefficient to the best model. We use the following models:

- deberta-base-long-nli [41] context length of 1280 trained for many tasks, including linguistics-oriented natural language inference (NLI) and zero-shot entailment-based classification tasks.
- deberta-v3-base-tasksource-nli [41] fine-tuned with multi-task learning on 600+ tasks of the tasksource collection. Performed as the best model in the separate evaluation.
- A2T_RoBERTa_SMFA_ACE [42] fine-tuned on multilingual NLI datasets.
- longformer-base-4096-bne-es-nli [43] fine-tuned on NLI-ES dataset¹⁶.

4.2. Matching

To address the matching subtask, an experimental strategy was designed that compares different language models based on zero-shot ICL and embedding representations models. The objective is to evaluate both approaches in terms of how they conceptualize and perform the matching task. LLMs offer dynamic, contextual reasoning, but with a higher computational cost, while embedding-based models allow for faster, similarity-based matching, but with more limited flexibility.

¹⁶<https://huggingface.co/datasets/somosnlp-hackathon-2022/nli-es>

4.2.1. Zero-shot ICL using LLMs

This approach relies on direct reasoning with LLM models, without the need for specific fine-tuning. The system interprets the instructions for each exercise and generates the answer in a single step, applying zero-shot ICL. We implemented the models presented in Table 5. In addition an ensemble version of gemma-3-12b-it + Qwen2.5-14B-Instruct-1M + phi-4 14B was implemented.

4.2.2. Embedding-Based Approaches

This line of experimentation relies on generating semantic vector representations of textual fragments using embedding models. The main objective is to measure the semantic similarity between source and target text, through cosine similarity, in order to identify the most likely matches. Both single-model and ensemble configurations were explored to enhance matching accuracy. Notably, ensemble strategies combined multiple embedding models to obtain more robust similarity estimates, often through weighted voting mechanisms. The implemented models are presented in Table 6, and an ensemble model was implemented (text-embedding-ada-002 + BAAI/bge-m3 + Open AI text-embedding-3-small).

4.3. Filling the gaps

Different systems were designed to address the "filling the gaps" task which included ICL in Section 4.3.1, modern advanced strategies like Retrieval Augmented Generation (RAG) in Section 4.3.2 and Agentic RAG in Section 4.3.3 as well as classical strategies such as ensemble models in Section 4.3.4 and semantic search strategies in Section 4.3.5. For all of these systems the evaluation dataset presented in Table 4 was used in order to compare results. The results are detailed in Figure 5.

4.3.1. Zero-shot ICL using LLMs

The classical ICL approach was initially employed to address the exercises. In this strategy, the task is described to the model through the system prompt, while the user prompt¹⁷ contains the text along with the fragment options. The selected state-of-the-art model generates a response in the correct JSON format without requiring any prior examples of the task, following a zero-shot strategy.

4.3.2. Zero-shot ICL and RAG

To investigate the effectiveness of modern approaches such as RAG, web search capabilities were integrated into the system using the Google Search API¹⁸. The query to the search engine is created from the first line of the text, previous to the first line break which usually corresponds to the title of the text. The API returns a list of links ranked by relevance to the query from which the system extracts the first four thousand characters of content. This retrieved information is then used to construct the user prompt¹⁹ providing the model with contextually relevant content that may enhance the accuracy of its response.

4.3.3. Zero-shot ICL and Agentic RAG

A more advanced RAG approach was designed to get more relevant context from the Search API. In this methodology, an Agentic RAG pipeline was designed to let the LLMs decide whether the contexts retrieved are relevant to complete the exercise or not.

In detail, once the Search API returns a list of links, an LLM is responsible for deciding if the links provided are relevant²⁰. The retrieved information is included in the context if relevant and ruled out if not. This process is repeated in a loop until the context is formed of 3 information sources or


¹⁷<https://github.com/Vicomtech/profe2025/tree/master/subtask3#zero-shot-icl-using-llms>

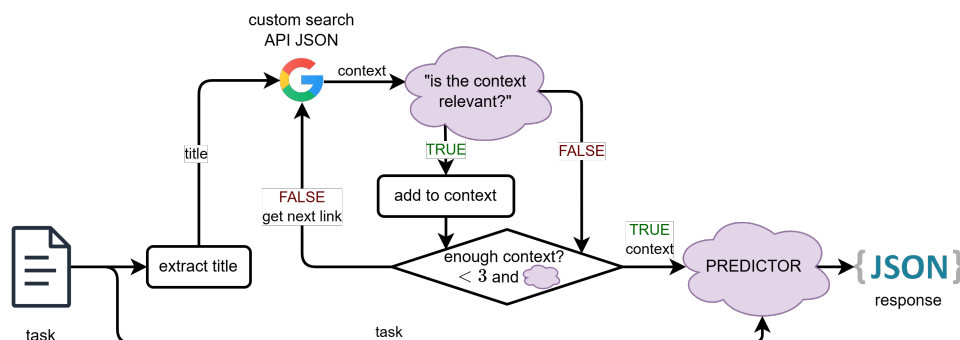
¹⁸<https://developers.google.com/custom-search>

¹⁹<https://github.com/Vicomtech/profe2025/tree/master/subtask3/#zero-shot-icl-and-rag>

²⁰<https://github.com/Vicomtech/profe2025/tree/master/subtask3/#zero-shot-icl-and-agentic-rag>

until another LLM decides that the current context is already enough to get an accurate response, as illustrated in Figure 1.

