Automated Redundancy Adaptation for Easy-to-Read in Spanish: Structuring a Pipeline Around Large Language Models

Anonymous ACL submission

Abstract

Linguistic redundancy can pose challenges for people with reading comprehension difficulties. According to the Easy-to-Read (E2R) Methodology, eliminating unnecessary words improves readability, particularly for individuals with cognitive disabilities. However, the E2R adaptation remains a manual and time-consuming process, highlighting the need for technological support. Thus, this study explores an automatic approach to adapt redundancies in Spanish, leveraging large language models (LLMs). We propose a pipeline that integrates redundancy detection, controlled adaptation through prompt engineering, and consensus-based verification. Given the lack of annotated redundancy datasets, we generate and validate synthetic data in Spanish to improve model performance. We evaluate two Spanish-pretrained LLMs, Salamandra-7B-Instruct and Llama-3.1-8B-Instruct, analysing their effectiveness in redundancy processing. Our results show that Llama-3.1-8B-Instruct performs better in detection and adaptation, whereas Salamandra-7B-Instruct excels in verification. This study demonstrates the feasibility of LLMs for E2R redundancy adaptation, offering a scalable and structured approach for automated text simplification and accessibility. Future work will focus on optimising computational efficiency, expanding training data, and refining adaptation techniques.

1 Introduction

Linguistic redundancy is a common phenomenon in natural language, where unnecessary repetition or superfluous expressions increase the length of a text without adding new meaning. While redundancy can sometimes serve rhetorical or stylistic purposes, excessive redundancy (often referred to as wordiness[1]) can obscure meaning, increase cognitive load, and reduce the efficiency of communication (Chandler and Sweller, 1991). According to the Easy-to-Read (E2R) Methodology (Inclusion Europe, 2009; Nomura et al., 2010; AENOR, 2018), the use of words that do not contribute essential information to the text should be avoided, as they make reading more difficult, particularly for individuals with cognitive disabilities. The goal of the E2R methodology is to present clear and easy to understand text by providing a set of guidelines on content, design and layout of written materials. This adaptation process is iterative and involves three key activities: analysis, adaptation and validation. Nevertheless, the E2R methodology is currently applied in a manual fashion. Such a manual process is labour-intensive and costly, and it would benefit from having a technological support.

Traditional redundancy reduction techniques have relied mainly on rule-based and syntactic approaches (Wilks et al., 1996; Jurafsky and Martin, 2009), which focus on detecting explicit repetition but struggle to capture semantic redundancy (i.e. cases where an expression is redundant in context rather than in form). The rise of large language models (LLMs) has introduced new possibilities for automating redundancy detection and adaptation, as these models are trained on vast corpora that inherently include various instances of redundancy in natural text. Recent advances in generative Natural Language Processing (NLP) have demonstrated the effectiveness of LLMs in paraphrasing (Raffel et al., 2020), text summarisation (Zhang et al., 2020), and text simplification (Xu et al., 2016; Martin et al., 2022) tasks.

Despite recent advancements, the automatic adaptation of redundancy according to the E2R guidelines still faces several challenges. One of the main issues is the lack of annotated datasets, as there is no large-scale corpus specifically designed for redundancy detection and adaptation in Spanish, making synthetic data generation a necessary

[1] https://www.wordreference.com/definition/wordiness

alternative. Furthermore, evaluation limitations remain an issue, since standard NLP metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily assess surface-level similarity and tend to penalise semantically valid reformulations. Finally, balancing precision and fluency is critical, as eliminating redundant expressions can sometimes unintentionally alter meaning.

To address these challenges, we introduce a pipeline for automated redundancy handling in Spanish, integrating several processes. First, redundancy detection and classification is performed using LLMs in a zero-shot setting to evaluate textual redundancy. Next, redundancy adaptation is achieved through prompt engineering techniques, including zero-shot, few-shot, and chain-of-thought (CoT) prompting, to enhance model performance. Additionally, synthetic dataset generation enables the construction of diverse training corpora, facilitating future fine-tuning and evaluation. Finally, consensus-based evaluation ensures reliability by requiring a high level of agreement between multiple LLM-generated responses, reinforcing the robustness of the adaptation process.

Specifically, in this research work we evaluate the effectiveness of two Spanish-pretrained LLMs, Salamandra-7B-Instruct and Llama-3.1-8B-Instruct, assessing their performance in redundancy detection, adaptation, and verification. Our findings provide insights into the capabilities of LLMs for E2R adaptation tasks and highlight the importance of structured evaluation pipelines for redundancy adaptation.

The rest of this paper is structured as follows: Section 2 reviews prior research on automatic approaches to identifying and adapting redundancy. In Section 3 we present the proposed methodology, detailing the stages of generation, adaptation, and classification, along with the selected models. Section 4 describes the experimental setup and assesses model performance across these tasks. Sections 5 and 6 introduce the proposed processing pipelines, integrating the previously analysed methods. Finally, we present some conclusions and future work, as well as the limitations of the research work.

2 State of the Art

The study of redundancy in texts has been approached from different perspectives within NLP. One common line of research considers redundancy in terms of how much information within a sentence is already contained in previously selected sentences. In this regard, Thadani and McKeown (Thadani and McKeown, 2008) proposed a graph-based algorithm for identifying redundancy at the sub-snippet level, where a snippet is defined as the smallest unit of text that can be removed to reduce redundancy. Similarly, Xiao and Carenini (Xiao and Carenini, 2020) explored redundancy reduction in neural summarisation of long documents, proposing two methods that explicitly address redundancy in the sentence selection phase. Bi and colleagues (Bi et al., 2021) also contributed to this area by introducing adaptive learning models, such as AREDSUM-SEQ and AREDSUM-CTX, which aim to balance salience and redundancy in extractive summarisation models.

