HumanAI-UCM at MER-TRANS 2026: Translating Easy-to-Read Standards into Prompting Strategies for Spanish Text Simplification

Juan Broto-Ortega[1,*,†], Virginia Francisco[1,2,†] and Raquel Hervás[1,2,†]

[1]Facultad de Informática, Universidad Complutense de Madrid, Spain
[2]Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Spain

Abstract

This paper presents the participation of the HumanAI-UCM team in the MER-TRANS 2026 shared task on Easy-to-Read (E2R) text translation. We explored different prompting strategies, including zero-shot, one-shot, and few-shot configurations, using several open Large Language Models (LLMs). Our prompts were based on Easy-to-Read guidelines extracted from the UNE 153101:2018 EX standard and corpus-derived simplification criteria. We also evaluated the impact of incorporating contextual information through a context window mechanism. Experimental results showed that Mistral-based configurations achieved the most competitive performance. The qualitative analysis revealed that many generated simplifications remained conservative, preserving much of the lexical and syntactic structure of the original sentences. Overall, the results suggest that lightweight open LLMs constitute a promising direction for automatic Easy-to-Read simplification.

Keywords

Easy-to-Read, Text Simplification, Large Language Models, Prompt Engineering, Accessibility

1. Introduction

Access to written information remains a challenge for many people with cognitive disabilities, particularly when texts contain complex vocabulary and syntactic structures. Text simplification aims to transform complex texts into more accessible versions while preserving their original meaning. This process is especially relevant for people with cognitive disabilities, low literacy levels, or reading comprehension difficulties, who may encounter significant barriers when accessing written information in digital services and online environments.

Easy-to-Read (E2R) guidelines provide a set of recommendations to improve textual accessibility through simpler vocabulary, syntax, and document structure. In Spain, the UNE 153101:2018 EX [1] standard constitutes one of the main references for producing cognitively accessible content. However, traditional E2R adaptation processes usually rely on domain experts and iterative manual revisions involving end users, making them costly and time-consuming.

Recent advances in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP) and Large Language Models (LLMs), have opened new opportunities for developing scalable and affordable text simplification systems. LLMs have shown strong capabilities for generating simplified content, although ensuring compliance with accessibility guidelines and preserving meaning remains challenging.

In this context, the MER-TRANS shared task [2], organized as part of IberLEF 2026 [3], promotes research on automatic E2R text simplification by providing a common evaluation framework for comparing simplification systems.

This paper presents the participation of the HumanAI-UCM team in the shared task. Our approach explores different LLMs, prompting strategies, and contextual generation settings for producing E2R simplifications in Spanish, analyzing their impact across multiple evaluation metrics.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
jbroto@ucm.es (J. Broto-Ortega); virginia@fdi.ucm.es (V. Francisco); raquelhb@fdi.ucm.es (R. Hervás)

2. Related Work

Easy-to-Read (E2R) has been widely adopted as an accessibility approach for textual information, constraining linguistic and structural properties of written texts through normative standards that define explicit requirements for how information should be written and presented. E2R has been formalized through a consolidated body of normative and professional guidelines, including the Spanish UNE standard [4] and international reference frameworks such as Inclusion Europe [5] and IFLA [6]. These guidelines define linguistic, structural, and editorial recommendations intended to make written information more accessible to people with reading comprehension difficulties. In practice, however, the production of E2R materials remains a highly specialized and time-consuming process, typically involving trained writers, domain experts, and validation by end users.

Recent advances in Large Language Models (LLMs) have opened new possibilities for supporting text rewriting tasks, including simplification, summarization, and audience-oriented adaptation. Their ability to generate fluent and contextually coherent text makes them promising tools for assisting E2R production. At the same time, their use in accessibility-oriented settings raises important challenges. LLMs may introduce unsupported information, omit relevant content, over-paraphrase the source text, or apply stylistic changes inconsistently. These risks are especially significant in E2R contexts, where accessibility, accuracy, and fidelity to the original content must be jointly preserved.