Figure 1: Zero-shot ICL with Agentic RAG pipeline.  icons are LLMs.



4.3.4. Ensemble models

The results concluded that proprietary models were slightly better than open-source small ones. Aiming to enhance results with open-source models, an ensemble model strategy was implemented, formed of Qwen3 (32B), Phi-4 (14B) and DeepSeek R1 Distill Qwen (14B). These models were selected due to their open-source availability, compatibility with our hardware constraints and for their reasoning capabilities for some of them. The task was completed by the three models using the zero-shot ICL strategy presented in Section 4.3.1 and performing a majority voting from the responses available for each gap.

4.3.5. Semantic matching

A more classic approach was also designed without the use of any modern LLM. We used embedding models to compare the text around the gaps with each of the available fragments using the embedding model paraphrase-multilingual-MiniLM-L12-v2. The gaps were then assigned the most semantically similar fragments. In addition, a more advanced algorithm was designed in order not to assign the same fragment more than once to different gaps.

5. Experimentation and Results

This section describes the experiments conducted for each subtask and the corresponding results.

5.1. Results based on our dataset

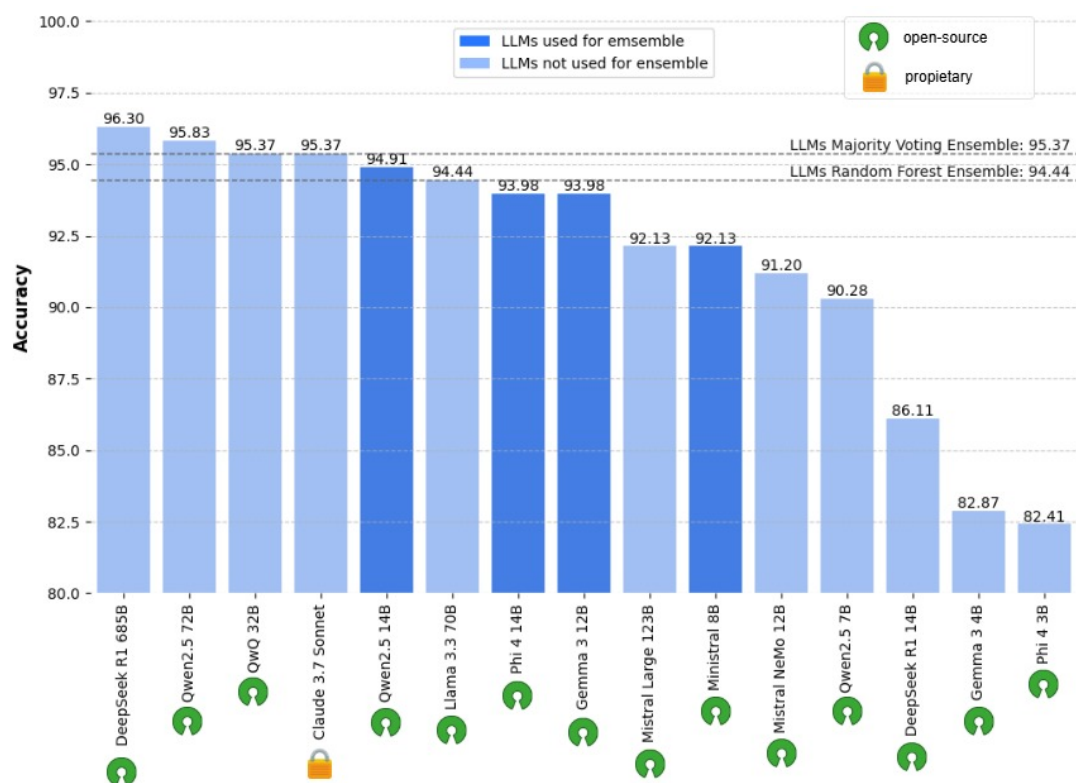
Here we present the results obtained using our collected dataset, as described in Section 3. These preliminary results guided our decision-making process regarding the selection of experiments, choice of models, and the five final runs submitted to the shared task.

5.1.1. Multiple-choice

This section presents the results obtained at each stage of the experimentation, using the collected test set described in Table 1. Figure 2 shows the performance of all LLMs in the zero-shot setting.

Although the largest models achieved the highest scores, the accuracy gain relative to model size is minimal. Models in the 12B–14B range scored only 2–3 points below much larger models such as DeepSeek R1, which has 685B parameters. In general, all models achieved remarkably high performance, with accuracy scores above 85%, except for the smaller models with 3-4 billion parameters, which manage to exceed 80% despite their reduced size.

Figure 2: LLMs and ensembles performance on the multiple-choice task.