An alternative perspective on redundancy focusses on words that do not contribute to the meaning of a sentence. That is, a phrase is considered redundant if its removal does not alter the sentence's meaning. However, determining redundancy from this stylistic viewpoint presents challenges. Xue and Hwa (Xue and Hwa, 2014) developed a computational model to detect sentence-level redundancies, combining language model scores with measures of word contribution to meaning.

Despite these advancements, current approaches still lack a systematic method for handling linguistic redundancies. Traditional techniques, such as syntactic parsing and rule-based systems (Wilks et al., 1996; Jurafsky and Martin, 2009), have proven effective for detecting explicit redundancies but struggle with semantic and contextual dependencies, limiting their applicability to more nuanced redundancy adaptation. These limitations have motivated the adoption of large language models (LLMs) as an alternative for redundancy detection, adaptation, and validation.

Recent advances in generative NLP have demonstrated that LLMs can successfully perform sentence transformation tasks, including paraphrasing (Raffel et al., 2020), summarising (Zhang et al., 2020), and text simplification (Xu et al., 2016; Martin et al., 2022). These tasks share similarities with redundancy adaptation, particularly in the need to restructure text while preserving meaning. However, while text adaptation research has focused primarily on accessibility for diverse audiences, it has not explicitly addressed the detection and adaptation of redundant structures in Spanish.

Another key aspect of redundancy processing is the automatic evaluation. Existing evaluation metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), have been widely used in paraphrase generation and text simplification, but they exhibit limitations in capturing subtle meaning changes (Akter and Karmaker, 2024). More recent approaches incorporate model-based evaluation strategies, where LLMs act as evaluators in zero-shot classification settings (Kocmi and Federmann, 2023). This methodology, often referred to as LLMs-as-a-judge, has been applied in text quality assessment and aligns with redundancy verification by determining whether an adaptation preserves the intended meaning and grammar.

Data generation is another critical process in redundancy processing. While manually curated corpora are standard in many linguistic tasks, redundancy adaptation requires high-variability datasets capturing pleonasm, tautology, and circumlocution. Large-scale text generation using LLMs has been explored in various domains (Yoo et al., 2021), particularly for synthetic dataset augmentation in low-resource NLP applications. Few-shot prompting techniques have been widely used for controlled data generation, especially in paraphrasing and reformulation tasks (Liu et al., 2022), providing a methodological foundation for structured redundancy generation and adaptation.

Furthermore, a modular approach to redundancy processing aligns with contemporary NLP trends. In this paradigm, successive models specialise in distinct subtasks, a strategy popularised by frameworks such as LangChain[2]. These frameworks decompose complex problems into manageable steps, leveraging methodologies such as prompt chaining, where each LLM call processes the output of the previous one. Recent work by Anthropic[3] discusses structured LLM workflows, including prompt chaining, which supports the methodological principles underlying redundancy processing. Even though these frameworks were not initially developed for redundancy detection and adaptation, they offer valuable insights into modular NLP architectures.

Although previous research has addressed redundancy in tasks such as summarising and paraphrasing, it has not systematically addressed its detection, adaptation, and validation in Spanish to make this phenomenon more accessible and easy-to-read. Moreover, the lack of annotated corpora limits progress in this area. This study thus addresses these gaps by proposing a structured LLM-based approach that integrates a pipeline for detection, adaptation, and verification of redundancies in Spanish according to the E2R guidelines.

3 Method[4]

This section details our methodology for addressing redundancy in Spanish text using LLMs. Our approach is composed around three key processes: (1) consensus-based redundancy detection and verification via specialized classification tasks; (2) redundancy adaptation; and (3) synthetic data generation. A crucial aspect of our work is model selection, balancing performance and computational cost.

The following subsections outline the model selection criteria, the specific models chosen, and the techniques used to evaluate the effectiveness of each model for each of the defined processes.

3.1 Selected Models: Salamandra and Llama-3.1

Model size is critical for optimising computational costs, as inference is computationally and temporally expensive, requiring specialised GPUs to achieve acceptable throughput for end users. This necessitates a trade-off between model size and performance. Our preliminary experiments demonstrated that mid-sized models, such as those with approximately 7B parameters, strike an optimal balance for the evaluated tasks as they deliver sufficient performance without requiring exceptional hardware resources.

We evaluated two models: Salamandra-7B-Instruct[5] and Llama-3.1-8B-Instruct[6]. Both models support Spanish, which was a key selection criterion.

As inference engine, we used Hugging Face's Transformers library[7].

[2] https://www.langchain.com

[3] https://www.anthropic.com/research/building-effective-agents

[4] To maintain anonymity during review, data and code repositories have not been included. To ensure full reproducibility, these repositories will be made public upon acceptance of the manuscript

[5] https://huggingface.co/BSC-LT/salamandra-7b-instruct

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/docs/transformers

3.2 Redundancy Detection and Verification: Classifying Redundancy Using LLMs

This section examines classification tasks that employ zero-shot prompting to detect redundancies and verify whether a redundant sentence has been properly adapted. Different data sets were created, destined to benchmark the performance of the models in each classification task.

3.2.1 Classification Tasks for Detection and Verification

Three complementary tasks have been evaluated, capturing the different aspects and properties of redundancies:

1. *Redundancy Detection:* Assesses whether the original sentence contains a redundancy relative to its adapted version.