A central open issue is how to steer LLMs toward outputs that comply with explicit accessibility requirements. Prompting has become one of the main mechanisms for controlling LLM behavior, through techniques such as role assignment, structured instructions, task decomposition, and the inclusion of examples. Previous work has shown that prompt design and specificity can have a substantial impact on model performance [7]. However, existing evidence remains fragmented across tasks and evaluation settings. In particular, there is still limited research on how normative E2R guidelines can be operationalized as prompts in a systematic, transparent, and reproducible way.

This gap motivates methodological work that connects E2R standards with LLM-based text generation under controlled experimental conditions. Such work is needed to determine which types of E2R requirements can be reliably addressed through prompting alone, which may require examples, revision loops, or human validation, and how compliance with accessibility guidelines can be assessed alongside content preservation.

3. Dataset and Evaluation Metrics

The dataset provided by the organizers of the MER-TRANS shared task consisted of parallel corpora for four languages: Italian, Catalan, Spanish, and Arabic. Each language included both a trial set and a test set. Our experiments focused exclusively on the Spanish subset.

The trial dataset contains only four original sentences and their corresponding Easy-to-Read (E2R) simplifications, all extracted from a single document. Each instance includes metadata fields such as document identifiers, sentence identifiers, and language labels, together with the original and simplified versions of the text. The simplification criteria used during dataset construction are described in the shared task dataset paper [8]. The Spanish test set follows the same structure but does not include reference simplifications. It contains sixteen documents comprising a total of 346 sentences.

The shared task evaluation was based on the following automatic metrics:

- **BLEU**: measures n-gram overlap between generated and reference texts, providing an estimate of surface-level similarity. Higher scores indicate greater similarity to the reference text.
- **SARI**: a metric specifically designed for text simplification that evaluates the quality of words added, deleted, and retained with respect to the source and reference texts. SARI scores range from 0 to 100, where higher scores indicate better simplification quality.
- **BERTScore**: evaluates semantic similarity between generated and reference texts using contextual embeddings, providing an estimate of meaning preservation. Higher values indicate stronger semantic preservation between the generated simplification and the reference text.

These metrics were used to compare the outputs generated by different model and prompt configurations explored in this work.

4. Methodology

This section describes the proposed methodology, including the prompt design process, the selected Large Language Models (LLMs), the experimental setup used to evaluate the different configurations, and the obtained results.

4.1. Prompting Strategies

Prompt design played a central role in our methodology, as instruction formulation strongly influences the behavior and performance of Large Language Models [9]. We explored different prompting strategies for E2R text simplification, combining two sources of simplification guidelines with several levels of in-context examples. Previous studies have shown that in-context examples can substantially improve model performance in text generation and translation tasks [10], including monolingual simplification scenarios such as E2R simplification [11].

The primary set of prompts was based on a subset of 53 guidelines extracted from the UNE 153101:2018 EX standard [1], covering aspects such as vocabulary, sentence structure, punctuation, and formatting. These recommendations provide a broad set of guidelines for producing cognitively accessible content.

In addition, we evaluated an alternative prompt configuration based on the simplification criteria described in the dataset creation paper [8]. Although these criteria overlap with several UNE recommendations, they do not cover all the accessibility guidelines included in the UNE standard. Since the reference simplifications used in the shared task were created following these corpus construction criteria, we included this configuration, referred to as Corpus-based, to analyze whether a prompt more closely aligned with the dataset creation process would improve performance.

All prompt configurations shared the same role definition, task description, output format, and general simplification objectives. Figure 1 presents the common prompt structure used across all configurations.

```
<ROL> Eres una herramienta de transformación de textos a Lectura Fácil según las
pautas UNE. Eres un experto en adaptación de textos en español a Lectura Fácil,
conforme a las pautas UNE. Tu función es transformar el texto original para que sea
claro, comprensible y accesible para personas con dificultades de lectura. Aplicas
las pautas de forma sistemática y rigurosa, manteniendo siempre el contenido y el
significado originales. </ROL>

<OBJETIVO> Adapta el texto proporcionado por el usuario a Lectura Fácil, siguiendo
las pautas UNE.
La adaptación debe conservar exactamente el mismo contenido y significado del texto
original. </OBJETIVO>

<CONSTRAINTS> Sigue las pautas incluso si el texto original no las cumple.
PROHIBIDO añadir títulos, notas, etiquetas, explicaciones o paréntesis.
PROHIBIDO explicar qué has cambiado o por qué. </CONSTRAINTS>

<SALIDA> Devuelve ÚNICAMENTE el texto adaptado a Lectura Fácil. </SALIDA>
```

Figure 1: Common prompt structure shared by all configurations.

