SINAI in SimpleText CLEF 2025: Simplifying Biomedical Scientific Texts and Identifying Hallucinations Using GPT-4.1 and Pattern Detection

Notebook for the SimpleText Lab at CLEF 2025

Jaime Collado-Montañez[1,*,†], Jenny Alexandra Ortiz-Zambrano[2,†], César Espin-Riofrio[2,†] and Arturo Montejo-Ráez[1,†]

[1]Department of Computer Science (University of Jaén), Campus Las Lagunillas, s/n, Jaén, 23071, Spain
[2]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Av. Delta S/N, 090514 Guayaquil, Ecuador

Abstract

This paper presents our participation in three different tasks of the CLEF 2025 SimpleText track. For tasks 1.1 and 1.2, we explore the application of advanced language models, specifically GPT-4.1, for the automatic simplification of biomedical texts in English using zero-shot learning. Two versions of prompts were designed and implemented on the Cochrane-auto dataset in Task 1.1 and Task 1.2, with the aim of generating texts that are more understandable for non-specialist audiences. The results show that the model successfully preserves the original semantic structure, identifying complex terms and providing clear restructuring, including brief explanations when necessary. Furthermore, an accurate listing of key elements and effective reorganization of difficult grammatical structures were observed. These characteristics indicate the adaptability of the model to facilitate access to technical information without affecting its accuracy. Finally, the results preliminarily support the effectiveness of prompt design as an approach to improve text comprehension in the biomedical field, without the need for additional supervised training.

In Task 2.1, we addressed the problem of detecting creative generation in simplification outputs at the document level. Our approach combined rule-based pattern matching—developed through exploratory analysis of the training set—with the use of the llama-3.1-8b-instruct language model. Surface-level patterns such as one-word sentences, double-space endings, and near-literal context matches were leveraged to pre-label data, while remaining cases were evaluated by the language model using a token-level confidence threshold. The highest-performing run in the sourced subtask achieved an F1-score of 0.976 whereas in the posthoc subtask it was 0.953.

Keywords

Lexical Complexity, Biomedical Scientific Texts, GPT-4.1, Zero-Shot learning, Pattern extraction, Synthetic generation

1. Introduction

The way a text is written can become a considerable barrier, especially when it contains rare or unfamiliar vocabulary, as well as complex lexical and semantic structures that make it difficult to access its content. [1, 2]. This situation is especially evident in diverse populations [3, 4], as there are numerous groups of readers such as foreign language students [5], people with cognitive disabilities [6], people with low levels of reading comprehension who face significant barriers to understanding written texts, even among university students who, despite their academic training and specialized knowledge, also experience difficulties reading and understanding complex texts [7], which poses a critical challenge to communicative equity.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.
†These authors contributed equally.
jcollado@ujaen.es (J. Collado-Montañez); jenny.ortiz@ug.edu.ec (J. A. Ortiz-Zambrano); cesar.espin@ug.edu.ec (C. Espin-Riofrio); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0002-9672-6740 (J. Collado-Montañez); 0000-0001-6708-4470 (J. A. Ortiz-Zambrano); 0000-0001-8864-756X (C. Espin-Riofrio); 0000-0002-8643-2714 (A. Montejo-Ráez)

The development of information technologies has radically transformed access to information, enabling the massive availability of content in diverse and key Field such as education, communication, health, public administration, and scientific research. In particular, digitization has exponentially boosted the production and dissemination of scientific literature, facilitating its consultation and analysis on a large scale. Despite advances in digitization and open access, scientific information continues to present a significant barrier for the general public: the high linguistic complexity of specialized texts. This difficulty limits direct access to knowledge from original sources, especially for people without prior training in the field, who face substantial challenges due to the lack of technical knowledge or specialized terminology [8], because understanding a text depends largely on the reader's prior knowledge of the meaning of words. In this context, ensuring access to linguistically accessible content not only responds to an educational and social need, but is also consolidated as a fundamental right increasingly supported by international regulations and institutions [9].

The SimpleText Lab is part of the CLEF 2025[1] initiative, which aims to promote the systematic evaluation of information access systems through shared tasks. This proposal focuses on the challenges associated with text simplification, with a particular emphasis on the accessibility of scientific information in recent years. In this context, SimpleText offers valuable resources and metrics for research, given that a large part of the general public avoids consulting reliable scientific sources due to their high linguistic complexity and lack of specialized knowledge. As a result, many people opt for simplified content available on the internet and social media, which often serves commercial or political interests rather than informational purposes [10].

The main objective of this research is to demonstrate the ability of the Transformer based GPT-4.1 linguistic model to perform lexical simplification. To this end, several variants of sentences without examples were created and evaluated. This approach allows for the simplification of sentences and entire documents extracted from Cochrane-auto corpus, derived from biomedical literature summaries and lay summaries of Cochrane systematic reviews, to facilitate the reader's understanding of scientific text.

The article is organized as follows: in Section 2 we present a brief description of the tasks we participated in. Sections 3 and 4 detail the data and methodology followed for tasks 1 and 2.1 respectively. Section 5 shows the results achieved during our experimentation and section 6 summarizes the main conclusions and proposes avenues for future research.

