HULAT-UC3M @ CLEARS2025 Subtask 1: Prompt-Based Simplification for Plain Language using Spanish Language Models

Lourdes Moreno[1], Jesus M. Sanchez-Gomez[1], Marco Antonio Sanchez-Escudero[1] and Paloma Martínez[1]

[1]*Universidad Carlos III de Madrid, Av. Universidad, 30, Leganés, 28911, Spain*

Abstract

This paper describes the participation of HULAT-UC3M in CLEARS 2025 Subtask 1: Adaptation of Text to Plain Language (PL) in Spanish. We explored strategies based on models trained on Spanish texts, including a zero-shot configuration using prompt engineering and a fine-tuned version with Low-Rank Adaptation (LoRA). Different strategies were evaluated on representative internal subsets of the training data, using the official task metrics, cosine similarity (SIM) and the Fernández-Huerta readability index (FH) to guide the selection of the optimal model and prompt combination. The final system was selected for its balanced and consistent performance, combining normalization steps, the RigoChat-7B-v2 model, and a dedicated PL-oriented prompt. It ranked first in semantic similarity (SIM = 0.75), however, fourth in readability (FH = 69.72). We also discuss key challenges related to training data heterogeneity and the limitations of current evaluation metrics in capturing both linguistic clarity and content preservation.

Keywords

Plain Language, Text Simplification, Spanish LLM

1. Introduction

Automatic text simplification using Natural Language Processing (NLP) methods is gaining increasing relevance in domains such as healthcare, public services, and finance. These tools aim to improve readability and comprehension, particularly for people with cognitive impairments or reading difficulties. However, to ensure real-world and social impact, simplification systems must not only align with international standards that guarantee the right to accessible information, but also involve end users in the NLP process itself. Within this framework, there are the initiatives of Plain Language (PL) and Easy Reading (ER), which promote the adaptation of texts to user needs and capabilities, especially people with intellectual disabilities or difficulties in reading comprehension, particularly in essential domains such as public administration, healthcare, education, and other key areas. The main international reference is ISO 24495-1:2023, which defines general principles applicable to any language and domain, stating that texts should be easy to find, understand, use, and evaluate by their target audience [1].

PL and ER initiatives constitute the methodological basis for textual simplification, as both aim to optimize vocabulary, sentence structure, and visual design to enhance message comprehension. UNE 153101:2018 EX in Spain, for example, provides guidelines for producing and validating documents in ER that explicitly require validation by people with intellectual disabilities [2]. Although PL and ER share the general objective of making the text understandable, there are important differences. PL is aimed at a broad audience, seeking to eliminate technicalities and complex expressions. Its principles

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

lmoreno@inf.uc3m.es (L. Moreno); jesusmsa@inf.uc3m.es (J. M. Sanchez-Gomez); marcoasa@inf.uc3m.es (M. A. Sanchez-Escudero); paloma.martinez@uc3m.es (P. Martínez)

https://hulat.inf.uc3m.es/ (L. Moreno); https://hulat.inf.uc3m.es/ (J. M. Sanchez-Gomez); https://hulat.inf.uc3m.es/ (M. A. Sanchez-Escudero); https://hulat.inf.uc3m.es/ (P. Martínez)

0000-0002-9021-2546 (L. Moreno); 0000-0002-6415-7467 (J. M. Sanchez-Gomez); 0009-0001-8163-5440 (M. A. Sanchez-Escudero); 0000-0003-3013-3771 (P. Martínez)

Table 1
Statistics of provided CLEARS Subtask 1 dataset

Description	Train dataset		Test dataset
	Original Texts	Adaptations	Adaptations
No. of samples	2,400	2,400	607
Avg. no. of words	431.48	189.31	326.43
Avg. no. of lines	6.63	5.30	10.77
Avg. word length (characters)	5.11	4.88	5.19

are based on short sentences, the use of inclusive and everyday language, and a structure that allows the key information to be located. It usually employs direct style, active voice, and segmentation into short paragraphs for easy reading. However, ER is specifically designed for people with intellectual disabilities and people with reading comprehension difficulties, and therefore requires even stricter guidelines for simplification. Additionally, some content presentation guidelines include bulleted lists, high-frequency words, abundant use of visual examples, and single-idea sentences, making it easier to read for people with cognitive barriers.

