Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text

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

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

Abstract

This paper presents an overview of the CLEF 2025 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 advanced the field of text simplification by creating new biomedical corpora that support true paragraph- and document-level simplification, capturing greater variation and complex discourse suited to LLMs. Second, our CLEF 2025 document-level corpus showed for the first time that document-level simplification models clearly outperformed sentence-level methods. Third, in addressing the biomedical domain, teams developed novel approaches to generate plain language summaries that overcome key barriers for consumers, enhancing accessibility to authoritative health information. 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 science-literate is more important than ever before. Objective scientific information helps any user navigate a world where misinformation, disinformation, or generated and unfounded information is only a single mouse click away. Everyone acknowledges the importance of objective scientific information, but the general public seldom consults scientific sources. The value of objective scientific information cannot be overstated. Biomedical research can directly impact people's decisions about health. However, the most reliable and up-to-date sources in biomedicine contain complex language and assume a high degree of background knowledge, making them difficult for the general public to understand.

To address these challenges, the CLEF 2025 Simple Track has three aims. First, we push the research frontier in text simplification by further expanding the scientific text simplification corpora, focusing on true paragraph-level and document-level simplification with greater variation, and considering the complex discourse structure. This setup fits current models, such as LLMs, that operate on a long input. This new biomedical corpus is constructed from aligned Cochrane abstracts and plain language summaries [1]. Second, we exploit the text simplification setup with aligned sources, references, and the output of generative models to detect, quantify, and avoid spurious information introduced gratuitously by the generative model. This is what is informally referred to as "hallucinations," addressing the remaining limitations of large generative models is crucial for the scientific use case, as current evaluation measures are "blind" and don't punish the unwarranted generation of additional content. This task addresses one of the main challenges in the Track, CLEF, and the fields of NLP and IR in general. Third, by popular demand, we will revisit and rerun some earlier tasks to ensure that the transition to the new track setup will retain the active track participants of earlier years.

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

CLEF 2025 Working Notes, 9–12 September 2025, Madrid, Spain
j.bakker@uva.nl (J. Bakker); benjamin.vendeville@univ-brest.fr (B. Vendeville); liana.ermakova@univ-brest.fr (L. Ermakova); kamps@uva.nl (J. Kamps)
0009-0002-9085-8491 (J. Bakker); 0009-0003-5298-147X (B. Vendeville); 0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)

Table 1
CLEF 2025 Simpletext official run submission statistics

Team	Task 1		Task 2			Task 1	Task 2	Total runs
	1.1	1.2	2.1	2.2	2.3			
AIIRLab [4]	4	2	5	5	0	6	10	16
ASM [5]	0	10	0	0	0	10	0	10
DSGT [6, 7]	2	1	6	6	3	3	15	18
DUTH [8]	3	0	2	2	0	3	4	7
EngKh (no paper)	2	0	0	0	0	2	0	2
Fujitsu [9]	19	0	0	0	0	19	0	19
LIA [10]	0	9	0	0	0	9	0	9
Mtest (no paper)	1	1	1	1	0	2	2	4
PICT [11]	1	1	0	0	0	2	0	2
RECAIDS [12]	1	1	1	1	0	2	2	4
Scalar [13]	10	1	0	0	1	11	1	12
SINAI [14]	2	2	15	15	0	4	30	34
THM [15]	22	0	0	0	0	22	0	22
UBO [16]	5	7	1	1	0	12	2	14
UM-FHS [17]	4	5	0	0	0	9	0	9
UvA [18]	5	9	0	0	0	14	0	14
Unknown (no paper)	2	0	0	0	0	2	0	2
Total	83	49	31	31	4	132	66	198

- ***Task 1: Text Simplification*** *simplify scientific text.*
- ***Task 2: Controlled Creativity*** *identify and avoid hallucination.*
- ***Task 3: SimpleText 2024 Revisited*** *selected tasks by popular request.*

This paper gives an overview of the CLEF 2025 SimpleText Task 1 on Text Simplification, which aims to simplify scientific text. Further detail on the entire track is in the CLEF 2025 SimpleText Track Overview [2]. Additional details on Task 2 on *Controlled Creativity* are in a companion Task 2 overview paper [3]. We also refer to the respective participants' papers for further details.

A total of 74 teams registered for our SimpleText track at CLEF 2025. A total of 18 teams submitted 198 runs in total for Tasks 1 and 2. The statistics for these runs submitted are presented in Table 1.[1] However, some runs had problems that we could not resolve. We do not detail them in the rest of the paper and leave out the 0-scoring runs. More details about individual runs and experiments can be found in the participants' papers, also shown in Table 1.

The rest of this paper is structured in the following way. 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.

2.1. Description

The *Text Simplification* task aims to *simplify scientific text*. We created a new CLEF 2025 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 was created by closely following

[1]The table includes submissions in the Tasks 1 and 2 Codabench evaluation platform, where we were privileged to have 29 (Task 1) and 13 (Task 2) participants.

Complex paragraph

Fifteen heterogeneous trials, involving 1022 adults with dorsally displaced and potentially or evidently unstable distal radial fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation especially in terms of patient characteristics and interventions. Methodological weaknesses among these trials included lack of allocation concealment and inadequate outcome assessment.

Simple paragraph

Fifteen trials, involving 1022 adults with potentially or evidently unstable fractures, were included. While all trials compared external fixation versus plaster cast immobilisation, there was considerable variation in their characteristics especially in terms of patient characteristics and the method of external fixation.

Figure 1: A complex-simple paragraph pair from Cochrane-auto (reproduced from [1])

Table 2
Statistics for the automatically aligned Cochrane-auto, Newsela-auto, and Wiki-auto datasets, versus the manual simplifications of the track in 2022-2024.

	Cochrane-auto	**Newsela-auto**	**Wiki-auto**	**SimpleText**
Domain	Biomedical	News	General	Science
# Document Pairs	5,585	18,820	138,095	278
# Sentence Pairs	35,800	813,972	685,769	1,536

the construction methods of the existing Wiki-auto and Newsela-auto datasets. It introduces a new domain of scientific text. The Cochrane-auto corpus is publicly available data, which is highly relevant and interesting to general and specialized users. It contains authentic parallel data produced by the same author. Cochrane-auto represents true document-level text simplification, incorporating greater variation, such as sentence merging and order rearrangements, while considering the discourse structure. Paragraph-level and sentence-level data are carefully realigned and restricted to matching paragraphs and sentences.

Table 2 compares the new scientific text simplification corpus to the main existing text simplification corpora. The aligned references also free our translation students and professionals from the task of manually simplifying texts. This allows them to concentrate on analyzing and annotating samples of the evaluation data for various types of information distortion, providing ground truth for Task 2.

Task Description This is the core NLP task of the track, and we continue with both sentence-level (*Task 1.1*) and document-level (*Task 1.2*) scientific text simplification. The main innovation is the very large new corpus we constructed in 2024, and the shift to the biomedical domain.

2.2. Data

As discussed above, we constructed a large scientific text simplification corpus, based on realigning abstracts and lay summaries at scale at the sentence, paragraph, and document levels. In 2025, we will use this 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 of the sentence input is included, allowing also for paragraph level text simplification submission to Task 1.2.

Test data The primary test data consists of 217 new Cochrane abstracts with paired plain English summaries, composed of 4,293 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 different ways.

- We process these as Cochrane-auto [1] to ensure a high-quality sentence and paragraph alignment. This results in a subset of 37 abstracts and 587 sentences, paired with 37 plain language summaries with 388 sentences. The processing is identical to Cochrane-auto and other text simplification data sets.
- 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 [19]. This results in 217 abstracts with 4,293 source sentences, paired with 217 plain language summaries with 3,641 sentences.

We use the aligned subset as the main evaluation and also report the scores over the whole subset.

Analysis data For further analysis, we extended the test data with the Cochrane-auto validation and test splits (part of the train data), Medline abstracts for which TREC PLABA references exist,[2] and SimpleText 2024 abstracts for which we have references. The combined test file, including additional data sources, contains 666 documents with 9,160 sentences.

2.3. Formats

This section outlines the format of the data used in the CLEF 2025 SimpleText Task 1.

