NIL-UCM at the CLEF 2026 SimpleText Track: Prompting, Fine-Tuning, and Multi-Agent Architectures for Biomedical Text Simplification

Sandra Conde[1], Pablo Folgueira[1] and Alberto Díaz[2]

[1]Facultad de Informática, Universidad Complutense de Madrid, Spain
[2]Facultad de Informática and ITC, Universidad Complutense de Madrid, Spain

Abstract

This paper describes the participation of NIL-UCM in SimpleText 2026 Task 1, addressing the automatic simplification of biomedical scientific texts drawn from the Cochrane-auto corpus. We explore a broad spectrum of approaches for both sentence-level (Task 1.1) and document-level (Task 1.2) simplification, including zero-shot and few-shot role-based prompting, a two-phase classification and simplification pipeline, plain language guidelines prompting, supervised fine-tuning via Low-Rank Adaptation (LoRA), and a multi-agent architecture. Our experiments are conducted across multiple open-source LLMs in the 7–8 billion parameter range and their quantized variants. Results show that zero-shot role-based prompting constitutes a strong and consistent baseline, while more complex strategies yield mixed results. The multi-agent pipeline achieves the highest overall SARI score (45.34), and supervised fine-tuning produces the most significant degradation. Quantized model variants exhibit minimal performance loss relative to their full-precision counterparts across both subtasks.

Keywords

text simplification, biomedical NLP, large language models, prompt engineering, multi-agent systems, plain language, Cochrane, LoRA fine-tuning

1. Introduction

The growing volume of scientific literature published annually presents a fundamental challenge to knowledge democratization: highly specialized texts remain largely inaccessible to non-expert audiences, including patients, policymakers, and the general public [1]. This accessibility gap is critical in the biomedical domain, where complex technical terminology, dense argumentation structures, and implicit domain knowledge constitute significant barriers to comprehension. Automatic scientific text simplification (STS) has emerged as a promising research direction to overcome these barriers, leveraging advances in Natural Language Processing (NLP) to reduce both linguistic and conceptual complexity without compromising factual accuracy.

Recent advances in Artificial Intelligence (AI), and particularly in Large Language Models (LLMs), have substantially shifted the landscape of text simplification research. While these generative models exhibit strong zero-shot and few-shot capabilities, critical challenges persist, foremost among these being the propensity to hallucinate or distort factual content, the difficulty of retaining essential domain-specific terminology, and the trade-off between maximizing readability gains and preserving semantic fidelity [2]. Furthermore, as source texts increase in length and linguistic complexity, mitigating these challenges becomes increasingly critical to ensure discourse coherence and accurate structural reorganization.

This paper describes our participation in SimpleText 2026 [4, 5, 6] Task 1, hosted within the CLEF 2026 [7] evaluation framework, addressing the automatic simplification of biomedical scientific texts drawn from the Cochrane-auto corpus. Our participation explores a broad spectrum of approaches, from prompting strategies of increasing complexity to supervised fine-tuning and a multi-agent architecture, evaluated across multiple open-source LLMs and their quantized variants. This work provides a

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
saconde@ucm.es (S. Conde); pablofol@ucm.es (P. Folgueira); adiazest@ucm.es (A. Díaz)
0009-0009-3535-1330 (S. Conde); 0009-0005-3725-3954 (P. Folgueira); 0000-0003-1966-3421 (A. Díaz)


systematic comparison of simplification quality, architectural overhead, and computational cost in the biomedical domain.

2. Task Description

SimpleText 2026 Task 1 focuses on the automatic simplification of biomedical scientific texts within the CLEF 2026 evaluation framework. The task is grounded in the **Cochrane-auto corpus**[9], a parallel dataset derived from Cochrane systematic review abstracts and their corresponding Plain Language Summaries (PLS). Because this corpus supplies authentic parallel texts generated by the original domain experts, it facilitates genuine document-level simplification driven by actual communicative goals. Furthermore, the dataset incorporates a broad range of discourse-level simplification strategies, including sentence merging, reordering, and structural realignment at the paragraph, sentence, and document levels, making it a particularly rich and challenging resource for the development and evaluation of automatic simplification systems.

2.1. Addressed Subtasks

While SimpleText 2026 track comprises multiple tasks, this work focuses on Task 1 and, within it, on the following two subtasks:

- **Task 1.1 – Sentence-level Scientific Text Simplification.** Participants are provided with individual sentences extracted from Cochrane biomedical abstracts and are required to produce a simplified version of each sentence in isolation.
- **Task 1.2 – Document-level Scientific Text Simplification.** Participants receive a complete Cochrane abstract as a single input document and must produce a simplified version of the entire text. Consequently, systems must maintain discourse coherence, manage cross-sentence dependencies, and produce output that reads as a unified, fluent plain language summary rather than a sequence of independently simplified sentences.

