NIL-UCM at PROFE 2025: Adapting QA Models to Multiple-Choice Tasks

Anna-Maria Winkler[1], Alberto Díaz[2]

[1] *Facultad de Filología, Universidad Complutense de Madrid, Spain*
[2] *Facultad de Informática and ITC, Universidad Complutense de Madrid, Spain*

Abstract

These working notes evaluate pre-trained transformer models—BERT-large, RoBERTa-base, and RoBERTa-large—for multiple-choice question answering in Spanish reading comprehension exams. Using the IC-UNED-RC-ES dataset, models were tested in a zero-shot setting with a token-matching approach. Exploratory experiments with semantic similarity, entailment, and generative prompts highlighted both limitations and future potential. The results underline the need for fine-tuning, robust prompting, and semantic alignment for reliable educational NLP applications.

Keywords

multiple choice, Spanish reading comprehension, NLP, transformer models, BERT, RoBERTa, generative language models

Introduction

In recent years, the exponential growth of large language models (LLMs) has revolutionized the field of Natural Language Processing (NLP), enabling major advancements across tasks such as machine translation, summarization, and question answering. This study focuses on the application of these models to multiple-choice question answering (MCQA) tasks that involve short fictional or simplified contexts, similar to those found in educational assessments. Specifically, we explore how different types of language models—encoder-only versus generative—perform in this setting, and what challenges arise from task design, model architecture, and prompt formula-tion.

The motivation for this work is rooted in the increasing relevance of AI-assisted tools in education. MCQA is widely used in academic and professional settings, but correcting such questions can be time-consuming. At the same time, students increasingly rely on AI systems for learning and test preparation. Understanding how current models perform in MCQA tasks is key to developing effective and fair educational technologies.

Our experiments compare the performance of BERT [1], RoBERTa [2], and Mistral [3] on a variety of MCQA formats. We examine the impact of prompt design, model architecture, and output interpretation on model accuracy and reliability.

1. Task description

1.1. Subtask: Multiple Choice Question Answering

This work is part of the PROFE 2025 shared task [4], proposed as part of IberLEF 2025 [5]. This task focuses on evaluating the ability of AI systems to solve language comprehension exercises originally designed for human learners. PROFE 2025 is divided into three subtasks, each corresponding to a different type of exercise commonly used in language assessment. In this paper, we focus exclusively on the multiple-choice subtask, where the objective is to correctly answer questions based on a short reading passage.

IberLEF 2025, September 2025, Zaragoza, Spain
annamawi@ucm.es (A.-M. Winkler); adiazest@ucm.es (A. Díaz)
0009-0001-8224-9437 (A.-M. Winkler); 0000-0003-1966-3421 (A. Díaz)

In this subtask, each instance consists of:

- A short context paragraph (text)
- A question based on the text
- A set of predefined answer choices (typically A–D)

Only one option is correct per question. The task requires systems to select the correct answer based on the information in the text. This mirrors real-world reading comprehension exercises used in language proficiency testing.

1.2. Dataset: IC-UNED-RC-ES

For this study, we use the IC-UNED-RC-ES dataset [6], which contains reading comprehension tasks extracted from real exams created by the Instituto Cervantes for the evaluation of Spanish as a Foreign Language (ELE). These exams are designed by human experts and cover levels from A1 to C2 according to the Common European Framework of Reference for Languages (CEFR). The exams span various difficulty levels (A1–C2) and were converted into machine-readable format as part of a collaboration between the Instituto Cervantes and UNED, under a formal agreement signed in May 2021 and funded by the DeepInfo Project (AEI PID2021-127777OB-C22).

In total, the full dataset comprises 282 exams, 855 exercises, and 6,146 annotated responses (from 16,570 possible options). For PROFE 2025, approximately 50% of the data is used, while the rest is reserved for future editions to prevent overfitting and contamination by public LLMs. The gold standard labels are not publicly distributed for the same reason.

1.3. Types of Questions Tackled

In this paper, we focus on standard multiple-choice questions where only one answer is correct. These questions are:

- Taken directly from ELE exams created by Instituto Cervantes.
- Designed to assess reading comprehension and general language understanding.
- Spanning levels from A1 to B2, where texts are short, often fictional or semi-authentic, and the questions require inference, reasoning, and detail recognition.
- A typical example includes a brief personal letter followed by four or five questions, each with four possible answers. The distractors are carefully constructed to appear plausible, making the task non-trivial for AI models.