2.3.1. Train data

The training data used was from the Cochrane-auto paper [1] as published on GitHub: https://github.com/JanB100/cochrane-auto/tree/main/data. This data format follows the earlier Wiki-auto and Newsela-auto corpora and includes additional fields for statistics. It also includes the specific transformation (copy, rephrase, split, merge, delete) for each sentence in the source abstract.

The training data already has train, validation, and test splits with references. We also included the validation and test splits of the training data [1] in the combined source data test set, so we also collected the predictions of each system on the training data.

2.3.2. Sources

Sentence-level simplification The source data is in a JSON format with the following fields:

1. *pair_id*: Unique ID for the complex-simple document pair.
2. *para_id*: Index of the source paragraph in the document.
3. *sent_id*: Index of the source sentence in the document.
4. *complex*: Complex sentence from the source document.

This format is a simplified version of the Cochrane-auto format in JSON. The combination of the pair_id (the id of the document or abstract) and the sent_id (the sentence order) determines the unique sentence unambiguously. As the abstracts can be quite lengthy, we retained the paragraph identifier. This third identifier enables paragraph-level text simplification approaches (submitted to Task 1.2).

An example of the Task 1.1 JSON source input is:


```
[
  {
    "pair_id": "CD012520",
    "para_id": 0,
    "sent_id": 0,
    "complex": "We included seven cluster-randomised trials with 42,489 patient participants from 129
    ↪  hospitals, conducted in Australia, the UK, China, and the Netherlands."
  },
  {
    "pair_id": "CD012520",
```


[2] https://bionlp.nlm.nih.gov/plaba2024/

```
    "para_id": 0,
    "sent_id": 1,
    "complex": "Health professional participants (numbers not specified) included nursing, medical
    ↪  and allied health professionals."
  },
  . . .
  {
    "pair_id": "CD012520",
    "para_id": 2,
    "sent_id": 12,
    "complex": "We are uncertain whether a multifaceted implementation intervention compared to no
    ↪  intervention improves adherence to evidence-based recommendations in acute stroke settings,
    ↪  because the certainty of evidence is very low."
  }, . . .
]
```

Document-level simplification The source data is in a JSON format with the following fields:

1. *pair_id*: Unique ID for the complex-simple document pair.
2. *source*: The origin of the data (only for reference).
3. *complex*: The complex document's content (all sentences).

This is again a simplified version of the Cochrane-auto format in JSON. The new Cochrane-auto data is the main evaluation of the task in 2025. However, the test input files also included the validation and train splits of the training data, as well as other scientific abstracts from PubMed and the earlier CLEF 2024 SimpleText corpus. The source field indicates the source of the abstracts. This enables comparative performance analysis against other existing text simplification corpora, in addition to the track's official evaluation scores on the new set of Cochrane-auto abstracts.

An example of the Task 1.2 JSON source input is:

```
[
  {
    "pair_id": "CD012520",
    "source": "Cochrane",
    "complex": "We included seven cluster-randomised trials with 42,489 patient participants from 129
    ↪  hospitals, conducted in Australia, the UK, China, and the Netherlands. Health professional
    ↪  participants (numbers not specified) included nursing, medical and allied health
    ↪  professionals.
    ...
    We are uncertain whether a multifaceted implementation intervention compared to no intervention
    ↪  improves adherence to evidence-based recommendations in acute stroke settings, because the
    ↪  certainty of evidence is very low."
  }, . . .
]
```

In this case, there are 13 sentences (Task 1.1) corresponding to the Cochrane document (Task 1.2).

2.3.3. Predictions

As output, participants were asked to add a column to the input data with their simplified sentence or abstracts (prediction) as well as a unique identifier for their run. So, in addition to the fields in the *Source* format above, submissions included the following fields:

- *prediction*: Your simplification of the complex sentence with pair_id and sent_id (for sentence-level simplification), or of the complex document with pair_id (for document-level simplification).
- *run_id*: <team-id>_<task-id>_<method-used>.

Sentence-level simplification An example of a sentence-level text simplification (Task 1.1) submission is the following:


```
[{"pair_id":"CD012520","para_id":0,"sent_id":0,"complex":"We included seven cluster-randomised
↪  trials with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK,
↪  China, and the Netherlands.","prediction":"We looked at seven studies that were done in groups.
↪  These studies involved 42,489 patients from 129 hospitals in Australia, the UK, China, and the
↪  Netherlands.","run_id":"UBOnlp_task11_gpt4o"},
{"pair_id":"CD012520","para_id":0,"sent_id":1,"complex":"Health professional participants (numbers
↪  not specified) included nursing, medical and allied health professionals.","prediction":"The
↪  group of health professionals included nurses, doctors, and other healthcare workers. The exact
↪  number of people in each group wasn't mentioned.","run_id":"UBOnlp_task11_gpt4o"},
. . .
{"pair_id":"CD012520","para_id":2,"sent_id":12,"complex":"We are uncertain whether a multifaceted
↪  implementation intervention compared to no intervention improves adherence to evidence-based
↪  recommendations in acute stroke settings, because the certainty of evidence is very
↪  low.","prediction":"We're not sure if using a special plan with many parts helps doctors follow
↪  the best advice for treating strokes, compared to not using any plan. This is because we don't
↪  have strong proof yet.","run_id":"UBOnlp_task11_gpt4o"},
. . . ]
```


Document-level simplification An example of a document-level text simplification (Task 1.2) submission is the following:


```
[{"pair_id":"CD012520","source":"Cochrane","complex":"We included seven cluster-randomised trials
↪  with 42,489 patient participants from 129 hospitals, conducted in Australia, the UK, China, and
↪  the Netherlands. Health professional participants (numbers not specified) included nursing,
↪  medical and allied health professionals. ... We are uncertain whether a multifaceted
↪  implementation intervention compared to no intervention improves adherence to evidence-based
↪  recommendations in acute stroke settings, because the certainty of evidence is very
↪  low.","prediction":"Researchers conducted studies in hospitals across Australia, the UK, China,
↪  and the Netherlands. They wanted to see if certain strategies could help healthcare workers
↪  provide better care. These strategies were aimed at doctors, nurses, and other health
↪  professionals. ... Overall, the evidence was not strong enough to say for sure if these
↪  strategies help healthcare workers follow best practices in treating stroke
↪  patients.","run_id":"UBOnlp_task12_gpt4o"},
. . . ]
```


2.3.4. References

The reference simplifications are based on pairs of Cochrane abstracts and plain language summaries. One possibility for document-level text simplification is to use the original pairs of abstracts and plain language summaries, using only the *results and conclusions* sections, similar to [19]. This results in a representative and realistic set of references and a large set of abstracts, but has the disadvantage that there may be many differences between the abstract and plain language summary. This version is included as *simple_original*.

We can also process these as Cochrane-auto [1] to ensure a high-quality sentence and paragraph alignment. This results in a smaller set of abstracts, but clear correspondence between the content in the pair of abstract and plain language summary. This is the main evaluation reference and included as *simple_auto*.

The reference data is also in JSON:


```
{
  "CD012520": {
    "simple_original": "Implementation interventions are designed to improve the delivery of
    ↪  'evidence-based' care, which is care that has been proven in research studies to help people
    ↪  with a particular health condition. ... What did we find? We included seven studies that
    ↪  involved 42,489 acute stroke patients and an unknown number of health professionals. The
    ↪  studies were conducted in 129 hospitals in Australia, the UK, China and the Netherlands. ...
    ↪  How up to date is this evidence? This review includes papers that we identified from
    ↪  searching in April 2022.",
```