2.2. Dataset

The evaluation data for CLEF 2026 comprises newly collected Cochrane systematic reviews published during the preceding year, curated and processed in accordance with the methodology established in prior SimpleText editions and the Cochrane-auto pipeline. The test set released for CLEF 2026 is structured around four complementary data sources. The primary evaluation resource, hereafter referred to as **Cochrane-auto 2026**, consists of 175 newly collected Cochrane abstracts comprising 3,679 sentences in total, of which 32 abstracts include high-quality sentence-level alignments with their corresponding PLS, enabling fine-grained sentence-level evaluation. The remaining abstracts are evaluated exclusively at the document level. Table 1 summarizes the key statistics of this evaluation corpus.

Table 1
Statistics of the Cochrane-auto 2026 evaluation dataset.

Property	Value
Total abstracts	175
Total sentences	3,679
Abstracts with sentence-level alignment	32
Evaluation granularities available	Sentence & Document

Beyond the main evaluation set, the test input incorporates the CLEF 2025 SimpleText data for cross-year comparability, a pilot set of fully structured Cochrane abstracts spanning all seven canonical sections, and a multilingual pilot covering abstracts in up to eleven languages. Our evaluation is

conducted mainly on the Cochrane-auto 2026 data, as it constitutes the primary benchmark for the 2026 edition and aligns directly with the sentence-level and document-level subtasks in which we participated.

System outputs are assessed using a comprehensive suite of automatic metrics covering complementary dimensions of simplification quality. Lexical fidelity and overlap with reference simplifications are measured through SARI [10] and BLEU [11], while readability is assessed via the Flesch-Kincaid Grade Level (FKGL) [12]. Additional operation-level metrics quantify the proportion of copied, added, and deleted content relative to the source, alongside sentence splitting rate and Levenshtein distance, providing a fine-grained characterization of the transformations applied by each system. Lexical complexity and overall comprehensibility scores complement this suite by capturing the accessibility of the output for a lay audience. This multi-dimensional evaluation framework reflects the inherent complexity of simplification assessment, where no single metric captures all relevant aspects of output quality.

3. Methodology

In this section, we describe the approaches proposed for Task 1.1 and Task 1.2. Our methodology is primarily driven by the use of open-source LLMs and explores various prompt engineering strategies, supervised fine-tuning, and a multi-agent architecture.

3.1. Task 1.1: Sentence-level Simplification

For the sentence-level subtask, we evaluated a set of open-source LLMs in the 7–8 billion parameter range, as well as their quantized variants, testing three prompting strategies of increasing structural complexity.

3.1.1. Strategy 1: Role-based Prompting (Zero-shot)

The first strategy employs a role-based zero-shot prompt in which the model is assigned the role of an expert medical writer specialized in adapting complex biomedical text for lay audiences. The system prompt explicitly constrains the output format to a single simplified sentence, prohibiting additional commentary or explanation. This baseline configuration tests the intrinsic simplification capacity of the model under minimal guidance, relying on the role assignment to orient the model's output register towards plain language.

3.1.2. Strategy 2: Role-based Prompting with Few-shot Examples

The second strategy extends Strategy 1 by incorporating few-shot examples into the prompt. The system prompt remains identical to the previous one, while the user prompt is augmented with six input–output pairs covering a diverse range of sentence types representative of the Cochrane-auto corpus, including qualitative conclusions, study descriptions, numerical results with statistical parameters, and research recommendations. This diversity was intentional, aiming to provide the model with in-context demonstrations of varied simplification operations rather than a homogeneous set of examples.

3.1.3. Strategy 3: Two-phase Classification and Simplification

The third strategy decomposes the simplification process into two sequential phases, directly motivated by the annotation scheme of the Cochrane-auto corpus itself. In this dataset, each complex sentence is automatically assigned one of four simplification operation labels derived from its alignment with the corresponding Plain Language Summary: sentences unaligned to any simple sentence are labeled DELETE; sentences aligned to a single simple sentence with high lexical similarity (Levenshtein similarity ≥ 0.92) are labeled COPY; those aligned to a single simple sentence with lower similarity are labeled REPHRASE; and sentences aligned to multiple simple sentences are labeled SPLIT [9]. This label taxonomy reflects the actual discourse-level transformations performed by domain experts when

producing plain language summaries, and our two-phase architecture operationalizes it directly: in the first phase, the model acts as a classifier replicating this labeling scheme, while in the second phase, the predicted label routes each sentence to the appropriate operation-specific prompt.

