David_ at PROFE 2026: Text-Only Prompting of a Quantized Instruction-Tuned LLM for the Spanish Reading-Comprehension Matching Subtask

David Alejandro Bojaca[1]

[1]*Universidad Carlos III de Madrid, Spain*

Abstract

This paper describes our participation in the Matching subtask of PROFE 2026 (Spanish Language Proficiency Evaluation) at IberLEF 2026. The subtask reuses real Instituto Cervantes proficiency exams (levels A1–C2) and provides no task-specific training data, encouraging transfer-learning and generative Large Language Model (LLM) approaches. A novelty of this edition is that some exercises contain images that may or may not carry information needed to answer. We frame matching as a zero-shot, structured-output problem and solve it with a single open-weights instruction-tuned model, `Qwen2.5-7B-Instruct`, loaded in 4-bit precision so that the whole pipeline runs on a single 16 GB GPU. The model receives a Spanish free-form reasoning prompt and emits a JSON object mapping every question to one option; a validated parser with fallbacks guarantees a well-formed submission. Our text-only system reaches **66.67%** accuracy on the hidden test set, **+15.6** points over the official ChatGPT baseline (0.51). We additionally ran a controlled comparison against a multimodal variant (`Qwen2.5-VL-7B-Instruct`); contrary to intuition, naively feeding the images *lowered* accuracy, because most images in this subtask are decorative rather than informative. Our main finding is that, for the matching subtask, a carefully prompted text-only model is a strong and resource-efficient baseline, and that visual information must be integrated selectively rather than indiscriminately.

Keywords

Large Language Models, Prompting, Reading Comprehension, Spanish, Quantization, Multimodal reasoning, IberLEF

1. Introduction

Deep language comprehension—the ability to recover semantic nuances and perform logical inference over a text—is a long-standing goal of Natural Language Processing (NLP). A recurring obstacle in evaluating comprehension is the construction of resources that genuinely test reasoning rather than memorisation: many benchmarks are English-centric and contain training data highly similar to the test set, which inflates apparent reasoning ability [1].

PROFE 2026 [2], part of the IberLEF 2026 evaluation campaign [3], addresses this by reusing Spanish language-proficiency exams developed over many years by the Instituto Cervantes to assess human students. Automatic systems are evaluated under the same conditions as humans: they receive the exercises and their instructions, with *no* specific training material. This setup deliberately favours transfer-learning and generative LLM approaches and penalises systems that rely on data contamination. As a novelty, this edition introduces exercises that contain images: some images are the candidate answers, some provide information needed to answer, and some are purely decorative. Systems must therefore decide *when* visual cues are relevant and when they should be ignored [4].

We participate in the **Matching subtask**, in which each exercise provides two sets of items: systems must find, for every item in the second set (the questions), the single item in the first set (the options) that best corresponds to it. The first set may contain extra, unnecessary options (distractors), and each option is used at most once.

Our work is guided by two research questions:

1. Can an open-weights, instruction-tuned LLM, used purely zero-shot through prompting, solve

IberLEF 2026, September 2026, León, Spain

davidbojaca1@gmail.com (D. A. Bojaca)

the Spanish matching subtask clearly above the official baseline—while remaining cheap enough to run on a single consumer-grade GPU?
2. Given that only a minority of exercises contain images and that many of those images are decorative, does adding a multimodal model help, or does indiscriminate visual input hurt?

Contributions. (i) We present a reproducible, resource-efficient text-only pipeline built on `Qwen2.5-7B-Instruct` in 4-bit precision that obtains **66.67%** accuracy on the hidden test set, **+15.6** points over the ChatGPT baseline (0.51). (ii) We design a robust structured-output strategy—a Spanish free-form reasoning prompt with a trailing JSON answer, a validated parser, and an automatic detector of degenerate (constant-letter) outputs—that guarantees a well-formed submission for every exercise. (iii) Through a controlled comparison we show that a multimodal variant (`Qwen2.5-VL-7B-Instruct`) *degrades* accuracy on this subtask, providing concrete evidence that visual relevance must be filtered rather than assumed.