For the UNE-based prompts, we evaluated three prompting strategies:

- **Zero-shot:** included the 53 selected E2R guidelines grouped by category (e.g., vocabulary, structure, punctuation, and formatting), without simplification examples.
- **One-shot:** incorporated a single simplification example for the guidelines that showed poor compliance in the zero-shot outputs.

- **Few-shot:** extended the previous configuration by adding multiple examples for the guidelines that continued to show poor compliance after the one-shot configuration.

Prompt refinement was carried out iteratively using a development corpus composed of twenty texts, including fifteen texts from the ClearSim Corpus [12] and five synthetically generated examples. To balance realism and experimental control, the fifteen ClearSim texts were intentionally selected from the corpus to maximize the diversity and number of E2R guideline violations while preserving authentic administrative writing styles and typical linguistic complexity. However, some guideline violations were underrepresented or absent in the available ClearSim texts. To ensure comprehensive coverage of the E2R guidelines, we therefore complemented the corpus with five synthetic texts generated by prompting the model to produce short fictional narratives containing specific targeted violations. All texts were manually analyzed and annotated for E2R violations, and the resulting corpus was used to iteratively identify shortcomings in the simplification outputs, specifically by determining which E2R guidelines were not satisfied by the previous prompt and refining the prompting strategies accordingly, adding one example when moving from zero-shot to one-shot and multiple examples when moving from one-shot to few-shot.

Finally, to avoid undesired conversational behaviors, such as answering questions instead of simplifying them, the input text was explicitly enclosed within the tag: <TEXTO_A_ADAPTAR> (English: <TEXT_TO_SIMPLIFY>).

4.2. Models

To evaluate the impact of our prompting strategies on E2R text simplification, we selected several Large Language Models (LLMs) with different sizes and instruction-following capabilities. The evaluated models were:

- **Llama 3.2 3B:** a multilingual instruction-tuned model developed by Meta with 3.2B parameters, designed for conversational and text generation tasks including summarization [13].
- **Phi-4 Mini 3.8B:** a lightweight 3.8B parameter model from Microsoft's Phi family, optimized for efficient inference and instruction following [14].
- **Mistral 7B:** a multilingual 7.2B open-weight model focused on efficient text generation and strong general-purpose performance[1].
- **Mistral 7B Instruct:** the instruction-tuned variant of Mistral 7.2B, also with 7.2B parameters, specifically optimized for instruction-following tasks[1].

These models were selected to compare architectures with different sizes and instruction-following capabilities under different prompting configurations for E2R simplification.

Table 1 presents a summary of the models evaluated in this task. It includes information such as the number of parameters and the context length, which may influence the ability of the models to process long prompts and contextual information.

Table 1
Summary of the evaluated LLMs.

Model	Parameters (B)	Context Length	Instruction-tuned	Organization
Llama 3.2	3.2B	131K	Yes	Meta
Phi-4 Mini	3.8B	128K	Yes	Microsoft
Mistral	7.2B	32K	No	Mistral AI
Mistral Instruct	7.2B	32K	Yes	Mistral AI

[1] https://mistral.ai/news/announcing-mistral-7b

4.3. Experimental Setup

All experiments were conducted using the trial and test datasets provided by the shared task organizers. Since only the trial dataset included reference simplifications, automatic evaluation metrics were computed exclusively on this subset. The final submitted runs were then generated on the test set, whose results were used for the shared task evaluation.

Because the shared task operated at sentence level, each simplification had to be generated independently for every row in the CSV file. However, processing sentences in isolation may reduce discourse coherence across consecutive sentences. To mitigate this limitation, we incorporated a context window mechanism in some configurations. During generation, the model received the previous N simplified sentences as additional contextual information when simplifying the current sentence. The prompt explicitly indicated that these previous sentences should be considered contextual information rather than target outputs. For the context-aware configurations, we set N = 5, considering it a reasonable compromise between providing sufficient local context and limiting the amount of potentially irrelevant information introduced into the prompt.