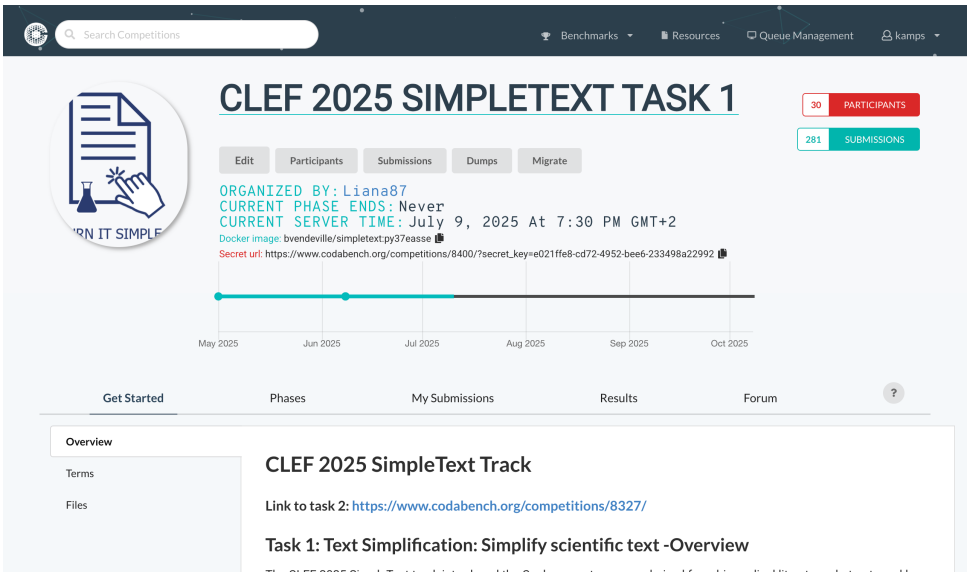

Figure 2: CLEF 2025 SimpleText Task 1 Codabench.

```
    "simple_auto": "We included seven studies that involved 42,489 acute stroke patients and an
    ↪  unknown number of health professionals. The studies were conducted in 129 hospitals in
    ↪  Australia, the UK, China and the Netherlands. ...  We do not know if implementation
    ↪  interventions delivered in acute stroke units lead to better delivery of evidence-based care."
},
```

To have identical ground truth and directly comparable evaluation scores over both sentence and document level submissions, in particular in the single leaderboard table in the Task's Codabench, we decided to evaluate both types of submissions at the document level. That is, a sentence-level submission is merged to a complete, simplified abstract, and evaluated against the plain language summary.

2.4. Codabench

Submissions were made through Codabench.[3] Due to the differences in the setup, each task had a designated separate competition on Codabench. The Task 1 runs were submitted at: https://www.codabench.org/competitions/8400/ (shown in Figure 2). The Task 2 runs were submitted at: https://www.codabench.org/competitions/8327/. The Codabench greatly facilitated running the track in 2025 and provided active participants (who had also registered at the Codabench) with full access to the competition, including the submission and leaderboard pages.

2.5. Evaluation

In 2025, 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 [20]. This automatic evaluation will be supplemented with a detailed human evaluation of other aspects, essential for deeper analysis.

Almost all participants used generative models for text simplification, yet existing evaluation measures are blind to potential hallucinations with extra or distorted content. In 2025, we will continue to provide further analysis of ways to detect and quantify spurious content in the output, potentially corresponding to what is informally called "hallucinations."

3. Participant's Approaches

A total of 18 teams submitted 132 runs in total. In the detailed results, we only include runs without

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

errors, which got a non-zero score.

AIIRLab Largey et al. [4] submitted six runs in total for Task 1. They submitted four runs for Task 1.1 and two runs for Task 1.2. They use a range of open-source models (Mistral, LLaMA), with extensive finetuning and exploration of effective prompts, for sentence- and document-level text simplification. Special precautions against noise and unwanted output were taken. The prompt instructions were directly focused on the desired outcome evaluation measures of the task.

ASM Djoudi et al. [5] submitted 10 runs in total for Task 1. They submitted no runs for Task 1.1 and 10 runs for Task 1.2. They created an extensive set of over 3,000 simplified medical definitions compiled from multiple public sources. A Mistral 7B model was used to detect jargon in the abstracts, and matching simplified definitions were added to the prompt. Three open-source models (Mistral 7B, Gemma 2-9B, Med42-v2) were used for text simplification, obtaining competitive performance.

DSGT Marturi and Elwazzan [6] submitted three runs in total for Task 1. They submitted two runs for Task 1.1 and one run for Task 1.2. The paper uses an open-source LLaMA 3.3 70b model with a few-shot prompt approach. For sentence-level simplification, they first prompt a plan to guide the simplification and, in a second stage, prompt the model to execute the plan at the sentence level. For document-level simplification, they first apply a summarization prompt and, in a second stage, prompt the model to simplify the summary. Related Task 2.3 experiments are in a separate paper [7].

DUTH Arampatzis and Arampatzis [8] submitted three runs in total for Task 1. They submitted three runs for Task 1.1 and none for Task 1.2. They use open-source models, such as FLAN-T5 and BART-SAMSum, with a zero-shot prompt for sentence-level and document-level text simplification.

EngKh (no paper) submitted two runs in total for Task 1. They submitted two runs for Task 1.1 and none for Task 1.2.

Fujitsu Agüero-Torales et al. [9] submitted 19 runs in total for Task 1. They submitted 19 runs for Task 1.1 and none for Task 1.2. They explore an in-context learning approach, with zero and three-shot prompting of GPT-3.5, o4-mini, and T5-small models, in an optimized pipeline for sentence-level scientific text classification.

LIA Gallina et al. [10] submitted 9 runs in total for Task 1. They submitted no runs for Task 1.1 and 9 runs for Task 1.2. The paper does interesting experiments with a range of open-source models (LLaMA-4, LLaMA-3.3, Mistral-Small, Gemma2, Helsinki). They use both generic prompts and specific guidance based on the Cochrane plain language summaries instructions. The specific instructions help the performance of their models.

Mtest (no paper) submitted two runs in total for Task 1. They submitted one run for Task 1.1 and 1 run for Task 1.2.

PICT Vora et al. [11] submitted two runs in total for Task 1. They submitted one run for Task 1.1 and one run for Task 1.2. They explore an advanced pipeline to create an abstract meaning representation of the text, focusing on lexical and phrase-level simplification, sentence-level structural simplification, and a final T5 model for generative text simplification.

RECAIDS (no paper) submitted two runs in total for Task 1. They submitted one run for Task 1.1 and one run for Task 1.2. They explore a T5 model for Tasks 1.1 and 1.2 with a straightforward T5 completion prompt, and with a model fine-tuned on each task.

Scalar Dongre et al. [13] submitted 11 runs in total for Task 1. They submitted ten runs for Task 1.1 and one run for Task 1.2. They perform an interesting experiment for Task 1.1, motivated by avoiding biomedical jargon or technical terminology. They deploy earlier generation models (BioBERT/BioBART, GPT-2), which are considerably more efficient than current LLMs and demonstrate reasonable performance.

SINAI Collado-Montañez et al. [14] submitted four runs in total for Task 1. They submitted two runs for Task 1.1 and two runs for Task 1.2. They uses a closed-source model, GPT-4.1, in a zero-shot prompt setting. They use tailored biomedical test simplification prompts for Tasks 1.1 and 1.2, and the model shows high performance.

THM Hofmann et al. [15] submitted 22 runs in total for Task 1. In fact, they submitted 22 runs for Task 1.1 and none for Task 1.2. They devote special interest to biomedical jargon or technical terminology. The main experiment uses five different prompts with advanced closed-source models (OpenAI and Gemini).

UBO Vendeville et al. [16] submitted 12 runs in total for Task 1. They submitted five runs for Task 1.1 and seven runs for Task 1.2. The submissions were mostly test submissions, which were not discussed in detail in the paper.

UM-FHS Kocbek and Stiglic [17] submitted 9 runs in total for Task 1. They submitted four runs for Task 1.1 and five runs for Task 1.2. They utilize closed-source models (GPT-4.1 standard, mini, and nano) in a zero-shot prompt setup with detailed prompts for Tasks 1.1 and 1.2. They also explore the value of fine-tuning the smaller models.

UvA Papandreou et al. [18] submitted 14 runs in total for Task 1. They submitted five runs for Task 1.1 and nine runs for Task 1.2. They submitted Cochrane-auto [1] trained BART models. These were either operating at the sentence level, including a plan-guided version, for Task 1.1, or the paragraph or document level for Task 1.2. They also experimented with jargon detection trained on MedReadMe [21] to create a jargon-aware prompt for a LLaMA 3.1-8b model, for both tasks.

Unknown team (no paper) submitted two runs in total for Task 1. They submitted two runs for Task 1.1 and none 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 37 abstracts, with 587 sentences aligned identically to the way Cochrane-auto and other collections are aligned. In this track overview paper, we decided to evaluate all submissions in Task 1.1 and Task 1.2 at the document level to ensure identical ground truth and comparable scores across tasks.