2. *Information Preservation:* Assesses whether the adapted sentence retains the essential message without loss of information.

3. *Semantic Coherence:* Assesses whether the adapted output is grammatically correct and semantically coherent in Spanish.

3.2.2 Benchmarking Language Models in Classification Tasks

To evaluate model performance in redundancy handling tasks, we constructed three specialised datasets, each containing 500 manually annotated entries, designed for zero-shot classification. The datasets were curated from linguistic resources, editorial guidelines, and expert-validated examples to reflect realistic redundancy patterns in Spanish.

The first dataset, *Redundancy Detection*, consists of 500 sentence pairs where one is a redundant version of the other. It includes both correctly (non-redundant) and incorrectly modified examples.

For *Information Preservation*, the second dataset evaluates meaning retention during simplification. It contains 250 cases with successful preservation and 250 instances of partial or total information loss, for a total of 500 data entries.

The third dataset, *Semantic Coherence*, is composed of 500 individual sentences to test semantic integrity. It encompasses both coherent and incoherent examples across varying complexity levels.

3.2.3 Consensus Based Classification

For each entry in the dataset, five identical queries are issued to the LLMs to obtain binary responses

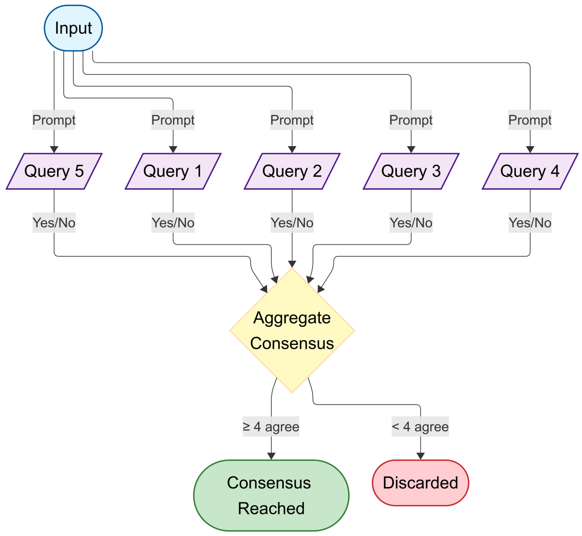

Figure 1: Consensus-based classification workflow.

(Yes/No). This approximates a multi-annotator agreement scenario where multiple reviewers might be consulted. A consensus is reached when at least 80% (4 out of 5 responses) agree. This threshold was chosen to balance precision and efficiency in decision-making, providing a robust measure of agreement while minimizing the impact of occasional anomalous responses from the model. Using more than 5 queries, while potentially desirable for increased reliability, is impractical due to the high temporal and computational cost required for inference with LLMs. This classification process is illustrated in Figure 1.

In consensus methodologies, thresholds for defining consensus vary widely, typically ranging from 51% to 100%. The median threshold for defining consensus in one review paper was found to be 75% (Gottlieb et al., 2023). In line with this, a threshold of 80% is considered reasonable to achieve consensus among experts. In the verification prompt, we specified that the model should only respond with ‘Yes’ or ‘No’.

3.3 Redundancy Adaptation with Prompting

Having established a method for detecting and verifying redundancies, we next focus on the process of adapting redundant sentences. Both Llama-3.1-8B-Instruct and Salamandra-7b-Instruct were prompted to generate adapted versions of redundant sentences while preserving the original meaning and grammatical correctness.

This section describes the prompt engineering techniques used to tackle redundancies and generate adapted sentences according to the E2R methodology. We also describe the evaluation criteria to select the most effective prompt to solve this issue.

3.3.1 Prompt Engineering to Adapt Redundant Sentences

Three core prompting paradigms, zero-shot, few-shot, and chain-of-thought (CoT), were systematically deployed across eight prompt configurations[8]. Zero-shot approaches featured two instruction levels: basic (concise redundancy adaptation commands) and elaborated (enhanced with the Easy-to-Read guidelines(AENOR, 2018)). Few-shot implementations diversified through exemplar quantity, instructional phrasing variants, and strategic inclusion of non-redundant control samples to mitigate over-correction tendencies. CoT prompts guided the model through a step-by-step reasoning process, variations included non-redundant sentences as control cases.

3.3.2 Evaluating Prompts Effectiveness: Evaluation Criteria

To evaluate adaptation, we assessed the models using the following metrics:

1. *Parsability:* This metric measures the percentage of generated outputs that match a predefined, regular-expression-based format determined in the prompt, ensuring reliable automated processing.

2. *Correct Adaptation:* Percentage of redundant sentences that were evidently adapted into fluent, natural and coherent sentences removing any repetitive structures (pleonasm, circumlocution, or redundancy).

3. *Inappropriate Modification:* Percentage of non-redundant sentences that were unnecessarily or incorrectly modified.

3.4 Redundancy Generation: Creating Quality Synthetic Data

Unlike simpler linguistic phenomena with easily definable rules, redundancies often rely on subtle semantic and contextual cues. Manually generating these data is an exhausting, time-consuming, and limited by the imagination task; and creating a comprehensive, rule-based system to generate these redundancies is extremely challenging, and has not ever been developed. Therefore, leveraging the generative capabilities of LLMs provides a valuable approach to building a diverse and representative dataset for further training and evaluating redundancy handling systems.

All synthetically generated sentence pairs underwent automated verification using a three-stage classification process composed by the tasks described in Section 3.2, employing the Llama-3.1-8B-Instruct model as-a-judge due to its strong performance in classification tasks. A sentence pair is verified after successfully passing the three verification stages, each requiring an 80% consensus: 4 out of 5 positive judgments.