Although the Cochrane-auto annotation scheme additionally defines a MERGE operation assigned to consecutive complex sentences that collectively align to a single simple sentence, this operation was excluded from our pipeline for two reasons. First, MERGE is structurally incompatible with the sentence-level processing paradigm of Task 1.1, which operates on individual sentences in isolation and therefore lacks access to the inter-sentence context required to identify mergeable sequences. Second, implementing MERGE would require an additional grouping mechanism prior to classification, redesigning the pipeline to operate over sentence windows rather than individual sentences, a complexity whose empirical benefit remains uncertain and which falls outside the scope of the present work.

- **COPY:** The sentence is already sufficiently simple and should be retained without modification.
- **DELETE:** The sentence contains purely technical details, excessive statistical parameters, or information deemed non-essential for a lay audience.
- **REPHRASE:** The sentence contains medical jargon that must be simplified through lexical substitution or paraphrase, producing a single output sentence.
- **SPLIT:** The sentence is too long or structurally complex and must be decomposed into two or more shorter sentences.

The classification prompt instructs the model to respond exclusively with one of the four labels, with no additional output, ensuring reliable downstream routing. In the second phase, the predicted label determines which operation-specific prompt is applied:

- **COPY** and **DELETE** sentences require no further model call so the sentence is either passed through unchanged or discarded.
- **REPHRASE** sentences are processed with a prompt that instructs the model to substitute or explain complex terminology using everyday language while preserving the core medical facts in a single continuous sentence.
- **SPLIT** sentences are processed with a prompt that instructs the model to decompose the input into two or more shorter sentences, separated by full stops, without introducing lists or line breaks.

Both operation-specific prompts enforce plain-text output constraints, explicitly prohibiting the use of bullet points, numbered lists, or line breaks, which are common failure modes of instruction-following LLMs in structured generation tasks. Figure 1 illustrates the overall architecture of this strategy.

3.2. Task 1.2: Document-level Simplification

Document-level simplification requires systems to produce coherent, discourse-aware plain language summaries from full biomedical abstracts, presenting greater challenges than sentence-level processing. To address this, we explored prompting, fine-tuning, and multi-agent strategies across multiple LLMs, each described in the following subsections.

3.2.1. Strategy 1: Role-based Prompting (Zero-shot)

The baseline system employs a zero-shot role-based prompt analogous to the approach described in Section 3.1.1, adapted to the document-level setting. The model is instructed to preserve all numerical findings and factual content without alteration, replacing domain-specific terminology with everyday language. An explicit constraint prohibits the generation of titles or meta-commentary, ensuring that the output consists solely of the simplified text.

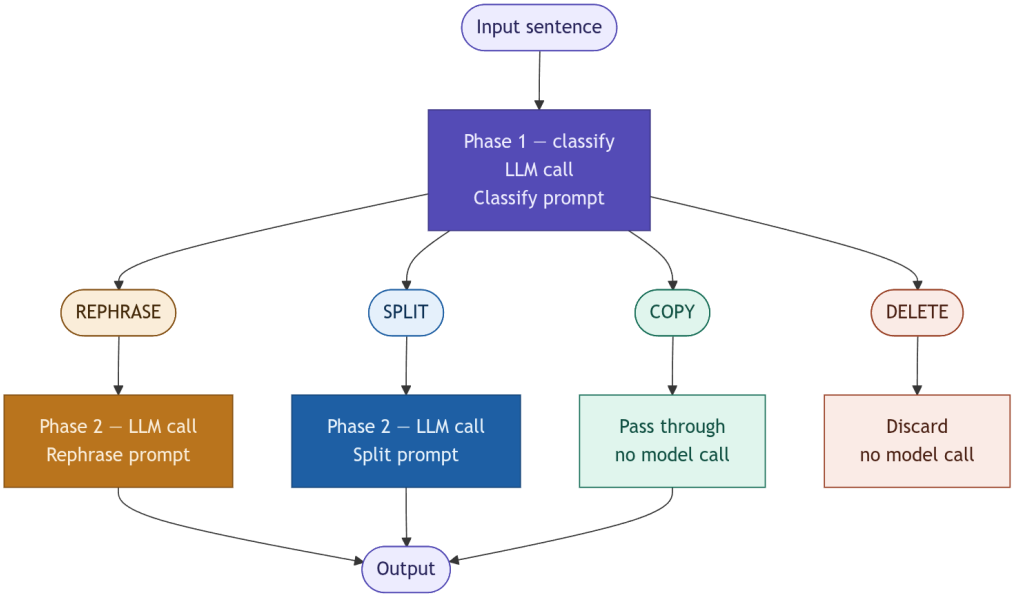

Figure 1: Two-phase classification and simplification pipeline for Task 1.1. In Phase 1, each input sentence is assigned one of four operation labels. In Phase 2, the label routes the sentence to the appropriate operation-specific prompt or to a pass-through/discard procedure.