Table 3 shows the Task 1.1 (sentence-level text simplification) results. The table is restricted to submissions without issues, and we show a maximum of five runs per team. We show several evaluation scores against the human reference simplifications, particularly SARI and BLEU. In addition, we provide additional text statistics on the system output, such as FKGL, and compare them to the source input.

We make a number of observations. First, the table is sorted on SARI, the primary automatic text simplification measure used in the track. We observe SARI scores above 30% for almost all systems and above 40% for the top-scoring systems. This high overlap with the plain language reference simplifications is encouraging, and it indicates that the effectiveness of text simplification approaches, traditionally trained on youth news reading corpora like Newsela, also extends to scientific text.

Second, in terms of the level of text complexity, readability measures like FKGL provide a rough indicator of lexical and grammatical complexity. The original sentences have an FKGL of 13-14 corresponding to university-level text, and most systems reduce this to an FKGL of 11-12 corresponding to the exit level of compulsory education. This is an encouraging result, as it indicates that the scientific text simplification approach can be a viable approach to lower the textual complexity of scientific text toward the range acceptable by a layperson. Although this indicator is positive, this approximate measure does not consider terminological complexities.

Third, the table includes various other scores that indicate that there is still considerable room for improvement in scientific text simplification. Throughout the table, the BLEU evaluation measure remains very low. It leads to a different ranking of systems, with some of the best systems on BLEU demonstrating superior overlap with the human reference simplifications. The table also reveals some runs with very high "compression" ratios, sentence splits, and high proportions of additions. While evaluation measures like SARI are essential for understanding important aspects of text simplification

Table 3
Results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 37 aligned Cochrane-auto abstracts, best five runs per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100	100	11.73	0.56	0.67	0.50	0.0	0.16	0.60	8.71
UM-FHS gpt-4.1-mini	37	43.34	13.93	7.46	0.78	1.58	0.63	0.00	0.28	0.50	8.50
UM-FHS gpt-4.1-mini-	37	42.83	20.85	12.29	0.71	0.86	0.62	0.00	0.15	0.46	8.67
DSGT plan_guided_lla	37	42.33	10.43	7.77	0.48	0.97	0.47	0.00	0.18	0.70	8.52
UvA o-bartsent-cochr	37	42.31	25.72	12.08	0.41	0.51	0.55	0.00	0.01	0.62	8.72
SINAI PRMZSTASK11V1	37	41.82	6.50	11.41	1.37	1.56	0.53	0.00	0.59	0.30	8.33
THM p2–gpt-4.1-nano	37	41.32	10.49	14.90	1.27	1.16	0.63	0.00	0.45	0.26	8.62
UvA bartsent-cochran	37	41.28	17.67	11.20	0.35	0.49	0.48	0.00	0.01	0.67	8.76
Scalar gpt_md_2_1	37	40.95	14.07	18.79	0.62	0.47	0.53	0.00	0.22	0.60	8.68
UBOnlp gpt4o	37	40.74	7.53	7.39	0.46	0.80	0.41	0.00	0.23	0.73	8.31
THM p1–gpt-4.1-nano	37	40.42	11.02	14.66	1.23	1.13	0.65	0.00	0.42	0.24	8.61
PICT S3Pipeline	37	40.15	12.96	7.61	0.71	1.53	0.62	0.00	0.21	0.49	8.84
Fujitsu llm_t5_rule	37	39.04	6.70	6.79	0.31	0.71	0.42	0.00	0.08	0.76	8.85
UM-FHS gpt-4.1	37	38.84	14.04	8.51	0.79	1.26	0.68	0.30	0.22	0.41	8.49
UvA llama31	37	38.76	2.83	8.30	0.93	1.58	0.46	0.00	0.60	0.66	8.34
DUTH Task11_flan-t5-	37	38.73	18.84	11.95	0.61	0.78	0.66	0.00	0.10	0.50	8.96
Fujitsu t5efficient	37	38.60	4.28	5.58	1.79	3.63	0.43	0.00	0.77	0.29	10.31
Fujitsu llm_gpt3.5-t	37	38.53	6.30	5.18	0.36	0.99	0.45	0.00	0.11	0.74	8.89
Fujitsu llm_45_judge	37	38.41	5.45	5.26	0.32	0.89	0.42	0.00	0.09	0.77	8.87
Fujitsu dummy60	37	38.37	14.50	1.19	0.37	2.74	0.52	0.00	0.08	0.67	8.74
SINAI PRMZSTASK11V2	37	37.84	5.93	12.97	1.64	1.63	0.56	0.00	0.59	0.17	8.47
THM pni1–gpt-4.1-na	37	37.60	8.24	15.21	1.84	1.63	0.56	0.00	0.57	0.12	8.61
UvA bartdoc-ca	37	37.25	19.54	11.97	0.51	0.61	0.62	0.00	0.02	0.52	8.77
EngKh biomedical_lla	37	36.68	11.47	10.62	1.14	1.51	0.65	0.00	0.37	0.28	8.69
UvA llama31	37	36.45	1.22	13.04	1.07	1.31	0.41	0.00	0.66	0.70	8.61
AIIRLab mistral	37	36.08	18.41	12.78	0.94	1.06	0.76	0.00	0.19	0.28	8.81
MTest bartfinetuned	37	34.98	26.52	11.94	0.74	0.98	0.83	0.00	0.01	0.30	8.78
THM pn1–gemini-2.0-	37	34.47	9.67	7.75	1.25	1.90	0.67	0.00	0.45	0.20	8.62
Scalar BioBart_1	37	33.95	25.69	12.19	0.78	1.00	0.86	0.00	0.01	0.27	8.80
Scalar BioBart	37	33.95	25.69	12.19	0.78	1.00	0.86	0.00	0.01	0.27	8.80
THM c–gpt-4.1-nano	37	33.94	5.81	21.56	1.49	0.99	0.63	0.00	0.44	0.22	9.22
DUTH Task11_bart-sam	37	32.18	12.28	7.69	1.43	2.75	0.68	0.00	0.41	0.14	8.71
RECAIDS T5	37	31.68	0.09	3.72	0.37	0.96	0.31	0.00	0.23	0.88	8.87
AIIRLab llama3.1-8b	37	31.27	19.59	11.44	0.85	1.09	0.83	0.00	0.09	0.25	8.83
UM-FHS gpt-4.1-nano	37	29.47	18.46	11.10	0.86	1.14	0.83	0.43	0.11	0.24	8.71
DUTH Task11_bart-lar	37	27.59	12.01	8.67	1.69	2.90	0.66	0.00	0.46	0.09	8.61
DUTH Task11_flan-t5-	37	22.75	21.95	13.15	0.91	0.95	0.94	0.00	0.01	0.11	8.89
DUTH Task11_gpt4	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89

output quality, they are also known to be relatively insensitive to content outside the intersection of manual text simplifications. Hence, high levels of content insertion can still lead to favorable SARI scores and even improve text statistics like FKGL without conveying key content of the original text.

4.2. Task 1.2: Document-level Scientific Text Simplification

Table 4 shows the results of Task 1.2 (document-level text simplification). Again, we restrict the table to submissions covering a maximum of five runs with non-zero scores per team.

We make a number of observations. First, in terms of evaluation measures like SARI, we see similar encouraging performance levels again when evaluating against the plain language reference

