Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text

Jan Bakker[1], Liana Ermakova[2] and Jaap Kamps[2]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]Université de Bretagne Occidentale, HCTI, France

Abstract

This paper presents an overview of the CLEF 2026 SimpleText Task 1 on Text Simplification. The task aims to simplify scientific text. We discuss the data and benchmarks provided for these tasks, along with preliminary insights and anticipated challenges. Our main findings are the following. First, we observe a strong consistency in system rankings across carefully sentence-aligned abstracts and the larger, unfiltered set of document-level-aligned abstracts, indicating that training and evaluation on document-aligned texts are promising and viable. Second, this perspective broadens the scope of text simplification beyond the traditional focus on lexical and grammatical adjustments, and brings new dimensions into play. These include handling complex discourse structures, and accounting for the background knowledge required to understand the text. Third, although the results are encouraging and the submitted runs generally show higher quality than several years ago, there remains considerable room for improvement, especially when models must handle scientific language and biomedical jargon. More generally, we hope and expect that the constructed corpora and evaluation data will be used by researchers to further advance text simplification approaches, both in general and specifically for the biomedical domain.

Keywords

Scientific text simplification, Biomedical AI, Generative AI, Information access, Natural language processing

1. Introduction

Becoming scientifically literate is increasingly crucial. Objective scientific information helps people navigate a world in which misinformation, disinformation, and fabricated claims are only a click away. Although the importance of such information is widely recognized, the general public rarely engages with primary scientific sources. This raises a central question: how can we make scientific texts more accessible to everyone?

The impact of objective scientific information is especially clear in biomedicine, where research findings directly influence health decisions. Yet, the most reliable and up-to-date sources use specialized language and assume substantial prior knowledge, making them inaccessible to non-experts. Our track therefore centers on a concrete, societally relevant biomedical use case and treats accessibility as a core objective, distinguishing our work from earlier text simplification approaches that considered lexical and grammatical complexity in isolation.

Scientific text simplification offers a way to make these sources more approachable, and recent advances in large language models have accelerated progress in this area. Nonetheless, key challenges remain: dealing with domain-specific jargon and preserving the accuracy of the original information. Addressing these challenges requires high-quality training data and rigorously designed benchmark datasets, informed by detailed human assessments and systematic analyses of the limitations of current summarization and simplification systems.

To address these challenges, the CLEF 2026 SimpleTrack has three main aims: First, to *advance scientific text simplification* by expanding existing corpora with more varied paragraph- and document-level simplifications that respect complex discourse structure. This new biomedical corpus is built from aligned Cochrane abstracts and plain language summaries [1, 2] and is tailored to models such as

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
j.bakker@uva.nl (J. Bakker); liana.ermakova@univ-brest.fr (L. Ermakova); kamps@uva.nl (J. Kamps)
0009-0002-9085-8491 (J. Bakker); 0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)


Source (Cochrane Abstract)

CD012019 *We included eight studies involving 646 participants, most of which were of poor methodological quality. |The urinary biomarkers were evaluated either in a specific phase of menstrual cycle or irrespective of the cycle phase. |Five studies evaluated the diagnostic performance of four urinary biomarkers for endometriosis, including three biomarkers distinguishing women with and without endometriosis (enolase 1 (NNE); vitamin D binding protein (VDBP); and urinary peptide profiling); and one biomarker (cytokeratin 19 (CK 19)) showing no significant difference between the two groups. |All of these biomarkers were assessed in small individual studies and could not be statistically evaluated in a meaningful way. |None of the biomarkers met the criteria for a replacement test or a triage test. |Three studies evaluated three biomarkers that did not differentiate women with endometriosis from disease-free controls. |There was insufficient evidence to recommend any urinary biomarker for use as a replacement or triage test in clinical practice for the diagnosis of endometriosis. |Several urinary biomarkers may have diagnostic potential, but require further evaluation before being introduced into routine clinical practice. |Laparoscopy remains the gold standard for the diagnosis of endometriosis, and diagnosis of endometriosis using urinary biomarkers should only be undertaken in a research setting.*

Reference (Aligned Cochrane-auto Plain Language Summary)

CD012019 *We included eight studies involving 646 participants. |[Deleted] |All studies evaluated reproductive-aged women who were undertaking diagnostic surgery to investigate symptoms of endometriosis or for other indications. |Five studies evaluated the diagnostic accuracy of four urinary biomarkers, including four biomarkers that were expressed differently in women with and without endometriosis and one showing no difference between the two groups. |Generally, the reports were of low methodological quality and urine tests were only assessed in small individual studies. |None of the tests were accurate enough to replace diagnostic surgery. |Three other studies just identified three biomarkers that did not distinguish the two groups. |There is not enough evidence to recommend any urinary biomarker for use in clinical practice for the diagnosis of endometriosis. |[Deleted] |[Deleted]*

Reference (Original Cochrane Plain Language Summary)

CD012019 *The evidence included in this review is current to July 2015. We included eight studies involving 646 participants. All studies evaluated reproductive-aged women who were undertaking diagnostic surgery to investigate symptoms of endometriosis or for other indications. Five studies evaluated the diagnostic accuracy of four urinary biomarkers, including four biomarkers that were expressed differently in women with and without endometriosis and one showing no difference between the two groups. Three other studies just identified three biomarkers that did not distinguish the two groups. None of the assessed biomarkers, including cytokeratin 19 (CK 19), enolase 1 (NNE), vitamin D binding protein (VDBP) and urinary peptide profiling have been evaluated by enough studies to provide a meaningful assessment of test accuracy. None of the tests were accurate enough to replace diagnostic surgery. Several studies identified biomarkers that might be of value in diagnosing endometriosis, but there are too few reports to be sure of their diagnostic benefit. There is not enough evidence to recommend any urinary biomarker for use in clinical practice for the diagnosis of endometriosis. Generally, the reports were of low methodological quality and urine tests were only assessed in small individual studies. More high quality research trials are needed to accurately assess the diagnostic potential of urinary biomarkers identified in small numbers of studies as having value in detecting endometriosis.*

Figure 1: Aligned Abstracts and Plain Language Summaries

LLMs that handle long inputs. Second, to *detect and reduce hallucinations* by exploiting aligned sources, references, and model outputs to identify and quantify spurious information introduced by generative models. This tackles a key limitation of current systems and evaluation measures, which often fail to penalize unwarranted additional content—an especially critical issue for scientific applications and for NLP/IR more broadly. Third, to *classify scientific text* by predicting the field and the discipline of a scientific paper based on a given taxonomy. This accelerates consistent metadata across collections (arXiv, OpenAIRE, MADOC).

Hence, the CLEF 2025 SimpleText track is based on three interrelated tasks:

- ***Task 1: Text Simplification****: simplify scientific text.*
- ***Task 2: Controlled Creativity****: identify and avoid hallucination.*
- ***Task 3: Research Area Classification****: classification of scientific articles by research area.*

This paper gives an overview of the CLEF 2026 SimpleText Task 1 on Text Simplification, which aims to simplify scientific text. Further details on the entire track are provided in the CLEF 2026 SimpleText Track Overview [3]. Additional details on Task 2 on *Controlled Creativity* are in a companion Task 2 overview paper [4], and details on Task 3 on *Research Area Classification* are in a companion Task 3 overview paper [5]. We also refer to the respective participants' papers for further details.

The rest of this paper is structured as follows. Section 2 describes the task, the data, the format, and the evaluation measures. Section 3 describes the participants' approaches. Section 4 provides detailed results for the task. Section 5 provides further analysis of the results. We end with a discussion and conclusions in Section 6.

2. Task 1: Simplify Scientific Text

This section details *Task 1: Text Simplification* on simplify scientific text.

Table 1
Statistics for the automatically aligned Cochrane Abstracts and Plain Language Summaries used in SimpleText 2024–2026.

	Cochrane-auto			Cochrane	Cochrane
	2024	**2025**	**2026**	**full**	**sections**
# Documents	5,585	217	175	893	6,335
# Sentences	35,800	4,293	3,679	35,970	35,970

2.1. Description

The *Text Simplification* task aims to *simplify scientific text*. For 2025, we created a new CLEF SimpleText corpus based on biomedical literature abstracts and lay summaries from Cochrane systematic reviews, called Cochrane-auto [1]. An example is shown in Figure 1. This corpus is constructed following the exact construction of the existing Wiki-auto and Newsela-auto [6, 7]. The test data at CLEF 2025 were based on new Cochrane Abstracts and Plain Language Summaries published in the last 12 months (April 2024-March 2025). We continue this practice at CLEF 2026, adding new Cochrane Abstracts and Plain Language Summaries published in the last 12 months (April 2025-March 2026). In this way, the test data is new data published after the knowledge cut-off for almost all models used.

For 2026, we add a range of additional data for analysis while keeping the main evaluation focused on the Cochrane-auto-aligned pairs of Cochrane Abstracts and Plain-Language Summaries. First, we include a section-aligned version of the new 2026 data, following [2]. Whereas Cochrane-auto reconstructs a sentence-level alignment, the original data contain no direct sentence-level simplifications, and the section-level alignment is usually more reliable. That is, each of the seven sections of the original abstract—Background, Objectives, Search Methods, Selection Criteria, Data collection and analysis, Main results, and Authors' conclusions—is added individually to the test set. Second, research up to now has used only the Main Results and Author's Conclusion sections of the full abstracts, as these exhibit the best sentence alignment. The Cochrane-sections alignment [2] preserves all seven sections of the original abstract, and we also add this (rather long) full abstract for all new data. Third, we also crawled and processed the new matching Cochrane abstracts and plain language summaries in all other available languages, for both the section-level and full abstract. In addition to English, this includes Spanish, Farsi, French, Korean, Thai, Portuguese, Chinese, Hindi, German, and Japanese. The additional data offers a range of monolingual text-simplification tasks in all these languages (for which, in many cases, no other corpora exist). We can also provide cross-language text simplification, both from English to another simplified language and from another complex language to simplified English.

Table 1 shows the CLEF 2026 data in relation to the earlier released data sets. The larger number of Cochrane abstracts in Cochrane full and Cochrane-sections is due to the inclusion of additional languages and to each abstract being a structured abstract with seven sections.

2.2. Data

As discussed above, we constructed a large scientific text simplification corpus by realigning abstracts and lay summaries at the sentence, paragraph, and document levels. In 2026, we used the original Cochrane-auto corpus as the training data.

Train data The specific train data for Task 1 consists of 1,085 documents, 4,171 paragraphs, and 14,719 sentences, with paired content from the abstract and the plain language summary. While the track distinguishes only between sentence- and document-level text simplification, the paragraph-level input is also included, allowing paragraph-level text-simplification submissions to Task 1.2.

Test data The primary test data consists of 175 new Cochrane-auto abstracts with paired plain English summaries, composed of 3,679 source sentences.

These are new systematic reviews published by Cochrane over the last year. We process these paired abstracts and plain-language summaries in two ways.

- We process these as Cochrane-auto [1] to ensure a high-quality sentence and paragraph alignment. This results in a subset of 32 abstracts and 346 sentences. The processing is identical to that for Cochrane-auto and other text simplification datasets.
- For document-level text simplification, we can also use the original pairs of abstracts and plain language summaries, using only the *results and conclusions* sections, similar to [8]. This results in 175 abstracts with 3,679 source sentences.