3.2.2. Strategy 2: Role-based Prompting with Few-shot Examples

Building on Strategy 1, this run incorporates two document-level input–output pairs drawn from the Cochrane-auto 2025 corpus into the prompt, providing the model with in-context demonstrations of the expected simplification structure.

3.2.3. Strategy 3: Plain Language Guidelines Prompting

The third strategy operationalises the plain language writing standards defined in the Cochrane Handbook for Systematic Reviews of Interventions [13], which provides official guidelines for the preparation of Plain Language Summaries accompanying Cochrane systematic reviews. These guidelines were directly integrated into the system prompt, structured around four explicit categories:

1. **Language and vocabulary:** Substitution of research jargon with everyday equivalents (e.g., *study* for *trial*, *medicines* for *drugs*); mandatory explanation of acronyms and technical terms, with the plain language form presented first, followed by the technical term in parentheses.
2. **Style and tone:** Short sentences averaging 20 words; preference for active voice and direct verbs; use of first-person pronouns (*we* for the research team, *you* for the reader); numerals for all quantities unless sentence-initial.
3. **Structure:** Question-based subheadings to organize content (e.g., *What did we find?*); use of dash-delimited bullet points for enumerated items; paragraph breaks at topic transitions.
4. **Factual constraints:** All numerical findings must be reproduced exactly; no paraphrase of statistical data is allowed.

3.2.4. Strategy 4: Plain Language Guidelines with Medical Term Definitions

This strategy extends Strategy 3 by enriching the prompt with domain-specific lexical grounding. Prior to each inference call, medical terms are identified in the source abstract through dictionary lookup against the University of Michigan Plain Language Medical Dictionary [14], an open-source lexical resource mapping technical biomedical terminology to the equivalent Plain Language definitions. The retrieved definitions are dynamically injected into the system prompt alongside the plain language

guidelines, providing the model with explicit definitions for the most technical terms present in each input document. This strategy aims to provide the model with reliable plain language equivalents for technical terms directly within the prompt, reducing the need to generate them from the model's own knowledge and thus limiting the risk of imprecise or hallucinated substitutions.

3.2.5. Strategy 5: Fine-tuning on Cochrane-auto 2025

To assess the contribution of supervised adaptation, Mistral-7B-Instruct and Llama-3.1-8B-Instruct were fine-tuned on the Cochrane-auto 2025 training split using Low-Rank Adaptation (LoRA) [15] with rank $r = 64$, over 3 epochs on 849 abstract–PLS pairs. This approach was intended to allow the models to internalize the stylistic and structural patterns characteristic of Cochrane plain language summaries without the computational overhead of full parameter updates, under the hypothesis that domain-specific adaptation would yield simplifications more closely aligned with the register and conventions of the reference corpus.

3.2.6. Strategy 6: SARI-focused Prompting

Informed by qualitative analysis of Cochrane-auto 2025 reference simplifications, a sixth strategy was designed to optimize for the lexical operations rewarded by the SARI metric, namely deletion of redundant content, vocabulary substitution, and structural preservation. The prompt explicitly instructs the model to summarize and delete complex statistical parameters (confidence intervals, interquartile ranges, exact p-values, risk ratios) unless essential, while retaining the overall paragraph structure and tone of the source. Formatting constraints prohibit the use of bullet points or question-based subheadings, preserving the prose character of the original. This design reflects the empirical observation that Cochrane plain language summaries often retain the rhetorical organization of the source abstract while substantially compressing its statistical content.

3.2.7. Strategy 7: Multi-agent Pipeline

The most complex system developed for Task 1.2 is a multi-agent pipeline designed to decompose document-level simplification into specialized roles handled by coordinated LLMs, as illustrated in Figure 2. The pipeline is designed around a division of responsibilities, where each agent addresses a single, well-defined task rather than attempting to jointly optimize style, factual accuracy, and fluency within a single model call.

Parallel Drafters. Four LLM agents generate candidate simplifications concurrently and independently. Specifically, we employed Gemini 2.5 Flash Lite [16], DeepSeek V4 Flash [17], Mistral Small 4 [18], and Llama 3.3 70B Versatile [19], diversifying the candidate pool across model families and reducing the risk of systematic simplification biases propagating to the final output.