Table 4
Results for CLEF 2025 SimpleText Task 1.2 document-level text simplification: Test data on 37 aligned Cochrane-auto abstracts, best five runs per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	37	12.03	20.53	13.54	1.00	1.00	1.00	1.00	0.00	0.00	8.89
Reference	37	100	100	11.73	0.56	0.67	0.50	0.0	0.16	0.60	8.71
LIA sumguid-all-w500	37	44.55	12.18	9.71	0.84	1.26	0.50	0.00	0.35	0.54	8.56
SINAI PRMZSTASK12V1	37	43.93	10.81	10.45	0.86	1.07	0.55	0.00	0.39	0.49	8.33
UM-FHS gpt-4.1	37	43.83	18.12	8.80	0.67	1.10	0.58	0.14	0.21	0.53	8.44
UM-FHS gpt-4.1-nano-	37	43.61	16.00	10.63	0.50	0.69	0.45	0.00	0.16	0.65	8.55
LIA sumguid-lang-w50	37	43.61	10.55	10.50	0.83	1.18	0.47	0.00	0.37	0.57	8.52
UM-FHS gpt-4.1-mini	37	43.53	14.11	7.48	0.72	1.49	0.62	0.00	0.25	0.52	8.52
ASM MistralMaxFRE	37	43.35	12.32	11.63	0.73	0.92	0.53	0.00	0.27	0.56	8.74
ASM MistralV0	37	43.31	12.41	11.65	0.73	0.92	0.53	0.00	0.27	0.55	8.74
ASM MistralMinFKGL	37	43.24	12.27	11.63	0.73	0.93	0.53	0.00	0.27	0.56	8.75
ASM MistralV7	37	42.95	11.34	12.53	0.78	0.94	0.51	0.00	0.30	0.55	8.80
ASM MistralV7CleanLi	37	42.93	11.38	13.77	0.78	0.84	0.51	0.00	0.29	0.56	8.80
UM-FHS gpt-4.1-mini-	37	42.82	22.94	11.93	0.60	0.76	0.60	0.03	0.10	0.52	8.73
AIIRLab Mistral_7b_b	37	42.40	12.98	8.82	0.58	0.94	0.52	0.00	0.21	0.61	8.48
UvA baseline-cochran	37	42.10	24.27	11.71	0.57	0.71	0.61	0.00	0.06	0.49	8.74
LIA sumguid-styl-w50	37	41.98	10.38	10.09	0.63	1.00	0.46	0.00	0.27	0.66	8.65
UBOnlp gpt4o	37	41.56	5.45	7.22	1.14	2.08	0.50	0.00	0.58	0.43	8.25
LIA sumguid-styl-w50	37	41.11	8.73	6.35	0.61	1.30	0.42	0.00	0.33	0.68	8.44
AIIRLab llama_3.1-8b	37	41.07	8.61	9.22	0.46	0.70	0.43	0.00	0.20	0.72	8.44
LIA testLlama33	37	40.79	8.42	10.74	0.46	0.65	0.42	0.00	0.18	0.73	8.64
DSGT llama_summary_s	37	40.32	7.63	9.56	0.59	0.86	0.42	0.00	0.31	0.70	8.49
PICT S3Pipeline	37	40.29	13.43	7.77	0.74	1.55	0.63	0.00	0.21	0.47	8.77
AIIRLab llama-8b	37	39.14	5.62	8.88	0.34	0.62	0.35	0.00	0.15	0.80	8.43
AIIRLab llama3.2-3b	37	39.14	5.62	8.88	0.34	0.62	0.35	0.00	0.15	0.80	8.43
DUTH task12_led-larg	37	39.11	9.83	12.41	0.37	0.47	0.45	0.00	0.06	0.70	8.80
SINAI PRMZSTASK12V2	37	38.50	10.30	11.55	1.09	1.16	0.63	0.00	0.43	0.29	8.44
UvA bartpara-cochran	37	37.89	27.43	12.22	0.62	0.77	0.74	0.00	0.01	0.41	8.78
Mtest bartdoc	37	37.62	20.42	11.79	0.50	0.61	0.62	0.00	0.01	0.51	8.76
UvA bartdoc-ca	37	37.25	19.54	11.97	0.51	0.61	0.62	0.00	0.02	0.52	8.77
UvA bartdoc-cochrane	37	37.25	19.54	11.97	0.51	0.61	0.62	0.00	0.02	0.52	8.77
UM-FHS gpt-4.1-nano	37	37.01	14.74	9.05	0.69	1.13	0.64	0.19	0.16	0.46	8.57
UvA llama31	37	36.98	3.99	7.61	0.79	1.59	0.39	0.00	0.46	0.77	8.48
DUTH task12_flan-t5-	37	36.65	3.75	12.08	0.24	0.27	0.33	0.03	0.00	0.77	8.76
DUTH task12_bart-sam	37	36.25	1.38	10.32	0.17	0.28	0.27	0.00	0.01	0.85	8.74
DUTH task12_flan-t5-	37	34.73	0.67	12.76	0.14	0.15	0.22	0.00	0.01	0.87	8.81
Scalar gpt_md_2_1	37	34.39	1.01	10.56	0.14	0.19	0.20	0.00	0.03	0.88	8.67
EngKh biomedical_lla	37	33.25	17.88	12.55	0.72	0.87	0.61	0.05	0.15	0.44	8.77
DUTH task12_flan-t5-	37	32.55	0.36	12.83	0.12	0.13	0.18	0.00	0.01	0.89	9.10
RECAIDS T5	37	31.49	0.00	10.08	0.06	0.07	0.10	0.00	0.00	0.95	8.12

simplifications. In earlier years of the track, this mainly resulted from using proven sentence-level text simplification models with the output merged back into the entire abstract. However, this year, we see almost exclusively large language models applied to the lengthy source abstract as a whole. This is a clear sign of the remarkable progress in models for text simplification and other complex NLP tasks. Second, there remains room for improvement in capturing the human simplifications more closely, as the BLEU score remains low throughout. Here, the more conservative approaches seem to obtain better scores. For scientific text simplification, we aim for a careful balance between simplicity and accuracy, and being conservative is a key strength to avoid unnecessary and potentially inaccurate changes. Third, we see less extreme values on the other indicators, but still considerable variation in

the compression ratio and number of splits, and proportions of additions and deletions. Generally, we see more compression and deletions, indicating summarization aspects such as reducing the number of sentences, which happens frequently.

It is encouraging to see solid performance for the approaches that perform text simplification on the entire abstract in one pass. This holds the promise to incorporate the discourse structure, use more complex text simplification operations such as deletions and merges, and deploy planner-based approaches to the text simplification of long documents. Traditional sentence-level simplification approaches and earlier evaluation data cannot capture these aspects. This demonstrates the value of the new test collections constructed during the CLEF 2025 SimpleText track.

4.3. Results on Plain Language Summaries

In this section, we provide additional evaluation on the larger set of 217 abstracts with 4,293 source sentences paired with 217 plain language summaries with 3,641 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.

Table 5 shows the results of Task 1.1 (sentence-level text simplification) against a larger set of 217 abstracts and plain language summaries without further alignment. Again, we restrict the table to submissions covering a maximum of five runs with non-zero scores per team. Note again that all submissions in Task 1.1 and Task 1.2 at the document level, to ensure identical ground truth and comparable scores across tasks.

We make a number of observations. First, in terms of evaluation measures like SARI, we see again similar encouraging performance levels when evaluating against the larger set of plain language reference simplifications. The ranking in Table 5 is similar to the subset of Table 3 before, with some notable shifts and upsets, particularly for run with a low BLEU score, but overall high agreement. Second, we see relatively low BLEU scores again, and even considerably lower than before. This is partly a result of the less clear source to reference alignment at the sentence and paragraph level for this larger set of references. But it also shows that document-level text simplification is a challenging task, even for current advanced models. Third, this also indicates that real-world plain language summaries are far removed from direct sentence-level simplifications. It also suggests that more conservative approaches, which may be desirable from an accuracy point of view, fail to capture the complex plain language adaptations.

Table 6 shows the results of Task 1.2 (document-level text simplification) against a larger set of 217 abstracts and plain language summaries without further alignment. Again, we restrict the table to submissions covering a maximum of five runs with non-zero scores per team.

We make a number of observations. First, in terms of evaluation measures like SARI, we see similar encouraging performance levels again when evaluating against the plain language reference simplifications. The tables show some swaps and upset, but generally good agreement between Table 6 and Table 4 shown before. One exception seems to be closed-source models, such as GPT-4, which perform less impressively on the larger set of plain language summaries. Second, the BLEU score remains low throughout again, and notably lower than on the subset of Table 4. This seems to be a result of the greater variation and discourse changes in the plain language summaries. However, this also immediately suggests that this is not yet captured well by the predictions of advanced NLP models for text simplification. Third, we see less extreme values on the other indicators for document-level text simplification approaches. The fraction of deletions remains very high throughout all systems. Interestingly, the better-scoring systems also seem to have more insertions. This can be an indication that some systems are finding valuable content to insert, such as explanations of jargon or other specialized terminology.