We use the aligned subset for the main evaluation and also report the scores across the entire subset.

Analysis data For further analysis, we extended the test data with the CLEF 2025 test data and variants of the new 2026 data. This includes section-level-aligned Cochrane sections for each of the seven sections, as well as the full Cochrane abstracts. We also include versions in other languages of these new Cochrane summaries, both in full and section-split versions. This paper focuses on the main Cochrane-auto 2026 data, and the additional data is meant to inform future editions of the CLEF SimpleText track.

2.3. Formats

This section outlines the format of the data used in the CLEF 2026 SimpleText Task 1.

2.3.1. Source

Sentence-level:


```
[
  {
    "pair_id": "CD015746",
    "source": "Cochrane-auto 2026",
    "language": "en",
    "para_id": 0,
    "sent_id": 0,
    "complex": "We included 19 trials (17 RCTs and two cluster-RCTs)."
  },
  {
    "pair_id": "CD015746",
    "source": "Cochrane-auto 2026",
    "language": "en",
    "para_id": 0,
    "sent_id": 1,
    "complex": "The 19 trials enrolled 395,650 participants, with ages ranging from six weeks to 60
    ↪  years."
  },
  {
    "pair_id": "CD015746",
    "source": "Cochrane-auto 2026",
    "language": "en",
    "para_id": 0,
    "sent_id": 2,
    "complex": "Vaccines were delivered as a single dose in 14 studies; two doses, ranging from four
    ↪  to 24 weeks apart, in six studies; and three doses, four weeks apart, in one study."
  },
  . . .
  {
    "pair_id": "CD015746",
    "source": "Cochrane-auto 2026",
    "language": "en",
    "para_id": 5,
```

```
    "sent_id": 17,
    "complex": "Vi-TT compared to another TCV may result in little to no difference in all-cause
    ↪  mortality or SAEs, and likely results in little to no difference in AEs."
  },
  . . .
]
```


Document-level:


```
[
  {
    "pair_id": "CD015746",
    "source": "Cochrane-auto 2026",
    "language": "en",
    "complex":"We included 19 trials (17 RCTs and two cluster-RCTs). The 19 trials enrolled 395,650
    ↪  participants, with ages ranging from six weeks to 60 years. Vaccines were delivered as a
    ↪  single dose in 14 studies; two doses, ranging from four to 24 weeks apart, in six studies;
    ↪  and three doses, four weeks apart, in one study. Comparators included: no vaccine, placebo
    ↪  and other vaccines. Seven studies compared TCV with non-conjugated typhoid vaccines. Six
    ↪  studies compared one TCV to another TCV. TCV compared to control may result in a large
    ↪  reduction in acute typhoid fever (risk ratio (RR) 0.20, 95% confidence interval (CI) 0.12 to
    ↪  0.32; I2 = 70%; 6 studies, 101,896 participants; low-certainty evidence) and probably results
    ↪  in little to no difference in all-cause mortality (RR 0.80, 95% CI 0.35 to 1.85; I2 = 52%; 4
    ↪  studies, 100,337 participants; moderate-certainty evidence). TCV results in little to no
    ↪  difference in AEs when compared to control (RR 0.91, 95% CI 0.76 to 1.09; I2 = 0%; 3 studies,
    ↪  29,465 participants; high-certainty evidence) and a slight reduction in SAEs compared to
    ↪  control (RR 0.82, 95% CI 0.71 to 0.95; I2 = 0%; 6 studies, 89,625 participants;
    ↪  high-certainty evidence). TCV compared to non-conjugated typhoid vaccines may result in
    ↪  little to no difference in acute typhoid fever (RR 0.90, 95% CI 0.48 to 1.69; 1 study, 78
    ↪  participants; low-certainty evidence). There were no deaths in the included studies. When
    ↪  compared to non-conjugated typhoid vaccines, TCV likely results in little to no difference in
    ↪  AEs (RR 1.00, 95% CI 0.77 to 1.31; I2 = 0%; 3 studies, 244 participants; moderate-certainty
    ↪  evidence) and likely results in a slight reduction in SAEs (RR 0.30, 95% CI 0.05 to 1.88; I2
    ↪  = 0%; 2 studies, 732 participants; moderate-certainty evidence). For TCV compared to another
    ↪  TCV, none of the studies reported on acute typhoid fever. Vi tetanus toxoid vaccine (Vi-TT)
    ↪  may result in little to no difference in all-cause mortality compared to a different TCV (RR
    ↪  5.19, 95% CI 0.54 to 49.80; I2 = 0%; 2 studies, 2422 participants; low-certainty evidence).
    ↪  Vi-TT likely results in little to no difference in AEs compared to another TCV (RR 1.18, 95%
    ↪  CI 0.92 to 1.51; I2 = 39%; 4 studies, 2916 participants; moderate-certainty evidence) and may
    ↪  result in little to no difference in SAEs (RR 2.48, 95% CI 0.74 to 8.36; I2 = 0%; 3 studies,
    ↪  2866 participants; low-certainty evidence). The certainty of evidence was consistently
    ↪  reduced due to imprecision, indirectness and bias. This review highlights that TCVs, compared
    ↪  to controls, are effective in preventing typhoid fever, and may confer protection for up to
    ↪  four years. TCVs compared to non-conjugated typhoid vaccines may result in little to no
    ↪  difference in acute typhoid fever and AEs, and likely result in a slight reduction in SAEs.
    ↪  Vi-TT compared to another TCV may result in little to no difference in all-cause mortality or
    ↪  SAEs, and likely results in little to no difference in AEs.",
  }, . . . ]
```


2.3.2. Prediction

Sentence-level:


```
[
  {
    "pair_id":"CD015746",
    "source":"Cochrane-auto 2026",
    "language":"en",
    "para_id":0,
    "sent_id":0,
    "prediction":"We included 19 studies: 17 regular randomized trials and 2 cluster randomized
    ↪  trials.",
    "run_id":"TeamName_Task11_MethodUsed"
  },
```

```
  {
    "pair_id":"CD015746",
    "source":"Cochrane-auto 2026",
    "language":"en",
    "para_id":0,
    "sent_id":1,
    "prediction":"The 19 studies included 395,650 participants aged from six weeks to 60 years.",
    "run_id":"TeamName_Task11_MethodUsed"
  }, . . . ]
```

Document-level:

```
[
  {
    "pair_id":"CD015746",
    "source":"Cochrane-auto 2026",
    "language":"en",
    "prediction":"This review included 19 studies (17 standard randomized trials and 2 cluster
    ↪  trials) with a total of 395,650 participants aged from six weeks to 60 years. Most studies
    ↪  used a single vaccine dose (14 studies), while 6 studies used two doses given 4 to 24 weeks
    ↪  apart, and 1 study used three doses given 4 weeks apart. The vaccines were compared with no
    ↪  vaccine, a placebo, or other typhoid vaccines. Compared to no vaccine or placebo, typhoid
    ↪  conjugate vaccines (TCVs) may greatly reduce cases of typhoid fever. They probably make
    ↪  little or no difference to overall death rates. TCVs also show little or no difference in
    ↪  general side effects, and may slightly reduce serious side effects. When compared with older
    ↪  (non-conjugated) typhoid vaccines, TCVs may make little or no difference in preventing
    ↪  typhoid fever. There were no deaths reported in these studies. TCVs likely cause similar
    ↪  rates of general side effects and may slightly reduce serious side effects compared to these
    ↪  older vaccines. When different TCVs were compared with each other, none of the studies
    ↪  reported typhoid cases. One type (Vi-TT) may make little or no difference in overall death
    ↪  rates compared to other TCVs. It likely causes similar general side effects and may make
    ↪  little or no difference in serious side effects. Overall, the certainty of the evidence was
    ↪  often reduced due to study limitations (such as imprecise results or possible bias). Despite
    ↪  this, the review suggests that TCVs are effective at preventing typhoid fever and may provide
    ↪  protection for up to four years.",
    "run_id":"TeamName_Task12_MethodUsed"
  }, . . . ]
```

2.4. Evaluation

In 2026, we emphasize large-scale automatic evaluation measures (SARI, BLEU, compression, readability) that provide a reusable test collection. For further details on these evaluation measures for scientific text simplification, see [9].

Almost all participants used generative models for text simplification, yet existing evaluation measures are blind to potential hallucinations with extra or distorted content. In 2026, we continue to provide further analysis of ways to detect and quantify spurious content in the output, potentially corresponding to what is informally called "hallucinations."

Submissions were made through Codabench.[1] Due to differences in setup, each task had a separate competition on Codabench. The Task 1 runs were submitted at: https://www.codabench.org/competitions/16108/ (shown in Figure 2). The Task 2 runs were submitted at: https://www.codabench.org/competitions/16065/. Codabench greatly facilitated running the track in 2026 and provided active participants (who had also registered at Codabench) with full access to the competition, including the submission and leaderboard pages.

3. Participant's Approaches

A total of 24 teams submitted 387 runs in total. In the detailed results, we only include runs without errors that got a non-zero score.

[1] https://www.codabench.org/

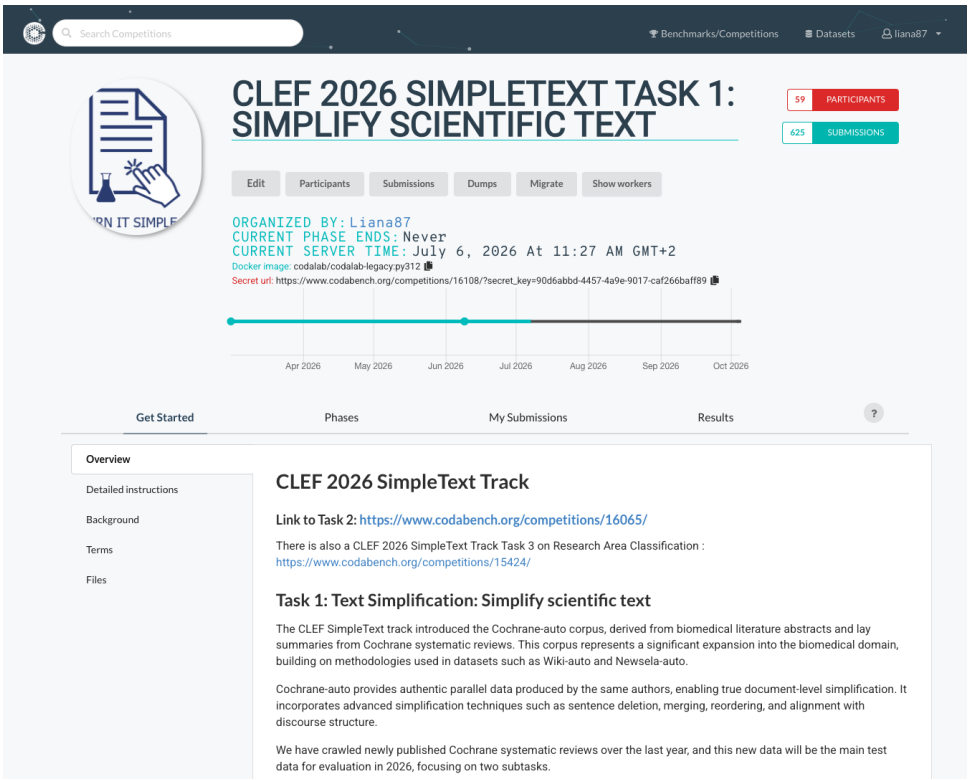

Figure 2: CLEF 2026 SimpleText Task 1 Codabench.