Judge. This agent (Llama 3.3 70B Versatile) evaluates the four candidate drafts and selects the most promising one based exclusively on style and readability criteria, namely vocabulary accessibility, sentence fluency, and natural tone for a lay audience. Crucially, the judge is explicitly instructed not to assess clinical accuracy, as factual verification is delegated to a dedicated downstream agent. Chain-of-thought reasoning is employed to produce a structured rationale analyzing the strengths and weaknesses of each candidate before issuing a verdict.

Evaluator Agents. Two specialized agents assess the selected candidate in parallel:

- **Fact Checker** (Gemini 2.5 Flash Lite): verifies numerical and clinical fidelity against the source abstract, flagging discrepancies in reported figures, effect sizes, or clinical conclusions, as well as any hallucinated content not present in the original. The agent is explicitly instructed to ignore style, tone, and formatting, focusing solely on factual and data accuracy.

- **Readability Evaluator** (Gemini 2.5 Flash Lite): assesses whether the simplified text is genuinely accessible and natural for a reader without medical training, identifying unexplained jargon, overly long sentences, or excessively academic phrasing. Conversely, this agent is instructed not to evaluate clinical correctness.

Both evaluators employ chain-of-thought reasoning and return structured feedback with specific, actionable instructions for the editor (e.g., *"The original states 42% efficacy but the simplification reports 24%"* or *"Simplify vocabulary in the second paragraph"*).

Editor. If either evaluator rejects the simplification and the maximum number of iterations has not been reached, this agent (Mistral Small 4) receives the selected draft together with the feedback from both evaluators and produces a revised version. The agent is explicitly instructed to correct only the issues identified in the feedback without rewriting the text from scratch, thereby preserving the style and structure selected by the judge. This iterative refinement cycle continues until both evaluators approve the output or the three-iteration limit is reached, at which point the best available version is returned as the final output.

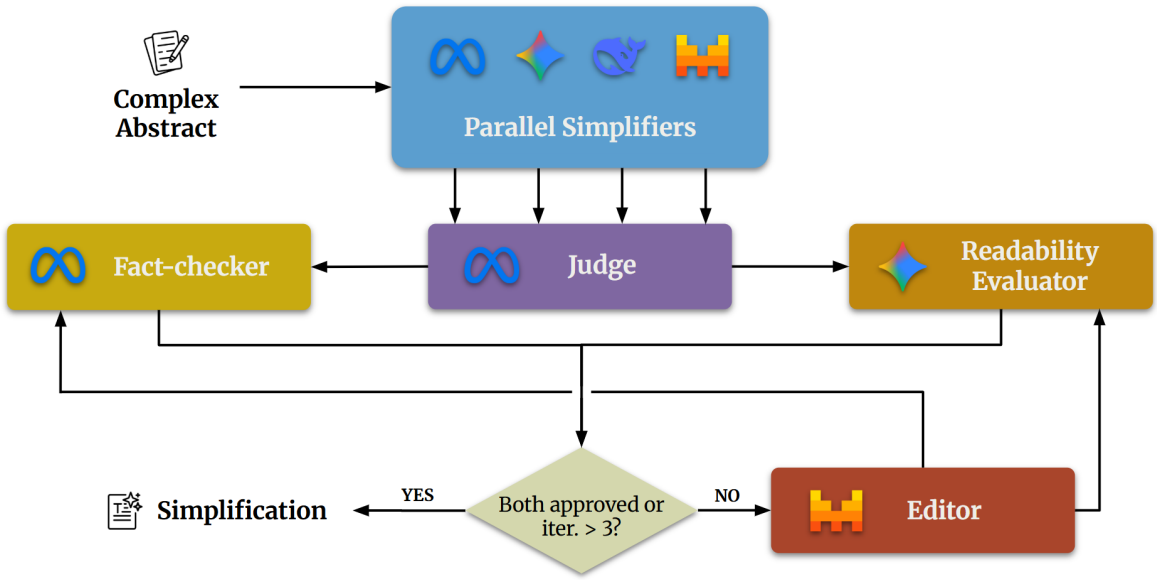

Figure 2: Multi-agent pipeline for Task 1.2. A complex biomedical abstract is processed by four parallel drafter agents, whose outputs are evaluated by a judge agent that selects the best candidate. The selected simplification then enters an iterative refinement loop in which a Fact Checker and a Readability Evaluator assess factual fidelity and linguistic accessibility respectively. An editor agent incorporates their feedback until both evaluators approve the output or the maximum number of iterations (3) is reached.