All models were executed locally using the Ollama[2] framework (version 0.30.10), maintaining the default generation parameters across all experiments.

4.4. Selection of Configurations

To identify the most suitable configurations for the shared task, we conducted several exploratory experiments combining the models and prompting strategies described in the previous sections. Since the trial dataset was the only subset containing reference simplifications, all automatic evaluations were performed exclusively on this dataset. Table 2 summarizes the results obtained across all evaluated configurations, including both sentence-independent and context-aware generation settings.

Overall, the Phi-4 Mini configurations achieved lower performance across most evaluation metrics compared to the remaining models. Consequently, not all prompting configurations were explored with this model.

For Llama 3.2, the Corpus-based configuration produced a substantial improvement in SARI compared to the UNE-based prompts, suggesting that simplification criteria derived from the dataset guidelines may better align with the evaluation framework of the shared task. However, context-aware experiments were not conducted with this model. Given its smaller size (3.2B parameters), we prioritized evaluating contextual generation with the Mistral-based models, which had already demonstrated stronger and more stable performance across the evaluated metrics.

The Mistral-based models exhibited the most consistent performance across the evaluated metrics, obtaining some of the highest SARI and BERTScore values. In general, the instruction-tuned variants showed more stable behavior across prompting configurations.

When comparing sentence-independent and context-aware generation settings, contextual information generally improved simplification quality and readability, particularly in terms of SARI score. This effect was especially noticeable for the Mistral-based models.

For the Mistral model, the UNE-based ZS and FS configurations achieved the highest SARI values when contextual information was incorporated, substantially outperforming their corresponding sentence-independent variants. These results suggest that access to previous simplified sentences may improve local coherence and support the generation of more accessible simplifications.

In the case of Mistral Instruct, the best readability scores were obtained by the context-aware configurations, particularly with the Corpus-based prompt. Additionally, the one-shot and few-shot variants achieved improved BERTScore values, indicating better semantic preservation with respect to the reference simplifications.

Since the trial dataset was extremely limited in size, the obtained results should be interpreted as exploratory rather than statistically robust. Consequently, the final runs were selected not only according to the best individual metric values, but also to explore complementary combinations of

[2] https://ollama.com/

Table 2
Experimental results.

Model	Context	Prompt	BLEU	SARI	BERTScore
Llama 3.2	No	UNE-based (ZS)	2.85	30.5	0.73
		UNE-based (1S)	4.23	22.59	0.73
		UNE-based (FS)	11.55	21.92	0.75
		Corpus-based	**12.92**	39.81	0.75
Phi-4 Mini	No	UNE-based (ZS)	0.03	35.02	0.62
		Corpus-based	0.29	29.02	0.55
Mistral	No	UNE-based (ZS)	8.46	40.49	0.81
		UNE-based (1S)	12.77	37.93	0.82
		UNE-based (FS)	9.26	40.47	0.74
		Corpus-based	12.15	39.87	0.76
	Yes	UNE-based (ZS)	10.75	**48.87**	0.81
		UNE-based (1S)	1.59	37.38	0.79
		UNE-based (FS)	10.78	47.5	0.79
		Corpus-based	11.77	43.44	0.78
Mistral Instruct	No	UNE-based (ZS)	10.27	41.17	0.81
		UNE-based (1S)	4.27	36.51	0.74
		UNE-based (FS)	8.1	38.81	0.75
		Corpus-based	5.62	40.53	0.78
	Yes	UNE-based (ZS)	9.96	43.24	0.81
		UNE-based (1S)	6.08	45.01	**0.83**
		UNE-based (FS)	8.18	47.87	0.81
		Corpus-based	3.43	44.71	0.81

models, prompting strategies, and contextual generation settings. The three submitted runs were defined as follows:

- Run 1 corresponded to the *Mistral UNE-based (1S)* configuration without contextual information, as it showed competitive BLEU and BERTScore values while maintaining stable overall performance.
- Run 2 corresponded to the *Mistral Instruct UNE-based (ZS)* configuration with contextual generation, allowing us to evaluate the behavior of an instruction-tuned variant under the same prompting strategy used in Run 3.
- Run 3 used the *Mistral UNE-based (ZS)* configuration with contextual information, since it achieved the highest SARI score (a metric specifically designed for text simplification evaluation) among the evaluated configurations.