AIIRLab Hawkins et al. [10] submitted 30 runs in total for Task 1. They submitted 17 runs for Task 1.1 and 13 runs for Task 1.2. Their approach is Mistral-based "gist"- preserving, based on semantic role labeling, SARI-targeting prompts for Gemma-4, and an LLM-based ensemble of models that tries to maximize SARI without references. They also deploy an agentic pipeline based on Gemma-4 personas, and sequential simplification based on LLama.

BioSimplex Maurya et al. [11] submitted 3 runs in total for Task 1. They submitted 3 runs for Task 1.1 and 0 runs for Task 1.2. Their approach is based on a plan-guided text simplification approach, using SciBERT to predict text simplification operations and FLAN-T5 to perform them.

CLNN Noutche and Kreutz [12] submitted 1 run in total for Task 1. They submitted 0 runs for Task 1.1 and 1 run for Task 1.2. Their approach experiments with multiple zero-shot prompts of a LLaMA model.

Chennai Kumar et al. [13] submitted 8 runs in total for Task 1. They submitted 3 runs for Task 1.1 and 5 runs for Task 1.2. They use a plan-guided, sentence-level text-simplification approach and a summary-guided, document-level text-simplification approach, both based on LLaMA-3.1-8b-instruct.

DPO-KTO (no paper) submitted 13 runs in total for Task 1. They submitted 7 runs for Task 1.1 and 6 runs for Task 1.2.

DUTH Arampatzis and Arampatzis [14] submitted 77 runs in total for Task 1. They submitted 10 runs for Task 1.1 and 67 runs for Task 1.2. Their approach is based on fine-tuned Gemma and LLaMA models, either alone or in an ensemble with parameter tuning.

ELiRF-UPV Hurtado et al. [15] submitted 33 runs in total for Task 1. They submitted 16 runs for Task 1.1 and 17 runs for Task 1.2. They use a planner-based approach, using LLMs to generate the simplification in a controlled way, either directly with a Gemma 4 model or with Flan-T5 or BART combined with supervised fine-tuning and DPO.

Elsevier Capari et al. [16] submitted 32 runs in total for Task 1. They submitted 9 runs for Task 1.1 and 23 runs for Task 1.2. Their approach is based on sentences, with or without full abstract context, and on one-pass and dual-pass document-level text simplification, using structure and terminology planning, with GPT-4.1 as the underlying model.

FOSU-CR Chen et al. [17] submitted 3 runs in total for Task 1. They submitted 1 run for Task 1.1 and 2 runs for Task 1.2. Their approach is based on reference-guided few-shot prompting of DeepSeek-V3, using examples extracted from Cochrane-auto data, along with specific heuristic prompts to remove detailed statistical information from the summary.

FOSU-YZH Yang et al. [18] submitted 61 runs in total for Task 1. They submitted 46 runs for Task 1.1 and 15 runs for Task 1.2. Their approach experiments with a novel statistics-aware prompting approach using DeepSeek, ensuring the main systematic review results are preserved in an accessible way.

FutureMakers Liu et al. [19] submitted 5 runs in total for Task 1. They submitted 5 runs for Task 1.1 and 0 runs for Task 1.2. Their Task 1 runs got submitted as *FOSU-YZH* Yang et al. [18].

HULAT-HAI Pirvu and Cardon [20] submitted 2 runs in total for Task 1. They submitted 2 runs for Task 1.1 and 0 runs for Task 1.2. They use rule-based syntactic analysis in combination with specific prompts using Gemma-3-27B.

HULAT-UC3M Núñez et al. [21] submitted 7 runs in total for Task 1. They submitted 4 runs for Task 1.1 and 3 runs for Task 1.2. Their approach is based on a candidate generator and evaluator that optimize the evaluation measures, both built on Qwen models.

Monkey Chen et al. [22] submitted 1 run in total for Task 1. They submitted 1 run for Task 1.1 and 0 runs for Task 1.2. They use a prompted DeepSeek model with in-context learning based on lexically similar sentences from the train corpus.

NIL-UCM Conde et al. [23] submitted 31 runs in total for Task 1. They submitted 11 runs for Task 1.1 and 20 runs for Task 1.2. Their approaches include role-based prompting with in-context learning, PLS guidelines prompting, SARI-focused prompting, and an agentic approach using Llama, Mistral, and Gemma models.

NUQLab Ibrahim and Zaghouani [24] submitted 2 runs in total for Task 1. They submitted 1 run for Task 1.1 and 1 run for Task 1.2. They follow a planner-based approach to text simplification using RoBERTa for plan generation and BART for plan-guided text simplification.

SPU Rabih et al. [25] submitted 7 runs in total for Task 1. They submitted 7 runs for Task 1.1 and 0 runs for Task 1.2. Their approach is based on terminology and task-based prompting of an LLaMA model for text simplification.

THM Dauenhauer et al. [26] submitted 5 runs in total for Task 1. They submitted 3 runs for Task 1.1 and 2 runs for Task 1.2. Their experiments focused on Task 2, but their approach is also detailed in [27].

TUKE-IAI Shevchuk et al. [28] submitted 4 runs in total for Task 1. They submitted 4 runs for Task 1.1 and 0 runs for Task 1.2. Their approach uses a range of models (FLAN-T5, BioBERT, Gemma QLoRA) and prompt variations.

Tokatrons M et al. [29] submitted 2 runs in total for Task 1. They submitted 2 runs for Task 1.1 and 0 runs for Task 1.2. They use a plan-guided BART model and zero-shot-prompted LLaMA 3 and LLaMA 4 models.

UBO Vendeville et al. [30] submitted 1 run in total for Task 1. They submitted 0 runs for Task 1.1 and 1 run for Task 1.2. Their approach is based on a zero-shot prompt for Gemma 4, with a quality check using Gemma 4 to optimize evaluation measures and another quality check using Gemma 4 to detect information distortion, resulting in an iterative text-simplification approach.

UPF Hayakawa and Saggion [31] submitted 10 runs in total for Task 1. They submitted 0 runs for Task 1.1 and 10 runs for Task 1.2. They use a generate-and-select approach over a range of models (Gemma, Qwen, GPT, and Gemini), optimizing for SARI.

UvA Mathas et al. [32] submitted 5 runs in total for Task 1. They submitted 0 runs for Task 1.1 and 5 runs for Task 1.2. Their approach experiments with plan-guided text simplification using Qwen3 to execute the plans [33], and a novel retrieval-augmented text simplification approach that provides the four most similar section-level abstract-PLS pairs as context.

Writerslogic Condrey [34] submitted 40 runs in total for Task 1. They submitted 15 runs for Task 1.1 and 25 runs for Task 1.2. Their approach is based on generating multiple candidates simplifications with GPT-4o-mini or Claude Sonnet 4 and selecting the best candidate based on reference-free proxy measures.

Unknown team(s) (no paper) submitted 5 runs in total for Task 1. They submitted 3 runs for Task 1.1 and 2 runs for Task 1.2.

4. Results

This section details the task results for sentence- and document-level test simplification subtasks.

4.1. Task 1.1: Sentence-level Scientific Text Simplification

The main evaluation concerns the 32 abstracts, with 346 sentences aligned identically to the way Cochrane-auto and other collections are aligned. In this track overview paper, we evaluated all submissions to Task 1.1 and Task 1.2 at the document level to ensure identical ground truth and comparable scores across tasks.

Table 2 shows the Task 1.1 (sentence-level text simplification) results. The table is limited to submissions without issues, and we show at most one run per team. We present several evaluation scores against the human-reference simplifications, particularly SARI and BLEU. In addition, we provide statistics on the system output, such as FKGL, and compare them with the source input.

We make a number of observations. First, we include only the best-scoring submission based on SARI and observe SARI scores above 40% for almost all submissions. This is an encouraging result, and highlights that text simplification approaches are effective on scientific text and the specific biomedical corpus. Second, the readability measure FKGL gives a rough indication of the resulting grammatical and lexical complexity. We observe that all submissions lower the reading level, indicating that the text becomes more accessible. Third, the low BLEU scores differ from SARI's ranking, and they indicate that there is still considerable discrepancy with the official plain language summaries. The high fraction