2. SimpleText@CLEF-2024 Tasks

2.1. Task 1: Text Simplification: Simplify scientific text

The objective of this task is to simplify scientific texts extracted from the Cochrane-auto corpus, both at the level of complete sentences and entire documents, in order to facilitate the understanding of the content by non-specialist readers. We have contributed to two subtasks:

1. Task 1.1 - Sentence-level Scientific Text Simplification. The goal of this task is to simplify whole sentences extracted from the Cochrane-auto dataset.
2. Task 1.2 - Document-level Scientific Text Simplification. The goal of this task is to simplify whole documents extracted from the Cochrane-auto dataset.

2.2. Task 2: Controlled Creativity: Identify and Avoid Hallucination

The objective of Task 2 focuses on identifying and evaluating creative generation and information distortion in text simplification. We have contributed to he subtask 2.1 Identify Creative Generation at Document Level.

This task aims to detect creative generation at the abstract or document level. Participants will analyze system outputs from previous years, along with deliberately generated outputs from known models.

[1]CLEF 2025 SimpleText - https://simpletext-project.com/2025/tasks

The goal is to identify which sentences are fully grounded in the source text, both without access to the original sentences and with access to them. Additionally, sentences that introduce significant new content must be labeled. This task serves as a post-hoc identification or explanation challenge.

3. Task 1: Experiments with Zero-Shot learning

3.1. Cochrane-auto Corpus

As part of the CLEF 2025 SimpleText program, the Cochrane-auto corpus was launched. It is composed of biomedical scientific abstracts and their corresponding simplified versions for non-specialist readers, derived from Cochrane systematic reviews. This resource represents a significant advance in the field of biomedical simplification, adopting approaches previously applied to datasets such as Wiki-auto and Newsela-auto.

Compared to other traditional corpora, Cochrane-auto provides novel parallel data written by the review authors themselves, enabling simplification processes at the whole-document level. Furthermore, Cochrane-auto incorporates advanced techniques such as sentence merging, text restructuring, and discourse alignment, allowing for a deeper and more coherent treatment of content. This design enables multi-scale realignment, encompassing paragraphs, sentences, and documents, and distinguishes it from conventional approaches focused solely on superficial simplification.

3.2. Proposed system

State-of-the-art deep learning architectures—including BERT [11], RoBERTa [12], GPT-3 [13], and GPT-4.1 [14] are significantly outperforming classical techniques. In particular, GPT-4.1, a large scale Transformers based text generation framework developed by OpenAI[2], reflects these advances. These solutions have achieved outstanding performance across a variety of natural language processing tasks, setting unprecedented levels of accuracy and performance in the field.

As part of our approach, we used this model through the OpenAI API, configured with a temperature of 0.0 and a maximum limit of 10,000 tokens per response, which allowed us to obtain detailed and deterministic results. This configuration can be seen in Table 1. The model was integrated into our workflow through Python code, which facilitated test automation. We also used the OpenAI Playground environment as a complementary resource to quickly validate different inputs and generate query prototypes.

Table 1
Configuration used for prediction generation with GPT-4.1

Parameter	Value
Model (engine)	`GPT-4.1`
Temperature	0.0
Maximum tokens	10,000
API used	OpenAI Python API
Applied technique	Zero-shot (no examples)
Corpus used	*Cochrane-auto Corpus*
Simplification level	Full document and sentence

3.3. Prompt Design

To run the experiments with the Cochrane-auto corpus, we designed two specific instructions corresponding to Task 1.1 and Task 1.2. Both experiments were conducted using a zero-shot learning

[2] https://openai.com/

approach, that is, without providing explicit training examples to the model. The prompts developed to guide the system's automatic generation of simplifications are presented in the Appendix.

4. Task 2.1: Detecting patterns in the data

4.1. Data description

For this sub-task, participants are provided with system outputs from previous years, along with deliberately generated outputs from known models. The objective is to analyze these outputs to identify which sentences are fully grounded in the source text both with and without access to the original sentences.

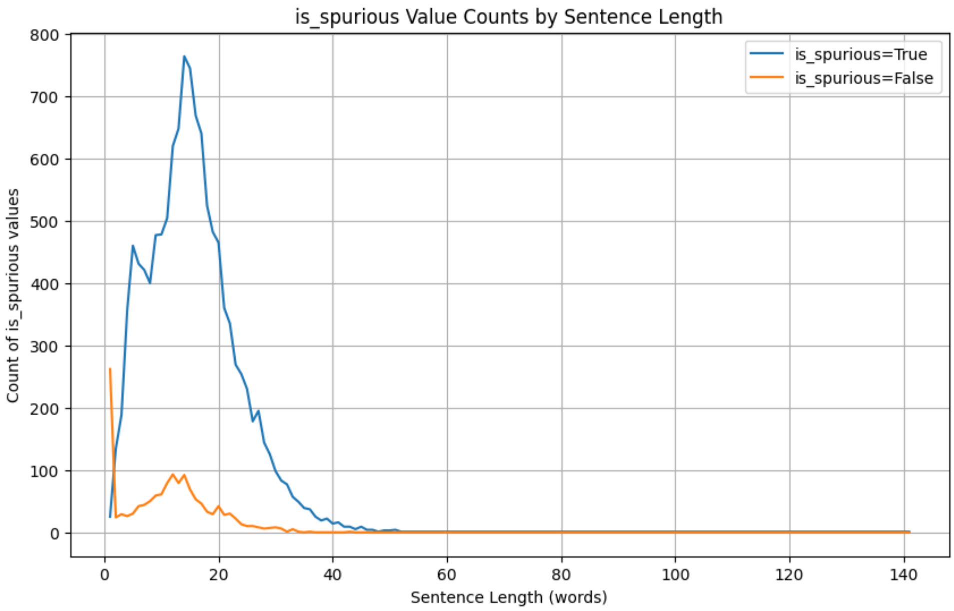

Figure 1: Sentence length distribution by label.