Table 5
Results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
DSGT plan_guided_lla	217	42.98	6.33	7.82	0.48	0.99	0.46	0.00	0.18	0.71	8.50
UBOnlp gpt4o	217	42.20	4.05	7.49	0.38	0.68	0.37	0.00	0.18	0.78	8.37
UM-FHS gpt-4.1-mini	217	42.13	9.52	7.56	0.74	1.52	0.61	0.00	0.26	0.53	8.54
SINAI PRMZSTASK11V1	217	41.25	4.59	12.39	1.44	1.56	0.51	0.00	0.61	0.30	8.44
UvA llama31	217	40.92	2.62	8.63	1.00	1.64	0.45	0.00	0.62	0.64	8.35
THM p2–gpt-4.1-nano	217	39.57	6.50	15.40	1.32	1.20	0.60	0.00	0.47	0.27	8.68
UM-FHS gpt-4.1-mini-	217	39.16	11.95	12.23	0.67	0.82	0.60	0.00	0.14	0.50	8.76
PICT S3Pipeline	217	39.11	8.30	6.52	0.69	1.65	0.60	0.00	0.21	0.52	8.85
Scalar gpt_md_2_1	217	38.96	8.25	19.45	0.62	0.43	0.52	0.00	0.23	0.60	8.77
Fujitsu llm_gpt3.5-t	217	38.84	3.05	5.04	0.35	1.02	0.44	0.00	0.11	0.75	8.96
UvA bartsent-cochran	217	38.71	6.01	11.34	0.31	0.46	0.45	0.00	0.00	0.72	8.81
Fujitsu llm_t5_rule	217	38.55	2.75	6.60	0.31	0.77	0.42	0.00	0.08	0.77	8.95
Fujitsu llm_45_judge	217	38.54	2.34	5.19	0.31	0.93	0.41	0.00	0.09	0.78	8.95
UvA o-bartsent-cochr	217	38.53	8.57	11.99	0.37	0.49	0.51	0.00	0.01	0.67	8.78
UvA llama31	217	38.50	1.13	13.66	1.09	1.23	0.40	0.00	0.66	0.71	8.65
Fujitsu llm_45	217	38.49	2.06	5.32	0.31	1.00	0.40	0.00	0.09	0.79	8.90
THM p1–gpt-4.1-nano	217	38.24	6.59	15.03	1.28	1.18	0.63	0.00	0.45	0.25	8.69
Fujitsu llm_45fewSho	217	38.20	1.87	3.51	0.28	0.88	0.37	0.00	0.12	0.81	8.82
UM-FHS gpt-4.1	217	37.93	9.46	8.82	0.76	1.22	0.64	0.23	0.22	0.46	8.54
UvA bartdoc-ca	217	37.14	7.23	11.43	0.39	0.49	0.52	0.00	0.01	0.63	8.85
SINAI PRMZSTASK11V2	217	35.95	4.03	14.00	1.76	1.64	0.54	0.00	0.61	0.15	8.56
DUTH Task11_flan-t5-	217	35.35	10.07	11.21	0.60	0.80	0.65	0.00	0.09	0.51	9.00
THM pni1–gpt-4.1-na	217	35.26	5.23	15.49	1.94	1.72	0.54	0.00	0.59	0.12	8.68
AIIRLab mistral	217	33.95	10.30	13.26	0.93	1.04	0.72	0.00	0.21	0.32	8.86
RECAIDS T5	217	33.89	0.03	3.72	0.37	0.98	0.31	0.00	0.23	0.89	8.87
EngKh biomedical_lla	217	33.16	7.30	10.76	1.18	1.53	0.65	0.00	0.37	0.25	8.75
THM c–gpt-4.1-nano	217	32.44	3.76	21.37	1.51	1.02	0.62	0.00	0.43	0.20	9.26
THM pn1–gemini-2.0-	217	32.27	5.80	7.92	1.28	1.94	0.66	0.00	0.46	0.20	8.68
MTest bartfinetuned	217	31.59	14.86	11.90	0.69	0.96	0.80	0.00	0.01	0.36	8.89
Scalar BioBart	217	30.35	14.26	12.04	0.74	0.99	0.83	0.00	0.01	0.32	8.88
Scalar BioBart_1	217	30.35	14.26	12.04	0.74	0.99	0.83	0.00	0.01	0.32	8.88
AIIRLab llama3.1-8b	217	29.80	11.32	11.19	0.83	1.10	0.80	0.00	0.10	0.29	8.93
DUTH Task11_bart-sam	217	29.68	7.32	7.50	1.38	2.73	0.67	0.00	0.41	0.16	8.79
UM-FHS gpt-4.1-nano	217	28.89	10.35	9.90	0.83	1.19	0.78	0.35	0.13	0.30	8.77
DUTH Task11_bart-lar	217	23.84	7.07	8.59	1.64	2.87	0.66	0.00	0.45	0.10	8.71
DUTH Task11_flan-t5-	217	18.78	11.48	12.89	0.89	0.94	0.93	0.00	0.01	0.13	9.03
DUTH Task11_gpt4	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
XXX method	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05

4.4. Findings

This concludes the results for the CLEF 2025 SimpleText Task 1: Text Simplification on simplify scientific text. Our main findings are the following: First, our analysis compared the results over the carefully sentence-aligned abstracts in Table 3 and Table 4, with the larger unfiltered set of document-level aligned abstracts in Table 5 and Table 6. It is encouraging to see the broad agreement in the ranking over both sets, as this suggests evaluation and training on document-aligned texts is a viable option. Similar to how machine translation was able to scale up due to the availability of parallel texts, this can help scale up text simplification by increasing the number of available corpora. Second, this also shifts the focus of the field of text simplification beyond the traditional aspects of lexical and grammatical simplification

Table 6
Results for CLEF 2025 SimpleText Task 1.2 document-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	217	7.84	10.55	13.29	1.00	1.00	1.00	1.00	0.00	0.00	9.05
Reference	217	100	100	11.28	0.72	0.97	0.40	0.00	0.29	0.63	8.65
LIA sumguid-all-w500	217	44.93	9.58	9.77	0.69	1.06	0.48	0.00	0.29	0.62	8.61
LIA sumguid-lang-w50	217	44.40	7.85	10.58	0.67	0.97	0.44	0.00	0.30	0.66	8.56
SINAI PRMZSTASK12V1	217	43.63	8.07	10.73	0.81	1.03	0.52	0.00	0.37	0.54	8.41
LIA sumguid-styl-w50	217	43.57	6.18	10.28	0.51	0.81	0.41	0.00	0.20	0.72	8.67
ASM MistralMinFKGL	217	43.51	8.26	11.85	0.63	0.82	0.48	0.00	0.22	0.62	8.78
ASM MistralV0	217	43.51	8.32	11.95	0.63	0.81	0.48	0.00	0.22	0.62	8.78
ASM MistralMaxFRE	217	43.50	8.27	11.87	0.63	0.82	0.48	0.00	0.22	0.62	8.78
LIA sumguid-styl-w50	217	43.17	5.92	6.87	0.49	1.03	0.39	0.00	0.25	0.75	8.50
UBOnlp gpt4o	217	43.37	4.55	7.55	1.20	2.16	0.48	0.00	0.60	0.43	8.31
ASM MistralV7	217	43.10	7.64	12.68	0.66	0.82	0.48	0.00	0.23	0.62	8.86
ASM MistralV7CleanLi	217	43.09	7.60	13.73	0.66	0.74	0.47	0.00	0.23	0.62	8.87
DSGT llama_summary_s	217	42.92	5.32	9.94	0.49	0.72	0.39	0.00	0.24	0.75	8.55
AIIRLab Mistral_7b_b	217	42.57	7.47	9.26	0.50	0.82	0.48	0.00	0.16	0.66	8.56
AIIRLab llama_3.1-8b	217	42.46	4.73	9.94	0.39	0.58	0.39	0.00	0.15	0.76	8.54
LIA testLlama33	217	42.35	4.70	11.19	0.39	0.54	0.39	0.00	0.14	0.76	8.73
UM-FHS gpt-4.1-mini	217	42.13	9.80	7.65	0.69	1.44	0.60	0.00	0.23	0.55	8.57
UvA baseline-cochran	217	41.83	10.85	11.10	0.44	0.60	0.49	0.00	0.06	0.63	8.75
UM-FHS gpt-4.1-nano-	217	41.01	7.15	10.64	0.48	0.66	0.41	0.00	0.15	0.69	8.58
UM-FHS gpt-4.1-mini-	217	40.81	10.67	11.69	0.55	0.72	0.55	0.01	0.10	0.58	8.74
AIIRLab llama3.2-3b	217	39.77	2.17	8.70	0.28	0.52	0.30	0.00	0.11	0.84	8.55
AIIRLab llama-8b	217	39.77	2.17	8.70	0.28	0.52	0.30	0.00	0.11	0.84	8.55
DUTH task12_led-larg	217	39.28	3.58	12.46	0.31	0.41	0.40	0.00	0.05	0.76	8.86
PICT S3Pipeline	217	39.11	8.23	6.46	0.71	1.69	0.60	0.00	0.22	0.50	8.76
UM-FHS gpt-4.1	217	38.88	10.00	8.97	0.67	1.07	0.59	0.18	0.20	0.52	8.53
UvA llama31	217	38.52	2.37	7.68	0.73	1.73	0.37	0.00	0.43	0.81	8.56
UM-FHS gpt-4.1-nano	217	37.60	10.07	8.56	0.65	1.11	0.61	0.12	0.16	0.51	8.62
SINAI PRMZSTASK12V2	217	37.34	7.21	11.85	1.05	1.11	0.63	0.00	0.40	0.31	8.55
UvA bartdoc-ca	217	37.14	7.23	11.43	0.39	0.49	0.52	0.00	0.01	0.63	8.85
UvA bartdoc-cochrane	217	37.14	7.23	11.43	0.39	0.49	0.52	0.00	0.01	0.63	8.85
Mtest bartdoc	217	37.08	7.25	11.50	0.39	0.50	0.52	0.00	0.01	0.63	8.86
DUTH task12_bart-sam	217	37.00	0.13	10.02	0.12	0.22	0.21	0.00	0.00	0.89	8.88
DUTH task12_flan-t5-	217	36.62	0.29	12.20	0.16	0.18	0.24	0.00	0.00	0.85	8.95
DUTH task12_flan-t5-	217	35.81	0.12	13.09	0.12	0.13	0.19	0.00	0.00	0.89	9.02
UvA bartpara-cochran	217	34.97	12.70	12.13	0.55	0.70	0.68	0.00	0.01	0.49	8.86
DUTH task12_flan-t5-	217	34.61	0.29	11.72	0.14	0.17	0.20	0.00	0.02	0.89	9.01
Scalar gpt_md_2_1	217	34.61	0.02	9.26	0.09	0.13	0.13	0.00	0.02	0.93	8.81
RECAIDS T5	217	33.14	0.00	8.79	0.04	0.06	0.07	0.00	0.00	0.96	8.24
EngKh biomedical_lla	217	28.19	8.55	11.95	0.69	0.79	0.57	0.04	0.12	0.46	9.00