Additionally, we observed a clear performance difference based on the release date of the LLMs. The most recently released models, within the past few months, show a considerable improvement in accuracy compared to earlier versions of similar size.

To construct the ensemble, we selected the best-performing medium-sized and small models. As described in Section 4.1.1, two strategies were used to determine the final answer: majority voting and a Random Forest classifier. As shown in Figure 2, both approaches achieved strong performance, with the majority voting ensemble standing out—it outperformed all individual models included in the ensemble.

To select the model for fine-tuning, we analyzed the trade-off between model size and accuracy. In Figure 3, we filtered models with 14B parameters or fewer, identifying three models with strong performance relative to their size: Qwen 2.5 (14B), Gemma 3 (12B), and Mistral (8B).

After analyzing the relationship between model size and performance, we observed the following: on one hand, despite its smaller size, the Mistral model performs only 2–3 points below Gemma and Qwen. On the other hand, although Qwen achieves the best results, it outperforms Gemma by just one point, despite being 2B parameters larger. Considering the fine-tuning cost and potential for improvement, we consider Gemma 3 to be a more viable option.

As described in Section 4.1.2, two fine-tuning stages were conducted: an initial stage using a small dataset (v1), followed by a second stage incorporating translated data (v2). Table 9 reports the results obtained from the fine-tuning experiments.

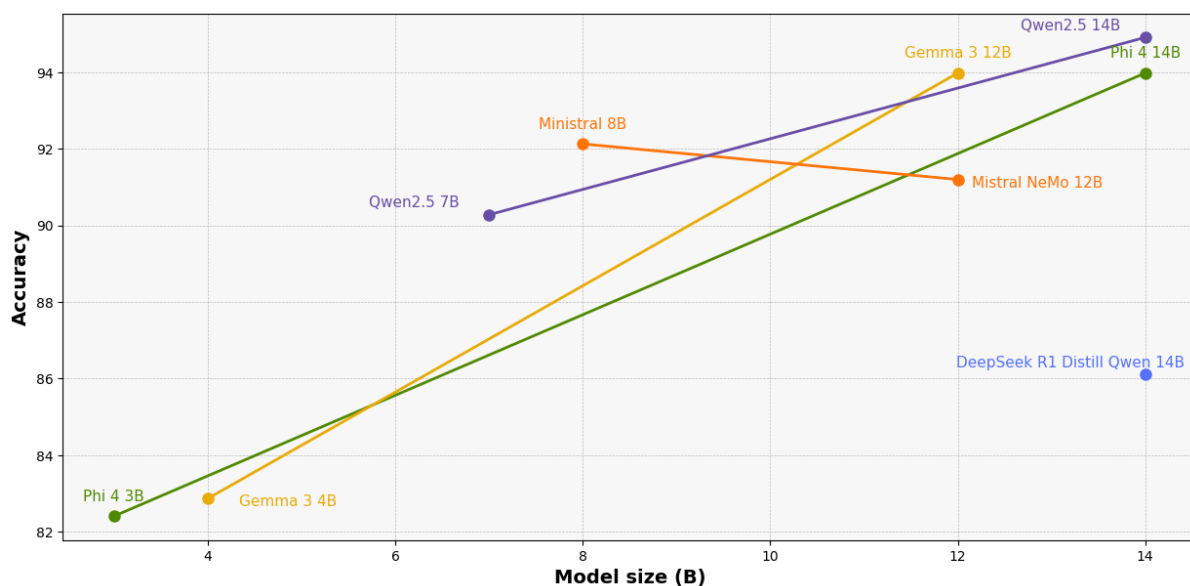
Table 9

Accuracy of different Gemma 3 12B model variants on a multiple-choice task

Model	Training Data	Accuracy
Gemma 3 12B (pre-trained)	–	93.98
Gemma 3 12B (finetuned, v1)	Belebele + RetrievalQA	88.89
Gemma 3 12B (finetuned, v2)	Belebele + RetrievalQA + RACE (es)	88.43

As shown in Table 9, in neither case did fine-tuning improve upon the performance of the base model.

Figure 3: Relation between zero-shot performance and LLMs size.



This may be attributed to the already strong zero-shot capabilities of the model, which achieved a very high score, as well as the possibility that the training data is not sufficiently representative of the test set, either due to translation errors. This is particularly critical given that the task aims to assess language proficiency in Spanish, and any deviation or inaccuracy in the linguistic input may propagate errors and hinder model performance.

It is also possible that the difficulty level of the training examples could be lower than that of the evaluation set. In fact, the only improvements were observed in the performance of the lower-level exams (A1 and A2).

Regarding the STS models, Table 10 shows the zero-shot results obtained with each of them. As can be seen, DeBERTa v3 clearly outperforms the other models. Nevertheless, its performance is far below that of the LLMs, as even the worst-performing LLM (Phi-4 3B) achieves better metrics than the best STS model.

Table 10

STS model and ensemble performance on the multiple-choice task.