An exploratory analysis over the sourced training dataset showed a high class imbalance: 12115 out of 13514 rows (90% of the total data) were labeled as spurious, while the remaining 1399 were not spurious. More interestingly, spurious sentences were significantly longer—more than 15 words per sentence in average—, whereas not spurious were only about 11 words long. This fact raised our interest to further analyze sentence length distribution, where we found a surprising high amount of one-word length sentences labeled as not spurious given the reduced number of examples in this class (see Figure 1). Specifically, 262 out of the 1399 not spurious examples are one-word sentences, 244 of which are simply a ‘.’. We also found all instances of the regex ‘#.’, where ‘#’ can be any digit, were always labeled as spurious, although these accounted just for 14 examples out of the 12115 in this subset.

Taking a look at some of the remaining not spurious sentences—i.e. those that were longer than 1 word—, we realized most of them appeared literally—or close to literally—within the context provided in the sourced dataset: 809 examples apply to this case, 790 of such were not spurious while only 19 were spurious.

Finally, we also noticed that all 1241 sentences ending with a double space character ‘ ’ were labeled as spurious.

Table 2 shows examples of all the patterns found during this exploratory analysis.

Table 2
Pattern examples in the sourced training set found during the exploratory analysis.

Pattern Type	Sentence example	Label	# of Examples
One-word ('#.')	3.	Spurious	14/12115
Literal match	Cookies can be used to support RBAC on the Web, holding users' role information.	Spurious	19/12115
Double space	We offer empirical evidence for better understanding of cryptocurrency adoption with practical implications in an e- .	Spurious	1241/12115
One-word ('.')	.	Not spurious	244/1399
Literal match	A significant problem of using deep learning techniques is the limited amount of data available for training.	Not spurious	790/1399

4.2. Proposed system

Based on the patterns detailed in the data description, we developed a pre-filtering system where sentences matching any of these patterns were automatically assigned its corresponding label. Specifically, our final submission relied on a filter to pre-anotate sentences matching the one-word, double space, and (close to) literal match patterns. This last filter includes a fuzzy matching that normalizes text and removes stopwords before comparing strings. For all sentences left, we prompted llama-3.1-8b-instruct [15] to output 'Yes' or 'No', meaning spurious or not spurious respectively, given the sentence and the sourced context.

In order to deal with class imbalance, we opted to trust the LLM output only if the probability of the 'No' token—not spurious, which is the minority class—was larger than a given threshold. In our submissions, we included 3 different ones: 95%, 99% and 100% (i.e. everything not matching a filtering pattern is labeled as spurious).

Regarding our post-hoc approach, we generated artificial contexts with the same LLM for each sentence and repeated the same experiments mentioned before. All prompts used during the experimentation are included in the appendices.

To summarize, we presented four different runs for both sourced and posthoc subtasks:

- Run 1: One-word and double space filters. All remaining sentences are labeled as spurious.
- Run 2: One-word, double space, and literal match filters. Remaining sentences where the 'No' token probability is greater than 95% are labeled as not spurious.
- Run 3: One-word, double space, and literal match filters. Remaining sentences where the 'No' token probability is greater than 99% are labeled as not spurious.
- Run 4: One-word, double space, and literal match filters. All remaining sentences are labeled as spurious.

5. Results

5.1. Results of Task 1.1

Table 3 shows the simplified prediction for a complex sentence extracted from the Cochrane-auto corpus. In this instance, it is evident that the model successfully transformed a technical sentence about cluster-randomized trials, preserving essential information (such as the number of patients and hospitals) and including a clarification about the acronym "UK"("United Kingdom"), as requested in the prompt. Similarly, Table 4 presents a second example in which the model reformulates a sentence describing types of healthcare professionals. The output generated by the model preserves the semantic structure and provides an explicit listing of the professionals, adding explanatory examples in parentheses (e.g., "physical therapists and dietitians"), as requested in the prompt.

Table 3
Example: Applying the First version of Prompt in the model with Zero-Shot learning for task 1.1

Field	Content
pair_id	CD012520
para_id	0
sent_id	0
complex	We included seven cluster-randomised trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and the Netherlands.
prediction	We included seven studies where groups of patients were randomly chosen. These studies had 42,489 patients from 129 hospitals in Australia, the United Kingdom (United Kingdom means UK), China, and the Netherlands.

Table 4
Individual entry of the par_id `CD012520` sentence 1

Field	Content
pair_id	CD012520
para_id	0
sent_id	1
complex	Health professional participants (numbers not specified) included nursing, medical and allied health professionals.
prediction	The people who took part in the study were nurses, doctors, and other health workers (such as physical therapists and dietitians).