Table 2
Results for CLEF 2026 SimpleText Task 1.1 sentence-level text simplification: Test data on 32 aligned Cochrane-auto abstracts, best three 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	32	9.12	14.41	13.81	1.00	1.00	1.00	1.00	0.00	0.00	8.78
Reference	32	100	100	10.82	0.44	0.63	0.45	0.00	0.11	0.69	8.50
ELiRF-UPV PromptingG	15	47.05	12.91	9.27	0.76	1.27	0.55	0.00	0.31	0.56	8.48
Writerslogic pls_pp	18	46.61	12.88	10.06	0.70	1.03	0.55	0.00	0.28	0.59	8.31
Writerslogic Task23_	18	46.61	12.88	10.06	0.70	1.03	0.55	0.00	0.28	0.59	8.31
Writerslogic pp_opus	18	46.55	12.64	9.28	0.67	1.06	0.53	0.00	0.27	0.62	8.35
FOSU-YZH complete_mi	52	46.26	12.18	10.86	0.70	1.05	0.55	0.00	0.25	0.58	8.46
FOSU-YZH Baseline_ne	52	46.26	12.18	10.86	0.70	1.05	0.55	0.00	0.25	0.58	8.46
FOSU-YZH hybrid_safe	52	46.23	12.42	10.59	0.67	1.05	0.55	0.00	0.23	0.59	8.47
FOSU-CR cr	1	45.62	16.82	10.33	0.68	1.02	0.63	0.00	0.19	0.51	8.42
AIIRLab EnsembleLLM	18	45.08	10.04	11.35	0.73	0.98	0.45	0.00	0.30	0.58	8.35
ELiRF-UPV PromptingG	15	45.07	9.36	7.72	0.90	1.81	0.54	0.00	0.43	0.52	8.41
SPU delete_fix	6	44.96	12.68	10.59	0.70	1.09	0.58	0.00	0.22	0.55	8.55
SPU executor	6	44.95	12.66	10.57	0.70	1.09	0.58	0.00	0.22	0.55	8.55
SPU raneem_PhaseId26	6	44.92	12.67	10.55	0.70	1.09	0.58	0.00	0.22	0.55	8.55
AIIRLab EnsembleLLM_	18	44.82	9.89	11.12	0.73	0.98	0.45	0.00	0.30	0.59	8.31
AIIRLab EnsembleLLM_	18	44.62	9.62	10.29	0.66	1.01	0.43	0.00	0.29	0.66	8.37
ELiRF-UPV PromptingG	15	44.55	8.10	7.38	0.92	1.81	0.54	0.00	0.46	0.53	8.44
DKSLA QWEN 3.5 9B	1	44.52	15.53	11.10	0.74	1.07	0.60	0.00	0.20	0.48	8.60
NIL-UCM Llama3Prompt	10	44.44	9.26	10.10	0.67	1.00	0.50	0.00	0.30	0.64	8.33
DUTH DeBERTaSelectV1	12	44.17	7.97	11.44	0.83	1.07	0.51	0.00	0.40	0.57	8.25
Elsevier FS3_BestMat	13	44.05	14.66	8.41	0.65	1.05	0.61	0.00	0.18	0.52	8.32
DUTH Gemma2_9B_Sente	12	44.02	8.11	10.90	0.72	0.97	0.50	0.00	0.34	0.62	8.24
DUTH Gemma3_12B_TopS	12	43.97	7.92	11.66	0.81	1.02	0.51	0.00	0.39	0.58	8.24
NIL-UCM Llama3Prompt	10	43.93	11.92	9.02	0.59	1.01	0.53	0.00	0.20	0.62	8.40
NIL-UCM MistralPromp	10	43.92	8.98	10.94	0.75	1.11	0.54	0.00	0.28	0.55	8.43
Monkey promptDeepsee	1	43.66	19.41	11.81	0.59	0.80	0.64	0.00	0.09	0.53	8.48
Elsevier SentencePro	13	43.27	13.92	8.21	0.65	1.07	0.61	0.00	0.18	0.52	8.36
CLNN Task1.1_P3	3	43.18	7.04	10.38	0.74	1.02	0.49	0.00	0.38	0.63	8.24
Elsevier SentencePro	13	43.15	13.04	9.90	0.73	0.98	0.62	0.00	0.23	0.47	8.35
Chennai Task1_1_llam	3	42.97	6.90	9.26	0.77	1.23	0.49	0.00	0.39	0.62	8.29
CLNN Task1.1_P1	3	42.53	6.13	13.76	1.01	1.05	0.52	0.00	0.50	0.49	8.36
TUKE RAG	6	42.49	11.08	7.11	0.67	1.23	0.59	0.00	0.22	0.53	8.47
HULAT-HAI with-orig	2	42.38	6.86	12.47	0.81	0.97	0.51	0.00	0.36	0.59	8.41
CLNN Task1.1_P2	3	42.19	5.71	14.39	1.11	1.13	0.51	0.00	0.54	0.46	8.40
HULAT-HAI no-orig	2	41.75	6.15	10.77	0.69	0.98	0.47	0.00	0.32	0.65	8.40
Chennai Task1_1_Mist	3	41.42	5.75	10.46	1.00	1.40	0.53	0.00	0.48	0.48	8.36
Chennai Task1_1_Llam	3	41.42	5.75	10.46	1.00	1.40	0.53	0.00	0.48	0.48	8.36
NUQLab janbakker_o_b	1	41.40	22.27	12.36	0.39	0.52	0.54	0.00	0.00	0.64	8.66
HULAT-UC3M task11_Qw	3	40.70	5.79	11.72	0.81	1.10	0.51	0.00	0.34	0.57	8.56
HULAT-UC3M task11_Qw	3	40.70	5.79	11.72	0.81	1.10	0.51	0.00	0.34	0.57	8.56
HULAT-UC3M task11_Qw	3	40.49	4.30	12.98	1.01	1.27	0.48	0.00	0.46	0.54	8.63
BioSimplex SciBERT_F	3	40.01	17.81	10.05	0.50	0.89	0.62	0.00	0.04	0.58	8.73
BioSimplex SciBERT_F	3	39.69	19.63	11.10	0.65	1.00	0.73	0.00	0.06	0.44	8.66
AUTH DPO on our synt	6	39.68	5.25	10.37	1.85	2.69	0.50	0.00	0.60	0.25	8.50
BioSimplex SciBERT_F	3	39.54	20.53	11.52	0.67	1.00	0.74	0.00	0.05	0.42	8.67
TUKE FlanT5Small_sen	6	39.35	19.15	11.77	0.37	0.49	0.51	0.00	0.01	0.65	8.74
AUTH KTO on our synt	6	38.60	6.64	9.74	1.04	1.56	0.60	0.00	0.42	0.40	8.54
AUTH SFT on open dat	6	38.12	2.32	8.75	4.45	7.22	0.35	0.00	0.77	0.15	8.66
TUKE Gemma2_9B_IT_Pr	6	36.52	21.14	12.11	0.70	0.99	0.79	0.00	0.03	0.37	8.73
Tokatrons task11_BAR	2	34.42	9.17	13.22	1.71	1.96	0.58	0.00	0.52	0.15	8.54
Tokatrons task11_LLa	2	14.70	13.84	13.61	0.98	1.00	0.96	0.31	0.04	0.06	8.73

Table 3
Results for CLEF 2026 SimpleText Task 1.2 document-level text simplification: Test data on 32 aligned Cochrane-auto abstracts, best three 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	32	9.12	14.41	13.81	1.00	1.00	1.00	1.00	0.00	0.00	8.78
Reference	32	80.82	45.86	10.32	0.96	1.40	0.47	0.00	0.39	0.48	8.51
FOSU-CR cr	1	47.97	16.96	8.54	0.55	0.97	0.48	0.00	0.18	0.63	8.47
ELiRF-UPV PromptingG	18	47.48	17.59	10.19	0.51	0.81	0.51	0.00	0.15	0.65	8.52
Writerslogic top3min	22	47.16	15.35	11.04	0.55	0.84	0.51	0.00	0.17	0.67	8.50
Writerslogic doc_v5	22	47.15	15.35	11.04	0.55	0.84	0.51	0.00	0.17	0.67	8.50
Writerslogic doc_v5	22	47.11	15.28	11.04	0.55	0.84	0.51	0.00	0.17	0.67	8.50
ELiRF-UPV PromptingG	18	46.98	14.57	7.91	0.57	1.14	0.51	0.00	0.20	0.64	8.40
ELiRF-UPV PromptingG	18	46.98	14.57	7.91	0.57	1.14	0.51	0.00	0.20	0.64	8.40
UvA RAG_BoN_k4_Q32	5	46.81	15.04	10.87	0.57	0.84	0.51	0.00	0.18	0.65	8.45
UvA PG_PCW_Q32	5	46.61	13.02	10.36	0.79	1.29	0.52	0.00	0.31	0.57	8.50
UvA RAG_BoN_k4_Q32L7	5	46.20	12.26	12.00	0.72	0.93	0.51	0.00	0.29	0.60	8.37
Elsevier FS3_BestMat	18	45.94	12.80	8.85	0.32	0.58	0.39	0.00	0.09	0.77	8.31
UPF MBR-SARI-ALL	10	45.79	11.60	9.89	0.70	1.11	0.51	0.00	0.26	0.58	8.51
Elsevier DocumentPro	18	45.56	11.36	9.16	0.29	0.52	0.36	0.00	0.07	0.79	8.38
Elsevier FS3_V3_Stru	18	45.52	13.06	8.99	0.37	0.62	0.40	0.00	0.12	0.75	8.28
NIL-UCM Agents	19	45.32	10.99	9.29	0.83	1.28	0.56	0.00	0.33	0.51	8.35
FOSU-YZH AgenticRAG4	15	45.31	13.91	8.10	0.37	0.68	0.43	0.00	0.09	0.73	8.42
TUKE Gemma2_9B_IT_Do	1	44.97	13.50	9.88	0.54	0.97	0.42	0.00	0.22	0.71	8.40
UPF MBR-SARI-Rank	10	44.85	11.19	10.01	0.69	1.04	0.49	0.00	0.26	0.59	8.52
FOSU-YZH AgenticRAG2	15	44.79	11.63	9.52	0.32	0.53	0.40	0.00	0.07	0.76	8.46
NIL-UCM Agents_Guide	19	44.59	9.48	6.54	0.89	1.77	0.51	0.00	0.42	0.50	8.38
NIL-UCM AgentsGuidel	19	44.59	9.48	6.54	0.89	1.77	0.51	0.00	0.42	0.50	8.38
UPF Random-Proprieta	10	44.59	10.28	9.40	0.72	1.17	0.50	0.00	0.28	0.57	8.46
DKSLA QWEN 3.6 27B	2	44.28	18.68	12.29	0.61	0.83	0.57	0.00	0.12	0.55	8.63
DUTH Gemma2_9B_Hybri	61	44.28	10.49	10.19	0.52	0.82	0.44	0.00	0.21	0.70	8.31
FOSU-YZH DocAgent_Co	15	44.28	10.05	9.62	0.28	0.50	0.36	0.00	0.05	0.79	8.60
DUTH Gemma2_9B_Hybri	61	44.20	8.78	11.49	0.68	0.90	0.48	0.00	0.32	0.64	8.23
AIIRLab concat	14	44.19	11.53	12.29	0.93	1.25	0.57	0.00	0.34	0.48	8.60
DUTH Gemma2_9B_Hybri	61	44.07	8.87	11.37	0.64	0.88	0.47	0.00	0.29	0.66	8.24
CLNN Task1.2_P5	3	43.91	12.17	11.32	0.42	0.55	0.41	0.00	0.14	0.72	8.56
CLNN Task1.2_P4	3	43.82	11.78	10.38	0.43	0.60	0.41	0.00	0.15	0.73	8.37
AIIRLab concat	14	43.27	11.01	11.89	0.88	1.20	0.57	0.00	0.32	0.50	8.60
UBO PASTIS	1	42.97	8.21	9.48	0.61	0.97	0.45	0.00	0.27	0.67	8.34
Chennai Task1_2_Mist	5	42.93	10.32	11.84	0.39	0.53	0.41	0.00	0.11	0.74	8.59
Chennai Task1_2_Llam	5	42.93	10.32	11.84	0.39	0.53	0.41	0.00	0.11	0.74	8.59
AIIRLab APIO_FewShot	14	42.21	13.06	13.49	0.86	1.05	0.66	0.00	0.24	0.45	8.48
CLNN Task1.2_P6	3	42.14	9.19	13.58	0.37	0.44	0.39	0.00	0.11	0.75	8.73
Chennai Task1_2_llam	5	41.46	7.57	10.28	0.42	0.62	0.39	0.00	0.19	0.77	8.48
NUQLab janbakker_o_b	1	41.40	22.27	12.36	0.39	0.52	0.54	0.00	0.00	0.64	8.66
HULAT-UC3M task12_Qw	3	41.09	18.63	12.59	0.70	0.90	0.68	0.00	0.11	0.45	8.66
AUTH KTO on our synt	6	39.13	9.14	8.61	0.47	0.74	0.48	0.00	0.14	0.68	8.68
AUTH DPO on our synt	6	39.06	8.10	9.36	0.59	0.86	0.53	0.00	0.20	0.61	8.59
AUTH SFT on our synt	6	38.59	8.70	9.28	0.53	0.79	0.50	0.03	0.18	0.66	8.65
HULAT-UC3M task12_Qw	3	36.92	2.24	14.49	0.40	0.48	0.31	0.00	0.16	0.83	8.69
DKSLA QWEN 3.5 9B	2	34.77	19.48	12.67	0.69	0.89	0.77	0.00	0.03	0.39	8.66
HULAT-UC3M task23d_Q	3	9.23	14.42	13.83	1.00	1.00	1.00	0.97	0.00	0.00	8.78
SIM Task23_Copy_Grou	2	9.12	14.41	13.81	1.00	1.00	1.00	1.00	0.00	0.00	8.78
SIM Task23_Copy	2	9.12	14.41	13.81	1.00	1.00	1.00	1.00	0.00	0.00	8.78

of insertions also indicates considerable edits, for many submissions exceeding those of the official references. This is of some concern to ensure no gratuitous edits or other unwarranted additions occur

in the predictions.