STS Model	Accuracy
DeBERTa Base Long NLI	71.95
DeBERTa v3 Long NLI	81.00
A2T RoBERTa SMFA ACE	41.35
Longformer Base	46.94
Ensemble (4 models)	81.00

As with the LLMs, an ensemble using all the STS models was also implemented to surpass their individual performance. As shown in Table 10, the ensemble yields the same results as DeBERTa v3. This suggests that DeBERTa v3 already captures most of the relevant semantic information encoded by the other models.

Considering all the results obtained, we decided to submit different types of solutions to the shared task. First, we observed that LLMs achieve remarkably high zero-shot performance on this task. Therefore, we selected the best-performing model (DeepSeek R1) along with a medium-sized LLM that also performed well (Qwen2.5 14B). This combination aimed to balance performance and computational efficiency across different use cases.

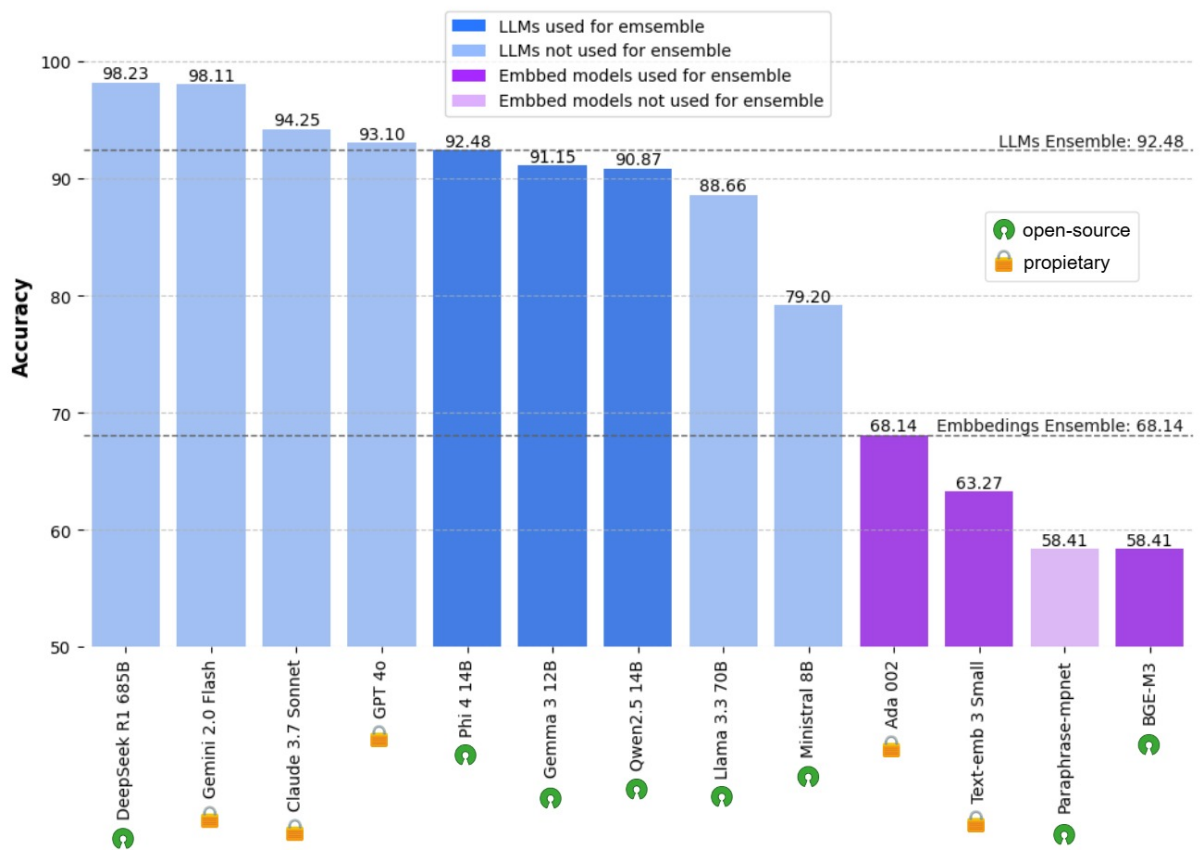
Regarding fine-tuning, since the base model outperformed the fine-tuned versions, we submitted

Gemma3 in its original (non-fine-tuned) form. On the other hand, due to the strong results obtained with LLM ensembling via majority voting, we also submitted this approach. Finally, we chose to include a more traditional solution based on an ensemble of STS models.

5.1.2. Matching

The evaluation metric used was accuracy, measured at each CEFR level and globally. Figure 4 shows results corresponding to LLMs and embedding models in matching task. Among the evaluated models, DeepSeek R1 achieved the highest global accuracy (98.23%), closely followed by Gemini 2.0 Flash (98.11%) and Claude 3.7 Sonnet (94.25%). These results suggest that advanced LLMs are highly capable of performing the matching task in a zero-shot setting, even in the absence of domain-specific fine-tuning. On the other hand, the evaluation of mid-sized models Gemma 3, Phi 4, and Qwen 2.5, reveals a consistent yet slightly lower performance range compared to state-of-the-art large-scale models. Despite their smaller parameter count and reduced inference cost, these models demonstrate a solid capacity for text matching across multiple CEFR levels.