The participation of HULAT-UC3M in CLEARS Subtask 1 [3], within the IberLEF 2025 framework [4], focused on the use of Spanish generative Large Language Models (LLMs) through prompt engineering. Our goal was to explore different prompt strategies to generate high-quality PL adaptations, evaluating their effectiveness in balancing semantic fidelity and readability.

2. Dataset and Evaluation Metrics

This section describes the training and evaluation datasets, along with the evaluation metrics provided by CLEARS Subtask 1 to guide system design.

2.1. CLEARS Subtask 1 Dataset

As a first step, we explored the training dataset [5, 6] provided by the organizers of the Subtask 1 at CLEARS 2025 task [3]. It is composed of 2,400 texts extracted from official communications published between 2022 and 2023 by various Spanish municipalities, including Madrid, Alicante, Benidorm, Orihuela, Torrevieja, and Elda. These texts include topics of public interest and administrative relevance, and constitute the base material for evaluating textual adaptation to PL (see Table 1).

The adaptation methodology or guidelines followed by the experts who created the PL versions in the training dataset were not provided. Therefore, our aim was to analyze how the adaptations have been carried out. Although our research methodology follows ISO 24495-1:2023, we did not find clear evidence of its application in the training dataset. As a result, we chose not to apply our trained models and prompt strategies and instead to adapt our approach to reflect the style of the observed adaptations.

We found various inconsistencies in the training dataset that posed challenges for developing and evaluating our systems. For example, some adapted texts included information not present in the original (e.g., entry ID=2346), while others omitted relevant content due to excessive simplification (e.g., ID=1075).

In the test dataset, several source entries lacked a coherent semantic unit suitable for PL adaptation, such as:

- "Las actividades veraniegas en Parques continúan esta semana con:" *("Summer activities in parks continue this week with:")* (ID=542)
- "Más actividades en www.agendacultural.org" *("More activities at www.agendacultural.org")* (ID=538)

These issues introduced ambiguity and hindered the consistent application of adaptation strategies (see Table 7 in the Appendix for representative examples).

2.2. Evaluation Metrics

Subtask 1 at CLEARS 2025 defined two evaluation metrics to assess the quality of the PL-adapted outputs: Semantic Similarity (SIM) to measure content preservation, and the Fernández-Huerta Readability Index (FH) to assess linguistic clarity in Spanish. These metrics were applied to all generated outputs in both development and official test phases.

- **Semantic Similarity.** Semantic alignment between each system output and its PL reference was evaluated using cosine similarity under two complementary approaches: Bag-of-Words (BoW) and sentence embeddings. For the BoW method, we used `CountVectorizer` to obtain frequency vectors. For the embedding-based method, we used the multilingual model `paraphrase-multilingual-mpnet-base-v2` from `SentenceTransformers`. The final SIM score was computed as the average of both methods. The cosine similarity ranges from -1 to 1, with higher values indicating stronger semantic overlap.
- **Fernández-Huerta Readability Index.** This index is a standard formula for Spanish readability, based on the number of syllables per word and the number of sentences per 100 words. We implemented it in Python using the original mathematical formulation, combining the `pyphen` library for syllable segmentation and regular expressions for sentence boundary detection. The index yields a score from 0 (very difficult) to 100 (very easy), providing a quantifiable estimate of textual clarity.

3. Proposed Architecture

The architecture proposed by HULAT-UC3M was based on the application of prompt engineering to pre-trained transformer models with Spanish texts, in order to analyze the quality of texts adapted to PL. The objective was to maximize the metrics used in the shared task: Cosine similarity measure [7] and Fernández-Huerta readability index [8].

Concerning the generative model to be integrated, there is a wide variety of generative models, both encoder-decoder and decoder types, such as Mistral-7B-Instruct [9], Gemma [10], and StableLM [11]. However, they did not provide the level of linguistic quality or instructional control required for adapting texts to PL in Spanish. Processing was performed on infrastructure equipped with 23 GB of GPU memory and sufficient disk storage.