3.2.8. Strategy 8: Multilingual Plain Language Guidelines Prompting

To explore the applicability of the Guidelines-based approach beyond the primary evaluation set, an additional run was conducted over the complete CodaBench test set, which encompasses not only the main Cochrane-auto 2026 English abstracts but also the pilot subsets covering multiple languages and fully structured abstracts. This run is based on Strategy 3 (Plain Language Guidelines), extended with a single additional instruction in the system prompt specifying that the simplification should be produced in the language indicated by the language field of each input instance. No other modifications were introduced relative to the original Strategy 3 configuration.

This strategy was not applied to the full dataset in earlier experiments due to its substantial computational cost. Execution over the complete test set required approximately 14 hours of inference on an NVIDIA A100-SXM4-40GB GPU, consuming around 4.84 kWh of energy and generating approximately

1.51 kg of CO_2 equivalent emissions as measured by CodeCarbon. Given that the primary evaluation set constitutes the main benchmark for the 2026 edition, the remaining configurations were deliberately restricted to that subset in the interest of computational efficiency.

Given that the complete test set includes instances across up to eleven languages, this extension tests the multilingual generalisation capacity of Llama 3 under a guidelines-based prompting regime without any language-specific adaptation. It should be noted that evaluation on CodaBench is publicly available only for the primary Cochrane-auto 2026 subset, meaning that results for the pilot language and structured abstract subsets are not reported here.

4. Results and Discussion

This section presents and discusses the results obtained by all submitted configurations for Task 1.1 and Task 1.2, evaluated using the SARI metric. For each subtask, we analyse the effect of the different prompting strategies, fine-tuning, and architectural choices across models, with the aim of identifying which factors contribute most to simplification quality and where the main limitations of each approach lie.

4.1. Task 1.1: Sentence-level Simplification

Table 2 reports SARI scores for all system configurations evaluated on Task 1.1. Across models, scores range from 40.16 to 45.51, a relatively narrow band that suggests the choice of prompting strategy has a modest but non-negligible effect on simplification quality.

Among the three strategies evaluated, the two-phase classification and simplification pipeline (Strategy 3) does not yield consistent improvements over the simpler baselines. For Mistral, it produces the lowest score observed across all Task 1.1 configurations (40.16), underperforming the zero-shot role-based baseline by approximately 4 points. This result suggests that the structural decomposition introduced by Strategy 3 may not be beneficial at the sentence level, at least for this model: the classification phase introduces an additional potential source of error, and any misassignment of operation labels propagates directly to the generation phase with no correction mechanism. For Llama 3, the two-phase strategy also underperforms relative to both the zero-shot and few-shot variants (42.25 vs. 45.51 and 45.43 respectively), reinforcing this pattern. The origin of this consistent degradation under Strategy 3 requires further investigation, particularly regarding the accuracy of the classification phase and its downstream impact on simplification quality.

Llama 3 achieves the highest individual score (45.51) under the role-based zero-shot prompt, and exhibits strong performance across strategies, with the quantized variant (45.38) incurring negligible degradation relative to the full-precision model. This robustness to quantization is consistent across Mistral as well (44.87 vs. 44.33), suggesting that for sentence-level simplification, quantized models represent a computationally efficient alternative without meaningful quality loss.

Gemma 2 obtains competitive results under the zero-shot setting (44.91), although the two-phase strategy yields a slight decrease (44.40), reinforcing the pattern observed across models whereby the added structural complexity of Strategy 3 does not translate into measurable SARI gains at the sentence level.

Overall, the zero-shot role-based prompt emerges as the most consistently competitive configuration for Task 1.1, with Llama 3 as the strongest model under this setting. The limited gains from more complex prompting strategies invite caution regarding the assumption that structural decomposition of the simplification process necessarily improves performance as measured by automatic metrics.

4.2. Task 1.2: Document-level Simplification

Table 3 reports SARI scores for all system configurations evaluated on Task 1.2. The overall score range (32.54–45.34) is considerably wider than that observed in Task 1.1, reflecting the greater variability introduced by the diversity of strategies explored and the added difficulty of document-level processing.

Table 2
SARI scores for Task 1.1 (sentence-level simplification). Q denotes quantized model variant. — indicates configuration not evaluated.

Model	Strategy 1 Role-based	Strategy 2 Few-shot	Strategy 3 Two-phase
Mistral	44.87	—	40.16
Mistral Q	44.33	—	—
Llama 3	45.51	45.43	42.25
Llama 3 Q	45.38	—	—
Gemma 2	44.91	—	44.40

The multi-agent pipeline achieves the highest SARI score across all Task 1.2 configurations (45.34 under the role-based prompt variant), outperforming all single-model systems. This result should be interpreted with some caution, however, as the multi-agent pipeline employs proprietary, non-open-source models as parallel drafters which likely possess greater parametric capacity than the 7–8 billion parameter open-source models used in the remaining configurations. Consequently, the performance advantage of the multi-agent system may reflect model capability differences as much as architectural design choices. The second multi-agent variant, configured with Guidelines and plain language term definitions, obtains a slightly lower score (44.89), suggesting that the more constrained prompting does not translate into additional metric gains despite its greater structural complexity.