Figure 4: Results for the matching task over our dataset.



While medium-sized models offer a reasonable balance between computational cost and performance, the best results in this task are clearly achieved with larger, higher-capacity LLMs. For high-impact or high-level educational applications, such as automated exam grading, state-of-the-art LLMs such as DeepSeek R1 and Gemini 2.0 Flash are the most reliable options. However, for applications with limited computational resources, Qwen 2.5 and the ensemble strategy represent viable alternatives, especially if refined with task-specific tuning or hybrid matching logic.

In contrast, embedding-based approaches offer an alternative paradigm, relying on semantic similarity metrics rather than direct reasoning. The most effective model was Ada (text-embedding-ada-002), which achieved an overall accuracy of 68.14%, excelling at basic levels but failing at advanced levels.

Models such as text-embedding-3-small and baai-m3 yielded comparable results (with global accuracy scores of 63.27%, and 58.41%, respectively), but exhibited limitations when faced with complex semantic relationships or distracting fragments. Overall, although efficient, embedding models lack the semantic depth necessary for accurate reading comprehension in contexts of greater linguistic complexity.

5.1.3. Filling the gaps

For the approaches detailed in Section 4.3, we employed several of the LLMs listed in Table 5. The corresponding results, evaluated using accuracy as the metric, are presented in Figure 5. Among the evaluated methods, the traditional semantic similarity-based approach yielded significantly lower results (30.05%). In contrast, all other approaches, which leveraged generative LLMs, achieved substantially higher accuracy scores, clearly outperforming the baseline.

Figure 5: Results for the filling the gaps task for different ICL strategies.



When comparing reasoning-enabled LLMs to those without explicit reasoning capabilities, the former demonstrated greater suitability for this task. Analyzing their reasoning outputs it reveals that these models consider multiple possible predictions and, throughout the reasoning process, less plausible options are ruled out. This behaviour closely mirrors the way a human would approach the task through an iterative process that ultimately converges on the most appropriate answer.

The use of a RAG approach also outperformed the results obtained using standalone LLMs. This suggests that LLMs benefit from incorporating external information retrieved from the web, enabling them to generate more accurate responses when provided with relevant context. However, the performance of the agentic RAG approach was not consistently reliable, likely due to the presence of irrelevant or unhelpful information in the retrieved context, which may have introduced noise and hindered task resolution.

The ensemble model did not yield particularly noteworthy results, as its performance did not surpass that of the best individual model (Qwen3-32B) included in the ensemble.

On the other hand, the different approaches and models did not exhibit significant variations in performance across the various exam levels.

In conclusion, for the submission of five selected approaches, we included the best non-RAG method (Gemini 2.5 Pro), the best-performing open-source model (DeepSeek R1), and the model that exhibited the largest performance variations across tasks (Gemini 2.0 Flash Thinking).

5.2. Results on the task dataset

The evaluation follows the official shared task protocol, which uses accuracy (proportion of correct answers) as the primary performance metric. Evaluation scores are reported from two complementary perspectives:

- Question-level accuracy (Acc.), where each question is evaluated independently, and the final score is the proportion of correctly answered questions.
- Exam level, where each exam is composed of multiple exercises spanning different task types. An exam is considered successfully passed if it achieves an accuracy score above 0.6. The overall exam-level score corresponds to the proportion of passed exams across the dataset.

The test set used comprises multiple exams, each consisting of several exercises. These exercises are categorized according to the corresponding subtask. Table 11 shows the number of exercises per subtask and proficiency level.

Table 11

Official evaluation dataset distribution for each task.

IC-UNED-RC-ES dataset										
Proficiency level	A1	A1E	A2	A2B1E	B1	B1E	B2	C1	C2	Total
Subtask 1 - Multiple-choice	18	8	245	4	206	36	97	6	54	674
Subtask 2 - Matching	18	5	52	4	8	0	5	3	20	115
Subtask 3 - Filling the gaps	0	0	0	0	8	0	5	3	2	18

Table 12 summarizes the performance of systems across the three subtasks as well as the overall exam-level accuracy. Notably, systems incorporating LLMs achieve substantially higher accuracy across all subtasks. For instance, DeepSeek R1, Gemini 2.0 Flash and Gemini 2.5 Pro achieves the best performances in each task, respectively.

Table 12

Results over the task dataset. “–” indicates unknown data. “Acc” indicates Accuracy (%). **Bold** values indicate the best performance per subtask. “*” indicates Agentic RAG.