5. Results and Discussion

This section presents both the quantitative results obtained across the evaluation metrics and a qualitative analysis of the generated simplifications, providing a better understanding of the strengths and limitations of the evaluated configurations. In addition, it discusses the main limitations of the experimental setup and evaluation methodology that should be considered when interpreting the reported results.

5.1. Quantitative Analysis

Table 3 presents the official results obtained by our submitted configurations on the test dataset. For readability, the table does not include all participating systems and runs. Instead, it reports the three HumanAI-UCM submissions together with the systems that achieved the best or worst performance in

at least one evaluation metric. Specifically, Do Nothing obtained the highest $BLEU_{orig}$, $BLEU_{gold}$, and BERTScore values, as well as the lowest SARI score; ClearText achieved the highest SARI score; HULAT2 obtained the lowest $BLEU_{orig}$ score; and NIL-UCM obtained the lowest $BLEU_{gold}$ and BERTScore values. Among the submitted systems, HULAT1 achieved the highest $BLEU_{gold}$ score, second only to the Do Nothing baseline. $BLEU_{orig}$ corresponds to the BLEU score computed against the original sentences, while $BLEU_{gold}$ is computed against the reference simplifications.

Table 3
Final results on the test dataset. Higher scores are highlighted in bold, while the lower scores are highlighted in red.

Team	Run	$BLEU_{orig}$	$BLEU_{gold}$	SARI	BERTScore
Do Nothing	RUN1	**100.00**	**23.21**	13.86	**0.929**
HULAT1	RUN1	37.15	18.20	43.38	0.927
HULAT2	RUN3	2.40	5.33	38.51	0.911
NIL_UCM	RUN1	6.05	2.57	39.68	0.878
ClearText	RUN1	18.00	15.44	**47.01**	0.922
HumanAI-UCM	RUN1	38.34	15.08	40.41	0.920
HumanAI-UCM	RUN2	26.98	11.84	41.62	0.911
HumanAI-UCM	RUN3	45.02	14.53	36.24	0.912

The obtained results by our submitted systems reveal a clear trade-off between preserving similarity with the original text and producing stronger simplifications. In particular, the *Mistral UNE-based (ZS)* (HumanAI-UCM RUN3) configuration achieved the highest $BLEU_{orig}$ score in the entire competition, indicating very high lexical and structural similarity with the original sentences. However, this same configuration obtained the lowest SARI score among our submitted runs. This behavior highlights an important limitation of BLEU-based metrics for text simplification tasks. High similarity with the original sentence does not necessarily correspond to better simplification quality. This tendency can also be observed in the lower-performance baseline reported by the organizers, which corresponds to leaving the original sentences unchanged. Although this baseline maximizes similarity with the source text, it obtained one of the lowest SARI scores in the competition.

Among our submissions, the *Mistral Instruct UNE-based (ZS)* (HumanAI-UCM RUN2) configuration achieved the best balance across the evaluated metrics, obtaining the highest SARI score among our runs while maintaining competitive $BLEU_{orig}$ and BERTScore values.

The results also suggest that contextual generation may provide greater benefits than increasing the number of in-context examples. While the incorporation of previous sentences generally improved SARI and readability scores, the one-shot and few-shot prompting strategies produced less consistent improvements across configurations.

Furthermore, the instruction-tuned variants generally exhibited more stable behavior across different prompting configurations, suggesting that instruction alignment may facilitate adherence to Easy-to-Read simplification requirements.

In general, our systems tended to adopt relatively conservative simplification strategies, preserving a substantial proportion of the lexical and syntactic structure of the original sentences. This behavior is reflected in the high $BLEU_{orig}$ scores and comparatively lower SARI values obtained by our submissions. Despite this conservative behavior, the proposed systems achieved competitive results across the evaluated metrics, remaining substantially closer to the highest-performing configurations than to the lower-performing configurations reported in the shared task. In particular, our submissions remained within less than three points of the best $BLEU_{gold}$ score and approximately six points below the highest SARI score achieved in the competition.

These results suggest that the proposed prompting strategies were generally more effective at preserving semantic and structural information than at performing aggressive lexical and syntactic simplifications.