After this review and prioritizing models trained with Spanish texts, Salamandra [12] and RigoChat [13, 14] were selected as the most promising for text adaptation to PL. The Salamandra family includes Salamandra-7B-Instruct, which has been fine-tuned through instruction tuning, enabling it to follow instructions in generative tasks such as simplification, summarization, or rewriting [12, 15]. RigoChat, developed by the Instituto de Ingeniería del Conocimiento (IIC), is a generative LLM specialized in Spanish, trained with an instruction-tuning strategy focused on useful tasks for Spanish-speaking users. One of its key features is that it was trained under limited computational conditions, which demonstrates its applicability in resource-constrained environments. Moreover, RigoChat has been fine-tuned to maintain a balance between accuracy, clarity, and instructional control, which are fundamental aspects for the task of text simplification and adaptation to PL [16, 17]. Additionally, we explored fine-tuning strategies like Low-Rank Adaptation (LoRA) in Spanish generative models. These strategies, together with the preprocessing steps and the system variants described below, are detailed in the next sections.

3.1. Text Normalization

To adapt the output of the model to the type of adaptations observed in the training data, a set of filters and rules (implemented in *Python* using custom *regex*) were applied as pre and postprocessing. These processes ensured that the output of the model was consistent with the analysis of the training data. Examples of representative transformations include the following.

- Dates and times: expressions such as “20:00 hours” are normalized to more natural forms such as “at 8 in the evening”.
- Conversion of monetary amounts: numeric expressions like “13.50 EUR” are transformed into simpler forms, for example, “13 with 50 euros”.
- Protection and normalization: special numeric elements such as percentages, years, date ranges, and large numbers are processed to prevent misinterpretations with the model.

3.2. Prompt Design

To identify the most effective configuration for Subtask 1, we evaluated both Salamandra and RigoChat models using various prompt strategies. Our approach focused on three complementary prompt strategies, progressively refined through experimentation:

P1 – Two-step pipeline strategy: This strategy uses two consecutive prompts. The first phase performs a structural reduction by removing non-essential content, such as metadata, secondary names, and repetitive introductions, without altering the writing style. The second phase rewrites the remaining content using PL rules: short sentences, direct style, no complex subordinate clauses or repetitions, and simplified vocabulary. This approach was motivated by early observations of the training data, where many adaptations showed excessive content reduction. This two-step structure was designed to control that trend, especially in longer texts.

P2 – Unified prompt strategy: This strategy applies a single prompt that combines reduction and rewriting instructions. The model processes the full text using rules for clarity, conciseness, and direct language. Specifically, it includes instructions to avoid technical terms, long sentences, and direct quotations. Examples of these instructions include:

- Mantén solo las personas e instituciones clave, como el alcalde, concejal de deportes u organización principal *(Keep only key people and institutions, such as the mayor, sports councilor, or lead organization).*
- Simplifica ubicaciones muy específicas que no añadan valor principal *(Simplify very specific locations that do not add primary value).*
- Resume listas de nombres o participantes secundarios utilizando expresiones como “otras autoridades” o “otros participantes” *(Summarize lists of names or secondary participants using expressions such as “other authorities” or “other participants”).*
- Mantén las ideas principales y conserva el vocabulario clave del original para asegurar alta coincidencia de palabras *(Keep the main ideas and retain key vocabulary from the original to ensure high word matching).*

This strategy aims to improve readability within acceptable ranges without fragmenting or omitting essential content.

P3 – Category-based strategy: This strategy uses automatic classification to assign the input text to one of four categories: event listings, economic notices, short news items, or institutional notes. A set of common rewriting rules is applied, followed by category-specific adjustments such as formatting dates and prices, refining titles, or controlling the use of connectors. This classifier-based approach ensures stylistic and structural consistency depending on the content type.

4. Experimental Setup and Evaluation

This section presents the experimental setup and the evaluation results of the systems described in Section 3. We report the performance of the two proposed system variants across both internal validation

Table 2
Software environment

Component	Version / Details
Python	3.10.12
PyTorch	2.6.0 + CUDA 12.4
Transformers (HuggingFace)	4.50.0
PEFT	0.15.1
CUDA Toolkit	12.2

sets and the official evaluation data from Subtask 1. Evaluation focused on two metrics specified by the shared task organizers: Semantic Similarity and the Fernández-Huerta Readability Index, as outlined in Section 3.

4.1. Experimental Setup

All experiments were conducted on a shared Linux server running Ubuntu 22.04. The hardware setup included one NVIDIA L4 GPU (23 GB VRAM), an Intel Xeon Silver processor, 64 GB of RAM, and 877 GB of NVMe SSD storage. The NVIDIA driver version was 535.230.02. All model runs were performed on GPU. The software environment is summarized in Table 2.