Subtask 1 Multiple-Choice		Subtask 2 Matching		Subtask 3 Fill the gap		Exam Level
System	Acc	System	Acc	System	Acc	
baseline	64.00	baseline	51.00	baseline	43.00	–
DeepSeek R1	95.54	DeepSeek R1	89.93	DeepSeek R1	88.89	97.83
Gemma 3 12B	87.32	Gemini 2.0 Flash	95.91	Gemini 2.5 Pro	93.52	98.55
Qwen 2.5 14B	90.90	Ada embeddings	57.35	Gemini 2.0 Flash	89.81	94.20
LLM Ensemble	89.67	LLM Ensemble	84.31	Gemini 2.0 Flash + RAG	92.59	98.55
NLI Ensemble	55.34	Embedding Ensemble	71.93	Gemini 2.0 Flash + RAG*	89.81	55.07

At the exam level, the top-performing systems consistently surpass 98% accuracy. This demonstrates that employing LLMs yields the most effective results for this task. Conversely, approaches based on embeddings or text similarity techniques, while generally competitive, fall notably short of the performance achieved by LLM-based models.

The pronounced performance gap between baseline and advanced systems highlights the substantial advantage of utilizing pretrained language models and ensemble strategies. Nevertheless, the variability in accuracy observed across the subtasks reflects the distinct challenges and complexities inherent to each task type, indicating areas that merit deeper exploration in future work.

6. Discussion

We present the results and resources of the PROFE shared task on solving a variety of exercises extracted from official Spanish language exams. For each subtask, only a single example per exercise type was available, limiting the viability of personalized solutions such as fine-tuning or any training-based approaches. Furthermore, the highly specific nature of the tasks made it difficult to identify appropriate open-source resources without relying on synthetic data generation or machine translation.

The results from this shared task clearly illustrate the significant progress made by LLMs in recent years. These models achieved remarkably high performance (often exceeding 85% accuracy) in zero-shot settings, especially those with advanced reasoning capabilities. This ability to self-evaluate and refine their outputs without requiring few-shot prompting makes them particularly appropriate for scenarios with limited training data.

Regarding the evaluation data, some issues were identified in the dataset, such as duplicated instances or incomplete text fragments. Interestingly, these irregularities were often only effectively handled by LLM-based approaches, which were capable of inferring the intended structure or context through their generative reasoning. Table 13 presents an example of the reasoning process exhibited by DeepSeek R1 when confronted with a multiple-choice question in which the correct answer was located within a missing portion of the input text.

Table 13

Example of DeepSeek’s reasoning process. Irrelevant information has been omitted and is indicated by “[...]”.

DeepSeek-R1 Reasoning (A1E_2017-11-24_E4)
<p>Okay, let’s see. The question is asking "Hay una fiesta el ____". So I need to find which day has a party mentioned in the schedule. Looking at the text, each day is listed with activities. Let me check each day one by one. Starting with LUNES (Monday). The activities are a bus departure, walking tour, lunch, free time, concert, dinner. No mention of a fiesta here. MARTES (Tuesday): Breakfast, boat ride, lunch, free time, football match, dinner. Again, no fiesta. MIÉRCOLES (Wednesday): Breakfast, museum visit, lunch, free time, cinema, dinner. Still no party. JUEVES (Thursday): Breakfast at the pool cafeteria, mountain excursion, lunch. Wait, the Lunch here has two entries? 14:00–15:00 Comida: Pescado con ensalada y Postre. Then another Comida: Sopa, carne y Fruta. That might be a typo, but after that, free time, then merienda (snack) at 17:00–19:00. Then at 19:30, Fiesta de despedida. Oh, that’s the farewell party. So Jueves (Thursday) has the fiesta. But wait, the options given are A) Lunes, B) Viernes, C) Sábado. Wait, Jueves is Thursday, which is not listed. The options are Monday, Friday, Saturday. Hmm. Wait, the days listed in the text are Lunes, Martes, Miércoles, Jueves, Sábado. Wait, where’s Viernes (Friday)? The schedule skips from Jueves (Thursday) to Sábado (Saturday).</p> <p>[...]</p> <p>Therefore, I’ll choose B) viernes. But I’m not confident. Alternatively, maybe the answer is C) sábado. But according to the text, it’s not. Therefore, I think the answer is just not present. But since we’re supposed to fill it, I’ll choose B) viernes.</p>
Correct Answer: B) viernes

Despite the lack of this fragment, the model was able to reconstruct plausible content based on the adjacent context and use this inferred information to select the correct option. A solution based solely on semantic similarity would be unable to resolve such cases correctly, as the target word does not appear anywhere in the text and must instead be intuited through higher-level reasoning. This highlights the LLM’s capacity for contextual extrapolation and implicit reasoning, which proves especially valuable in scenarios where input data is incomplete, noisy, or partially corrupted.

Moreover, given that this is a general-domain task where the only requirement is a strong command and understanding of the Spanish language, LLMs—trained on vast multilingual corpora—demonstrate a clear advantage over traditional approaches such as semantic similarity methods. Their deep understanding of linguistic nuances, combined with robust reasoning and contextual interpretation, positions them as far superior in these types of exercises without the need for domain-specific tuning.

Additionally, analysis of the results revealed no consistent performance drop at higher proficiency levels of the exam. Instead, the accuracy metrics varied regularly across models, regardless of difficulty. This suggests that LLMs already have a high level of competence in Spanish, and that remaining errors are more likely attributable to factors such as complexity, ambiguity or longer contextual dependencies rather than lack of linguistic understanding.