3.4.1 Prompts for Redundancy Generation

To generate redundant structures in Spanish, we evaluated the following two techniques:

- *Few-Shot Generation:* A standard few-shot prompt was used, providing the model with representative examples of redundant sentences and their adapted versions. Each model was requested 10 times to generate new pairs of redundant and adapted sentences.

- *Few-Shot Paraphrasing with References:* This approach utilised a few-shot prompt that iterated once through a dataset of 150 human-verified reference pairs. The model was tasked with generating three variations for each pair in the dataset, while preserving semantic correctness and adhering to the E2R guidelines.

4 Model Performances in Classification, Adaptation, and Generation

This section presents our experimental results comparing Llama-3.1-8B-Instruct and Salamandra-7B-Instruct across the three key processes of our methodology: (1) classification tasks, (2) adaptation capabilities, and (3) generation of redundant Spanish sentences. In the following subsections, we describe and interpret the experimental results, focusing on model performance, failure patterns, and linguistic challenges.

4.1 Benchmark Results for Classification Tasks: Detection, Information Preservation, and Coherence

Using the methodology described in Section 3.2 (Redundancy Detection and Verification: Classifying Redundancy Using LLMs), both models were evaluated, as shown in Table 2. For instance, in redundancy detection, Llama-3.1-8B-Instruct

[8]Detailed prompt design is shown in Appendix A.

scored 0.946 for redundant cases versus 0.930 for Salamandra-7B-Instruct. In Information Preservation, both models achieved high precision (0.971 for Llama-3.1 and 0.984 for Salamandra), and in Semantic Coherence, Llama reached a perfect 1.000 when identifying coherent sentences compared to 0.966 for Salamandra. Overall, both models are competent in all tested categories, with their performance remaining very close across all classification tasks.

Category	LLaMa-3.1-8B-Instr	Salamandra-7B-Instr
Redundancy Detection		
Redundant	0.946	0.930
Non-Redundant	0.669	0.448
R-W-A*	0.912	0.873
N-R-W-A**	0.976	0.929
Information Preservation		
Preserved	0.844	0.992
Not-Preserved	0.976	0.984
Semantic Coherence		
Coherent	1.000	0.966
Incoherent	0.933	0.773

Table 1: Performance Comparison in Detection and Verification Tasks: Classification Success Rate per Class.

*Redundant-Wrong-Adaptation class
**Non-Redundant-Wrong-Adaptation class

Measure	LLaMa-3.1-8B-Instr	Salamandra-7B-Instr
Redundancy Detection*		
Weighted Precision	0.904	0.866
Weighted Recall	0.875	0.795
Weighted F1-score	0.880	0.808
Information Preservation		
Precision	0.971	0.984
Recall	0.844	0.992
F1-score	0.904	0.987
Semantic Coherence		
Precision	0.937	0.810
Recall	1.000	0.966
F1-score	0.967	0.880

Table 2: Precision, Recall, and F1-Score for Each Classification Task (N=500).

*Note: Weighted metrics are used for the redundancy detection category due to class are imbalance (125 positive vs. 375 negative). Weighting ensures that each class's performance is fairly represented according to its frequency, providing a more accurate overall assessment of the model.

4.2 Redundancy Adaptation with Prompting: Key Findings

After determining the models' performance in redundancy classification tasks, we investigated their capacity to adapt redundant sentences. For this purpose, multiple prompting strategies were tested. Each prompt was evaluated on three metrics detailed in Section 3.3.2. Higher percentages of *Parsability* and *Correct Adaptation* are considered positive, while lower percentages of *Inappropriate Modification* are desired. Results for each tested prompt are shown in Table 3.

The Llama-3.1-8B-Instruct model demonstrated robust performance in adapting redundant Spanish sentences, achieving a 90% correct adaptation rate and 100% parsability. In stark contrast, Salamandra-7B-Instruct exhibited critical shortcomings, with $\leq 20\%$ correct adaptation rates, due to recurring errors such as unjustified semantic substitutions, orthographic inaccuracies, and arbitrary modifications that distorted the original meaning.

We observed that a lower rate of inappropriate modifications correlates directly with fewer correct adaptations, showcasing a intrinsic tradeoff. Prompting strategies effective at removing redundancies simultaneously increase the risk of over-editing non-redundant sentences. This phenomenon stems from the inherent difficulty of distinguishing structural repetition from semantically meaningful content in morphologically rich languages like Spanish.

Failures predominantly occurred when models struggled to preserve contextually essential qualifiers or domain-specific details during text adaptation, underscoring the challenge of balancing redundancy elimination with semantic fidelity. A granular error analysis is detailed in Appendix B.

4.3 Generation of Redundant Sentences: Model Performances

For generating synthetic data, we evaluated two approaches: few-shot generation and reference-guided paraphrasing (Section 3.4). Llama-3.1-8B-Instruct demonstrated functional utility, producing 26 valid redundant-adapted pairs across 10 iterations via direct few-shot generation, albeit with repetitive syntactic structures and thematic overlap. Its performance improved markedly in the reference-guided paradigm: using 150 example pairs, it generated 78 valid variations (52% yield rate), showcasing context-sensitive paraphrasing capabilities.

Salamandra-7B-Instruct failed in basic few-shot generation, producing zero parsable outputs. Even with reference examples, it achieved only two valid pairs due to syntactic errors and semantic drift.