In both cases, the model demonstrates accurate interpretation of the biomedical domain and applies lexical simplification strategies, such as the replacement of technical terms, and restructures complex sentences, generating more accessible versions of the texts. This provides preliminary validation of the prompt design and the model's usefulness for automatic simplification tasks without the need for prior supervised training.

Table 5
Example: Applying the Second version of Prompt in the model with Zero-Shot learning for task 1.1

Field	Content
pair_id	CD012520
para_id	0
sent_id	1
complex	Health professional participants (numbers not specified) included nursing, medical and allied health professionals.
prediction	The people who took part in the study and work in health care (the exact number is not given) included nurses, doctors, and other health workers (such as physical therapists or dietitians).

In the second version of the prompt applied to the model using zero-shot learning for Task 1.1, we observed a significant improvement in the clarity and accessibility of the original content. In sentence 1, the model retains the general semantic structure but provides an explicit explanation of the professional context ("people who took part in the study and work in health care"), in addition to providing a more detailed and explanatory list of participants, including specific examples ("such as physical therapists or dietitians"). Regarding sentence 2, the model develops an outstanding approach, that is, it simplifies and reorganizes a complex grammatical structure, transforming a sentence full of technical terminology into a more accessible expression. The use of explanations in parentheses to aid understanding ("people who work in health care") stands out, as does the paraphrasing of technical expressions such as "delivery arrangements" or "financial arrangements."

Table 6
Entrada individual del par_id CD012520, sentence 2.

Field	Content
pair_id	CD012520
para_id	0
sent_id	2
complex	Interventions in all studies included implementation strategies targeting healthcare workers; three studies included delivery arrangements, no studies used financial arrangements or governance arrangements.
prediction	All the studies used ways to help healthcare workers (people who work in health care) do their jobs better. Three studies also changed how care was given to patients. No studies used changes in money or rules to improve care.

These results initially confirm that the design of the second prompt allows for greater adaptability to readers with low literacy levels, without affecting semantic fidelity. Likewise, the model's greater sensitivity to the application context is evident, suggesting an improvement in the effectiveness of automatic simplification strategies compared to an unsupervised model.

Table 7
Results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team

Team/Method	Count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
DSGT plan_guided_lla	217	42.98	6.33	7.82	0.48	0.99	0.46	0.00	0.18	0.71	8.50
UM-FHS gpt-4.1-mini	217	42.13	9.52	7.56	0.74	1.52	0.61	0.00	0.26	0.53	8.54
SINAI PRMZSTASK11V1	217	41.25	4.59	12.39	1.44	1.56	0.51	0.00	0.61	0.30	8.44
UvA llama31	217	40.92	2.62	8.63	1.00	1.64	0.45	0.00	0.62	0.64	8.35
THM p2–gpt-4.1-nano	217	39.57	6.50	15.40	1.32	1.20	0.60	0.00	0.47	0.27	8.68

According to the official results presented in Table 7 [10], the solution submitted by our team for Task 1.1, SINAI PRMZSTASK11V1, ranks among the top performers in the global CLEF 2025 SimpleText Task 1.1 ranking, reaching the third overall place when all runs are sorted by the main SARI metric, which evaluates the quality of simplification. This result indicates that SINAI's solution achieved a SARI of 41.25, outperforming many of the variants submitted by other teams. Furthermore, its BLEU score of 4.59 is the highest among the top four results, demonstrating that the simplified version maintains a high similarity to the human reference. Taken together, these indicators reflect that SINAI PRMZSTASK11V1 achieved a favorable balance between simplification, fidelity to the original content, and readability, standing out as one of the best solutions to the challenge compared to the best executions submitted by other teams.

Table 8
Example: Applying the First version of Prompt in the model with Zero-Shot learning for task 1.2