System 1. Salamandra-7B and RigoChat-7B-v2 in zero-shot mode, guided only by prompt engineering strategies P1, P2, and P3 were used. The generation process was configured with `do_sample=False` to avoid sampling variability and promote output consistency. Input texts were first preprocessed using the normalization criteria described in Section 3.1. After generation, outputs were postprocessed to correct residual errors and preserve structural clarity, as models occasionally reversed or ignored earlier preprocessing operations.

We ran prompt-based generation and fine-tuning experiments on three representative subsets of the training data —*smallest*, *random*, and *categories*— to assess the effectiveness of each prompt strategy (P1, P2, P3) and identify the best-performing setup. These subsets, used for internal validation, are detailed in Section 4.2.

System 2. Fine-tuning of RigoChat-7B-v2 was performed using LoRA technique, allowing efficient training with minimal computational cost. We used the same training set composed of original texts and their PL adaptations, split into 85% training and 15% validation sets. Each fine-tuning experiment used a different prompt strategy (P1, P2, or P3), and all data were semantically normalized beforehand.

Fine-tuning was executed on an NVIDIA L4 GPU using the `transformers` and `PEFT` libraries. Training lasted approximately 2.5 hours per variant, with the following configuration: `r = 8`, `alpha = 32`, `dropout = 0.05`, `bias = "none"`, `task_type = "CAUSAL_LM"`, batch size = 16, evaluation at the end of each epoch, and between 3 and 5 epochs. All generated outputs were postprocessed to correct errors, ensure clarity, and remove non-essential content.

4.2. Training Data Subsets

To evaluate the performance of the Salamandra and RigoChat models before the official test set was released, we created three reduced but representative subsets from the training data. This preliminary evaluation enabled us to refine our prompt strategies efficiently and select the best-performing configuration.

The subsets were designed with distinct selection criteria:

- **Smallest subset:** Composed of 50 of the shortest texts in the dataset, characterized by low word count and brief structure (up to 90 words, with an average of 53.18 words). This subset was used to assess the structural simplification capabilities of the models.

Table 3
Internal evaluation results (System 1) for Salamandra and RigoChat across prompt strategies

Prompt	Salamandra		RigoChat	
	FH	SIM	FH	SIM
P1	74.57	0.69	**76.80**	**0.81**
P2	71.98	0.49	**81.90**	**0.85**
P3	83.22	0.50	**84.75**	**0.80**

Table 4
Performance scores on internal subsets (150 texts) using RigoChat (zero-shot) and RigoChat (fine-tuned)

Prompt	RigoChat (zero-shot)		RigoChat (fine-tuned)	
	FH	SIM	FH	SIM
P2	81.90	**0.85**	**83.22**	**0.81**
P3	**84.75**	0.80	83.33	0.78

- **Random subset:** Composed of 50 texts randomly sampled from the training set, representing a broad range of document lengths and topics (with an average word count of 447.78).
- **Categories subset:** Composed of 50 texts manually grouped by origin (e.g., Alicante, Elche, etc.), topic (e.g., Culture, Sport, etc.) and size (between 77 and 2,674 words, with an average of 818.34 words). This allowed us to explore how the models performed across different semantic domains.

These subsets were used for iterative prompt testing, qualitative review, and quantitative comparison, as shown in Section 4.3. This preliminary evaluation allowed us to calibrate our approach and ensure the reliability of the metrics applied to PL adaptations.

4.3. Internal Evaluation on Training Subsets

To evaluate the behavior of both models under different prompts and preprocessing conditions, we first conducted experiments on a subset of 150 texts (50 from each: *smallest*, *random*, and *categories*). We tested System 1 in zero-shot mode using the Salamandra and RigoChat models combined with each prompt strategy (P1, P2, P3), applying the preprocessing and postprocessing pipeline described in Section 3.1.

We evaluated outputs using the official task metrics: Fernández-Huerta for readability and Cosine Similarity for semantic preservation. The results are shown in Table 3.

As shown in Table 3, RigoChat outperformed Salamandra in both metrics across all prompts, with the highest SIM from P2 (0.85) and best FH from P3 (84.75). Prompt P1 showed the weakest performance and was discarded. Based on these results, we selected RigoChat for System 2 (fine-tuning) and retained only prompt strategies P2 and P3 for further experiments.