Technique	Parsability		Correct Adaptation		Inappropriate Change	
	Llama-3.1	Salamandra	Llama-3.1	Salamandra	Llama-3.1	Salamandra
Zero-shot 1	N/A	N/A	40%	20%	10%	10%
Zero-shot 2	100%	95%	90%	20%	70%	10%
Few-shot 1	95%	5%	80%	10%	90%	90%
Few-shot 2	5%	85%	50%	0%	60%	80%
Few-shot 3	100%	0%	50%	0%	80%	90%
Few-shot 4	100%	80%	20%	0%	90%	90%
CoT 1	75%	70%	70%	0%	80%	90%
CoT 2	85%	80%	60%	0%	40%	80%

Table 3: Adaptation Performance Comparison: Llama-3.1-8B-Instruct vs. Salamandra-7B-Instruct.

5 A Pipeline for Redundancy Adaptation Leveraging LLMs

We propose the use of a detection-adaptation-parsing-verification pipeline (see Figure 2) to automate redundancy adaptation. This pipeline coordinates the following tasks:

1. *Detection:* This leverages the redundancy detection task from Section 3.2.1, which accurately identified 94.6% of redundant phrases with Llama-3.1-8B-Instruct.

2. *Adaptation:* If a redundancy is detected, the model is prompted to generate an adapted version of the sentence. This uses the best-performing prompt, Zero-shot 2, with a 90.0% success rate with Llama-3.1-8B-Instruct.

3. *Parsing:* The system parses the adapted sentence from the previous step to extract the adapted sentence itself. This step prepares the output for the verification stage. Zero-shot 2 has a 100.0% parsability rate, which means that the adapted sentence can always be reliably extracted.

4. *Verification:* The verification chain is composed of the methods described in Section 3.2.1. For each verification criterion, we select the model that demonstrated the highest accuracy on that specific task. The selected methods exhibit the following classification rate on correctly modified sentence pairs: (1) 94.6% for Redundancy Detection, (2) 99.2% for Information Preserve, and (3) 100.0% for Semantic Coherence.

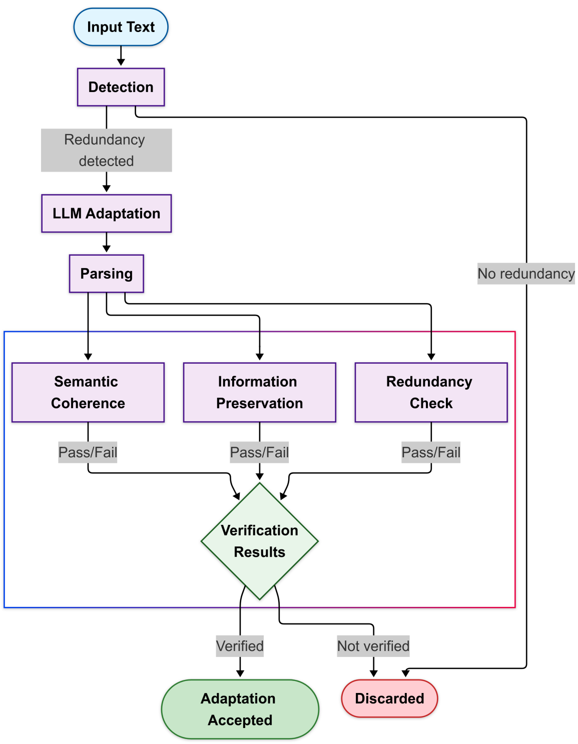

Figure 2: Pipeline for redundancy adaptation.

5.1 Redundancy Adaptation Pipeline Performance[9]

The pipeline achieved a 79.9% end-to-end success rate, considering an input phrase successfully identified and adapted.

False positive analysis revealed critical safeguards: when the system correctly detected a redundancy but generated flawed adaptations, erroneous outputs passed verification with the following probability: $4.52 \times 10^{-6}\%$. Conversely, if the system incorrectly detects redundancy (i.e. there is no redundancy in the original phrase), and subsequently generates a modified phrase, the probability of passing the verification chain is: $4.93 \times 10^{-7}\%$.

These low probabilities indicate that the verification chain is highly conservative. It is much more likely to reject a valid adaptation than to accept an erroneous one. This low false positive rate is a desirable characteristic in applications where preserving the original meaning is critical, even at the cost of potentially missing some redundancies.

6 Automated Synthetic Data Generation Pipeline: Using LLMs to Generate Redundancies

To address the limitations of scarce, manually created redundancy examples, we propose a modular pipeline leveraging Llama-3.1-8B-Instruct to synthesise linguistically complex examples. Building on its demonstrated paraphrasing capabilities (Section 4.1), the workflow begins with the injection of reference data, seeding the model with human-verified redundant adapted pairs.

The model then performs context-aware paraphrasing, generating structurally diverse variations while preserving redundancy patterns. A multistage verification chain, combining redundancy detection, semantic preservation checks, and coherence validation, automatically filters the output, achieving the near-deterministic rejection of flawed samples (Section 4.3). While fully automated operation is feasible, an optional human curation layer further polishes high-value edge cases, ensuring dataset integrity for critical applications.

This hybrid approach balances scalability with precision, enabling iterative dataset expansion while mitigating hallucination risks inherent to LLM-based generation. The following list summarises the process:

1. *Reference Data Input:* Supply the model with a set of human-verified pairs of redundant phrases and their adapted forms.

2. *Few-shot Paraphrasing with References:* Utilize this prompt to instruct Llama-3.1-8B-Instruct to generate variations based on the reference data.

3. *Automated Verification:* Apply the verification chain (Redundancy Detection, Information Preservation, Semantic Coherence) to automatically filter the generated pairs.

4. *[Optional] Human Review:* Incorporate a human review stage to further ensure the quality and correctness of the generated data.