Field	Content
pair_id	CD012520
source	Cochrane corpus
complex	We included seven cluster-randomised trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and the Netherlands. Health professional participants (numbers not specified) included nursing, medical and allied health professionals. Interventions in all studies included implementation strategies targeting healthcare workers; three studies included delivery arrangements, no studies used financial arrangements or governance arrangements. Five trials compared a multifaceted implementation intervention to no intervention, two trials compared one multifaceted implementation intervention to another multifaceted implementation intervention. No included studies compared a single implementation intervention to no intervention or to a multifaceted implementation intervention. Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies. All studies had low risks of selection bias and reporting bias, but high risk of performance bias. Three studies had high risks of bias from non-blinding of outcome assessors or due to analyses used. We are uncertain whether a multifaceted implementation intervention leads to any change in adherence to evidence-based recommendations compared with no intervention (risk ratio (RR) 1.73; 95% confidence interval (CI) 0.83 to 3.61; 4 trials; 76 clusters; 2144 participants, I2 =92%, very low-certainty evidence). Looking at two specific processes of care, multifaceted implementation interventions compared to no intervention probably lead to little or no difference in the proportion of patients with ischaemic stroke who received thrombolysis (RR 1.14, 95% CI 0.94 to 1.37, 2 trials; 32 clusters; 1228 participants, moderate-certainty evidence), but probably do increase the proportion of patients who receive a swallow screen within 24 hours of admission (RR 6.76, 95% CI 4.44 to 10.76; 1 trial; 19 clusters; 1,804 participants; moderate-certainty evidence). Multifaceted implementation interventions probably make little or no difference in reducing the risk of death, disability or dependency compared to no intervention (RR 0.93, 95% CI 0.85 to 1.02; 3 trials; 51 clusters ; 1228 participants; moderate-certainty evidence), and probably make little or no difference to hospital length of stay compared with no intervention (difference in absolute change 1.5 days; 95% CI -0.5 to 3.5; 1 trial; 19 clusters; 1804 participants; moderate-certainty evidence). We do not know if a multifaceted implementation intervention compared to no intervention result in changes to resource use or health professionals' knowledge because no included studies collected these outcomes. We are uncertain whether a multifaceted implementation intervention compared to no intervention improves adherence to evidence-based recommendations in acute stroke settings, because the certainty of evidence is very low.
prediction	We looked at seven studies that included 42,489 patients from 129 hospitals in Australia, the United Kingdom, China, and the Netherlands. The studies also included nurses, doctors, and other health workers (the exact number was not given). In all the studies, the main action was to use different ways to help healthcare workers follow guidelines. Three studies also changed how care was given, but none used changes in money or hospital rules. Five studies compared using several methods to help healthcare workers with not using any methods. Two studies compared different sets of methods with each other. No studies compared using just one method with not using any, or with using several methods. All studies measured how many patients got care based on the best available evidence. The studies were not likely to have problems with how patients were chosen or how results were reported, but there was a high chance that knowing which group patients were in could have affected the results (performance bias). Three studies also had a high chance of mistakes because the people checking the results knew which group patients were in, or because of the way the results were analyzed. We are not sure if using several methods to help healthcare workers follow guidelines makes a difference in how well they follow these guidelines compared to not using any methods (the evidence is not strong). When looking at two specific types of care, using several methods probably does not change how many patients with a type of stroke called ischaemic stroke (a stroke caused by a blocked blood vessel) get a treatment called thrombolysis (a medicine to break up blood clots), but it probably does increase how many patients get a test to check if they can swallow safely within 24 hours of arriving at the hospital. Using several methods probably does not make much difference in lowering the risk of death, disability, or needing help from others, or in how long patients stay in the hospital. We do not know if using several methods changes how resources are used or how much health workers know, because the studies did not measure these things. Overall, we are not sure if using several methods to help healthcare workers follow guidelines improves care for patients with sudden stroke, because the evidence is not strong.

5.2. Results of Task 1.2

Table 8 shows the result achieved after applying the first version of the prompt (specified in Table 2) to the zero-shot learning-based model on the pair identified as CD012520. This pair represents a complex biomedical document extracted from Cochrane, with dense and highly specialized content, typical of the clinical-academic domain. The original text (complex field) presents several particularities characteristic of lexical and structural complexity. Once the model was implemented, the result generated in the prediction field shows a restructured text that significantly reduces technical complexity. We analyzed the evaluation of the results using the following criteria: The system replaced specialized terms with simpler vocabulary, and long sentences in the original text were organized into simple, consecutive sentences. Despite the simplification, the model preserves the key ideas of the original text in the result; The simplified text meets the main objective of the proposed system, obtaining an accessible version of it.

Table 9 shows the results after applying the second version of the prompt designed for the automatic lexical simplification task in biomedical texts, specifically on a complex fragment of the Cochrane corpus for Task 1.2. This version of the prompt includes clear instructions regarding preserving meaning, handling acronyms, and inserting parenthetical explanations. The resulting model output reflects several notable improvements: This instruction guides the model to provide summarized definitions in parentheses after potentially complex words that should not be replaced, but should be clarified. Examples of this are "RR (risk ratio, a way to compare groups)" and "CI (confidence interval, a range that shows uncertainty)," which provide comprehension without eliminating technical terms.

The model faithfully applies the prompt's stipulation of expanding acronyms at least upon their first mention, which benefits readers with little familiarity with biomedical jargon. Furthermore, it omits simplifying expressions that constitute a specialized lexical unit, preserving the semantic integrity of concepts such as "thrombolysis" or "swallow screen.". In relation to previous results, here we observe a more precise fragmentation that preserves the flow of ideas from the source text. Despite the greater ease of understanding of the text, no critical information is omitted; thus, the text moves from technical writing to an interpretation geared toward the general reader: "risk ratio (RR)... very low certainty in the results" instead of simply presenting the data as it appears in the source.

Table 10 presents the official results of the document-level text simplification task 1.2, evaluating 217 plain language summaries [10]. Participating teams were compared on multiple metrics. SINAI PRMZSTASK12V1 obtained an outstanding score according to the SARI evaluation metric, which was 43.63, placing them among the top five teams. The SINAI team outperformed several popular LLM-based models, such as GPT-4.1 (UM-FHS gpt-4.1-mini) and LLaMA (AIIRLab llama-8b), highlighting the effectiveness of their approach on this specific task. The best results were obtained by LIA sumguid-all-w500 (44.93) and LIA sumguid-lang-w50 (44.40), closely followed by SINAI PRMZSTASK12V1. Later versions and other equipment had variable performance, with some results noticeably lower.