Among single-model configurations, Llama 3 with Guidelines prompting (Strategy 3) achieves the best individual result (44.99), marginally surpassing the role-based zero-shot baseline (44.40). The Guidelines with plain language definitions variant (Strategy 4) yields a comparable score (44.91), indicating that the dynamic injection of medical term definitions provides negligible additional benefit over the guidelines alone as measured by SARI, a noteworthy finding given the additional computational overhead that dictionary lookup introduces at inference time.

The SARI-focused prompt (Strategy 6) produces competitive results for Llama 3 (44.18) and Mistral (42.01), but does not consistently outperform the role-based baseline in either case, raising questions about the extent to which explicit metric-oriented prompt design translates into gains on the target metric itself. For Gemma 2, Strategy 3 (Guidelines, 43.98) slightly outperforms the role-based baseline (43.29), representing one of the few cases where a more complex prompting strategy yields a consistent improvement.

The few-shot strategy (Strategy 2) was only evaluated on Llama 3, where it yields a decrease relative to the role-based baseline (42.58 vs. 44.40). This underperformance mirrors the pattern observed in Task 1.1 and suggests that the two document-level examples incorporated into the prompt may constrain the model's output in ways penalised by the lexical operations underlying SARI, though this interpretation requires further empirical validation.

Regarding supervised adaptation, Mistral fine-tuned with Strategy 5 obtains 35.95, while Llama 3 produces the most pronounced degradation across all configurations with a SARI score of 32.54, a decrease of nearly 12 points relative to its zero-shot baseline. As noted in Section 3.2.5, the hypothesis that LoRA fine-tuning would align model outputs with the stylistic conventions of Cochrane plain language summaries was not confirmed under the experimental conditions employed. This degradation is likely attributable to overfitting to the patterns of the training corpus, resulting in a loss of generalization capacity: the fine-tuned models tend to produce outputs that closely mirror the lexical and syntactic structure of the complex source text rather than generating genuinely simplified reformulations, which is precisely the behaviour penalized by SARI's deletion and substitution components.

Strategy 8, which extends the Guidelines-based prompt to the complete CodaBench test set by instructing the model to produce each simplification in the language of the source instance, obtains a SARI score of 44.11 on the primary Cochrane-auto 2026 subset — which comprises exclusively English abstracts — with Llama 3, marginally below that of Strategy 3 on the same subset (44.99). This suggests that the additional multilingual instruction does not meaningfully affect performance on English

instances.

Quantized model variants exhibit minimal performance degradation relative to their full-precision counterparts in the role-based setting (Llama 3: 44.40 vs. 44.34; Mistral: 43.84 vs. 42.59), consistent with the findings from Task 1.1 and supporting the use of quantized models as computationally efficient alternatives for document-level simplification.

Table 3
SARI scores for Task 1.2 (document-level simplification). Q denotes quantized model variant. — indicates configuration not evaluated.

Model	Strategy 1 Role-based	Strategy 2 Few-shot	Strategy 3 Guidelines	Strategy 4 Guidelines+PL	Strategy 5 Fine-tuning	Strategy 6 SARI-focused	Strategy 8 Multilingual Guidelines
Mistral	43.84	—	41.43	—	35.95	42.01	—
Mistral Q	42.59	—	—	—	—	—	—
Llama 3	44.40	42.58	44.99	44.91	32.54	44.18	44.11
Llama 3 Q	44.34	—	—	—	—	—	—
Gemma 2	43.29	—	43.98	—	—	—	—
Multi-agent (Role-based): 45.34							
Multi-agent (Guidelines+PL): 44.89							

5. Conclusions

This paper has described the participation of NIL-UCM in SimpleText 2026 Task 1, presenting a broad range of systems for both sentence-level (Task 1.1) and document-level (Task 1.2) biomedical text simplification. Our experiments explored prompting strategies of increasing complexity and supervised fine-tuning via LoRA across multiple open-source LLMs and their quantized variants, as well as a multi-agent architecture combining both open-source and proprietary models.