7 Conclusions and Future Work

This study presents a structured detection, adaptation, and verification pipeline as an effective approach for handling redundancy in Spanish according to the Easy-to-Read guidelines. The results obtained highlight its potential for NLP tasks such as text adaptation. Furthermore, the structured design of this method suggests its applicability to broader NLP challenges that are difficult to address with traditional rule-based techniques. A key contribution of this study is the potential for the generation of synthetic data, where the method based on few-shot paraphrasing with references, combined with a multistage verification process, extended a reference dataset by 52%. These automatically validated datasets could also support fine-tuning of models tailored to redundancy adaptation. In terms of model performance, Llama-3.1-8B-Instruct proved to be more efficient in most tasks, while Salamandra-7B-Instruct excelled in verification and information preservation, highlighting the importance of task-specific model selection in redundancy adaptation. However, despite these results, the study also reveals areas that require further refinement. Future work should focus on optimizing computational efficiency, expanding high-quality datasets, and developing specialized models to enhance redundancy processing further. Additionally, separating generation and verification models could mitigate biases and improve evaluation robustness, contributing to the development of more reliable and scalable NLP solutions.

[9] See Appendix C for detailed calculations.

8 Limitations

Despite the promising results achieved in this study, several limitations must be acknowledged. One of the primary constraints is the computational and temporal cost of inference. The use of LLMs requires specialised GPU resources to ensure acceptable throughput, making real-time application in large-scale systems a challenge. The trade-off between model size and performance remains a critical consideration, as larger models typically provide better accuracy but at the expense of significantly higher computational demands. Another limitation is the scarcity of reference methodologies for redundancy adaptation. Redundant structures are inherently complex, deeply rooted in semantic and contextual factors, making it difficult to rely on purely syntactic or rule-based approaches. This lack of established benchmarks and methodologies restricts the ability to systematically compare different approaches and evaluate progress in the field. Without a clear standard for redundancy handling, assessing the effectiveness of LLM-driven approaches remains a challenge. Additionally, the bottleneck in the adaptation process is a key limitation. While our study leverages pre-trained LLMs with prompt-based strategies, achieving a fully automated system without human supervision would require the development of fine-tuned models specifically optimised for redundancy processing. The absence of such specialised models limits the potential for seamless real-time redundancy adaptation, since general-purpose LLMs are not explicitly trained for this task and may introduce inconsistencies in their outputs.

References

AENOR. 2018. *Easy-to-Read. Guidelines and recommendations for the production of documents (UNE 153101:2018 EX)*. Asociación Española de Normalización.

Mousumi Akter and Santu Karmaker. 2024. Redundancy aware multiple reference based gainwise evaluation of extractive summarization. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 182–195, Vienna, Austria. Association for Computational Linguistics.

Keping Bi, Rahul Jha, Bruce Croft, and Asli Celikyilmaz. 2021. AREDSUM: Adaptive redundancy-aware iterative sentence ranking for extractive document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 281–291, Online. Association for Computational Linguistics.

Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332.

Michael Gottlieb, Holly Caretta-Weyer, Teresa M. Chan, and Susan Humphrey-Murto. 2023. Educator's blueprint: A primer on consensus methods in medical education research. *AEM Education and Training*, 7(4):e10891.

Inclusion Europe. 2009. *Information for All. European standards for making information easy to read and understand*. Inclusion Europe.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edition. Prentice Hall.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaochen Liu, Yang Gao, Yu Bai, Jiawei Li, Yinan Hu, Heyan Huang, and Boxing Chen. 2022. PSP: Pre-trained soft prompts for few-shot abstractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6355–6368, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

M. Nomura, G. S. Nielsen, International Federation of Library Associations and Institutions, and Library Services to People with Special Needs Section. 2010. *Guidelines for easy-to-read materials*. IFLA Headquarters, The Hague.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kapil Thadani and Kathleen McKeown. 2008. A framework for identifying textual redundancy. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 873–880, Manchester, UK. Coling 2008 Organizing Committee.

Yorick A Wilks, Brian M Slator, and Louise M Guthrie. 1996. *Electric words : dictionaries, computers and meanings*. ACL-MIT Press series in natural language processing. The MIT Press, Cambridge, Massachusetts ;.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Huichao Xue and Rebecca Hwa. 2014. Redundancy detection in ESL writings. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 683–691, Gothenburg, Sweden. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

A Full Prompts Reference

These prompts have been adapted from Spanish into English.

A.1 Consensus Based Classification Prompts

A.1.1 Redundancy Detection Prompt

"Given two redundant phrases with repetitive structures, pleonasms, or circumlocutions in relation to each other in the format phrase1 -> phrase2, where phrase1 is a redundant or circumlocutory version of phrase2: Verify whether phrase1 is a redundancy of phrase2 by answering Yes or No."