4.2. Task 1.2: Document-level Scientific Text Simplification

Table 3 shows the results of Task 1.2 (document-level text simplification). Again, we restrict the table to submissions that cover at most one run with non-zero scores per team. The used reference run is based on the 175 references and is not scoring optimally due to deletions and reordering of the aligned sentences in Cochrane-auto.

We make a number of observations. First, we observe a high SARI of above 40% again for almost all submissions. The document-level text simplification approaches score generally slightly higher than the sentence-level approaches, which is an encouraging development over traditional lexical and grammatical text simplification approaches. Second, the BLEU scores remain low but higher than those of the sentence-level text simplification approaches. This strengthens support for document-level approaches that take advantage of the full abstract's discourse structure. Third, the document-level text simplifications are generally shorter and exceed in the number of deletions

5. Analysis

In this section, we provide sentence-level evaluation over the set of 346 sentence pairs in the 32 Cochrane abstracts with high-quality sentence alignment. In addition, we provide additional evaluation on the larger set of 175 abstracts with 3,679 source sentences Unlike the subset discussed above, high-quality sentence alignment is not possible for this data. However, our primary interest is in document-level text simplification and evaluation, and our analysis explores the value of using parallel text directly as evaluation.

5.1. Sentence-level evaluation of Task 1.1 submissions

Table 4 evaluates the sentence-level submissions at the sentence level, against only the pairs of Cochrane-auto-aligned sentences in the abstract-PLS pairs. Here, we directly evaluate on the paired sentences, with the 32 aligned abstracts having 346 sentence pairs with the plain language summaries. Again, we restrict the table to the best-scoring runs per team.

We make a number of observations. First, we see broad agreement with the document-level evaluation of the same 32 abstracts in Table 2 before, with very similar SARI scores and slightly higher BLEU scores. Second, at the sentence level, the text statistics show more variation. We now see some fractions of exact sentence copies, and sentences retained in the PLS are occasionally split. Third, we generally see that sentence-level text simplification approaches perform well on sentence pairs with high alignment quality, achieving SARI scores similar to those observed in other collections. This suggests the validity of Cochrane-auto and the general approach of reusing biomedical abstract-PLS pairs as references for text simplification.

5.2. Evaluation on Plain Language Summaries

Table 5 shows the results of Task 1.1 (sentence-level text simplification) against a larger set of 175 abstracts and plain language summaries without further alignment. Again, we restrict the table to submissions that cover at most one run with non-zero scores per team. The used reference run is based on the 32 references and is not scoring optimally due to deletions and reordering of the aligned sentences in Cochrane-auto. Note again that all submissions in Task 1.1 and Task 1.2 are at the document level, to ensure identical ground truth and comparable scores across tasks.

We make a number of observations. First, we observe a similar, encouraging performance when evaluated against the larger set of plain-language reference simplifications. The ranking in Table 5 is very similar to the subset in Table 2 before, clearly supporting the claim that sentence alignment is not necessary even for the evaluation of sentence-level text simplification approaches. Second, we again

Table 4
Results for CLEF 2026 SimpleText Task 1.1 sentence-level text simplification: Sentence-level evaluation on 32 Plain Language Summaries, best run per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	346	10.76	17.92	14.12	1.00	1.00	1.00	1.00	0.00	0.00	8.54
Reference	346	100	100	10.77	0.93	1.10	0.55	0.03	0.33	0.50	8.24
FOSU-CR cr	1	47.08	19.05	10.11	0.80	1.03	0.67	0.03	0.24	0.42	8.19
ELiRF-UPV PromptingG	15	46.66	14.28	9.18	0.97	1.44	0.59	0.14	0.38	0.44	8.26
Writerslogic V4	18	45.86	12.69	9.02	0.59	1.00	0.58	0.00	0.17	0.57	8.27
Writerslogic V4	18	45.86	12.69	9.02	0.59	1.00	0.58	0.00	0.17	0.57	8.27
FOSU-YZH Agenticdeep	52	45.72	14.68	11.17	0.78	0.99	0.62	0.01	0.29	0.51	8.09
Writerslogic Task23_	18	45.25	14.13	10.05	0.80	1.06	0.58	0.01	0.35	0.53	8.11
FOSU-YZH Agenticgemi	52	45.11	13.20	6.77	0.96	2.08	0.65	0.01	0.38	0.42	8.13
FOSU-YZH NoContext	52	44.82	12.43	10.34	0.70	0.99	0.59	0.01	0.28	0.57	8.11
DKSLA QWEN 3.5 9B	1	44.81	17.39	10.58	0.93	1.12	0.64	0.14	0.26	0.40	8.41
SPU executor	6	44.58	13.34	10.57	0.82	1.10	0.63	0.08	0.30	0.49	8.32
SPU delete_fix	6	44.58	13.34	10.57	0.82	1.10	0.63	0.08	0.30	0.49	8.32
SPU copy_fix	6	44.58	13.34	10.57	0.82	1.10	0.63	0.08	0.30	0.49	8.32
AIIRLab llama	18	44.29	12.88	12.37	1.15	1.23	0.58	0.06	0.40	0.45	8.35
ELiRF-UPV PromptingG	15	44.01	10.21	7.73	1.03	1.83	0.56	0.01	0.46	0.46	8.23
AIIRLab Task11b_Qwen	18	43.47	19.87	10.36	0.76	1.12	0.75	0.08	0.11	0.38	8.37
AIIRLab EnsembleLLM	18	43.46	11.03	11.36	0.87	0.99	0.57	0.00	0.40	0.53	8.14
ELiRF-UPV PromptingG	15	43.31	9.23	7.36	1.06	1.84	0.55	0.00	0.50	0.47	8.12
Elsevier FS3_BestMat	13	43.11	15.91	8.43	0.74	1.06	0.63	0.00	0.26	0.48	8.06
Elsevier SentencePro	13	43.09	15.61	9.87	0.81	1.00	0.64	0.00	0.31	0.45	8.11
BioSimplex SciBERT_F	3	43.09	23.05	11.27	0.79	1.01	0.78	0.20	0.07	0.30	8.39
NIL-UCM Llama3Prompt	10	42.68	9.84	10.19	0.80	1.03	0.53	0.00	0.42	0.60	8.14
NIL-UCM Gemma2Prompt	10	42.61	11.80	11.01	0.84	1.16	0.58	0.23	0.32	0.49	8.48
Elsevier ZS_V3_Struc	13	42.45	15.42	10.16	0.83	1.00	0.65	0.00	0.31	0.43	8.10
NIL-UCM MistralPromp	10	42.43	9.94	11.01	0.86	1.11	0.56	0.00	0.40	0.54	8.19
DUTH Gemma2_9B_Sourc	12	42.39	9.11	11.60	0.92	1.03	0.55	0.03	0.45	0.53	8.10
BioSimplex SciBERT_F	3	42.31	20.75	10.68	0.76	1.01	0.77	0.15	0.08	0.32	8.36
DUTH PlainSelectorV2	12	42.18	8.39	11.74	0.97	1.03	0.52	0.00	0.50	0.55	8.09
DUTH PlainBestV1	12	42.17	8.38	11.70	0.97	1.03	0.52	0.00	0.50	0.55	8.07
TUKE RAG	6	41.84	12.76	6.92	0.81	1.28	0.62	0.00	0.31	0.46	8.22
Monkey promptDeepsee	1	41.51	18.50	10.64	0.72	0.98	0.69	0.15	0.13	0.42	8.21
CLNN Task1.1_P2	3	41.10	6.03	14.61	1.23	1.14	0.52	0.00	0.58	0.46	8.17
TUKE Gemma2_9B_IT_Pr	6	41.07	25.00	11.73	0.81	1.00	0.82	0.42	0.04	0.25	8.46
CLNN Task1.1_P1	3	41.05	6.57	13.89	1.13	1.06	0.54	0.00	0.55	0.47	8.14
CLNN Task1.1_P3	3	40.83	7.37	10.58	0.89	1.04	0.51	0.00	0.49	0.57	8.07
BioSimplex SciBERT_F	3	40.82	12.60	9.17	0.60	1.00	0.66	0.03	0.06	0.48	8.40
Chennai Task1_1_llam	3	40.60	7.11	9.04	0.86	1.28	0.50	0.00	0.50	0.64	8.11
HULAT-HAI with-orig	2	40.41	7.26	12.43	0.93	0.99	0.52	0.00	0.47	0.58	8.25
NUQLab janbakker_o_b	1	40.20	13.72	11.95	0.54	0.65	0.57	0.29	0.01	0.48	9.27
TUKE FlanT5Small_sen	6	39.72	10.64	10.75	0.51	0.57	0.51	0.38	0.01	0.51	9.47
HULAT-HAI no-orig	2	39.58	6.15	10.94	0.82	1.00	0.49	0.00	0.46	0.64	8.20
Chennai Task1_1_Llam	3	39.49	5.88	10.25	1.09	1.48	0.53	0.00	0.57	0.51	8.17
Chennai Task1_1_Mist	3	39.49	5.88	10.25	1.09	1.48	0.53	0.00	0.57	0.51	8.17
AUTH SFT on open dat	6	39.45	2.05	8.41	5.73	7.16	0.46	0.05	0.63	0.31	8.50
Tokatrons task11_BAR	2	38.87	11.58	13.22	2.50	2.00	0.54	0.00	0.54	0.14	8.43
AUTH DPO on our synt	6	38.38	4.52	10.82	3.26	3.04	0.54	0.01	0.58	0.33	8.35
HULAT-UC3M task11_Qw	3	37.48	3.26	12.69	1.34	1.31	0.48	0.00	0.53	0.57	8.48
AUTH KTO on our synt	6	37.47	6.99	9.73	1.05	1.60	0.61	0.00	0.47	0.46	8.26
HULAT-UC3M task11_Qw	3	37.42	4.43	11.74	1.11	1.15	0.52	0.00	0.44	0.56	8.41
HULAT-UC3M task11_Qw	3	37.42	4.43	11.74	1.11	1.15	0.52	0.00	0.44	0.56	8.41
Tokatrons task11_LLa	2	15.49	16.81	13.85	0.99	1.00	0.95	0.91	0.05	0.06	8.49