The results indicate that zero-shot role-based prompting constitutes a strong baseline that is difficult to improve upon consistently. More complex prompting strategies yielded mixed results: the Guidelines-based prompt proved to be the most effective single-model configuration for Task 1.2 (44.99 SARI with Llama 3), while few-shot prompting and the two-phase classification pipeline systematically underperformed relative to the zero-shot baseline across both subtasks. Supervised fine-tuning produced the most significant degradation, with scores dropping by up to 12 points, likely due to overfitting leading to outputs that mirror the source text rather than simplify it.

The multi-agent pipeline achieved the highest overall score (45.34 SARI), though this result should be interpreted considering that the pipeline incorporates proprietary models of greater capacity than the open-source systems used elsewhere. Quantized variants exhibited minimal degradation relative to full-precision models across both subtasks, supporting their use as computationally efficient alternatives.

5.1. Limitations

The primary limitation of this work is the reliance on a single automatic metric (SARI) for system comparison, which captures only certain dimensions of simplification quality and may not correlate with human judgements of readability or factual fidelity. Additionally, the multi-agent pipeline incorporates proprietary models, limiting its reproducibility and making it difficult to isolate the contribution of the architectural design from that of model capacity. Finally, the fine-tuning experiments were conducted under constrained conditions (limited training data, a single LoRA rank, and a fixed number of epochs) which restricts the conclusions that can be drawn about the potential of supervised adaptation for this task.

5.2. Future Work

Several directions emerge naturally from the limitations identified above. Human evaluation of a representative subset of system outputs would provide a more reliable assessment of simplification quality across the dimensions of readability, factual accuracy, and fluency. Regarding fine-tuning, future work should explore larger training sets, alternative LoRA configurations, and training objectives more directly aligned with simplification metrics. Finally, extending the multi-agent architecture to operate exclusively with open-source models of comparable capacity would allow a cleaner assessment of the architectural contribution independently of model provenance.

Acknowledgments

This publication is part of the R&D&I project HumanAI-UI, Grant PID2023-148577OB-C22 (Human-Centered AI: User-Driven Adaptative Interfaces-HumanAI-UI) funded by MICI-U/AEI/10.13039/501100011033 and by FEDER/UE. We also thank the organizers of the CLEF 2026 SimpleText track for designing the evaluation framework, providing valuable datasets, and offering timely support throughout the competition. We appreciate the reviewers' feedback, which helped improve our methods and analysis. Additionally, we acknowledge the contributions of colleagues and mentors who provided insights during model development and experimentation.

Declaration on Generative AI

During the preparation of this work, the authors used **Claude Sonnet 4.6** in order to: **Grammar and spelling check**, **Paraphrase and reword**. After using this tool, the authors reviewed and edited the content as needed, and they take full responsibility for the publication's content.

References

[1] M. Shardlow, A survey of automated text simplification, International Journal of Advanced Computer Science and Applications 4 (2014). URL: https://api.semanticscholar.org/CorpusID:6068649.

[2] A. Devaraj, I. Marshall, B. Wallace, J. J. Li, Paragraph-level simplification of medical texts, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4972–4984. URL: https://aclanthology.org/2021.naacl-main.395/. doi:10.18653/v1/2021.naacl-main.395.

[3] S. Conde González, P. Folgueira Galán, A. Díaz Esteban, NIL-UCM at the CLEF 2026 SimpleText Track: Prompting, Fine-Tuning, and Multi-Agent Architectures for Biomedical Text Simplification, in: [8], 2026.

[4] L. Ermakova, H. Azarbonyad, J. Bakker, B. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: [7], 2026.

[5] N. Hofmann, J. Dauenhauer, N. O. Dietzer, I. D. Idahor, C. K. Kreutz, Lexical Simplification for Scientific Texts: Revisiting SARI in the Era of Modern LLMs, in: [7], 2026.

[6] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text, in: [8], 2026.

[7] M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[8] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[9] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.

[10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029/. doi:10.1162/tacl_a_00107.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[12] P. Kincaid, R. P. Fishburne, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, 1975. URL: https://api.semanticscholar.org/CorpusID:61131325.

[13] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, et al. (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, version 6.5 ed., Cochrane, 2024. Updated August 2024. Available from www.cochrane.org/handbook.

[14] University of Michigan Library, Medical dictionary for plain language, https://github.com/mlibrary/medical-dictionary, 2024.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[16] Google Cloud, Gemini 2.5 flash-lite documentation, https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-flash-lite?hl=es-419, 2025.

[17] DeepSeek AI, Deepseek-v4 flash api documentation and release notes, https://api-docs.deepseek.com/news/news260424#deepseek-v4-flash, 2024.

[18] Mistral AI, Mistral small model card, https://docs.mistral.ai/models/model-cards/mistral-small-4-0-26-03, 2026.

[19] Meta Llama Team, Llama 3.3 model card, https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md, 2024.