As shown in Table 4, RigoChat with prompt P2 offered the best overall performance. It achieved the highest cosine similarity (0.85 in zero-shot, 0.81 after fine-tuning) and strong readability (FH = 83.22). Although P3 reached the best FH in zero-shot (84.75), it showed higher variability and reduced SIM after fine-tuning. This issue in P3 is likely due to heterogeneity in the training data —differences in length, format, and structure that affected classification accuracy. P2, in contrast, consistently preserved key content while improving readability, without over simplifying or fragmenting the text. During the refinement phase, we noted that improving FH often required reformulation, which tended to reduce semantic similarity. P2 was the prompt to reach a stable equilibrium, preserving core content while raising readability to acceptable levels.

Based on these findings, we selected RigoChat with prompt P2 and no fine-tuning (System 1) as our final configuration. Using this setup, we generated PL outputs for the 607 official test texts and submitted them to the CLEARS 2025 Subtask 1 evaluation.

Table 5
Ranking based on Average Cosine Similarity

Subtask 1	Team	Avg. Cosine Similarity
1	**HULAT-UC3M**	**0.75**
2	VICOMTECH	0.71
3	NIL_UCM and CARDIFFNLP	0.70

Table 6
Ranking based on Average Fernández-Huerta readability index

Subtask 1	Team	Avg. Fernández-Huerta index
1	VICOMTECH	**82.98**
2	CARDIFFNLP	78.87
3	NIL_UCM	70.42
4	**HULAT-UC3M**	69.72

5. Results

The official evaluation results for Subtask 1 are presented in Tables 5 and 6. Our system, HULAT-UC3M, ranked 1st in Semantic Similarity, achieving a score of 0.75, and 4th in readability, with a Fernández-Huerta score of 69.72.

Our system achieved the best result in semantic preservation, reflecting its capacity to retain core content meaning. Although it ranked last in FH readability, this outcome stemmed from a strategic decision. Preliminary analysis revealed that the training data contained limited reformulation and focused on deletion or surface-level simplification, diverging from the normative principles of ISO 24495-1:2023 (Plain Language).

Applying our standard reformulation methods would have produced clearer outputs but at the cost of lower similarity scores due to misalignment with the reference. To balance both dimensions, we chose an intermediate prompt strategy (P2), which maintains full content coverage while applying minimal, clarity-oriented changes. Since the evaluation protocol did not specify how SIM and FH would be weighted, we prioritized content fidelity and semantic completeness.

It is also worth noting that our FH score of 69.72 lies within the "normal to somewhat easy" readability range according to the FH scale, corresponding to secondary education (ESO 2-3) or B1 CEFR level-acceptable for general public communications with functional literacy.

6. Error Analysis

This section presents an error analysis focused on two aspects: (1) limitations in the task-provided resources and (2) issues observed in our system's output.

Task-related limitations and insights. Several issues were identified during our participation in the task:

- **Low semantic similarity in training pairs:** Initial inspection showed that many (original, PL) pairs in the training dataset had moderate-to-low semantic similarity, likely due to poor segmentation and unclear alignment. Some adaptations included additional content (e.g., entry ID=2346) or omitted elements from the original (e.g., ID=1075), which may have affected semantic coherence.
- **Impact on system performance:** These inconsistencies affected system behavior. In **System 2 (fine-tuning)**, conflicting examples limited the model's ability to learn consistent simplification patterns. In **System 1 (zero-shot)**, the model sometimes under- or overgenerated content

depending on the prompt, with limited reformulation, which reduced readability from very good to moderate-good levels.
- **Lack of adaptation methodology:** No adaptation methodology or PL guidelines were provided with the dataset. As a result, we had to adjust our standards-based approach. This was evident in some of the results; the lack of consistent simplification guidelines introduced variability that weakened model generalization.

Reflections on our own approach. Within the constraints of the task, we identified areas in our system that could be improved:

- Managing the trade-off between readability and similarity remained challenging. Enhancing clarity often required reformulation, which decreased cosine similarity, especially in the absence of reference adaptations aligned with ISO 24495-1:2023.
- A more structured pipeline could help address this issue by incorporating text classification based on complexity, length, or target audience before generating simplifications based on LLM. This would allow models to adapt their behavior to specific input types and user needs, improving both clarity and content preservation.