A.1.2 Information Preservation Prompt

"Given two redundant phrases with repetitive structures, pleonasms, or circumlocutions in relation to each other in the format phrase1 -> phrase2, verify whether phrase2 retains the essential message and information of phrase1 by answering Yes (retains the message) or No (alters the message)"

A.1.3 Semantic Coherence Prompt

"Given a sentence in Spanish, verify whether it is well-formed and semantically coherent. Answer with Yes (well-formed and coherent) or No (poorly formed or incoherent)"

A.2 Redundancy Adaptation Prompts

A.2.1 Zero-shot 1

"Rewrite the following text to remove any redundancies if present. Text: 'input'"

A.2.2 Zero-shot 2

"The Easy-to-Read Methodology aims to create more accessible texts for people with reading comprehension difficulties. According to this methodology, one guideline for improving accessibility is avoiding words or structures that add no information and unnecessarily lengthen the text. These are redundancies or pleonasms. Rewrite the following text by correcting and removing any redundancies, pleonasms, or repetitive structures if detected. Limit your response to the corrected text only, without additional commentary. Text: 'input'"

A.2.3 Few-shot 1: Positive Examples (10) with Context

"The Easy-to-Read Methodology aims to create more accessible texts for people with reading comprehension difficulties. According to this methodology, one guideline is to avoid redundant words or structures. Below are examples of sentences with redundancies and their corrected versions:

original_text: To smell with the nose
adapted_text: To smell
original_text: Shut up your mouth
adapted_text: Shut up
original_text: He said to himself (Se dijo a sí mismo)
adapted_text: He said (Se dijo)
original_text: Free gift
adapted_text: Gift
original_text: To cry with tears
adapted_text: To cry
original_text: Drink liquids
adapted_text: Drink
original_text: To go inside into the building.

adapted_text: To go inside the building.
original_text: They went outside to enjoy the fresh air.
adapted_text: They went out to enjoy the fresh air.
original_text: The police found a lifeless corpse nearby.
adapted_text: The police found a corpse nearby.
original_text: The glass had cold ice and condensation.
adapted_text: The glass had ice and condensation.
original_text: ’input’

adapted_text:"

A.2.4 Prompt 2: Positive Examples (10) with Context and Instructions

"The Easy-to-Read Methodology aims to create more accessible texts for people with reading comprehension difficulties. According to this methodology, one guideline is to avoid redundant words or structures. Below are examples of sentences with redundancies and their corrected versions:
original_text: To smell with the nose
adapted_text: To smell
original_text: Shut up your mouth
adapted_text: Shut up
original_text: He said to himself (Se dijo a sí mismo)
adapted_text: He said (Se dijo)
original_text: Free gift
adapted_text: Gift
original_text: To cry with tears
adapted_text: To cry
original_text: Drink liquids
adapted_text: Drink
original_text: To go inside into the building.
adapted_text: To go inside the building.
original_text: They went outside to enjoy the fresh air.
adapted_text: They went out to enjoy the fresh air.
original_text: The police found a lifeless corpse nearby.
adapted_text: The police found a corpse nearby.
original_text: The glass had cold ice and condensation.
adapted_text: The glass had ice and condensation.

Given a sentence, adapt it to Easy-to-Read standards by removing repetitive structures if present. ’input’"

A.2.5 Prompt 3: Positive Examples (10) with Context and Extra Instructions

"The Easy-to-Read Methodology aims to create more accessible texts for people with reading comprehension difficulties. According to this methodology, one guideline is to avoid redundant words or structures. Below are examples of sentences with redundancies and their corrected versions:
original_text: To smell with the nose
adapted_text: To smell
original_text: Shut up your mouth
adapted_text: Shut up
original_text: He said to himself (Se dijo a sí mismo)
adapted_text: He said (Se dijo)
original_text: Free gift
adapted_text: Gift
original_text: To cry with tears
adapted_text: To cry
original_text: Drink liquids
adapted_text: Drink
original_text: To go inside into the building.
adapted_text: To go inside the building.
original_text: They went outside to enjoy the fresh air.
adapted_text: They went out to enjoy the fresh air.
original_text: The police found a lifeless corpse nearby.
adapted_text: The police found a corpse nearby.
original_text: The glass had cold ice and condensation.
adapted_text: The glass had ice and condensation.
Given a sentence, adapt it to Easy-to-Read standards by removing repetitive structures if present. Limit your response to the corrected text only, without additional commentary.

’input’"

A.2.6 Prompt 4: Positive (10) and Negative (5) Examples with Context and Extra Instructions

"The Easy-to-Read Methodology aims to create more accessible texts for people with reading comprehension difficulties. According to this methodology, one guideline is to avoid redundant words or structures. Below are examples of sentences with redundancies and their corrected versions:
original_text: To smell with the nose
adapted_text: To smell
original_text: Shut up your mouth
adapted_text: Shut up
original_text: He said to himself (Se dijo a sí mismo)
adapted_text: He said (Se dijo)
original_text: Free gift
adapted_text: Gift
original_text: To cry with tears
adapted_text: To cry
original_text: Drink liquids
adapted_text: Drink
original_text: To go inside into the building.
adapted_text: To go inside the building.
original_text: They went outside to enjoy the fresh air.
adapted_text: They went out to enjoy the fresh air.
original_text: The police found a lifeless corpse nearby.
adapted_text: The police found a corpse nearby.
original_text: The glass had cold ice and condensation.
adapted_text: The glass had ice and condensation.
Examples of sentences without redundancies (no correction needed):
original_text: Walk in the park.
adapted_text: Walk in the park.
original_text: Look carefully.
adapted_text: Look carefully.
original_text: He climbed the stairs to the second floor.
adapted_text: He climbed the stairs to the second floor.
original_text: She drank a coffee with milk for breakfast.
adapted_text: She drank a coffee with milk for breakfast.
original_text: They watched a sci-fi movie last night.
adapted_text: They watched a sci-fi movie last night.
Given a sentence, adapt it to Easy-to-Read standards by removing repetitive structures if present. Limit your response to the corrected text only, without additional commentary.