2. Related Work

Reading-comprehension benchmarks from human exams. Reusing human examinations to build comprehension benchmarks is an established idea: RACE [1] derives multiple-choice questions from English exams for middle- and high-school students. PROFE follows the same philosophy but for Spanish, drawing on the IC-UNED-RC-ES corpus of real Instituto Cervantes exams [4], which spans CEFR levels A1 to C2 and several exercise types (multiple choice, matching, and fill the gap). Because the gold standard is not distributed, the benchmark is resistant to contamination and to overfitting in post-campaign experiments.

Large Language Models and prompting. Modern LLMs exhibit strong zero- and few-shot behaviour when steered through natural-language prompts [5, 6, 7]. Eliciting intermediate reasoning—chain-of-thought prompting [8]—and aggregating multiple reasoning paths via self-consistency [9] are known to improve performance on tasks that require multi-step inference. We adopt a free-form reasoning prompt in Spanish and a structured (JSON) final answer, and we experiment with self-consistency as an optional decoding strategy.

Efficient inference and multimodality. We use the open-weights `Qwen2.5` family [10] and its vision-language counterpart `Qwen2.5-VL` [11]. To fit a 7B model and its activations within a single 16 GB GPU, we rely on 4-bit NF4 quantization in the style of QLoRA [12], served through the Hugging Face Transformers library [13]. Unlike QLoRA, however, we perform *no* fine-tuning: the model is used strictly in inference, consistent with the no-training-data spirit of the task.

3. Task and Dataset Description

The matching subtask. Each matching exercise contains a *set 1* of option texts (identified by letters, e.g. A, B, C, …) and a *set 2* of question items (identified by integers). The system must assign to every question the single option that best matches it. There is exactly one correct option per question, options are not reused, and set 1 may include extra options that match no question. Some instructions explicitly warn about these distractors (e.g. *"HAY TRES TEXTOS QUE NO DEBE RELACIONAR"*).

Statistics of the released test set. The matching test set comprises **189** exercises drawn from **126** exams, with a total of **1,394** questions (matching points). On average an exercise offers 8.88 options (set 1, range 3–11) for 7.38 questions (set 2, range 6–10), so most exercises contain a small number of distractor options. Exercises span the full CEFR range from A1 to C2; Figure 1 shows their distribution. A1 and A2 dominate the collection, while B2 and C1 are the rarest levels.

Images. Out of 189 exercises, **71 (37.6%)** contain at least one image. Most of these images appear on the *question* side: there are 308 image references in set 2 but only 81 in set 1, and only **27 exercises (14.3%)** have an image among the options. Manual inspection shows that a large share of the question-side images are decorative (e.g. topic icons, horoscope ornaments) rather than carriers of answer-relevant