7. Conclusions

The organization of tasks like this one fosters research on Plain Language (PL) adaptation, an increasingly relevant topic for accessibility in domains such as healthcare and public services. This initiative contributes to the development of systems that support people with cognitive impairments and reading comprehension difficulties.

Despite the complexity of the task, our system achieved competitive results and yielded valuable insights into the challenges of PL adaptation. Our approach combined two complementary strategies: System 1, based on zero-shot inference with large language models (LLMs) using three prompt variants (P1, P2, P3); and System 2, a fine-tuned model trained on the task-provided data. The system finally selected was System 1. Prompt P2 was ultimately chosen for generation due to its robust and generalizable behavior across diverse input types. Unlike Prompt P3, which classified the input text by type (e.g., informative, event, and others) before applying specific strategies, often resulting in rigid outputs, P2 followed a simpler, generic instruction aimed at enhancing comprehension while preserving meaning. This balance, System 1 with prompt P2, proved effective, as deeper reformulations to optimize the Fernández-Huerta index (readability) would have decreased similarity scores due to misalignment with reference texts. Given that the evaluation protocol did not specify the weighting between readability and similarity, we prioritized semantic preservation.

It would be advisable for future shared tasks to define a clear annotation methodology aligned with the regulatory context. Given the growing importance of PL in European and Spanish legislation, adopting standards like ISO 24495-1:2023 would help reduce variability and improve data quality.

Concerning evaluation, metrics remain a known limitation in the NLP community. Cosine similarity and the Fernández-Huerta index capture only surface-level features and may penalize meaningful reformulation. Future frameworks should combine metric-based and user-centered approaches, covering aspects such as factual consistency, sentence restructuring, and lexical simplification, among others.

Looking ahead, we plan to obtain and analyze the adapted versions of the task test dataset to conduct a more targeted error analysis. Future work will explore approaches that not only consider the user, but actively involve them in the process. We also aim to design richer evaluation frameworks that combine automatic metrics with user validation, ensuring alignment with real-world accessibility needs and regulatory standards such as ISO 24495-1:2023 for Plain Language and UNE 153101:2018 EX for Easy-to-Read.

Acknowledgments

This work has been supported by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

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

References

[1] International Organization for Standardization (ISO), Plain language – Part 1: Governing principles and guidelines, 2023. URL: https://www.iso.org/standard/78907.html, last accessed: 09-June-2025.

[2] Asociación Española de Normalización (AENOR), Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos, 2018. URL: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060036, last accessed: 09-June-2025.

[3] 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).

[4] 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, pp. 1–14.

[5] 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. Saggion, 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/.

[6] B. Botella-Gil, I. Espinosa-Zaragoza, P. Moreda, M. Palomar, Corpus ClearSim, 2024. URL: http://hdl.handle.net/10045/151688, last accessed: 09-June-2025.

[7] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[8] J. Fernández Huerta, Medidas sencillas de lecturabilidad, Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS) 1959 214 (1959) 29–32.

[9] Mistral AI, Mistral-7B-Instruct-v0.1, 2023. URL: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1, last accessed: 09-June-2025.

[10] Google DeepMind, Gemma-7B-it, 2024. URL: https://huggingface.co/google/gemma-7b-it, last accessed: 09-June-2025.

[11] Stability AI, StableLM, 2024. URL: https://huggingface.co/stabilityai, last accessed: 09-June-2025.

[12] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra Technical Report, 2025. URL: https://arxiv.org/abs/2502.08489. doi:10.48550/arXiv.2502.08489. arXiv:2502.08489, last accessed: 09-June-2025.

[13] G. S. Gómez, G. G. Subies, P. G. Ruiz, M. G. Valero, N. Fuertes, H. M. Zamorano, C. M. Sanz, L. R. Plaza, N. A. García, D. B. Sánchez, K. Sushkova, M. G. Nieto, Álvaro Barbero Jiménez, RigoChat 2: an adapted language model to Spanish using a bounded dataset and reduced hardware, 2025. URL: https://arxiv.org/abs/2503.08188, last accessed: 09-June-2025.

[14] Instituto de Ingeniería del Conocimiento, RigoChat-7b-v2, 2025. URL: https://huggingface.co/IIC/RigoChat-7b-v2. doi:10.57967/hf/4158, last accessed: 09-June-2025.