Table 4
Example of generated simplifications for document ES_070, sentence 3.

Run	Sentence
Original	Finalmente, de entre todos los ciudadanos registrados, 150 serán seleccionados aleatoriamente para representar la diversidad demográfica de la UE y serán invitados a viajar a Bruselas.
RUN1	Finalmente, 150 personas de entre todos los registrados serán elegidas al azar para representar la diversidad demográfica de la UE y serán invitadas a viajar a Bruselas.
RUN2	Por último, de entre todos los ciudadanos registrados, 150 serán seleccionados al azar para representar la diversidad demográfica de la UE y serán invitados a viajar a Bruselas.
RUN3	De entre todos los ciudadanos registrados, 150 serán seleccionados aleatoriamente para representar la diversidad demográfica de la UE. ¡Serán invitados a viajar a Bruselas!

5.2. Qualitative Analysis

To complement the quantitative results, we conducted a qualitative analysis of the generated simplifications for the provided test set. This analysis allowed us to better understand how the different configurations approached the simplification task and to identify common strengths and limitations that are not fully captured by automatic evaluation metrics. We manually inspected some examples from the different runs in order to identify recurrent simplification behaviors and common failure patterns across configurations. The qualitative analysis revealed three recurrent behaviors across the evaluated configurations: overly conservative simplifications, unnecessary semantic modifications, and context-induced hallucinations.

Table 4 shows an example of the simplifications generated by the evaluated configurations. Although all runs preserved the original meaning of the sentence, they applied different simplification strategies:

- RUN2 produced a very conservative simplification, introducing only small lexical changes such as replacing “Finalmente” with “Por último” and “aleatoriamente” with “al azar”. However, most of the original sentence structure and several complex expressions were maintained, limiting the overall simplification despite obtaining the highest SARI score.
- RUN1 introduced additional lexical simplifications, replacing “ciudadanos registrados” with “personas registradas” and “seleccionados” with “elegidas”. However, the sentence remained structurally similar to the original version.
- RUN3 applied a more aggressive strategy by splitting the content into two sentences. Nevertheless, some complex expressions were still preserved, preventing the simplification from becoming fully accessible. This behavior is consistent with the combination of high BLEU and comparatively lower SARI scores observed for this configuration.

Table 5 also illustrates one limitation of automatic simplification metrics. In this example, the original sentence (“¿Por qué?”) is already short and easily understandable. However, some configurations introduced unnecessary modifications that altered the original meaning of the sentence:

- RUN1 added new semantic content.
- RUN2 generated a completely different question unrelated to the source sentence. This behavior may be partially explained by the default temperature parameter used in the Ollama framework, which can encourage more creative or generative outputs even in cases where no simplification is required.
- RUN3 preserved the original sentence unchanged. Although this behavior may be penalized by metrics such as SARI, which reward lexical and structural modifications, it arguably represents the most appropriate simplification strategy for this particular example.

Table 5
Example of generated simplifications for document ES_051, sentence 13.

Run	Sentence
Original	¿Por qué?
RUN1	Por qué es interesante saberlo.
RUN2	¿Por qué es importante no estar cerca del tráfico en las ciudades?
RUN3	¿Por qué?

These examples illustrate that effective E2R simplification requires balancing simplification, meaning preservation, and the avoidance of unnecessary modifications, a trade-off that is only partially captured by current automatic evaluation metrics.

5.3. Limitations

During the development of this work, several limitations were identified that should be considered when interpreting the reported results. First, the trial dataset provided by the organizers was extremely small, limiting the robustness of the configuration evaluation and selection process. Consequently, the experimental results should be interpreted as exploratory rather than statistically conclusive.

Additionally, the evaluation relied exclusively on automatic metrics such as BLEU, SARI, and BERTScore. Although these metrics are widely used in text simplification tasks, they may not fully reflect the actual compliance of Easy-to-Read simplifications with the established guidelines, particularly regarding adherence to the recommendations defined in the UNE standard, which was taken as the main reference framework in this work. As illustrated in the qualitative analysis, automatic metrics may reward lexical or structural modifications even in cases where preserving the original sentence would constitute the most appropriate simplification strategy.