1.4. Task Challenges

This task presents several challenges:

- Contextual dependence: Many questions require precise understanding of the context and implicit reasoning.
- Ambiguity: Distractor options can be semantically close to the correct answer.
- Varying difficulty: Questions span CEFR levels from A1 (basic) to B2 (upper-intermediate), introducing variation in vocabulary complexity and required inference.
- Answer format: Models must return the correct letter label (e.g., “B”).

In summary, this task provides a realistic benchmark for evaluating how well current NLP models—both encoder-based and generative—can perform in educational language assessment contexts.

2. Methodology

2.1. Models Explored

To evaluate the effectiveness of different pre-trained language models on contextual multiple-choice question answering (MCQA), we tested three transformer models:

- **RoBERTa-base** (deepset/roberta-base-squad2)
- **RoBERTa-large** (deepset/roberta-large-squad2)
- **BERT-large** (bert-large-uncased-whole-word-masking-finetuned-squad)

All three models are pretrained on the SQuAD dataset and follow an extractive QA paradigm, returning a text span from the context that best answers the given question. None of these models were fine-tuned on the PROFE dataset; instead, they were used in a zero-shot evaluation setup.

2.2. Adapting QA Models to Multiple-Choice Tasks

As these models are trained to return free-text answers, an answer-matching strategy was required to map their outputs to one of the predefined multiple-choice options.

- For each question, the model receives the context and question as input and returns an answer span. This span is then normalized using the following steps:
- Lowercasing and Unicode normalization (removing diacritics).
- Removal of punctuation and splitting into tokens.
- Filtering out Spanish stopwords (via NLTK).

Each multiple-choice option is normalized in the same way. We then calculate the token overlap score between the model's answer and each of the options. The option with the highest score is selected as the model's final prediction for that question. This heuristic provides a lightweight but effective way to adapt extractive QA models to the classification nature of MCQA tasks.

2.3. Prompting Strategy for Generative Models (Future Work)

Although the focus of this work is on encoder-based models, we also explored the use of prompt-based methods with decoder-only generative models, such as Mistral, as a complementary approach. The goal was to evaluate how generative language models perform in multiple-choice question answering when prompted in natural language. We designed a set of dynamic prompts, automatically generated for each example, that included the reading comprehension text, the question, and all answer choices.

However, due to time constraints and the additional complexity involved in integrating and evaluating generative models (e.g., handling variations in output, rate limits, prompt tuning), this approach was not fully implemented or evaluated within the current phase of the project. It remains as future work, with the potential to compare generation-based answers to the results of extractive models.

2.4. Technical Implementation

All experiments were implemented in Python using Google Colab as the execution environment. The dataset was loaded from Google Drive, and predictions were saved back in .json format. The following tools and libraries were used:

- Hugging Face Transformers: loading and executing QA pipelines for all models.
- NLTK: Spanish stopword removal for normalization.
- Standard Python libraries: json, unicodedata, re, and os for file and text handling.

Each model was executed using the pipeline() utility from Hugging Face: Model outputs were processed in a loop across all questions in the test dataset, and final predictions were stored in a dictionary indexed by questionId.

3. Results and Discussion

3.1. Final Evaluation Results

The final evaluation results were significantly lower than expected. On the test set, all three models—bert-large-uncased-whole-word-masking-finetuned-squad, deepset/roberta-base-squad2, and deepset/roberta-large-squad2—achieved nearly identical accuracy scores. Both BERT and RoBERTa-base reached 32.28%, while RoBERTa-large performed only marginally better with 32.51%. These figures are surprisingly low given earlier development experiments and indicate a possible mismatch between the development and test distributions.

One particularly noticeable pattern was the disproportionately high number of predictions for option A across questions. This suggests that the overall accuracy may partly reflect the natural distribution of correct answers rather than actual model understanding. In other words, the models might have been defaulting to option A when uncertain or when the token-matching heuristic failed, which would artificially inflate scores if A happened to be correct more often.

3.2. Exploratory Experiments on the Development Set

During an earlier experimental phase, several alternative methods were tested on a separate development set. An overview of the models and their main results is presented in Table 1. These approaches yielded considerably better results and provided insights into model behaviour under different configurations. The most successful method combined RoBERTa-base with a straightforward token-matching heuristic, reaching an accuracy of 70%. Despite its simplicity, this method proved stable and precise across a variety of input structures.

Other experiments explored semantic similarity through embeddings. A version using MiniLM to compare the cosine similarity between the predicted span and the answer options reached 50% accuracy. However, this approach performed less reliably when the context was short or when answers shared overlapping vocabulary.

Generative models also showed promise. Using google/flan-t5-large in a free-text generation setup led to 60% accuracy on the development set. The model demonstrated good semantic understanding, but its performance was sensitive to input length and prompt truncation. Another approach using facebook/bart-large-mnli to compute textual entailment between the context-question pair and each answer option also achieved 50%, particularly effective when the distractors were clearly distinguishable.