[15] Barcelona Supercomputing Center, Salamandra-7B-Instruct, 2024. URL: https://huggingface.co/BSC-LT/salamandra-7b-instruct, last accessed: 09-June-2025.
[16] Instituto de Ingeniería del Conocimiento, El IIC libera un modelo de lenguaje generativo en español: RigoChat v2, 2024. URL: https://www.iic.uam.es/noticias/iic-libera-modelo-lenguaje-generativo-espanol-rigochat-v2/, last accessed: 09-June-2025.
[17] Instituto de Ingeniería del Conocimiento, RigoChat v2: adaptando LLMs al español con fines prácticos y recursos limitados, 2024. URL: https://www.iic.uam.es/noticias/rigochat-v2-adaptando-llms-al-espanol-con-fines-practicos-y-recursos-limitados/, last accessed: 09-June-2025.

A. Appendix - Examples of Inconsistencies in Training Dataset

Table 7
Examples of Inconsistencies in Training Data

ID	Original Text	PL Adaptation
2346	"La Concejalía de Igualdad les invita a participar en el programa de actividades "Igualdad y diversidad en los barrios" durante los meses de octubre a diciembre de 2022. Si desea más información consulte los documentos en .pdf anexos al final de la noticia."	"La concejalía de Igualdad presenta el programa de actividades de otoño del 2022. La concejalía de Igualdad invita a todas las personas a participar en el programa de actividades de igualdad y diversidad en los barrios de Alicante de octubre a diciembre del 2022. Programación Cuentacuentos Mili, Reportera superstar y los 3 cerditos: - El miércoles 26 de octubre en la biblioteca Florida Babel - El miércoles 23 de noviembre en la biblioteca El cabo - El lunes 12 de diciembre en la biblioteca Virgen del Remedio La verdad de las princesas: - El lunes 17 de octubre en la biblioteca Diagonal - El martes 18 de octubre en la biblioteca Villafranqueza - El miércoles 19 de octubre en la biblioteca San Blas Malena quiere volar: - El miércoles 5 de octubre en la biblioteca Benalúa - El lunes 10 de octubre en la biblioteca - El miércoles 9 de noviembre en la biblioteca Diagonal ¡Me lo pido! - El lunes 14 de noviembre en la biblioteca Virgen del Remedio - El martes 22 de noviembre en la biblioteca Carolinas - El miércoles 30 de noviembre en la biblioteca Benalúa ... **[NOTE: Over 277 words words omitted due to space constraints]**
1075	"El alcalde valora la trayectoria personal y profesional del futbolista de Sanlúcar de Barrameda, su capacidad goleadora y su intuición para situarse en el área del equipo contrario. Luis Barcala ha afirmado este domingo en el Salón Azul del Ayuntamiento de Alicante que "este emblemático lugar ha sido testigo de recepciones a personalidades de distintos ámbitos de la vida política, social y cultural. También los Alicantinos de Adopción, lo han visitado. Lo que no ha ocurrido nunca es que en él se haya homenajeado a una leyenda, en blanco y azul, como es el caso de Eduardo Rodríguez". Barcala ha agradecido a la Federación de Casas Regionales en Alicante, que preside Miguel Beano, que haya pensado en Rodríguez como Alicantino de Adopción 2022 "en el año del centenario del Hércules, y como reconocimiento al máximo goleador de su historia". El alcalde ha recordado cómo llegó Rodríguez al Hércules, procedente del Badajoz, pagando un traspaso de ocho millones. Ha subrayado, dirigiéndose al homenajeado, que "si eres hombre de pocas palabras, sí hablabas en el campo. Tu lenguaje era el de los goles". ... **[NOTE: Over 227 words omitted due to space constraints]**	"La Federación de Casas Regionales en Alicante elige al futbolista Eduardo Rodríguez como Alicantino de Adopción 2022. El alcalde de Alicante destacó en el acto de celebración, la capacidad del futbolista del Hércules para marcar goles y su intuición en el área del rival. Es la primera vez que se hace un reconocimiento a una leyenda blanca y azul y máximo goleador de la historia del club. El alcalde le felicitó por el reconocimiento como Alicantino en Adopción y por su carrera profesional. Al acto acudieron representantes de las Casas Regionales en Alicante, el ex capitán del Hércules Paquito y el ex presidente de la Diputación Provincial. El acto se celebró en el Salón Azul."