Prompting Large Language Models for Spanish Easy-to-Read Text Generation: UNED-INEDA at CLEARS 2025

Laura Plaza[1,*,†], Lourdes Araujo[1,†], Fernando López-Ostenero[1,†] and Juan Martinez-Romo[1,†]

¹Universidad Nacional de Educación a Distancia, Calle Juan del Rosal, 16, 28040, Madrid, Spain

Abstract

This paper describes the participation of the UNED-INEDA team in Subtask 2 of the CLEARS shared task at IberLEF 2025, which focuses on the automatic adaptation of texts into the Easy-to-Read (E2R) format in Spanish, following the UNE 153101:2018 EX guidelines. We explore three prompting strategies—zero-shot, random few-shot, and similarity-based few-shot—applied to three instruction-tuned language models: Dolphin Mistral, Starling-LM, and LLaMA 3.2–1B–Instruct. Our official submission employs Dolphin Mistral in a zero-shot setting, which obtained competitive results, ranking fourth in both semantic similarity and readability according to the official evaluation. These results highlight the potential of large language models and prompt-based strategies for controlled text simplification in Spanish. We discuss the trade-offs between semantic fidelity and accessibility, and outline future work aimed at improving rule compliance and linguistic simplification through adaptive prompting and fine-tuning.

Keywords

Text simplification, Easy-to-read, Text Accessibility, UNE 153101:2018 EX, Prompt Engineering

1. Introduction

Ensuring that written content is understandable to all individuals is a key factor of an inclusive society. Yet, a substantial proportion of everyday texts—ranging from public service announcements to news articles and administrative documents—remain inaccessible to people with cognitive or reading difficulties. This challenge has generated interest in developing strategies that aim to improve text comprehensibility and ensure equitable access to information.

Movements such as Plain Language (PL)[1] have also emerged in response to the growing recognition that complex and intricate language can exclude large segments of the population from understanding essential information. The PL movement advocates for the use of clear, concise, and well-structured language that allows readers to easily find, understand, and use the information they need [2]. Initially driven by legal and governmental bodies seeking to improve transparency and public trust, PL has since expanded into fields such as healthcare, education, and finance [3, 4].

Another strategy to enhance the accessibility of written content is Easy-to-Read (E2R). E2R is a structured approach designed to support readers with conditions such as dyslexia, autism spectrum disorders, intellectual disabilities, or acquired brain injuries. E2R promotes clear, concise, and well-organized language, guided by specific formatting and content rules, such as those outlined in the UNE 153101:2018 standard for the Spanish language. These adaptations are intended to enhance readability without compromising the essential meaning of the original content, thus enabling individuals with cognitive challenges to fully understand written information.

While both Plain Language (PL) and Easy-to-Read (E2R) approaches aim to improve the accessibility

---

of written information, they differ in scope, audience, and degree of simplification. Plain Language is intended for a broad audience, including people with lower literacy levels or limited language proficiency, and promotes clarity, conciseness, and well-structured information. In contrast, Easy-to-Read is designed specifically for individuals with intellectual or cognitive disabilities, such as autism or acquired brain injuries, and adheres to stricter guidelines concerning sentence length, vocabulary, syntax, and formatting—such as those defined in the UNE 153101:2018 standard. While all E2R texts should comply with the principles of Plain Language, the reverse does not hold true: not all PL texts meet the cognitive accessibility criteria required for E2R. E2R can be seen as a highly specialized and standardized form of Plain Language, and as such, the automation of E2R poses unique challenges that go beyond standard text simplification.

In this paper, we describe our participation in the CLEARS Challenge for Plain Language and Easy-to-Read Adaptation for Spanish texts; in particular, Subtask 2: Adaptation of texts to E2R. Traditionally, converting standard texts into E2R versions has relied heavily on human expertise—linguists, educators, and accessibility specialists—who apply the guidelines manually [5]. However, given the volume of content generated daily and the growing demand for accessible communication, this manual approach is increasingly unsustainable. Advances in Natural Language Processing (NLP) and Machine Learning (ML) offer effective means to automate this transformation, helping save time and costs while making the process more consistent and scalable [6].