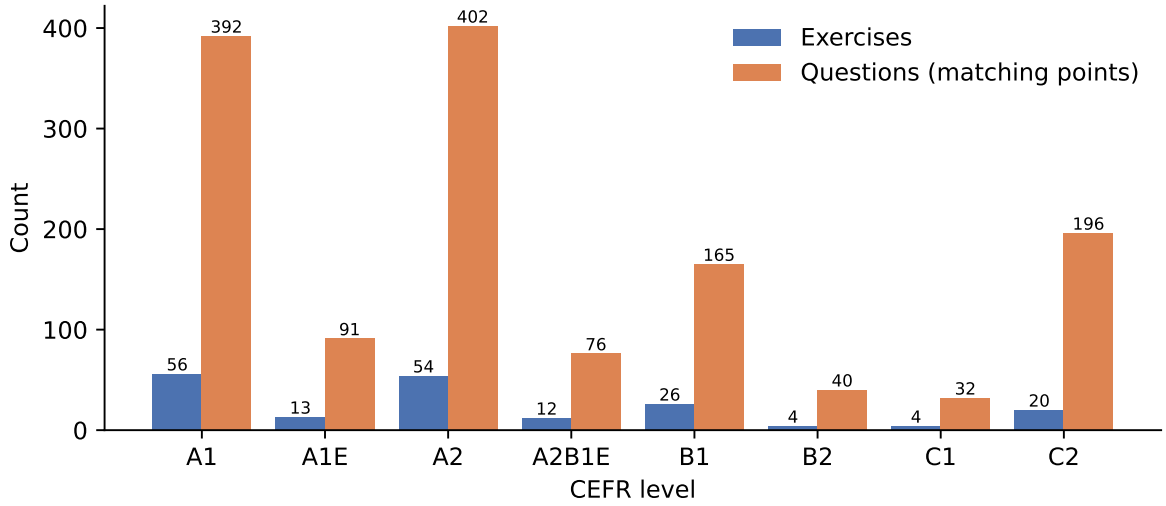

Figure 1: Distribution of matching exercises and questions per CEFR level in the test set. Lower levels (A1/A2) dominate; B2 and C1 are scarce.

information. This observation directly motivates our second research question and our selective treatment of visual input (Section 4).

Evaluation. The official metric is accuracy: the proportion of questions for which the predicted option matches the gold option. This is the same measure applied to human examinees, enabling a direct human–machine comparison. The organisers report a preliminary baseline obtained with ChatGPT of **0.51** for the matching subtask, which we adopt as our reference point.

4. Methodology

We frame matching as a single-shot, structured-prediction problem: for one exercise, the model reads the instructions, all options, and all questions, and produces a JSON object that maps each question identifier to an option letter. No gradient updates are performed at any point; the system is purely zero-shot. We justify each design decision below.

4.1. Model and quantization

We use `Qwen2.5-7B-Instruct` [10], an open-weights instruction-tuned model with strong multilingual (including Spanish) and instruction-following abilities. We chose a 7B model—rather than a larger one—because the task constraints (free-tier Colab) cap available memory, and because preliminary experiments with a 14B model did not improve over the 7B configuration while being substantially slower. The model is loaded in 4-bit NF4 precision with double quantization and a `bfloat16` compute dtype [12], which is what makes a 7B model fit and run on a single 16 GB GPU.

4.2. Prompt design

Each query consists of a system message that casts the model as an expert in Spanish reading comprehension and instructs it to prefer literal, specific matches over vague paraphrases, followed by a user message that lays out the exercise. The user message lists the valid option identifiers, the option texts, the question texts, and a strict instruction to reason briefly and then output a single-line JSON object covering every question. The exact JSON template (with the real identifiers of each exercise) is injected into the prompt so the model knows precisely which keys and which letters are valid. Listing 1 shows the template.

Listing 1: Prompt template (system message and the skeleton of the user message). Curly-brace fields are filled per exercise.

```
SYSTEM:
Eres un experto en comprension lectora del espanol, especializado en
examenes de proficiencia linguistica (Instituto Cervantes). Resuelves
ejercicios de tipo "matching" leyendo cuidadosamente cada texto y
eligiendo la mejor correspondencia. Prioriza coincidencias literales y
especificas sobre parafrasis vagas.

USER:
INSTRUCCIONES DEL EJERCICIO:
{instructions}

TEXTOS DEL SET 1 (identificadores validos: {set1_ids}):
{set1_block}

PREGUNTAS DEL SET 2 (identificadores: {set2_ids}):
{set2_block}

TAREA: Para cada pregunta del Set 2, decide que texto del Set 1 le
corresponde mejor. Razona brevemente y, al FINAL de tu respuesta,
escribe SOLO una linea con un JSON que asocie cada identificador del
Set 2 con un identificador del Set 1.

REGLAS ESTRICTAS:
- Cada respuesta DEBE ser una de estas letras: {set1_ids}.
- Nunca uses '-', 'ninguno' ni vacios. Si dudas, elige la opcion mas
  probable.

Formato exacto del JSON final (una sola linea, sin texto extra despues):
{"q1": "<letra>", "q2": "<letra>", ...}
```

For the text-only system, options or questions that consist of (or include) an image are rendered with an explicit note stating that the element contains an image that the model cannot see and that it should rely on the surrounding text. This keeps the prompt honest about missing modality instead of silently dropping the item.