These findings suggest that LLMs, even without task-specific adaptation, are capable of handling complex linguistic tasks with surprising robustness.

7. Conclusions and Future work

The difficulty of the Spanish exams has been reflected in the different subtasks. Nevertheless, our evaluation confirms that state-of-the-art LLMs are now capable of solving these tasks with accuracy scores approaching the upper bound using ICL techniques. In contrast, traditional approaches lag significantly behind and fail to achieve comparable performance. These tasks are designed to require students to reason over the available options, a demand that is mirrored in the superior performance of LLMs with reasoning capabilities.

On the other hand, current LLMs have demonstrated the ability to handle these complex tasks in Spanish effectively. Notably, this capability is not limited to large proprietary models; smaller, open-source LLMs can also attain comparable results on this type of task.

Although fine-tuning has traditionally been an effective strategy for adapting models to specific tasks, in this case, it has proven ineffective. As such, future work may focus on enhancing ICL methods—such as prompt engineering—and refining the application of modern techniques for LLM-based reasoning.

Acknowledgements

This work has been supported by the project MASTERMIND ZL-2025/00267 funded by the government of the Basque Country, Spain.

Declaration on Generative AI

Generative AI (ChatGPT with GPT-4.5) was utilised to: Improve writing style, Abstract drafting, Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] 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).
- [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] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding comprehension dataset from examinations, in: M. Palmer, R. Hwa, S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 785–794. URL: <https://aclanthology.org/D17-1082/>. doi:10.18653/v1/D17-1082.
- [4] A. Rogers, O. Kovaleva, M. Downey, A. Rumshisky, Getting closer to ai complete question answering: A set of prerequisite real tasks, *Proceedings of the AAAI Conference on Artificial*

- Intelligence 34 (2020) 8722–8731. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6398>. doi:10.1609/aaai.v34i05.6398.
- [5] C. P. Carrino, M. R. Costa-jussà, J. A. R. Fonollosa, Automatic Spanish translation of SQuAD dataset for multi-lingual question answering, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5515–5523. URL: <https://aclanthology.org/2020.lrec-1.677/>.
 - [6] A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, R. Sutcliffe, R. Morante, Qa4mre 2011–2013: Overview of question answering for machine reading evaluation, in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 303–320.
 - [7] E. Sánchez Salido, R. Morante, J. Gonzalo, G. Marco, J. Carrillo-de Albornoz, L. Plaza, E. Amigo, A. F. García, A. Benito-Santos, A. Ghajari Espinosa, V. Fresno, Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 6184–6200. URL: <https://aclanthology.org/2025.coling-main.413/>.
 - [8] M. A. S. C. y Diego Diestra y Rodrigo López y Erasmo Gómez y Arturo Oncevay y Fernando Alva-Manchego, Overview of recorres at iberlef 2022: Reading comprehension and reasoning explanation for spanish, *Procesamiento del Lenguaje Natural* 69 (2022) 281–287. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6448>.
 - [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 1–11. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
 - [10] D. de Fitero-Dominguez, A. Garcia-Cabot, E. Garcia-Lopez, Automated multiple-choice question generation in spanish using neural language models, *Neural Computing and Applications* 36 (2024) 18223–18235. URL: <https://doi.org/10.1007/s00521-024-10076-7>. doi:10.1007/s00521-024-10076-7.
 - [11] G. E. y Álvaro Rodrigo y Anselmo Peñas, Cross-lingual training for multiple-choice question answering, *Procesamiento del Lenguaje Natural* 65 (2020) 37–44. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6274>.
 - [12] M. Felice, S. Taslimipour, Ø. E. Andersen, P. Buttery, Cepoc: The cambridge exams publishing open cloze dataset, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 4285–4290.
 - [13] X. Kong, V. Gangal, E. Hovy, SCDE: Sentence cloze dataset with high quality distractors from examinations, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5668–5683. URL: <https://aclanthology.org/2020.acl-main.502/>. doi:10.18653/v1/2020.acl-main.502.
 - [14] B. Moharana, V. K. Singh, T. Sarkar, D. Singh, M. Rakhra, V. K. Pandey, Automated questions answering generation system adopting nlp and t5, in: 2024 International Conference on Cybernation and Computation (CYBERCOM), 2024, pp. 363–369. doi:10.1109/CYBERCOM63683.2024.10803238.
 - [15] S. K. Bitew, J. Deleu, A. S. Doğruöz, C. Develder, T. Demeester, Learning from partially annotated data: Example-aware creation of gap-filling exercises for language learning, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 598–609. URL: <https://aclanthology.org/2023.bea-1.51/>. doi:10.18653/v1/2023.bea-1.51.