2. Related Work

Prior research on accessible communication has focused on defining and applying guidelines for making texts understandable to people with cognitive or reading difficulties. In the Spanish-speaking context, the UNE 153101:2018 standard serves as a cornerstone for Easy-to-Read (E2R) adaptation, specifying rules related to sentence structure, vocabulary, visual formatting, and content organization [7]. These rules are commonly applied manually by trained experts such as educators, linguists, and accessibility professionals. Studies have shown that such adaptations can significantly improve comprehension and usability of written content among users with intellectual disabilities, autism, or acquired brain injuries [8].

Early attempts at automating text simplification focused on rule-based systems and syntactic transformations, such as sentence splitting, clause reordering, and removal of subordinate structures [9]. These methods relied heavily on hand-crafted grammars and often suffered from limited scalability and generalizability. Lexical simplification, which involves identifying complex words and replacing them with simpler alternatives, was typically addressed using frequency-based word lists or manually curated dictionaries [10].

With the advent of supervised learning, statistical and machine learning methods became increasingly prominent. Parallel corpora of original and simplified texts (e.g., Simple Wikipedia) enabled the development of phrase-based machine translation models for simplification tasks [11]. These models improved fluency and coverage but struggled to preserve meaning.

More recently, deep learning—particularly encoder-decoder architectures based on recurrent neural networks and transformers—has revolutionized the field. Models such as T5, BART, and GPT have been fine-tuned on simplification datasets to generate fluent and coherent simplified texts [12, 13] . Some approaches also incorporate control tokens or complexity constraints to steer the generation process, allowing for different levels of simplification.

Despite these advances, current systems often fall short of meeting the strict structural and stylistic requirements of E2R guidelines. For instance, sentence independence, the avoidance of metaphors or idioms, and the use of explicit references are rarely addressed explicitly in existing models. Moreover, most simplification datasets are not aligned with the cognitive accessibility criteria needed for E2R users.

In the Spanish language, the availability of high-quality corpora and NLP tools tailored to E2R remains limited. Recent initiatives have begun to explore annotation schemes and simplification pipelines that

conform to UNE 153101:2018, but publicly available benchmarks and evaluation methodologies are still scarce. These gaps hinder the development of robust, trustworthy, and equitable systems for automated E2R generation. To help address this challenge, the CLEARS shared task [14] at IberLEF 2025 [15] has been launched to promote research in this area by providing annotated datasets, well-defined evaluation protocols, and a competitive framework that fosters innovation and comparability across systems.

3. Task Definition and Dataset

Our team participates in Subtask 2, titled "Adaptation of Texts to Easy-to-Read (E2R)". According to the task description provided by the organizers, participants are expected to adapt the supplied texts into the E2R format, following the relevant UNE 153101 EX guidelines and applying accessibility principles to enhance comprehension for target audiences with cognitive or reading difficulties. To this end, NLP techniques—including algorithmic and machine learning approaches—may be used to generate simplified versions that align with E2R standards. While adherence to all UNE rules is encouraged, full compliance is not mandatory. Evaluation considers the number of UNE-compliant rules effectively applied in the generated output. An example from the task website is shown below.

Example of E2R Adaptation

The original text *"Del 17 al 23 de febrero, Madrid se convertirá en la capital del atletismo con tres citas destacadas."*

is adapted into the following E2R version:

Madrid será la ciudad del atletismo
en los próximos días.
3 pruebas importantes de atletismo
se harán en Madrid del 17 al 23 de febrero.

The dataset used for this task consists of 3,000 news articles from various municipalities in the province of Alicante (Spain), covering topics such as sports, culture, leisure, and local festivities. It was developed and validated by a team of domain experts in text accessibility. Each article in the corpus has two adapted versions: (a) a Plain Language (PL) version, referred to as the "facilitated version," which applies basic adaptation criteria but is more flexible in terms of layout and structure (related to Subtask 1); and (b) an Easy-to-Read (E2R) version, which strictly adheres to the UNE 153101:2018 EX guidelines, including both linguistic and formatting requirements (related to Subtask 2). A detailed description of the dataset can be found in [16].