Table 5
Results for CLEF 2026 SimpleText Task 1.1 sentence-level text simplification: Test data on 175 Plain Language Summaries, best three 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	175	8.88	12.40	13.02	1.00	1.00	1.00	1.00	0.00	0.00	8.85
Reference	175	46.84	0.01	10.82	0.08	0.11	0.08	0.00	0.02	0.94	10.40
Writerslogic pls_pp	18	47.43	14.21	10.19	0.73	1.02	0.55	0.00	0.29	0.58	8.38
Writerslogic Task23_	18	47.43	14.21	10.19	0.73	1.02	0.55	0.00	0.29	0.58	8.38
Writerslogic pp_opus	18	47.36	13.51	9.66	0.70	1.03	0.54	0.00	0.28	0.59	8.41
FOSU-YZH hybrid_safe	52	47.01	11.97	10.55	0.70	1.04	0.55	0.00	0.26	0.58	8.50
FOSU-YZH hybrid_safe	52	47.00	11.96	10.55	0.71	1.04	0.55	0.00	0.26	0.58	8.50
FOSU-YZH selector_ch	52	46.95	11.98	10.71	0.73	1.04	0.55	0.00	0.27	0.57	8.50
ELiRF-UPV PromptingG	15	46.36	9.91	7.87	0.95	1.77	0.54	0.00	0.45	0.50	8.44
ELiRF-UPV PromptingG	15	46.35	11.54	9.17	0.78	1.24	0.52	0.00	0.33	0.56	8.47
AIIRLab EnsembleLLM	18	46.24	10.80	11.36	0.77	0.98	0.45	0.00	0.33	0.57	8.40
AIIRLab EnsembleLLM_	18	46.13	9.85	10.40	0.70	1.01	0.43	0.00	0.31	0.64	8.44
DUTH PlainBestV1	12	46.09	8.85	11.75	0.87	1.02	0.51	0.00	0.43	0.56	8.28
DUTH GroundedBestV1	12	46.04	8.84	11.87	0.87	1.02	0.51	0.00	0.43	0.56	8.30
DUTH Gemma2_9B_Sente	12	46.04	8.84	11.87	0.87	1.02	0.51	0.00	0.43	0.56	8.30
ELiRF-UPV PromptingG	15	46.00	8.76	7.51	0.96	1.78	0.53	0.00	0.48	0.51	8.48
AIIRLab EnsembleLLM_	18	45.70	10.70	11.18	0.77	0.98	0.45	0.00	0.33	0.57	8.39
NIL-UCM Llama3Prompt	10	45.51	8.99	10.04	0.70	0.99	0.50	0.00	0.32	0.63	8.34
NIL-UCM Llama3Prompt	10	45.43	9.80	9.04	0.61	1.00	0.52	0.00	0.22	0.62	8.42
NIL-UCM Llama3QPromp	10	45.38	8.69	11.03	0.78	0.99	0.50	0.00	0.38	0.61	8.40
FOSU-CR cr	1	45.07	14.92	10.49	0.72	1.01	0.63	0.00	0.21	0.50	8.42
SPU DeepSimplify_tas	6	45.00	11.46	10.19	0.74	1.10	0.57	0.00	0.25	0.54	8.57
CLNN Task1.1_P3	3	44.93	7.74	10.41	0.79	1.02	0.49	0.00	0.41	0.61	8.26
SPU delete_fix	6	44.92	11.60	10.27	0.73	1.10	0.58	0.00	0.24	0.54	8.58
SPU executor	6	44.91	11.61	10.22	0.73	1.10	0.58	0.00	0.24	0.53	8.58
Chennai Task1_1_llam	3	44.89	7.51	9.31	0.80	1.21	0.48	0.00	0.42	0.62	8.32
CLNN Task1.1_P1	3	44.50	6.73	13.95	1.08	1.04	0.52	0.00	0.53	0.46	8.38
CLNN Task1.1_P2	3	44.41	6.52	14.53	1.18	1.12	0.51	0.00	0.55	0.42	8.41
HULAT-HAI with-orig	2	44.20	7.80	12.71	0.88	0.98	0.51	0.00	0.40	0.55	8.49
HULAT-HAI no-orig	2	43.86	7.16	11.06	0.76	0.98	0.48	0.00	0.36	0.63	8.46
Chennai Task1_1_Mist	3	43.25	6.27	10.75	1.04	1.32	0.53	0.00	0.48	0.45	8.40
Chennai Task1_1_Llam	3	43.25	6.27	10.75	1.04	1.32	0.53	0.00	0.48	0.45	8.40
DKSLA QWEN 3.5 9B	1	42.92	13.46	11.25	0.79	1.08	0.61	0.00	0.22	0.46	8.65
HULAT-UC3M task11_Qw	3	42.86	6.94	11.64	0.85	1.11	0.51	0.00	0.36	0.56	8.58
HULAT-UC3M task11_Qw	3	42.86	6.94	11.64	0.85	1.11	0.51	0.00	0.36	0.56	8.58
Monkey promptDeepsee	1	42.79	13.07	11.37	0.59	0.78	0.62	0.00	0.10	0.54	8.54
HULAT-UC3M task11_Qw	3	42.79	5.72	12.80	1.04	1.24	0.48	0.00	0.48	0.54	8.59
Elsevier SentencePro	13	42.07	12.81	8.12	0.69	1.07	0.62	0.00	0.20	0.49	8.42
TUKE RAG	6	41.76	9.90	7.14	0.69	1.23	0.59	0.00	0.23	0.51	8.50
Elsevier FS3_BestMat	13	41.73	12.98	8.20	0.69	1.05	0.62	0.00	0.20	0.48	8.39
Elsevier SentencePro	13	41.47	11.76	9.71	0.78	0.99	0.63	0.00	0.25	0.44	8.42
AUTH DPO on our synt	6	41.11	5.65	10.74	2.09	2.87	0.49	0.00	0.61	0.22	8.52
NUQLab janbakker_o_b	1	38.96	6.98	12.00	0.37	0.48	0.51	0.00	0.00	0.66	8.72
AUTH SFT on open dat	6	38.96	3.29	9.00	3.80	6.27	0.35	0.00	0.77	0.11	8.67
AUTH KTO on our synt	6	38.95	6.25	3.20	1.13	4.98	0.61	0.00	0.43	0.36	8.58
BioSimplex SciBERT_F	3	37.98	8.31	9.81	0.51	0.88	0.63	0.00	0.04	0.57	8.80
BioSimplex SciBERT_F	3	37.68	14.19	11.04	0.67	1.00	0.74	0.00	0.06	0.43	8.69
BioSimplex SciBERT_F	3	37.25	14.74	11.21	0.68	0.99	0.75	0.00	0.05	0.41	8.70
TUKE FlanT5Small_sen	6	36.91	5.70	11.44	0.36	0.47	0.50	0.00	0.01	0.67	8.75
TUKE biobart-large-d	6	35.79	12.12	12.82	0.96	1.09	0.72	0.00	0.26	0.37	8.80
Tokatrons task11_BAR	2	34.30	8.76	12.84	1.78	2.01	0.58	0.00	0.53	0.13	8.55
Tokatrons task11_LLa	2	17.92	12.18	12.62	0.97	1.01	0.93	0.28	0.06	0.09	8.76

Table 6
Results for CLEF 2026 SimpleText Task 1.2 document-level text simplification: Test data on 175 Plain Language Summaries, best three 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	175	8.88	12.40	13.02	1.00	1.00	1.00	1.00	0.00	0.00	8.85
Reference	175	100	100	11.26	0.79	1.05	0.44	0.00	0.34	0.59	8.63
FOSU-CR cr	1	48.85	12.22	8.80	0.52	0.88	0.45	0.00	0.18	0.66	8.47
UvA RAG_BoN_k4_Q32L7	5	47.57	12.80	12.34	0.75	0.90	0.51	0.00	0.31	0.60	8.44
TUKE Gemma2_9B_IT_Do	1	47.18	10.42	9.91	0.52	0.85	0.42	0.00	0.22	0.73	8.52
Writerslogic top3min	22	46.91	10.20	11.13	0.54	0.78	0.49	0.00	0.17	0.67	8.57
Writerslogic doc_v5	22	46.90	10.18	11.13	0.54	0.78	0.49	0.00	0.17	0.67	8.57
Writerslogic doc_v5	22	46.89	10.17	11.13	0.54	0.78	0.49	0.00	0.17	0.67	8.57
UvA RAG_BoN_k4_Q32	5	46.67	11.07	11.00	0.58	0.82	0.50	0.00	0.19	0.64	8.52
UvA PG_PCW_Q32	5	46.50	12.51	10.37	0.71	1.13	0.50	0.00	0.27	0.60	8.56
ELiRF-UPV PromptingG	18	46.22	9.69	7.77	0.56	1.10	0.50	0.00	0.20	0.65	8.43
ELiRF-UPV PromptingG	18	46.22	9.69	7.77	0.56	1.10	0.50	0.00	0.20	0.65	8.43
DUTH PlainBestV1	61	46.16	8.93	11.75	0.86	1.01	0.51	0.00	0.43	0.57	8.28
DUTH PlainBestV2Stat	61	46.15	8.93	11.75	0.86	1.01	0.51	0.00	0.43	0.57	8.28
DUTH Gemma27_ClsGuar	61	46.13	8.92	11.77	0.86	1.01	0.51	0.00	0.43	0.57	8.28
UPF MBR-SARI-Rank-Pr	10	45.84	11.25	10.17	0.76	1.11	0.51	0.00	0.28	0.55	8.56
ELiRF-UPV PromptingG	18	45.81	9.17	7.42	0.57	1.10	0.49	0.00	0.23	0.66	8.46
UPF MBR-SARI-Proprie	10	45.67	11.33	10.23	0.78	1.12	0.51	0.00	0.29	0.54	8.60
UPF Random-Proprieta	10	45.52	10.30	9.99	0.71	1.06	0.49	0.00	0.27	0.58	8.58
NIL-UCM Agents	19	45.34	10.86	9.33	0.79	1.18	0.54	0.00	0.32	0.53	8.42
Elsevier FS3_V3_Seed	18	45.33	6.92	9.59	0.40	0.61	0.43	0.00	0.10	0.71	8.42
Elsevier FS3_V3_Stru	18	45.20	4.54	9.13	0.33	0.53	0.37	0.00	0.11	0.78	8.32
NIL-UCM Llama3Prompt	19	44.99	8.68	7.93	0.59	1.07	0.45	0.00	0.25	0.65	8.59
NIL-UCM Llama3Prompt	19	44.91	8.83	7.55	0.62	1.16	0.45	0.00	0.28	0.65	8.53
Elsevier FS3_BestMat	18	44.82	4.10	8.82	0.30	0.51	0.36	0.00	0.09	0.79	8.31
UBO PASTIS	1	44.25	6.84	9.53	0.60	0.92	0.45	0.00	0.27	0.68	8.36
FOSU-YZH AgenticRAG4	15	44.09	5.22	7.98	0.36	0.64	0.41	0.00	0.09	0.74	8.44
AIIRLab concat	14	43.80	10.53	12.34	0.95	1.21	0.56	0.00	0.36	0.48	8.59
CLNN Task1.2_P5	3	43.80	4.91	11.38	0.39	0.51	0.39	0.00	0.13	0.75	8.60
AIIRLab concat	14	43.75	10.95	12.45	0.98	1.23	0.57	0.00	0.36	0.46	8.61
DKSLA QWEN 3.6 27B	2	43.59	10.43	11.80	0.55	0.73	0.51	0.00	0.13	0.60	8.65
CLNN Task1.2_P4	3	43.56	4.97	10.62	0.40	0.54	0.39	0.00	0.14	0.75	8.44
FOSU-YZH AgenticRAG2	15	43.50	3.54	9.42	0.30	0.49	0.37	0.00	0.07	0.78	8.49
FOSU-YZH AgenticRAG	15	43.33	3.14	8.97	0.28	0.48	0.35	0.00	0.08	0.80	8.46
CLNN Task1.2_P6	3	42.72	3.39	14.22	0.35	0.38	0.37	0.00	0.09	0.77	8.78
AIIRLab LLaMABaseLin	14	42.67	5.48	11.06	0.60	0.84	0.46	0.00	0.27	0.71	8.63
Chennai Task1_2_llam	5	42.62	3.99	10.27	0.40	0.56	0.38	0.00	0.18	0.78	8.46
Chennai Task1_2_llam	5	42.61	4.02	10.36	0.40	0.56	0.38	0.00	0.18	0.78	8.50
Chennai Task1_2_Mist	5	42.59	4.05	11.78	0.37	0.49	0.40	0.00	0.11	0.75	8.61
AUTH KTO on our synt	6	39.06	3.73	9.05	0.42	0.64	0.45	0.00	0.12	0.71	8.69
NUQLab janbakker_o_b	1	38.97	6.97	12.01	0.36	0.48	0.51	0.00	0.00	0.66	8.72
AUTH SFT on our synt	6	38.87	4.80	9.43	0.51	0.74	0.49	0.01	0.17	0.67	8.69
AUTH SFT on open dat	6	38.78	8.84	9.44	0.65	0.94	0.51	0.02	0.21	0.59	8.58
HULAT-UC3M task12_Qw	3	37.73	13.94	12.39	0.65	0.80	0.67	0.00	0.07	0.46	8.72
HULAT-UC3M task12_Qw	3	36.79	0.78	13.94	0.33	0.40	0.28	0.00	0.14	0.86	8.58
DKSLA QWEN 3.5 9B	2	32.01	15.82	12.58	0.75	0.91	0.80	0.01	0.04	0.34	8.78
HULAT-UC3M task23d_Q	3	8.97	12.41	13.02	1.00	1.00	1.00	0.98	0.00	0.00	8.85
SIM Task23_Copy_Grou	2	8.88	12.40	13.02	1.00	1.00	1.00	1.00	0.00	0.00	8.85
SIM Task23_Copy	2	8.88	12.40	13.02	1.00	1.00	1.00	1.00	0.00	0.00	8.85