4.3. Decoding

The submitted system uses **greedy decoding with a single sample** ($N = 1$) and a generation budget of up to 1,500 new tokens, which is enough for short Spanish reasoning followed by the JSON line. We also implemented self-consistency [9] ($N = 3$: one greedy sample plus two sampled at temperature 0.4, with a per-question majority vote). Self-consistency roughly triples inference time; under the free-tier time budget the single-sample greedy configuration was the one used for the best official submission.

4.4. Robust output parsing

LLM free-form output is not guaranteed to be valid JSON, and an ill-formed answer would forfeit an entire exercise. We therefore parse defensively in three layers:

1. **Validated JSON extraction.** We scan the generated text for JSON objects (last to first) and accept the first one that contains *all* question identifiers; each value is upper-cased and kept only if it is a valid option letter for that exercise. This step rejects malformed answers and spurious values such as `"-"` or `"ninguno"`.
2. **Regex fallback.** For any question still unresolved, we search the text near that question identifier for the first valid option letter.
3. **Default fallback.** If both fail, we assign the first option letter so that every question receives an answer.

Table 1
Accuracy on the PROFE 2026 matching test set (question level). The text-only system is our official submission.

System	Accuracy
ChatGPT baseline (organisers)	0.5100
Ours — Qwen2.5-7B-Instruct, text-only (N=1)	**0.6667**
Ours — Qwen2.5-VL-7B, multimodal on image exercises	< 0.6667 †

† The multimodal variant scored below the text-only system on the image-bearing exercises; see Section 6 for the analysis.

After generation we additionally flag *degenerate* exercises—those where (almost) all questions received the same letter, a typical symptom of the model refusing to answer and the parser repeatedly hitting the default fallback—and re-run them. Three exercises were detected and re-run this way.

4.5. Multimodal variant (ablation)

To answer the second research question we built a parallel pipeline with `Qwen2.5-VL-7B-Instruct` [11], also in 4-bit precision. The prompt structure is identical, but real images are passed to the model instead of the "cannot see" note. To control memory, each image is down-scaled so that its longest side is at most 512 pixels, generation is capped at 600 tokens, and out-of-memory exercises fall back to the text-only prediction. We applied this variant only to the exercises that contain images and compared it against the text-only predictions.

4.6. Reproducibility

All experiments ran in Google Colab on a single **NVIDIA T4 (16 GB)** GPU under the free tier. The model is `Qwen2.5-7B-Instruct` served with Hugging Face Transformers [13] and `bitsandbytes` 4-bit NF4 quantization (double quantization, `bfloat16` compute). Decoding is greedy ($N = 1$), `max_new_tokens=1500`. Predictions are written incrementally (one exercise at a time) so that long runs can be interrupted and resumed without losing work. The submission is a JSON file mapping each `exerciseID` to a {question: option} dictionary. The code (a single Colab notebook) is available at:

https://github.com/AlejandroBojaca/matching-task-PROFE

5. Results

Table 1 reports accuracy on the hidden test set. Our text-only system reaches **66.67%**, clearly surpassing the official ChatGPT baseline of 0.51 by **15.6** points. The multimodal variant did *not* improve over the text-only system; on the contrary, replacing text-only predictions with multimodal ones on the image-bearing exercises *reduced* overall accuracy.

On the small set of labelled *example* exercises released by the organisers (27 questions across three exercises—the only matching data with a public gold standard), the text-only system answered 24 correctly (88.9%). This is consistent with the test-set result being lower, since the hidden set is larger and includes harder, higher-level (C1/C2) exercises that are absent from the examples.

6. Error Analysis

Because the organisers do not release the test gold standard, our error analysis combines (a) the labelled example exercises, (b) the structure of the model's raw outputs, and (c) the controlled text-only vs. multimodal comparison.

Structured-output failures. Early runs occasionally produced invalid answer tokens (e.g. a literal `"-"`) or refused to commit to a letter. These never reach the submission thanks to the validated parser, but they reveal that the model is uncertain on some items. The degenerate-output detector flagged

three exercises in which nearly all questions collapsed to a single letter; re-running them produced better-differentiated answers.