and introduces new and interesting aspects. Examples include dealing with the discourse structure, particular background knowledge needed to understand the text, and avoiding or explaining jargon or specialized terminology. Third, while the results are encouraging and the submitted predictions are generally high quality compared to some years ago, there remains also clear room for improvement, in particular when dealing with the scientific vernacular and specific biomedical jargon. This demonstrates the value of the new test collections constructed during the CLEF 2025 SimpleText track.

Table 7
Results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on the 363 sentence pairs of the 37 aligned Cochrane-auto abstracts, best five runs per team

Team/Method	count	SARI	BLEU	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	363	15.01	27.71	13.46	1.00	1.00	1.00	1.00	0.00	0.00	8.62
Reference	363	100.00	100.00	11.71	1.05	1.11	0.60	0.03	0.33	0.42	8.43
UM-FHS gpt-4.1-mini-	363	42.65	19.83	12.03	0.85	0.91	0.62	0.17	0.19	0.41	8.74
AIIRLab llama3.1_gro	363	42.32	13.09	10.90	0.75	0.99	0.66	0.02	0.20	0.47	8.50
AIIRLab llama3.1_cro	363	42.05	13.10	10.97	0.75	0.99	0.66	0.02	0.19	0.46	8.50
UM-FHS gpt-4.1-mini	363	42.00	14.78	7.20	0.87	1.69	0.68	0.04	0.33	0.44	8.30
THM p2–gpt-4.1-nano	363	41.43	12.25	14.58	1.37	1.19	0.65	0.01	0.47	0.26	8.40
THM p1–gpt-4.1-nano	363	41.16	13.29	14.72	1.35	1.15	0.67	0.01	0.45	0.25	8.39
THM p1–gpt-4.1-nano	360	40.59	12.27	14.81	1.36	1.14	0.67	0.01	0.46	0.24	8.39
THM pni1–gpt-4.1-na	360	40.59	12.27	14.81	1.36	1.14	0.67	0.01	0.46	0.24	8.39
SINAI PRMZSTASK11V1	363	39.89	6.79	11.15	1.49	1.63	0.54	0.00	0.63	0.32	8.19
UvA o-bartsent-cochr	363	39.84	17.92	11.64	0.60	0.70	0.61	0.31	0.02	0.43	9.18
PICT S3Pipeline	363	39.21	12.47	8.05	0.76	1.52	0.62	0.01	0.24	0.48	8.46
UvA bartsent-cochran	363	38.98	10.72	10.85	0.51	0.63	0.53	0.26	0.02	0.51	9.34
UBO gpt4o	363	38.58	5.44	7.17	1.22	2.16	0.51	0.00	0.65	0.48	8.09
DSGT plan_guided_lla	363	38.56	5.23	7.65	0.59	1.00	0.51	0.00	0.29	0.68	8.26
EngKh biomedical_lla	363	38.25	14.03	10.19	0.98	1.45	0.67	0.07	0.32	0.39	8.35
MTest bartfinetuned	363	38.01	27.40	11.51	0.82	1.00	0.87	0.40	0.01	0.20	8.53
Fujitsu dummy90	363	37.64	15.08	3.35	0.65	2.58	0.75	0.20	0.07	0.38	8.51
Fujitsu dummy60	363	37.61	8.15	1.12	0.49	2.84	0.61	0.06	0.10	0.54	8.38
SINAI PRMZSTASK11V2	363	37.53	6.57	12.82	1.71	1.67	0.56	0.00	0.63	0.21	8.24
UM-FHS gpt-4.1	363	37.38	14.40	8.21	0.87	1.38	0.69	0.25	0.30	0.41	8.26
AIIRLab mistral	363	37.14	22.50	13.03	1.07	1.07	0.77	0.42	0.20	0.24	8.60
Scalar gpt_md_2_1	363	37.12	10.51	7.60	0.80	1.30	0.45	0.01	0.27	0.63	8.55
Fujitsu dummy50	363	36.95	5.67	0.13	0.43	2.92	0.54	0.02	0.11	0.61	8.35
Fujitsu t5efficient	363	36.62	4.53	4.45	2.34	4.49	0.43	0.00	0.75	0.33	9.73
DUTH Task11_flan-t5-	363	36.58	15.17	10.08	0.71	0.99	0.61	0.12	0.18	0.48	8.60
Fujitsu dummy45	363	36.44	4.49	0.00	0.39	2.94	0.51	0.01	0.11	0.64	8.34
Scalar BioBart	363	36.30	28.57	11.86	0.85	1.01	0.89	0.43	0.02	0.18	8.52
Scalar BioBart_1	363	36.30	28.57	11.86	0.85	1.01	0.89	0.43	0.02	0.18	8.52
THM p1–gpt-4.1-nano	360	35.37	14.27	15.85	1.42	1.05	0.73	0.03	0.39	0.16	8.44
UvA jargons_part1	363	34.61	2.05	9.17	12.05	14.07	0.19	0.00	0.90	0.02	8.75
UvA llama31	363	34.47	2.48	7.94	0.97	1.62	0.46	0.00	0.67	0.72	8.16
DUTH Task11_bart-sam	363	34.46	15.37	7.36	1.92	2.96	0.64	0.01	0.45	0.11	8.53
UvA llama31	363	31.95	0.84	10.75	1.35	1.61	0.36	0.00	0.67	0.75	8.40
AIIRLab llama3.1-8b	363	31.59	24.19	10.85	0.89	1.16	0.84	0.44	0.10	0.22	8.58
UM-FHS gpt-4.1-nano	363	31.47	24.21	10.77	0.91	1.14	0.85	0.60	0.12	0.20	8.48
RECAIDS T5	363	30.21	0.05	3.72	0.50	0.99	0.31	0.00	0.37	0.88	8.87
DUTH Task11_bart-lar	363	29.41	14.96	8.26	2.35	3.14	0.63	0.01	0.48	0.06	8.44
DUTH Task11_flan-t5-	363	25.86	29.65	12.55	0.93	1.00	0.95	0.66	0.01	0.08	8.59
DUTH Task11_flan-t5-	363	23.21	27.42	12.34	0.93	1.00	0.93	0.60	0.03	0.10	8.59
XXX method	363	15.01	27.71	13.46	1.00	1.00	1.00	1.00	0.00	0.00	8.62

5. Analysis

This section details further analysis of the submissions to the track. We focus in particular on a sentence-level evaluation of the Task 1.1 submissions.