5.3. Results of Task 2.1

In this section we report all metrics provided by the organizers through Codabench for our final submission and several ablation runs to evaluate different filters and prompts. Table 11 shows these results.

The results indicate that predicting non-spurious sentences using large language models (LLMs) is particularly challenging in this dataset. The highest-performing runs (1 and 4) achieved strong results by labeling all sentences that did not match specific predefined patterns as spurious. This approach appears effective largely due to the evaluation metric used: performance is assessed primarily on the overrepresented class (spurious), rather than through a micro or macro-averaged score across classes. As a result, attempts to correctly identify non-spurious instances have little impact on the final evaluation score.

In the posthoc subtask, our strategy of generating synthetic contexts did not yield the expected improvements. Many generated contexts included the original sentence verbatim, causing our filtering

Table 9
Example: Applying the Second version of Prompt in the model with Zero-Shot learning for task 1.2

Field	Content
pair_id	CD012520
source	Cochrane corpus
complex	We included seven cluster-randomised trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and the Netherlands. Health professional participants (numbers not specified) included nursing, medical and allied health professionals. Interventions in all studies included implementation strategies targeting healthcare workers; three studies included delivery arrangements, no studies used financial arrangements or governance arrangements. Five trials compared a multifaceted implementation intervention to no intervention, two trials compared one multifaceted implementation intervention to another multifaceted implementation intervention. No included studies compared a single implementation intervention to no intervention or to a multifaceted implementation intervention. Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies. All studies had low risks of selection bias and reporting bias, but high risk of performance bias. Three studies had high risks of bias from non-blinding of outcome assessors or due to analyses used. We are uncertain whether a multifaceted implementation intervention leads to any change in adherence to evidence-based recommendations compared with no intervention (risk ratio (RR) 1.73; 95% confidence interval (CI) 0.83 to 3.61; 4 trials; 76 clusters; 2144 participants, I2 =92%, very low-certainty evidence). Looking at two specific processes of care, multifaceted implementation interventions compared to no intervention probably lead to little or no difference in the proportion of patients with ischaemic stroke who received thrombolysis (RR 1.14, 95% CI 0.94 to 1.37, 2 trials; 32 clusters; 1228 participants, moderate-certainty evidence), but probably do increase the proportion of patients who receive a swallow screen within 24 hours of admission (RR 6.76, 95% CI 4.44 to 10.76; 1 trial; 19 clusters; 1,804 participants; moderate-certainty evidence). Multifaceted implementation interventions probably make little or no difference in reducing the risk of death, disability or dependency compared to no intervention (RR 0.93, 95% CI 0.85 to 1.02; 3 trials; 51 clusters ; 1228 participants; moderate-certainty evidence), and probably make little or no difference to hospital length of stay compared with no intervention (difference in absolute change 1.5 days; 95% CI -0.5 to 3.5; 1 trial; 19 clusters; 1804 participants; moderate-certainty evidence). We do not know if a multifaceted implementation intervention compared to no intervention result in changes to resource use or health professionals' knowledge because no included studies collected these outcomes. We are uncertain whether a multifaceted implementation intervention compared to no intervention improves adherence to evidence-based recommendations in acute stroke settings, because the certainty of evidence is very low.
prediction	We included seven studies where groups of patients were randomly assigned to different treatments. These studies involved 42,489 patients from 129 hospitals in Australia, the United Kingdom, China, and the Netherlands. The studies also included health workers (such as nurses, doctors, and other healthcare staff), but the exact number of health workers was not given. In all the studies, the treatments focused on changing how healthcare workers do their jobs. Three studies also changed how care was delivered. None of the studies changed how money was used or how hospitals were managed. Five studies compared a complex treatment (which used several methods to help healthcare workers follow guidelines) to no treatment. Two studies compared one complex treatment to another complex treatment. None of the studies compared a single method to no treatment or to a complex treatment. All studies measured how many patients received care based on the best available evidence. All studies were unlikely to have problems with how patients were chosen or how results were reported, but there was a high chance that knowing which treatment was given could have affected the results (performance bias). In three studies, there was a high chance of bias because the people checking the results knew which treatment was given, or because of the way the results were analyzed. We are not sure if using a complex treatment changes how well healthcare workers follow evidence-based guidelines compared to no treatment (risk ratio (RR, a way to compare groups) 1.73; 95% confidence interval (CI, a range that shows uncertainty) 0.83 to 3.61; 4 studies; 76 groups; 2144 patients; I2 =92%, very low certainty in the results). Looking at two specific parts of care, complex treatments compared to no treatment probably make little or no difference in the number of patients with ischaemic stroke (a type of stroke caused by a blocked blood vessel) who received thrombolysis (a treatment to break up blood clots) (RR 1.14, 95% CI 0.94 to 1.37, 2 studies; 32 groups; 1228 patients, moderate certainty in the results), but probably do increase the number of patients who get a swallow screen (a test to check if a person can swallow safely) within 24 hours of arriving at the hospital (RR 6.76, 95% CI 4.44 to 10.76; 1 study; 19 groups; 1,804 patients; moderate certainty in the results). Complex treatments probably make little or no difference in lowering the risk of death, disability, or needing help from others compared to no treatment (RR 0.93, 95% CI 0.85 to 1.02; 3 studies; 51 groups; 1228 patients; moderate certainty in the results), and probably make little or no difference to how long patients stay in the hospital (difference in absolute change 1.5 days; 95% CI -0.5 to 3.5; 1 study; 19 groups; 1804 patients; moderate certainty in the results). We do not know if complex treatments compared to no treatment change how resources are used or improve health workers' knowledge, because none of the studies measured these outcomes. Overall, we are not sure if complex treatments compared to no treatment help healthcare workers follow evidence-based guidelines in the care of patients with acute stroke (sudden stroke), because the certainty of the evidence is very low.

mechanism—designed to detect literal matches—to misclassify a large number of these as non-spurious.