Distractors. The hardest configurations are exercises with several extra options (set 1 larger than set 2), and especially the 13 exercises that explicitly contain options that must *not* be matched. Here the model must not only find the best option but also avoid being attracted by a plausible but unused distractor; this is where literal-match prompting ("prefer literal, specific matches over vague paraphrases") helps most.

Paraphrase vs. literal matching. Qualitatively, correct matches tend to hinge on a specific lexical or factual anchor shared between question and option (a name, a number, a concrete fact), whereas errors cluster on questions phrased as broad paraphrases of a whole text, where two options are semantically close.

Why multimodality hurt. The multimodal variant lowered accuracy for three compounding reasons. First, most images are on the *question* side and are decorative (icons/ornaments); feeding them adds distracting tokens without adding signal. Second, our 512-pixel down-scaling, needed to fit the vision model in memory, degrades any text embedded in an image, so even the genuinely informative images are read poorly. Third, and most importantly, the variant *overwrote* text-only predictions on the image-bearing exercises—many of which the text-only model had already answered correctly using the option texts alone—trading correct answers for noisier ones. The lesson is that visual input on this subtask must be *gated* by an estimate of visual relevance, not applied indiscriminately.

7. Conclusion and Future Work

We described our system for the Matching subtask of PROFE 2026. A single open-weights, instruction-tuned model (`Qwen2.5-7B-Instruct`), used zero-shot in 4-bit precision with a Spanish free-form reasoning prompt and a validated JSON output, reaches **66.67%** accuracy on the hidden test set, **+15.6** points over the official ChatGPT baseline, while running on a single free-tier 16 GB GPU. This answers our first research question affirmatively: careful prompting of a modest, quantized LLM is a strong and cheap approach to Spanish reading-comprehension matching.

Our second finding is a cautionary one. A multimodal model did not help and in fact hurt, because the majority of images in this subtask are decorative and because indiscriminately replacing text-only answers discards good predictions. This validates the task organisers' premise that systems must *discern when to integrate visual cues and when to disregard them.*

Future work. The natural next step is a visual-relevance gate: a cheap classifier (or a prompt-based judgement) that decides, per exercise, whether an image is informative before invoking the vision model, so that multimodal reasoning is applied only where it can help. Other promising directions are OCR or higher-resolution encoding for images that contain text, self-consistency or small ensembles within a larger time budget, and exploiting the publicly available PROFE 2025 exams as in-context examples or for light fine-tuning of the option-selection behaviour.

Declaration on Generative AI

During the preparation of this work, the authors used a generative AI assistant in order to: draft and refine the experimental pipeline, format results, and perform grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale ReAding Comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 785–794. doi:10.18653/v1/D17-1082.

[2] A. Peñas, Á. Rodrigo, et al., Overview of PROFE 2026: Spanish language proficiency evaluation at IberLEF 2026, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[3] L. Chiruzzo, S. M. Jiménez-Zafra, et al., Overview of IberLEF 2026: Natural language processing challenges for Spanish and other Iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[4] A. Peñas, Á. Rodrigo, J. Fruns-Jiménez, I. Soria-Pastor, S. Moreno-Álvarez, A. Pérez García-Plaza, J. Reyes-Montesinos, A Spanish language proficiency dataset for AI evaluation, Information 17 (2026) 159. doi:10.3390/info17020159.
[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), 2020. URL: https://arxiv.org/abs/2005.14165.
[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, arXiv preprint arXiv:2107.13586 (2021). URL: https://arxiv.org/abs/2107.13586.
[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023). URL: https://arxiv.org/abs/2303.18223.
[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems (NeurIPS), 2022. URL: https://arxiv.org/abs/2201.11903.
[9] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations (ICLR), 2023. URL: https://arxiv.org/abs/2203.11171.
[10] Qwen Team, Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024). URL: https://arxiv.org/abs/2412.15115.
[11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, et al., Qwen2.5-VL technical report, arXiv preprint arXiv:2502.13923 (2025). URL: https://arxiv.org/abs/2502.13923.
[12] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems (NeurIPS), 2023. URL: https://arxiv.org/abs/2305.14314.
[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.