5.1. Task 1.1: Sentence-level Scientific Text Simplification

As detailed above, we made particular efforts in Cochrane-auto [1] to ensure alignment at the document, paragraph, and sentence level. Hence, we have Cochrane-auto aligned references for 37 abstracts, with a total of 388 source sentences, carefully aligned with 363 sentences in the plain language summaries due to deletions.

Table 7 shows the evaluation of the Task 1.1 submissions against the aligned sentence-level references. We make several observations. First, we see solid agreement between the sentence-level evaluation in Table 7 and the earlier document-level evaluation of the same runs in Table 3. This is not surprising since both use the same ground truth references, but it still adds to the confidence in the evaluation setup using aligned data at scale. Second, we see the same divergence between reasonable SARI scores and relatively low BLEU scores that seem to favor more conservative approaches. Many of the high BLEU scoring approaches also have a relatively high fraction of exact copies. This is interesting, as a conservative approach feels appropriate for the scientific text simplification use case. Conservative approaches may promote accuracy, faithfulness, and correctness of the simplifications. This may be desirable approach even if the readability and accessibility are not as high. Third, at the fine-grained sentence level over a large set of sentences, the text statistics give a more detailed and informative picture of the text simplification approaches. At the sentence level, we see varying and even high fractions of copies for more conservative methods. We see quite a significant variation on almost every indicator, highlighting great differences between each approach, despite the relatively similar SARI scores.

6. Discussion and Conclusions

This paper describes the setup of the CLEF 2025 SimpleText track, which contains the following three tasks. Task 1 on *Text Simplification*: *simplify scientific text.* Task 2 on *Controlled Creativity*: *identify and avoid hallucination.* Task 3 on *SimpleText 2024 Revisited*: *selected tasks by popular request.* These tasks address some of today's main NLP/IR challenges. This Task overview focuses on the CLEF 2025 SimpleText Track's Task 1 on sentence-level and document-level text simplification. The main aim of our track, and the CLEF evaluation forum as a whole, is i) to construct corpora and evaluation resources to stimulate research on scientific text summarization and simplification, and ii) to foster a community of IR, NLP, and AI researchers working together on the important task of making science more accessible for everyone.

Within the CLEF 2025 SimpleText Task 1, we have constructed extensive corpora and new references for evaluation data. First, we pushed the research frontier in text simplification by creating new scientific text simplification corpora for biomedical literature. We focused on true paragraph-level and document-level simplification with greater variation and took the complex discourse structure into account. This fits current models such as LLMs, which operate on long input. Second, the document-level text simplification corpus created at CLEF 2025 is a major advance of the field, as earlier data was typically based on direct human simplifications at the sentence level. As a result, sentence-level text simplification approaches were very effective, and typically outcompeted true document-level approaches, while our models have been able to cope with long context for long. In 2025, we saw for the first time, that document-level text simplification approaches clearly outcompeted sentence-level text simplification. Third, the move to the biomedical domain presents many important challenges to current technology. The abstracts can be quite long, with a complex discourse structure, and the plain language summaries avoid or explain medical jargon and provide additional background information to make the key points of the scientific abstracts understandable for consumers. Several teams experimented with specific approaches and models for the health and biomedical domain. In particular, several teams explored novel ways to guide the model into producing output that address barriers consumers face when directly accessing the biomedical literature. These advances hold the promise to automatically provide plain language summaries of biomedical literature, and thereby greatly enhance the scope and impact of authoritative health information.

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 researching scientific text summarization and simplification, the track saw a record attendance in 2025, with significant changes in the tasks and the move to Codabench, more runs were submitted, and with 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 2025 chairs for hosting us, and the CLEF 2025 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 [22] for hosting the competition. Post-competition experiments are ongoing at https://www.codabench.org/competitions/8400/ (Task 1.1, Task 1.2, and Task 2.3) and https://www.codabench.org/competitions/8327/ (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.

Benjamin Vendeville and Liana Ermakova are partly funded by the French National Research Agency (ANR-22-CE23-0019-01, *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 (NWO CI # CISC.CC.016), 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.

Disclosure of Interests

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] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText track: Simplify scientific texts (and nothing more), in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[3] B. Vendeville, J. Bakker, H. Azarbonyad, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 2: Identify and Avoid Hallucination, in: [23], 2025.
[4] N. Largey, D. Wu, B. Mansouri, AIIRLab Systems for CLEF 2025 SimpleText: Cross-Encoders to Avoid Spurious Generation, in: [23], 2025.
[5] A. N. Djoudi, S. Nouali, M. Aabid, I. Badache, A.-G. Chifu, P. Bellot, LIS at the SimpleText 2025: Enhancing Scientific Text Accessibility with LLMs and Retrieval-Augmented Generation, in: [23], 2025.
[6] K. C. Marturi, H. H. Elwazzan, Hallucination Detection and Mitigation in Scientific Text Simplification using Ensemble Approaches: DS@GT at CLEF 2025 SimpleText, in: [23], 2025.
[7] K. C. Marturi, H. H. Elwazzan, LLM-Guided Planning and Summary-Based Scientific Text Simplification: DS@GT at CLEF 2025 SimpleText, in: [23], 2025.
[8] G. Arampatzis, A. Arampatzis, DUTH at CLEF 2025 SimpleText Track: Tackling Scientific Text Simplification and Hallucination Detection, in: [23], 2025.
[9] M. M. Agüero-Torales, C. Rodríguez-Abellán, C. A. C. Moraga, Sentence-level Scientific Text Simplification With Just a Pinch of Data, in: [23], 2025.
[10] Y. Gallina, T. Jiménez, S. Huet, University of Avignon at SimpleText 2025: Guided Medical Abstract Simplification, in: [23], 2025.
[11] A. Vora, T. Chaudhari, S. Hotha, S. Sonawane, S-3 Pipeline by PICT/Pune for Biomedical Text Simplification, in: [23], 2025.
[12] S. Eugin, A. Ms.Beula, V. Sathvikha, V. Sangamithra, SimpleText: Simplify Scientific Text, in: [23], 2025.
[13] A. A. Dongre, A. Vaadiraaju, A. K. Madasamy, NITK SCaLAR Lab at the CLEF 2025 SimpleText Track: Transformer-Based Models for Biomedical Sentence Simplification (Task 1.1), in: [23], 2025.
[14] J. Collado-Montañez, J. A. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, SINAI in SimpleText CLEF 2025: Simplifying Biomedical Scientific Texts and Identifying Hallucinations Using GPT-4.1 and Pattern Detection, in: [23], 2025.
[15] N. Hofmann, J. Dauenhauer, N. O. Dietzler, I. D. Idahor, C. K. Kreutz, THM@SimpleText 2025 Task 1.1: Revisiting Text Simplification based on Complex Terms for Non-Experts, in: [23], 2025.
[16] B. Vendeville, L. Ermakova, P. D. Loor, J. Kamps, UBONLP Report on the SimpleText lab, in: [23], 2025.
[17] P. Kocbek, G. Stiglic, UM-FHS at the CLEF 2025 SimpleText Track: Comparing No-Context and Fine-Tune Approaches for GPT-4.1 Models in Sentence and Document-Level Text Simplification, in: [23], 2025.
[18] T. Papandreou, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2025 SimpleText Track, in: [23], 2025.
[19] 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.
[20] 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. Di 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, volume 15177 of *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.
[21] C. Jiang, W. Xu, MedReadMe: A systematic study for fine-grained sentence readability in medical domain, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17293–17319. URL: https://aclanthology.org/2024.emnlp-main.958/.

doi:10.18653/v1/2024.emnlp-main.958.
[22] 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.
[23] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.