The dataset is split into 70% (2,100 texts) for training and 30% (900 texts) for testing. Spanish was chosen as the target language because, as noted by Instituto Cervantes, despite being the world's fourth most spoken language and the second most widely spoken mother tongue, there is a shortage of corpora specifically designed for PL and E2R adaptation. Addressing this gap is essential for developing NLP tools capable of effectively improving textual accessibility and comprehension. The dataset is publicly available at [17].

4. Approaches

We experimented with different prompts issued to three different Large Language Models (LLM).

1. **Instruction-based prompting (Zero-shot):** The prompt includes a description of the task, along with a summary of the UNE 153101:2018 guidelines for Easy-to-Read (E2R) texts, as described by the AENOR Spanish agency for standardization.

2. **Instruction with randomly selected examples (Few-shots):** In addition to the task description and UNE guidelines, the prompt incorporates a set of three randomly chosen training examples. These examples are distinct from those used in the test set.
3. **Instruction with similarity-based examples (Few-shots):** Similar to the previous strategy, but instead of selecting the examples randomly, they are retrieved based on their semantic similarity to the test instance being processed.

To evaluate these prompting strategies, we issued them to three large language models, each with distinct characteristics:

- **Dolphin Mistral:** A fine-tuned variant from the Mistral model family, optimized for instruction-following tasks. Dolphin Mistral is designed to generate coherent and contextually appropriate responses, particularly in tasks requiring structured reasoning and linguistic control.
- **Starling-LM:** A large language model trained using Reinforcement Learning from Human Feedback (RLHF). Starling-LM is tailored for open-ended instruction-based tasks, prioritizing factual accuracy, helpfulness, and safe outputs, which makes it especially suitable for accessibility-oriented applications.
- **LLaMA 3.2–1B–Instruct:** A lightweight instruct-tuned model from Meta's LLaMA 3 family, comprising approximately 1 billion parameters. It offers an efficient trade-off between performance and computational cost, making it ideal for scenarios with limited resources.

All experiments were conducted using the `Ollama` [18] framework to run the selected instruction-tuned models locally. Ollama enables streamlined model deployment and inference, providing a practical solution for executing large language models with minimal setup while ensuring reproducibility and data privacy.

5. Evaluation and Results

This section presents the evaluation methodology and experimental results obtained using both the training set provided by the task organizers and the official test set evaluated by the competition organizers. The goal is to assess the performance of the three instruction-tuned language models under the different prompting strategies described in Section 4.

5.1. Evaluation Metrics

To quantify system performance, we employ two complementary evaluation metrics. The first is a semantic similarity measure that captures both surface-level and deeper contextual correspondence between the automatically simplified texts and the reference texts created by experts. The second metric focuses on readability, with particular attention to Spanish-language texts, to ensure alignment with the Easy-to-Read (E2R) guidelines. These metrics are detailed below.

- **Cosine similarity/ Cosine similarity using embeddings**: This is a composite metric defined as the average of two cosine similarity measures. The first computes cosine similarity between the references and the automatically simplified texts based on term frequency vectors, thus reflecting lexical overlap. The second computes cosine similarity between sentence embeddings obtained from a pre-trained language model, capturing deeper semantic relationships beyond surface-level word matches.
- **The Fernández Huerta Readability Index**: This readability metric, designed specifically for Spanish texts, is based on the Flesch-Kincaid readability formula for English. It evaluates texts by considering average sentence length and average syllable length, assigning a readability score. A higher readability score suggests a text that is easier to understand.

5.2. Results

Tables 1 and 2 present the results obtained using the training set provided by the organizers, under the three prompting strategies described in Section 4. In the case of the few-shot experiments (Prompts 2 and 3), the examples included in the prompts were excluded from the evaluation to ensure fairness.