6. Conclusions and Future Work

For the text simplification results (Tasks 1.1 and 1.2), the model demonstrates remarkable adaptability in identifying and replacing complex terms with clear didactic descriptions, preserving the original meaning of the text and providing an accurate list of key aspects, such as the healthcare professionals

Table 10
Results for CLEF 2025 SimpleText Task 1.2 document-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team

Team/Method	Count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	1.00	1.00	9.05
Reference	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
LIA sumguid-all-w500	217	44.93	9.58	9.77	0.69	1.06	0.48	0.00	0.29	0.62	8.61
LIA sumguid-lang-w50	217	44.40	7.85	10.58	0.67	0.97	0.44	0.00	0.30	0.66	8.56
SINAI PRMZSTASK12V1	217	43.63	8.07	10.73	0.81	1.03	0.52	0.00	0.37	0.54	8.41
LIA sumguid-styl-w50	217	43.57	6.18	10.28	0.51	0.81	0.41	0.00	0.20	0.72	8.67
ASM MistralMinFKGL	217	43.51	8.26	11.85	0.63	0.82	0.48	0.00	0.22	0.62	8.78

Table 11
Results obtained in Codabench for every run detailed in Section 4. Highest scores per metric and subtask are highlighted in **bold**.

Run	Subtask	accuracy	precision	recall	f1_score	roc_auc
1	sourced	0.914	0.912	**1**	0.954	0.578
2	sourced	0.808	**1**	0.786	0.88	0.893
3	sourced	0.933	**1**	0.926	0.961	0.961
4	sourced	**0.958**	**1**	0.953	**0.976**	**0.975**
1	posthoc	**0.912**	0.911	**1**	**0.953**	0.555
2	posthoc	0.29	**0.957**	0.222	0.36	0.565
3	posthoc	0.345	0.948	0.289	0.443	**0.572**
4	posthoc	0.367	0.942	0.317	0.474	0.569

in the analyzed examples. This implies that the automatic simplification system is capable of producing clear results without affecting the quality and accuracy of the information provided.

For Task 2.1, our pattern-based approach combined with LLM evaluation proved highly effective for detecting hallucinations in text simplification outputs. The strategy of identifying surface-level patterns (one-word sentences, double spaces, and literal matches) followed by LLM-based classification achieved outstanding performance in the sourced subtask, with our best run (Run 4) achieving an F1-score of 0.976, precision of 1.0, and accuracy of 0.958.

The key insight from Task 2.1 was that simple rule-based pattern matching could effectively pre-label a significant portion of the data, while the LLM (llama-3.1-8b-instruct) provided reliable classification for remaining cases. However, the posthoc subtask proved more challenging, with our synthetic context generation approach yielding lower performance (best F1-score of 0.953), primarily due to the generated contexts including original sentences verbatim, which caused misclassification by our literal matching filters.

As future work, we propose the development of adaptive multi-specialty simplification approaches, which allow the system to adjust its text simplification strategies to the conditions, needs or changes in the environment according to the thematic domain, preserving terminological precision and communicative clarity. Additionally, improvements to synthetic context generation for hallucination detection could enhance post-hoc evaluation capabilities, particularly by developing more sophisticated methods that avoid verbatim inclusion of target sentences in generated contexts.

Acknowledgments

A sincere thank you to the organizers of the CLEF 2025 Conference for their motivation, dedication, and efforts in promoting research in the search for solutions that contribute to reducing barriers to reading.

Our thanks to Engineer Anthony Arteaga Burgos, a graduate of the Faculty of Mathematical and Physical Sciences at the University of Guayaquil, for his valuable contribution to the development of the material.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. It has also been funded by the scholarship (FPI-PRE2022-105603) from the Ministry of Science, Innovation and Universities of the Spanish Government.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini and ChatGPT in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Transactions on Accessible Computing (TACCESS) 6 (2015) 1–36.
[2] K. North, M. Zampieri, M. Shardlow, Lexical complexity prediction: An overview, ACM Computing Surveys 55 (2023) 1–42.
[3] I. Segura-Bedmar, P. Martinez, Simplifying drug package leaflets written in spanish by using word embedding, Journal of Biomedical Semantics 8 (2017). doi:10.1186/s13326-017-0156-7.
[4] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Trans. Access. Comput. 6 (2015). URL: https://doi.org/10.1145/2738046. doi:10.1145/2738046.
[5] I. Rets, J. Rogaten, To simplify or not? facilitating english l2 users' comprehension and processing of open educational resources in english using text simplification, Journal of Computer Assisted Learning 37 (2021) 705–717.
[6] M. Licardo, N. Volčanjk, D. Haramija, Differences in communication skills among elementary students with mild intellectual disabilities after using easy-to-read texts, The new educational review 64 (2021) 236–246.
[7] R. Alarcón, L. Moreno, P. Martínez, Hulat-alexs cwi task-cwi for language and learning disabilities applied to university educational texts, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.
[8] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the clef 2023 simpletext task 2: Difficult concept identification and explanation (2023).
[9] S. Bott, H. Saggion, N. P. Rojas, M. S. Salazar, S. C. Ramirez, Multils-sp/ca: Lexical complexity prediction and lexical simplification resources for catalan and spanish, 2024. URL: https://arxiv.org/abs/2404.07814. arXiv:2404.07814.
[10] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Clef 2025 simpletext track: Simplify scientific text (and nothing more), in: European Conference on Information Retrieval, Springer, 2025, pp. 425–433.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[14] OpenAI, Introducing gpt-4.1 in the api, https://openai.com/index/gpt-4-1/, 2025. Accessed: June 16, 2025.
[15] A. a. M. Llama Team, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