Finally, the proposed systems were not evaluated with end users. Future work should incorporate human evaluations involving people with cognitive disabilities in order to assess not only linguistic simplification, but also the real comprehensibility and accessibility of the generated texts.

6. Conclusions and Future Work

In this work, we explored the use of different prompting strategies and Large Language Models for Easy-to-Read (E2R) text simplification in Spanish within the framework of the MER-TRANS shared task. We evaluated several zero-shot, one-shot, and few-shot configurations based on both UNE guidelines and corpus-derived simplification instructions.

The experimental results showed that prompting strategies and model selection substantially affect simplification quality. In particular, the Mistral-based configurations achieved the most competitive results across the evaluated metrics, while context-aware configurations generally improved SARI scores.

The results obtained in the shared task indicate that current LLM-based approaches constitute a promising direction for automatic Easy-to-Read simplification. Although there is still considerable room for improvement, especially regarding accessibility-oriented evaluation, the evaluated systems were capable of generating coherent and partially simplified texts using relatively lightweight open models and simple prompting strategies.

The qualitative analysis revealed that many generated simplifications remained overly conservative, preserving complex lexical and syntactic structures from the original texts. It also highlighted limitations of current automatic evaluation metrics, which may reward unnecessary modifications even when preserving the original sentence constitutes the most appropriate simplification strategy.

Future work should therefore explore mechanisms that encourage deeper simplification transformations while maintaining semantic fidelity and readability, as well as evaluation methodologies that

better capture the accessibility requirements of Easy-to-Read content. Additionally, future research should assess efficiency and ecological considerations, as these factors are particularly relevant to the practical motivation for using lightweight open-source models.

Acknowledgments

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 FEDER/UE.

Declaration on Generative AI

The authors used ChatGPT for grammar and spelling revision.

References

[1] Asociación Española de Normalización y Certificación (AENOR), UNE 153101:2018 EX. Lectura fácil: Pautas y recomendaciones para la elaboración de documentos, Technical Report UNE 153101:2018 EX, Asociación Española de Normalización y Certificación (AENOR), 2018. URL: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma/?c=N0060036, Spanish Easy-to-Read standard.

[2] H. Saggion, M. Tareh, N. Khallaf, D. Adanza, S. Bott, N. P. Rojas, A. Rascón, S. Szasz, Overview of MER-TRANS at IberLEF 2026: First Shared Task on Multilingual Easy-to-Read Translation, Procesamiento del Lenguaje Natural 77 (2026).

[3] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, 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-WS.org, 2026.

[4] Asociación Española de Normalización y Certificación (AENOR), Lectura fácil: Pautas y recomendaciones para la elaboración de documentos, Technical Report UNE 153101:2018 EX, AENOR, 2018.

[5] Inclusion Europe, Information for All: European standards for making information easy to read and understand, Inclusion Europe, 2009.

[6] Nomura, M. and Nielsen, G.S. and Tronbacke, B., Guidelines for Easy-to-Read Materials, Technical Report 120, International Federation of Library Associations and Institutions (IFLA), 2010.

[7] 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, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815.

[8] S. Bott, V. Riegler, H. Saggion, A. Rascón-Alcaina, N. Khallaf, A multilingual human annotated corpus of original and easy-to-read texts to support access to democratic participatory processes, in: Proceedings of the Language Resources and Evaluation Conference (LREC-COLING 2026), Palma, Mallorca, Spain, 2026.

[9] 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, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815. doi:10.1145/3560815.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, et al., Language Models are Few-Shot Learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[11] S. Wubben, A. van den Bosch, E. Krahmer, Sentence simplification by monolingual machine translation, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, Association for Computational Linguistics, USA, 2012, p. 1015–1024.

[12] B. Botella-Gil, I. Espinosa-Zaragoza, P. Moreda, M. Palomar, ClearSim: Parallel Spanish Corpus of Original and Easy-to-Read Administrative Texts, 2024. URL: https://rua.ua.es/entities/publication/34d696a3-8322-4d66-87ac-d0852671ecda, corpus description and access; selected texts from public administration with Easy-to-Read adaptations.

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024).

[14] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, et al., Phi-4 Technical Report, arXiv preprint arXiv:2412.08905 (2024).