In terms of semantic similarity (see Table 1), measured as the mean of cosine similarity between terms and embeddings, both Dolphin Mistral and Starling-LM achieve the highest scores (0.80) under Prompt 1 (zero-shot setting). These results remain relatively stable under Prompt 2 (random few-shot), but show a slight drop with Prompt 3 (similarity-based examples), particularly for Llama-3.2-1B-Instruct, whose performance decreases more markedly across prompts (from 0.77 to 0.70).

Regarding the Fernández-Huerta readability index (see Table 2), the results are consistently high across all models and prompts, indicating that all systems are capable of generating lexically simple texts. Interestingly, Dolphin Mistral achieves its best score with Prompt 2 (82.33), suggesting that providing examples may help guide generation in this model. In contrast, Starling-LM shows a slight drop in Prompt 2, while Llama-3.2-1B-Instruct remains relatively stable but below the other models.

Overall, these results suggest that zero-shot prompting yields competitive performance, particularly in terms of semantic fidelity, while the impact of few-shot strategies is model-dependent. Larger or more instruction-tuned models may benefit more from structured examples, although careful selection appears to be critical.

Table 1
E2R_similarity_mean scores by model and prompting strategy

Modelo	Prompt 1	Prompt 2	Prompt 3
Dolphin Mistral	0.80	0.80	0.77
Starling-lm	0.80	0.79	0.78
Llama-3.2-1B-Instruct	0.77	0.73	0.70

Table 2
E2R_FH_index_mean scores by model and prompting strategy

Modelo	Prompt 1	Prompt 2	Prompt 3
Dolphin Mistral	81.85	82.33	81.55
Starling-lm	81.68	79.92	80.86
Llama-3.2-1B-Instruct	81.79	80.03	80.19

The official evaluation results for Subtask 2 are summarized in Tables 3 and 4, ranking participating teams according to two complementary criteria: semantic similarity and readability.

In terms of semantic adequacy, as measured by the average of cosine similarity metrics (lexical and embedding-based), the system submitted by UNED-INEDA ranks fourth, with a score of 0.68. Our system used the Dolphin Mistral model with Prompt 1, corresponding to the zero-shot setting. While the top-performing team, NIL-UCM, achieved a slightly higher score of 0.72, the margin separating the top four systems is relatively small, suggesting competitive performance across the board. The results also confirm the capacity of zero-shot prompting to achieve strong semantic alignment between original and simplified texts without relying on examples.

Regarding readability, assessed using the Fernández-Huerta index, our system again ranks fourth, with a score of 72.39. This indicates that the texts generated by our model were considerably more readable than those of NIL-UCM (69.40), but did not reach the simplicity achieved by Vicomtech (85.44) or UR (85.12).

Taken together, the results suggest that our approach offers a solid balance between semantic preservation and readability, particularly considering the use of a zero-shot strategy. While there is room for improvement in terms of enhancing readability, the performance of our system confirms the viability of instruction-based prompting with powerful language models such as Dolphin Mistral in the context of E2R text generation.

Table 3
Ranking based on the average of cosine similarities (Subtask 2)

Rank	Team	Average
1	NIL-UCM	0.72
2	CARDIFFNLP	0.71
3	UR	0.70
4	UNED-INEDA	0.68
5	Vicomtech	0.66

Table 4
Ranking based on the Fernández-Huerta index (Subtask 2)

Rank	Team	Average
1	Vicomtech	85.44
2	UR	85.12
3	CARDIFFNLP	77.85
4	UNED-INEDA	72.39
5	NIL-UCM	69.40

6. Conclusions and Future Work

This paper has presented the participation of the UNED-INEDA team in the CLEARS shared task at IberLEF 2025, specifically in Subtask 2: the adaptation of Spanish texts to the Easy-to-Read (E2R) format. Our approach explored three prompting strategies applied to three instruction-tuned language models, with a focus on evaluating their performance in terms of semantic preservation and readability.

Among the different strategies, the system submitted officially employed the Dolphin Mistral model with a zero-shot prompt that provided the model with the definition of the task to accomplish and the specifications of the UNE 153101:2018 EX standard. This configuration delivered competitive results: our system ranked fourth in both semantic similarity and FH index, and not far from the top ranked system (particularly in the similarity metric).