A. Prompts

I want you to act as an automatic simplification system for texts written in English in the context of medicine. Your task is to process the input text and make it clearer, that is, generate a simplified version, identifying difficult expressions that should be the most complex ones in the text. Remember that the concept of lexical simplification, in terms of natural language processing (NLP), refers to the process of replacing complex words with simpler alternatives, preserving the original meaning of the text. Try to reorganize paragraphs of the original text that contain difficult-to-understand grammatical structures so that the simplified text is easy to understand for the general public, second language learners, people with low literacy levels, and non-native speakers. You should keep in mind that if two or more words in the text form a single concept (for example, "artificial intelligence"), treat them as a unit and do not separate them; never simplify or replace words within a title; regarding acronyms, you should replace them with the corresponding meaning. Finally, when there are complex words in the simplified text, I require you to include a brief explanation in parentheses in all cases. We do not need you to include explanations of difficult words once the text has been simplified. Finally the result must be in JSON format Example:\n{ \n: "simplification" \n}

Figure 2: First version of the prompt used for the automatic generation of simplified texts in the biomedical domain for Task 1.1 and Task 1.2

It acts as an advanced automatic lexical simplification system for medical texts written in English. Its main objective is to process the original text and generate a significantly clearer and more understandable version, specifically designed for audiences with low literacy levels, students of English as a second language, non-native speakers, and the general public. The task of the advanced automatic lexical simplification system is to identify lexical complexity, that is, to detect the most difficult words or expressions in the text. These should be technical medical terms, uncommon words, or idiomatic phrases that make understanding difficult. It should then replace each difficult word or expression with a simpler and more understandable alternative, maintaining the same meaning. In cases where relatively complex words are identified, include a brief explanation in parentheses immediately after the word. It is necessary to preserve the meaning by ensuring that the original meaning of the source text is preserved at all times. If a sentence or paragraph has a complex grammatical structure, reorganize it to facilitate understanding without altering its message. The advanced automatic lexical simplification system should not alter lexical items that represent a single concept (e.g., "artificial intelligence"), nor should it simplify or modify words that are part of a title. In the case of acronyms, write their full meaning in parentheses next to them at least the first time they appear in the text. Finally, the result must be in JSON format. Example: "simplification": "simplificación"

Figure 3: Second version of the prompt used for the automatic generation of simplified texts in the biomedical domain for Task 1.1 and Task 1.2.

You are evaluating whether a sentence contains hallucinated information based on a given context.
A sentence is considered a **hallucination** if **any part** of it presents information that is not **explicitly and clearly stated** in the context.
- Do not assume or infer any facts.
- If the sentence goes beyond the given context, even slightly, mark it as a hallucination.
- If you are uncertain, err on the side of caution and mark it as a hallucination.
Sentence: {sentence}
Context: {context}
Answer with only one word: 'Yes' if it is a hallucination, or 'No' if it is fully supported. Do not explain.

Figure 4: Task 2.1. Prompt that classifies a sentence as spurious or not given a context, that is, classify whether a given sentence is hallucinated or not based on a context

You are an AI assistant specializing in academic research paper abstracts. Your task is to generate a full, plausible abstract for a scientific or technical paper.
The abstract's core content and findings MUST directly support, or clearly imply, the following simplified core idea/sentence:** {sentence}
This means the provided sentence should either be present verbatim, or the abstract's content should make that sentence a straightforward, accurate, and concise summary or conclusion that could be drawn from it.
Construct a complete abstract that typically includes:
1. **Background/Problem:** Introduce the context or problem addressed by the research.
2. **Approach/Methodology:** Briefly describe how the research was conducted, what system/design was proposed, or what data was analyzed.
3. **Key Findings/Results:** Present the main outcomes, discoveries, or the core functionality, ensuring this section is where the provided sentence's idea is most strongly rooted.
4. **Conclusion/Implications:** Summarize the significance, benefits, or future outlook derived from the findings.
The abstract should be between 150-250 words, maintain a formal, objective, and scientific tone, and ensure smooth, logical transitions between sections. It should read as if it were a genuine abstract from a published paper.
Generate only the abstract text.

Figure 5: Task 2.1. Prompt that generates a new synthetic context given a sentence.