see relatively low BLEU scores, even considerably lower than before, due to reduced alignment between the source and the references at the sentence level in the larger set of abstracts. Third, changes in

real-world plain-language adaptations exhibit considerable variation and pose an important challenge for sentence-level approaches to replicate.

Table 6 shows the results of Task 1.2 (document-level text simplification) against a larger set of 175 abstracts and plain language summaries without further alignment. Again, we restrict the table to the best-scoring runs per team.

We make a number of observations. First, in terms of SARI, we see again encouraging performance. The tables show some swaps and upsets, but generally good agreement between Table 6 and the previously shown Table 3. Second, the BLEU score remains low throughout again, and notably lower than on the subset of Table 3. This seems to be a result of greater variation and change in discourse in the plain language summaries. Third, processing an entire abstract again yields a high compression ratio and a high deletion rate. The addition rate is high in the references, and some of the better-scoring systems also seem to have more insertions. This may indicate that some systems are finding valuable content to insert, such as explanations of jargon or other specialized terminology.

5.3. Overgeneration analysis

We analyzed the entire test dataset, comprising 8,069 documents (Task 1.2) and 48,809 sentences (Task 1.1). This analysis assumes that there is always word overlap between a pair of complex-simple sentences or abstracts. Moreover, we specifically look for overgenerating output at the end of a sentence or an abstract. We indeed observe overgeneration/text-completion issues at the end of the sources/predictions. There are also cases in which systematic errors occur during output extraction, with additional content. Increasingly, there is additional LLM commentary other than the requested output. Accurately removing such additional content can be more challenging for the document-level submissions than for the sentence-level submissions, as some abstracts are very long.

Table 7 shows an overgeneration analysis of the Task 1.1 runs. This is done by aligning the source input with the prediction output based on their token sequences. If all the source sentence(s) have been aligned to some prediction sentence(s), we assume the prediction covers all the content of the sources. If there is still an additional sentence in the prediction, we regard this as spurious content for that specific input.

At the sentence level, alignment is fairly straightforward, allowing us to detect additional content in the prediction with relatively high reliability. Additional content may be useful clarifications or explanations, but may also be gratuitous additions not supported by the source text. This is an imperfect proxy, but it serves as a good indicator of spurious content in predictions and overgeneration issues in runs. For 2026, we estimate that 17 out of 179 sentence-level submissions (9.5%) contain entire spurious sentences in at least 25% of the 48,809 input sentences. Moreover, 69 submissions (38.5%) exhibit such over-generation in at least 10% of the input sentences.

For 2026, we estimate that 88 out of 210 document-level submissions (41.9%) contain entire spurious sentences in at least 25% of the 8,069 input documents. Moreover, 37 submissions (17.6%) exhibit such over-generation in at least 50% of the input documents.

For 2026, we estimate that 17 out of 179 sentence-level submissions (9.5%) contain entire spurious sentences in at least 25% of the 48,809 input sentences. Moreover, 88 out of 210 document-level submissions (41.9%) contain entire spurious sentences in at least 25% of the 8,069 input documents.

Table 8 shows an overgeneration analysis of the Task 1.2 runs. We again align the source input with the prediction output based on their token sequences. Aligning lengthy documents can be non-trivial, and the number of detected cases quickly grows to a large fraction of the total due to the far smaller set of documents relative to the number of sentences. Detecting overgeneration may indicate helpful additions such as explanations of medical jargon, but very often indicates output that is unfounded by the sources. Although an imperfect proxy, it serves as a good indicator of spurious content in predictions and overgeneration issues in runs.

We make several observations. First, despite competitive performance in terms of text overlap with the references, we observe widely varying rates of overgeneration, ranging from a few percentage points to large fractions of the output. Second, this difference in additional content is not at all reflected

Table 7
Analysis of SimpleText Task 1.1: Spurious generation at the sentence level

Run	count	Source	Spurious Content	
		Number	Number	Fraction
ELiRF-UPV PromptingG	15	48 809	47 231	0.97
AUTH DPO on our synt	6	48 809	39 536	0.81
AUTH DPO on our synt	6	48 809	38 002	0.78
Tokatrons task11_BAR	2	48 809	37 291	0.76
TUKE Gemma2_9B_IT_se	6	48 809	31 311	0.64
AUTH DPO on our synt	6	48 809	20 714	0.42
AIIRLab GSI_gemma	18	48 809	20 117	0.41
AIIRLab Mistral8B_GS	18	48 809	18 020	0.37
AIIRLab Mistral_GSI	18	48 809	18 020	0.37
CLNN Task1.1_P1	3	48 809	15 749	0.32
FOSU-YZH DocAgent_Co	52	47 515	14 091	0.30
CLNN Task1.1_P2	3	48 809	13 893	0.28
ELiRF-UPV PromptingG	15	48 809	12 854	0.26
Monkey promptDeepsee	1	42 563	10 984	0.26
TUKE biobart-large-d	6	48 809	12 545	0.26
Chennai Task1_1_llam	3	48 569	11 887	0.24
Chennai Task1_1_Llam	3	48 809	11 898	0.24
Chennai Task1_1_Mist	3	48 809	11 898	0.24
TUKE RAG	6	48 809	10 745	0.22
NIL-UCM MistralPromp	10	3679	786	0.21
CLNN Task1.1_P3	3	48 809	10 402	0.21
HULAT-UC3M task11_Qw	3	48 809	9991	0.20
FOSU-YZH refill269_m	52	48 809	9216	0.19
FOSU-YZH selector_no	52	48 809	9172	0.19
Writerslogic task11_	18	48 809	9066	0.19
HULAT-HAI no-orig	2	48 809	9022	0.18
HULAT-HAI with-orig	2	48 809	8956	0.18
Writerslogic sonnet_	18	48 809	8710	0.18
FOSU-CR cr	1	48 809	8682	0.18
Writerslogic V4	18	48 809	8338	0.17
Tokatrons task11_LLa	2	48 809	8217	0.17
BioSimplex SciBERT_F	3	48 809	7635	0.16
BioSimplex SciBERT_F	3	48 809	7571	0.16
BioSimplex SciBERT_F	3	48 809	6948	0.14
DKSLA QWEN 3.5 9B	1	48 809	6131	0.13
SPU DeepSimplify_tas	6	20 259	1782	0.09
ELiRF-UPV PromptingG	15	48 809	3563	0.07
SPU executor	6	20 259	1470	0.07
HULAT-UC3M task11_Qw	3	48 809	2455	0.05
SPU copy_fix	6	20 259	977	0.05
NIL-UCM MistralPromp	10	3679	164	0.04
HULAT-UC3M task11_Qw	3	48 809	1913	0.04
NIL-UCM Llama3QPromp	10	3679	118	0.03
Elsevier Legacy_ZS_S	13	3679	51	0.01
Elsevier Legacy_ZS_S	13	3679	47	0.01
Elsevier FS3_V3_Seed	13	3679	37	0.01
DUTH Gemma2_9B_RefSt	12	48 809	435	0.01
SrcRef Task1_Referen	3	48 809	80	0.00
DUTH DeBERTaSelectV1	12	48 809	56	0.00
DUTH Gemma3_12B_Guar	12	48 809	27	0.00
NUQLab janbakker_o_b	1	48 809	12	0.00
SrcRef Task1_Source	3	48 809	0	0.00
SrcRef Task1_Source	3	48 809	0	0.00

in the evaluation scores, as some of the top-performing runs still exhibit larger fractions of "extra" content. Third, some of these may be easily spotted as "noise," such as systematically left-in prompts.