’input’"

A.3 Redundancy Generation Prompts: Section 3.4

A.3.1 Few-Shot Generation

"You are an expert in identifying and adapting redundancies, repetitive structures, pleonasms, and circumlocutions in sentences. Your task is to generate pairs of redundant or circumlocutory phrases and their respective adaptations in order to make them easier to understand. Below are some representative examples:
Input: The small and tiny cat hides under the bed
Output: The tiny cat hides under the bed
Input: Show the documents of the registration documentation
Output: Show the registration documentation
Input: María conducted a complete and thorough review
Output: María conducted a complete review
Input: It fell, and he picked it up from the ground

Output: It fell, and he picked it up

Generate new redundant phrases along with their adaptations following the same format:"

A.3.2 Few-Shot Paraphrasing with References

"Your task is to review a dataset of sentences to apply easy read guidelines: redundancies, repetitive structures, or circumlocutions should be avoided in order to make the sentences easier to understand. Given a redundant phrase or one with a repetitive structure (original_input) and its adaptation (original_output), generate three variations of the phrase using the format shown in the examples:

original_input: He returned back home
original_output: He returned home
Input1: She went back to her house
Output1: She went to her house
Input2: They repeated the same mistake again
Output2: They repeated the same mistake
Input3: I entered inside the building
Output3: I entered the building
original_input: The autopsy performed on the corpse
original_output: The autopsy performed
Input1: The autopsy was performed on the lifeless body
Output1: The autopsy was performed
Input2: The doctor did the autopsy on the lifeless corpse
Output2: The doctor did the autopsy
Input3: He saw his wife's lifeless corpse
Output3: He saw his wife's corpse

The generated variations must be semantically correct and coherent. Generate variations according to the above format:
original_input: 'input_text'
original_output: 'output_text'"

B Error Analysis in Failed Adaptations

B.1 Salamandra-7B-Instruct

- *Spelling Errors:* The model introduces errors such as changing letters or entire words.
  - Original: *"She painted a picture full of vibrant colors and contrasts."*
  - Failed Adaptation: *"she painted a picture full of vibranrt colorss and contrates."*
- *Semantic Substitutions:* Changes words for similar ones without justification.
  - Original: *"She wrote a letter to her friend who lives in another country."*
  - Failed Adaptation: *"She sent a letter to her friend who lives in another country."*
- *Arbitrary Changes:* Changes the verb and details of phrases, altering the original meaning.
  - Original: *"The computer has been on for several hours."*
  - Failed Adaptation: *"The computer has been on for a while."*
- *Hallucinations (Radical Changes):* Completely changes the meaning or content of the phrase.
  - Original: *"The autopsy performed on the corpse."*
  - Failed Adaptation: *"The body was examined after being found."*

B.2 Llama-3.1-8B-Instruct

- *Over-Simplification:* Removal of critical qualifiers ("analyzed in detail"), resulting in loss of nuance.
  - Original: *"The researcher examined and analyzed each experiment sample in detail."*
  - Failed Adaptation: *"The researcher examined the samples."*
- *Context Misinterpretation:* Loss of specific contextual information ("rooftop"), altering spatial precision.
  - Original: *"He went up to the building's rooftop to see the stars."*
  - Failed Adaptation: *"He went up to the building to see the stars."*
- *Semantic Drift:* Subtle semantic shift ("postponed" vs. "delayed") changes temporal implications.
  - Original: *"The meeting was postponed to a later date due to unforeseen circumstances."*
  - Failed Adaptation: *"The meeting was delayed to a later date due to unforeseen circumstances."*
- *Incomplete Redundancy Removal:* Retained redundant modifier ("small" + "child") while removing "young".
  - Original: *"The small young child ran quickly toward the door."*
  - Failed Adaptation: *"The small child ran quickly toward the door."*

C Calculations and Formulae

The numerical values applied to the formulas are provided in tables 1 and 3, or are derived directly from the data presented in these tables—for example, complementary probability: $1 - P(A)$.

C.1 Pipeline Success Rate (SR)

$$SR = P(D) \times P(A|D) \times P(R|A) \times P(V|A \cap R)$$

$$SR = 0.946 \times 0.900 \times 1.000 \times (0.946 \times 0.992 \times 1.000)$$

$$SR = 0.799$$

Where: $P(D)$ is the probability of correct redundancy detection; $P(A|D)$ is the probability of successful adaptation given detection; $P(R|A)$ is the probability of successful parsing; and $P(V|A \cap R)$ is the joint probability of passing all three verification steps.

C.2 False Positive Rate Case 1: Correctly Detected a Redundancy but Generated Flawed Adaptation

$$FP1 = P(D) \times P(W|D) \times P(R|W) \times P(V|W \cap R)$$

$$FP1 = 0.946 \times 0.100 \times 1.000 \times (0.087 \times 0.016 \times 0.034)$$

$$FP1 = 0.00000452$$

Where: $P(D)$ is the probability of correct redundancy detection; $P(W|D)$ is the probability of wrong adaptation given correct detection; $P(R|W)$ is the probability of successful parsing; and $P(V|W \cap R)$ is the joint probability of passing all three verification steps given wrong adaptation.

C.3 False Positive Rate Case 2: Incorrectly Detected a Redundancy and Subsequently Generates a Modified Phrase

$$SR = P(N) \times P(W|N) \times P(R|N) \times P(V|W \cap R)$$

$$SR = 0.054 \times 0.700 \times 1.000 \times (0.024 \times 0.016 \times 0.034)$$

$$FP2 = 0.000000493$$

Where: $P(N)$ is the probability of incorrect redundancy detection; $P(W|N)$ is the probability of wrong adaptation given incorrect detection; $P(R|N)$ is the probability of successful parsing; and $P(V|W \cap R)$ is the joint probability of passing all three verification steps given wrong detection and adaptation.