3.3. Prompting Limitations and Generative Model Issues

Initial attempts to include a generative model such as Mistral in the final evaluation were ultimately not successful. The prompt templates used in early tests did not explicitly include the letter labels (A, B, C, D) associated with the answer choices. As a result, the model frequently defaulted to generating "A" regardless of the content, or returned ambiguous outputs such as “Not detected.” These issues underline the importance of prompt structure when working with generative language models.

Due to time constraints and the need for more robust prompt engineering and post-processing logic, generative models were excluded from the final comparison. However, the exploratory results suggest that, with better formatting and output handling, generation-based models could become competitive alternatives to extractive methods for this task.

3.4. Interpretation and Future Directions

The contrast between the final test results and earlier development experiments highlights several important considerations. First, pre-trained QA models—even those fine-tuned on SQuAD—do not generalize well to multiple-choice tasks without task-specific adaptation. Second, naive matching strategies like token overlap are insufficient in cases where distractor options are semantically similar. Third, the design of prompts and answer formats plays a critical role in how models interpret and respond to inputs.

Future work should explore fine-tuning encoder-based models on a subset of the PROFE dataset, implementing semantic similarity measures beyond token matching, and developing more reliable prompting strategies for generative models. With these improvements, it may be possible to significantly raise the performance ceiling on this type of contextual MCQA task.

Table 1
Results from the exploratory experiments

Method	Main model	Accuracy	Observations
RoBERTa + text matching	deepset/roberta-base-squad2	70%	Most stable and precise configuration.
RoBERTa + embeddings (cosine similarity)	deepset/roberta-base-squad2 + MiniLM	50%	Performed poorly in short and ambiguous contexts.
Generative (FLAN-T5-LARGE)	google/flan-t5-large	60%	Better semantic reasoning but affected by input truncation.
Zero-shot multiple-choice (NLI-based)	facebook/bart-large-mnli	50%	Work best with explicit and well-phrased distractors.

4. Conclusions and Future Work

The main objective of this study was to evaluate the performance of different NLP approaches in answering multiple-choice reading comprehension questions in Spanish, using the IC-UNED-RC-ES dataset provided by the ProFE 2025 shared task. Our experiments show that RoBERTa outperformed the other tested models in terms during the development phase, but the final test performance was notably low, with all models achieving just over 32% accuracy. Therefore, this study shows the challenges of applying extractive QA models to multiple-choice tasks without fine-tuning or more robust adaptation strategies.

Several areas for improvement were identified. First, performance could likely be enhanced through few-shot prompting, fine-tuning on a representative subset of the PROFE dataset, or more advanced semantic matching techniques. The evaluation also highlighted the limitations of using a static development/test split. Future iterations should explore a dynamic or stratified sampling approach to ensure consistency in task difficulty and domain distribution, minimizing the risk of overfitting to a particular data slice.

Although generative models such as Mistral were not fully integrated into the final evaluation for the competition due to time constraints, an experiment using an improved prompt, with a subset of 49 questions from the test set, has shown a promised 62.81% accuracy. Future work should continue exploring different prompting strategies and output formats to improve performance and stability in generation-based setups.

Beyond experimental refinements, this research has broader implications for educational applications. Accurate automatic MCQA systems could support large-scale assessment, adaptive learning environments, and AI-based tutoring systems. However, ensuring fairness, robustness, and transparency remains essential for their responsible deployment.

Acknowledgements

This publication is part of the R&D&I project HumanAI-UI, Grant PID2023-148577OB-C22 (Human-Centered AI: User-Driven Adaptative Interfaces-HumanAI-UI) funded by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

During the preparation of this work, the authors used **ChatGPT** in order to: **Grammar and spelling check**, **Paraphrase and reword**. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. arXiv. https://arxiv.org/abs/1810.04805

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach [Preprint]. arXiv. https://arxiv.org/abs/1907.11692

[3] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B [Preprint]. arXiv. https://arxiv.org/abs/2310.06825

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

[5] González-Barba, J. Á., Chiruzzo, L., & Jiménez-Zafra, S. M. (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.

[6] Alcántara, C., & Pérez, F. (2025). IC-UNED-RC-ES: Recursos para comprensión de lectura y preguntas de opción múltiple en español [Dataset]. In Proceedings of the IberLEF and ProFE 2025 Workshop. Association for Computational Linguistics. https://www.aclweb.org/portal/content/profe-2025-iberlef-2025-call-participation