- [16] M. P. P. Jadhav, M. M. D. Laddha, An automatic gap filling questions generation using nlp, *Ijcset. Com* 8 (2017).
- [17] C. Y. Yeung, J. S. Lee, B. K. Tsou, Difficulty-aware distractor generation for gap-fill items, in: *Proceedings of the 17th annual workshop of the Australasian language technology association*, 2019, pp. 159–164.
- [18] Instituto Cervantes, Página oficial del instituto cervantes, <https://www.cervantes.es>, 2025. Consultado el 26 de mayo de 2025.
- [19] A. Mullooly, O. Andersen, L. Benedetto, P. Buttery, A. Caines, M. J. F. Gales, Y. Karatay, K. Knill, A. Liusie, V. Raina, S. Taslimipoor, The Cambridge Multiple-Choice Questions Reading Dataset, Cambridge University Press and Assessment, 2023. URL: <https://www.repository.cam.ac.uk/handle/1810/358683>. doi:10.17863/CAM.102185.
- [20] Anthropic, Claude 3.5 sonnet model card addendum, 2025. URL: <https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf>.
- [21] Google, Gemini 2.0 flash, 2024. URL: <https://deepmind.google/technologies/gemini/flash/>.
- [22] Google, Gemini 2.0 flash exp. 01-21, 2025. URL: <https://ai.google.dev/gemini-api/docs/models?hl=es-419>.
- [23] Google, Gemini 2.5 flash preview, 2025. URL: <https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash-preview.pdf>.
- [24] Google, Gemini 2.5 pro preview, 2025. URL: <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro-preview.pdf>.
- [25] OpenAI, Openai o3 and o4-mini system card, 2025. URL: <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.
- [26] OpenAI, Hello gpt-4o, 2024. URL: <https://openai.com/index/hello-gpt-4o/>.
- [27] Q. Team, Qwq-32b: Embracing the power of reinforcement learning, 2025. URL: <https://qwenlm.github.io/blog/qwq-32b/>.
- [28] Q. Team, Qwen2.5: A party of foundation models!, 2024. URL: <https://qwenlm.github.io/blog/qwen2.5/>.
- [29] Q. Team, Qwen3 technical report, 2025. URL: <https://arxiv.org/abs/2505.09388>. arXiv:2505.09388.
- [30] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: <https://arxiv.org/abs/2501.12948>. arXiv:2501.12948.
- [31] M. Llama Team, The llama 3 herd of models, 2024. URL: <https://arxiv.org/abs/2407.21783>. arXiv:2407.21783.
- [32] G. Team, Gemma 3 technical report, 2025. URL: <https://arxiv.org/abs/2503.19786>. arXiv:2503.19786.
- [33] M. A. team, Mistral large 2 (2407), 2024. URL: <https://mistral.ai/news/mistral-large-2407>.
- [34] M. A. team, Mistral nemo, 2024. URL: <https://mistral.ai/news/mistral-nemo>.
- [35] M. A. team, Ministral, 2024. URL: <https://mistral.ai/news/ministraux>.
- [36] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024. URL: <https://arxiv.org/abs/2412.08905>. arXiv:2412.08905.
- [37] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: <https://aclanthology.org/D19-1410/>. doi:10.18653/v1/D19-1410.
- [38] OpenAI, Text embedding ada 002, 2022. URL: <https://openai.com/index/new-and-improved-embedding-model/>.
- [39] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL*

- 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. URL: <https://aclanthology.org/2024.findings-acl.137/>. doi:10.18653/v1/2024.findings-acl.137.
- [40] OpenAI, Text embedding 3-small, 2024. URL: <https://openai.com/index/new-embedding-models-and-api-updates/>.
- [41] D. Sileo, tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 15655–15684. URL: <https://aclanthology.org/2024.lrec-main.1361>.
- [42] O. Sainz, H. Qiu, O. Lopez de Lacalle, E. Agirre, B. Min, ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations, in: H. Hajishirzi, Q. Ning, A. Sil (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 27–38. URL: <https://aclanthology.org/2022.naacl-demo.4/>. doi:10.18653/v1/2022.naacl-demo.4.
- [43] Sleoruz, Huggingface model, 2022. URL: <https://huggingface.co/Sleoruz/longformer-base-4096-bne-es-nli>.
- [44] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, M. Khabsa, The belebele benchmark: a parallel reading comprehension dataset in 122 language variants, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2024, p. 749–775. URL: <http://dx.doi.org/10.18653/v1/2024.acl-long.44>. doi:10.18653/v1/2024.acl-long.44.
- [45] W. Wang, Retrievalqa: A benchmark dataset for retrieval-augmented question answering, https://huggingface.co/datasets/lnwang/retrieval_qa, 2023. https://github.com/wln20/Retrieval_QA.
- [46] T. Etchegoyhen, E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, E. Calonge, Neural machine translation of Basque, in: J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. Van den Bogaert, M. L. Forcada (Eds.), Proceedings of the 21st Annual Conference of the European Association for Machine Translation, Alicante, Spain, 2018, pp. 159–168. URL: <https://aclanthology.org/2018.eamt-main.14>.

A. Online Resources

All detailed technical information—including per-proficiency-level metrics for all models, the prompts used with all LLM approaches, and sample data for each subtask—is available in our GitHub repository [Vicomtech/profe2025](https://github.com/Vicomtech/profe2025) to facilitate the reproducibility of our experiments.