Table 8
Analysis of SimpleText Task 1.2: Spurious generation at the document level

Run	count	Source	Spurious Content	
		Number	Number	Fraction
ELiRF-UPV PromptingG	18	8069	7924	0.98
NIL-UCM Gemma2Prompt	19	175	166	0.95
Elsevier Legacy_FS3_	18	175	153	0.87
UBO PASTIS	1	175	152	0.87
FOSU-CR cr	1	6928	6006	0.87
Elsevier Legacy_ZS_S	18	175	151	0.86
NIL-UCM Llama3Prompt	19	175	143	0.82
NIL-UCM Llama3Prompt	19	175	143	0.82
FOSU-YZH DocAgent_Co	15	175	143	0.82
AIIRLab APIO_FewShot	14	175	138	0.79
Elsevier ZS_V3_Struc	18	175	133	0.76
AIIRLab GSI_doc_gemm	14	8069	5765	0.71
AIIRLab GSI_doc_gemm	14	7484	4580	0.61
FOSU-YZH AgenticRAG3	15	175	105	0.60
FOSU-YZH DocAgent_Co	15	175	101	0.58
TUKE Gemma2_9B_IT_Do	1	8069	4429	0.55
CLNN Task1.2_P5	3	6928	3518	0.51
CLNN Task1.2_P4	3	6928	3423	0.49
AUTH SFT on open dat	6	8069	3974	0.49
Chennai Task1_2_Mist	5	8069	3891	0.48
Chennai Task1_2_Llam	5	8069	3891	0.48
UPF MBR-SARI-Proprie	10	8069	3787	0.47
AUTH DPO on our synt	6	8069	3680	0.46
UPF Random-Proprieta	10	8069	3610	0.45
UPF MBR-SARI-Rank-Pr	10	8069	3605	0.45
AUTH DPO on our synt	6	8069	3566	0.44
SrcRef Task1_Referen	3	8069	3402	0.42
ELiRF-UPV PromptingG	18	8069	3185	0.39
UvA RAG_BoN_k4_Q32L7	5	8069	3162	0.39
CLNN Task1.2_P6	3	6928	2695	0.39
UvA RAG_BoN_k4_Q32	5	8069	2833	0.35
UvA LLM_Q32	5	8069	2770	0.34
Chennai Task1_2_llam	5	8069	2761	0.34
Writerslogic task12_	22	8069	2575	0.32
DKSLA QWEN 3.6 27B	2	8069	1926	0.24
Writerslogic gemini	22	8069	1864	0.23
Writerslogic doc_v5	22	8069	1320	0.16
ELiRF-UPV PromptingG	18	8069	992	0.12
HULAT-UC3M task12_Qw	3	8069	861	0.11
DUTH Strategy4C_Qwen	61	8069	767	0.10
DUTH Strategy6_Ensem	61	8069	755	0.09
DUTH Strategy5B_Qwen	61	8069	755	0.09
HULAT-UC3M task12_Qw	3	8069	613	0.08
DKSLA QWEN 3.5 9B	2	8069	594	0.07
HULAT-UC3M task23d_Q	3	8069	574	0.07
NUQLab janbakker_o_b	1	8069	113	0.01
SIM Task23_Copy_Grou	2	8069	0	0.00
SIM Task23_Copy	2	8069	0	0.00
SrcRef Task1_Source	3	8069	0	0.00
SrcRef Task1_Source	3	8069	0	0.00

Other cases may be challenging to detect in the output by users of text simplification systems.

6. Discussion and Conclusions

This concludes the results for the CLEF 2026 SimpleText Task 1: Text Simplification on simplify scientific text. Our main findings are as follows: First, we compared results obtained on the carefully sentence-aligned abstracts in Table 2 and Table 3 with those from the larger, unfiltered set of document-level-aligned abstracts in Table 5 and Table 6. It is reassuring to observe a strong consistency in system rankings across both settings, indicating that training and evaluation on document-aligned texts are promising and viable. This opens up opportunities to scale text simplification by substantially increasing the amount of usable data. Second, this perspective broadens the scope of text simplification beyond the traditional focus on lexical and grammatical adjustments, and brings new dimensions into play. These include handling complex discourse structures, accounting for the background knowledge required to understand the text, and reducing or explaining jargon and specialized terminology. Third, although the results are encouraging and the submitted runs generally show higher quality than several years ago, there remains considerable room for improvement, especially when models must handle scientific language and biomedical jargon. This underlines the importance and added value of the new test collections developed within the CLEF 2026 SimpleText track.

For the CLEF 2026 SimpleText Task 1, we have constructed extensive corpora and new reference data for evaluation. These reusable corpora and evaluation resources are available to participants and other researchers who want to work on the important problem of making scientific information open and easily accessible for everyone. In terms of building a community for research on scientific text summarization and simplification, the track saw record attendance in 2026, with significant changes to the tasks and a move to Codabench. More runs were submitted, and the largest number of participating teams ever.

Acknowledgments

We are incredibly thankful to the master's students in translation and technical writing from the University of Brest for participating in data annotation. We also thank each of the individual track participants for their effort in submitting a record number of submissions to CodaBench and documenting these in their papers.

We thank the CLEF 2026 chairs for hosting us, and the CLEF 2026 Labs and Proceedings chairs for their excellent assistance and flexibility. It is heartwarming to be part of such a great CLEF family. We thank CodaBench [35] for hosting the competition. Post-competition experiments are ongoing at https://www.codabench.org/competitions/16108/ (Task 1.1, Task 1.2, and Task 2.3) and https://www.codabench.org/competitions/16065/ (Task 2.1 and Task 2.2). We hope and expect that these "living test collections" remain in active use until the next iteration of the track.

Liana Ermakova is partly funded by the French National Research Agency (ANR-22-CE23-0019-01, *SimpleText: Automatic Simplification of Scientific Texts*). Liana Ermakova is further supported by the CNRS research group MaDICS (https://www.madics.fr/ateliers/simpletext/).

Jan Bakker and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is further supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] J. Bakker, J. Kamps, Section-level simplification of biomedical abstracts, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13819–13833. URL: https://aclanthology.org/2025.emnlp-main.697/. doi:10.18653/v1/2025.emnlp-main.697.

[3] 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: [36], 2026.

[4] B. Chakar, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 2: Identify and Avoid Hallucination, in: [37], 2026.

[5] G. K. Shahi, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 3: Research Area Classification, in: [37], 2026.

[6] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://aclanthology.org/Q15-1021/. doi:10.1162/tacl_a_00139.

[7] C. Jiang, M. Maddela, W. Lan, Y. Zhong, W. Xu, Neural CRF model for sentence alignment in text simplification, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7943–7960. URL: https://aclanthology.org/2020.acl-main.709/. doi:10.18653/v1/2020.acl-main.709.

[8] 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.

[9] D. Davari, L. Ermakova, R. Krestel, Comparative analysis of evaluation measures for scientific text simplification, in: A. Antonacopoulos, A. Hinze, B. Piwowarski, M. Coustaty, G. M. D. Nunzio, F. Gelati, N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries - 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24-27, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Springer, 2024, pp. 76–91. URL: https://doi.org/10.1007/978-3-031-72437-4_5. doi:10.1007/978-3-031-72437-4_5.

[10] E. Hawkins, K. Zbinden, E. Fitzgerald, B. Mansouri, AIIRLab Systems at SimpleText CLEF 2026: Ensemble and Multi-Agent Pipelines for Scientific Text Simplification, in: [37], 2026.

[11] S. Maurya, R. Ladda, G. Tupe, R. Kumar, BioSimplex at the CLEF 2026 SimpleText Task 1.1: A Two-Stage Operation-Aware Framework for Scientific Text Simplification, in: [37], 2026.

[12] C. L. N. Noutche, C. K. Kreutz, clnn88@SimpleText 2026 - Task 1: Llama 3-based Simplification for Biomedical Text, in: [37], 2026.

[13] M. Kumar, P. Balasundaram, D. Kumar, Multilingual Scientific Text Simplification Using a Locally Deployed Open-Source LLM: Plan-Driven Sentence Simplification and Summary-Guided Document Simplification with Llama 3.1 8B at the CLEF 2026 SimpleText Track — Task 1, in: [37], 2026.

[14] G. Arampatzis, A. Arampatzis, DUTH at CLEF 2026 SimpleText: Grounded Biomedical Text Simplification and Evidence-Guided Overgeneration Detection, in: [37], 2026.

[15] M. Hurtado, V. Ahuir, A. Molina, M.-J. Castro-Bleda, ELiRF-UPV at SimpleText 2026: Decoupling Action Prediction and Controlled Generation for Text Simplification, in: [37], 2026.

[16] A. Capari, H. Azarbonyad, Z. Afzal, G. Tsatsaronis, Elsevier at CLEF 2026 SimpleText: Deconstructing Scientific Text Simplification with Direct and Multi-Stage LLM Approaches, in: [37], 2026.
[17] R. Chen, Y. Yang, H. Zhang, L. Kong, FOSU_cr at the CLEF 2026 SimpleText Track: A Reference-Guided Few-Shot Prompting Approach for Biomedical Text Simplification, in: [37], 2026.
[18] Z. Yang, H. Qi, Y. Yi, FOSU-YZH-Group at the SimpleText 2026 Track: Statistics-Aware Prompting and Candidate Selection for Task 1.1 and Exploratory Overgeneration Detection for Task 2, in: [37], 2026.
[19] W. Liu, S. Qi, Y. Yi, R. Lin, FutureMakers at the CLEF 2026 SimpleText Track: Retrieval-Augmented Evidence-Focused Overgeneration Detection for Task 2.2, in: [37], 2026.
[20] N. Pirvu, R. Cardon, HULAT-HAI @ SimpleText 2026 Task 1: Prompting Based on Linguistic Analysis for Text Simplification, in: [37], 2026.
[21] P. A. Núñez, P. Martínez, L. Moreno, HULAT-UC3M at CLEF 2026 SimpleText: Controlled Generation for Scientific Text Simplification with Hallucination Control, in: [37], 2026.
[22] F. Chen, L. Kong, Y. Lu, MONKEY at SimpleText 2026: Static Few-Shot Prompting for Cochrane Review Scientific Text Simplification, in: [37], 2026.
[23] S. Conde, P. Folgueira, A. Díaz, NIL-UCM at the CLEF 2026 SimpleText Track: Prompting, Fine-Tuning, and Multi-Agent Architectures for Biomedical Text Simplification, in: [37], 2026.
[24] S. Ibrahim, W. Zaghouani, NUQLab at SimpleText 2026: Plan-Guided Biomedical Scientific Text Simplification, in: [37], 2026.
[25] R. Rabih, R. Sonbol, W. Alsohle, SPU_DeepSimplify at the SimpleText 2026 Track: A Planning-Based Approach to Scientific Text Simplification, in: [37], 2026.
[26] J. Dauenhauer, N. Hofmann, C. K. Kreutz, THM@SimpleText 2026 - Task 2: Ensemble-based Hallucination Classification in Text Simplification, in: [37], 2026.
[27] 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: [36], 2026.
[28] D. Shevchuk, N. Tymochko, O. Petsa, TUKE at CLEF 2026 SimpleText: Biomedical Text Simplification with Seq2Seq Models, BioBART, and Gemma QLoRA, in: [37], 2026.
[29] S. M, S. K. S, V. K. James, P. Balasundaram, SSN Tokatrons at the CLEF 2026 SimpleText Track: Plan-Guided BART and Zero-Shot LLM Approaches to Biomedical Text Simplification, in: [37], 2026.
[30] B. Vendeville, L. Ermakova, P. De Loor, UBONLP Report on the SimpleText Track at CLEF 2026, in: [37], 2026.
[31] A. Hayakawa, H. Saggion, UPF-TALN at the CLEF 2026 SimpleText Track: Ensemble Approach for Scientific Text Simplification and Hallucination Classification, in: [37], 2026.
[32] P. Mathas, B. Chakar, D. Carranza Navarrete, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2026 SimpleText Track, in: [37], 2026.
[33] P. Mathas, J. Bakker, J. Kamps, Plan-guided text simplification with extended contexts, in: M. Shardlow, T. François, R. Amaro, J. Baptista, R. Cardon, E. Ribeiro, H. Saggion, R. Stodden, A. Todirascu, R. Wilkens (Eds.), Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, ELRA and ICCL, Palma de Mallorca, Spain, 2026, pp. 121–129.
[34] D. Condrey, Writerslogic at the CLEF 2026 SimpleText Track: Multi-Candidate LLM Simplification and Stacked Complexity Spotting, in: [37], 2026.
[35] Z. Xu, S. Escalera, A. Pavão, M. Richard, W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022) 100543. URL: https://doi.org/10.1016/j.patter.2022.100543. doi:10.1016/J.PATTER.2022.100543.
[36] 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.
[37] 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.