Nonetheless, the gap in readability scores between our system and top-performing teams suggests opportunities for improvement. Future work will focus on integrating hybrid prompting strategies, leveraging few-shot examples selected for both lexical and syntactic similarity. Additionally, we plan to explore the use of constraint-based generation and fine-tuning approaches aligned with the UNE 153101:2018 EX standard to enhance compliance with E2R guidelines.

7. Acknowledgments

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to check grammar and spelling.

References

[1] M. Adler, The plain language movement, in: The Oxford Handbook of Language and Law, Oxford University Press, 2012.

[2] R. Petelin, Considering plain language: issues and initiatives, Corporate Communications: An International Journal 15 (2010) 205–216.

[3] Stableford, S., Mettger, W, Plain language: A strategic response to the health literacy challenge, 2007.

[4] A. Rossetti, P. Cadwell, S. O'Brien, "the terms and conditions came back to bite":, in: HCI International 2020 – Late Breaking Papers: Universal Access and Inclusive Design, 2020, pp. 699–711.

[5] M. Nomura, G. S. Nielsen, B. Tronbacke, Guidelines for Easy-to-Read Materials, IFLA Professional Reports 120, International Federation of Library Associations and Institutions (IFLA), Library Services to People with Special Needs Section, 2010. URL: https://repository.ifla.org/handle/123456789/31, revision on behalf of the IFLA/Library Services to People with Special Needs Section.

[6] Valentina Ospina-Henao, Sebastián López Flórez, V. Juan M. Núñez, Óscar Loureda Lamas, Fernando De la Prieta, Generative ai: Simplifying text for cognitive impairments and non-native speakers, methodologies and intelligent systems for technology enhanced learning, in: 14th International Conference, 2024, pp. 33–44.

[7] UNE 153101:2018 EX. Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos, Norma UNE 153101:2018 EX, Asociación Española de Normalización (UNE), Madrid, España, 2018. URL: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma/?c=N0060036, norma experimental que especifica las pautas y recomendaciones para la adaptación, creación y validación de documentos en Lectura Fácil.

[8] T. I. Rebekah Joy Sutherland, The evidence for easy-read for people with intellectual disabilities: A systematic literature review, Journal of Policy and Practice in Intellectual Disabilities 13 (2016) 297–310.

[9] R. Chandrasekar, B. Srinivas, Automatic induction of rules for text simplification, Knowledge-Based Systems 10 (1997) 183–190.

[10] S. S. Al-Thanyyan, A. M. Azmi, Automated text simplification: A survey, ACM Comput. Surv. 54 (2021).

[11] Z. Zhu, D. Bernhard, I. Gurevych, A monolingual tree-based translation model for sentence simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics, 2010, p. 1353–1361.

[12] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 584–594.

[13] T. Wang, P. Chen, J. Rochford, J. Qiang, Text simplification using neural machine translation, in: AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, p. 4270–7271.

[14] B. Botella-Gil, I. Espinosa-Zaragoza, A. Bonet-Jover, M. Madina, L. Molino Piñar, P. Moreda, I. Gonzalez-Dios, M. T. Martín Valdivia, Ureña, Overview of clears at iberlef 2025: Challenge for plain language and easy-to-read adaptation for spanish texts, Procesamiento del Lenguaje Natural 75 (2025).

[15] 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.

[16] I. Espinosa-Zaragoza, J. Abreu-Salas, P. Moreda, M. Palomar, Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project, in: S. Štajner, H. Saggio, M. Shardlow, F. Alva-Manchego (Eds.), Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 68–77. URL: https://aclanthology.org/2023.tsar-1.7/.

[17] B. Botella-Gil, I. Espinosa-Zaragoza, P. Moreda, M. Palomar, Corpus ClearSim, 2024. URL: http://hdl.handle.net/10045/151688.

[18] Ollama, Ollama: Run large language models locally, 2023. URL: https://ollama.com, accessed: 2025-06-01.