# Overview of the CLEF 2023 SimpleText Task 3

*Simplification of Scientific Texts*

Ermakova, L.; Bertin, S.; McCombie, H.; Kamps, J.

**Publication date**
2023

**Document Version**
Final published version

**Published in**
Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)

# Overview of the CLEF 2023 SimpleText Task 3: Simplification of Scientific Texts

Liana Ermakova[1], Sarah Bertin[1], Helen McCombie[2] and Jaap Kamps[3]

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

Abstract

This article provides a comprehensive summary of the CLEF 2023 SimpleText Task 3, which focuses on simplifying scientific text based on specific queries. The paper begins by explaining the motivation behind the task and providing an overview of the overall setup. It then proceeds to describe the test collection in detail, which includes a training set of sentences extracted from scientific abstracts along with corresponding simplified sentences created by human annotators. Additionally, a comprehensive test corpus of sentences is introduced, accompanied by meticulous annotations of lexical and syntactic complexity. The article concludes with an in-depth analysis, including information distortion and LLM hallucinations, of the simplified sentences submitted by participants and the resulting evaluation scores.

**Keywords**

automatic text simplification, science popularization, information distortion, error analysis, lexical complexity, syntactic complexity, LLMs hallucination

## 1. Introduction

The advent of digitization and open access has facilitated the accessibility of scientific literature to the general public. While this represents a significant milestone, there are still numerous obstacles hindering non-experts from obtaining unbiased scientific information from these texts. Specifically, scientific literature can be challenging to comprehend due to its reliance on specialized knowledge and the use of complex terminology.

Despite recent attempts at text simplification (e.g. [1]) to address this issue, the automatic removal of comprehension barriers between scientific texts and the general public remains an ongoing challenge. The paper highlights that even the most advanced language models currently available face difficulties when it comes to simplifying scientific texts. The described results demonstrate the limitations of these models in effectively tackling the task of simplification in the scientific domain.

The CLEF 2023 SimpleText track brings together researchers and practitioners working on the generation of simplified summaries of scientific texts. It is an evaluation lab that follows up on the CLEF 2021 SimpleText Workshop [2] and CLEF 2022 SimpleText Track [3]. The track

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*
liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)

provides data and benchmarks for discussion of the challenges of automatic text simplification by bringing together the following interconnected tasks:

**Task 1: What is in (or out)?** Selecting passages to include in a simplified summary.

**Task 2: What is unclear?** Difficult concept identification and explanation (definitions, abbreviation deciphering, context, applications,..).

**Task 3: Rewrite this!** Given a query, simplify passages from scientific abstracts.

This paper focuses on the third task of text simplification proper. For a comprehensive understanding of the other tasks, the overview papers of Task 1 [4] and Task 2 [5], as well as the Track overview paper [6], provide detailed information and insights.

This introduction is followed by Section 2 presenting the text simplification task with the datasets and evaluation metrics used. Section 3 gives an overview of text simplification approaches for scientific text as deployed by the participants. In Section 4, we present and discuss the results of the official submissions. In Section 5, a thorough analysis of the results is carried out, covering several important aspects. This includes examining the relationship between difficult scientific terms and the simplification process, investigating information distortion that may occur during simplification, and exploring instances of language models (LLMs) generating hallucinations and producing inaccurate information. The analysis delves into these topics to provide a comprehensive understanding of the findings and insights derived from the study. We end with Section 6 summarizes the findings and draws perspective for future work.

## 2. CLEF 2023 SimpleText Task 3 Test Collection

The main objective of this task is to generate simplified versions of sentences taken from scientific abstracts. To assess the performance of participants, a comprehensive evaluation was conducted using various automatic measures such as SARI, ROUGE, compression, and readability. Additionally, a smaller-scale, but more detailed, human evaluation was carried out to assess other aspects of the simplifications, including the presence of information distortion.

### 2.1. Data

Similar to the previous year, a parallel corpus comprising 648 manually simplified sentences was provided as training data for the participants [3].

To ensure a certain level of overlap in the partial runs submitted by participants, three test sets were provided: *small*, *medium*, and *large*. The expectation was that participants would opt to use LLMs, which could result in the generation of partial runs due to the efficiency constraints associated with these models. By offering different test sets, we aimed to account for the varying computational limitations and facilitate the participation of a wide range of approaches, including those leveraging LLMs.

The *small* dataset was included in the *medium* one, while the latter is included in the *large* one. By evaluating the systems on the *small* test sets, it helped ensure that there would be some common ground among the partial runs generated by different participants. This approach

facilitated the comparison and analysis of the system outputs, even though they were based on different test sets.

The *small* dataset included the train data, which facilitates the comparison of system performance on both the training and testing data. Including the train data in the *small* dataset enables the evaluation of how well the systems generalize to unseen test data by examining their performance on familiar training examples. This comparison provides valuable insights into the effectiveness and robustness of the systems across different datasets.

In this year's evaluation, the submitted runs were assessed by comparing them to a new set of 245 manually simplified sentences from the *small* dataset extracted from relevant passages for Task 1.

### 2.1.1. Input format

The train and the test data are provided in JSON and TSV formats with the following fields:

**snt_id** a unique passage (sentence) identifier

**doc_id** a unique source document identifier

**query_id** a query ID

**query_text** difficult terms should be extracted from sentences with regard to this query

**source_snt** passage text

Input example:

```
{"snt_id":"G11.1_2892036907_2",
 "source_snt":"With the ever increasing number of unmanned aerial vehicles getting
↪  involved in activities in the civilian and commercial domain, there is an
↪  increased need for autonomy in these systems too.",
 "doc_id":2892036907,
"query_id":"G11.1",
"query_text":"drones"}
```

### 2.1.2. Output format

Results should be provided in a TREC-style JSON or TSV format with the following fields:

**run_id** Run ID starting with (team_id)_(task_3)_(method_used), e.g. UBO_BLOOM

**manual** Whether the run is manual $\{0, 1\}$.

**snt_id** a unique passage (sentence) identifier from the input file.

**simplified_snt** simplified passage .

Output example (JSON format):

```
{"run_id":"BTU_run1",
"manual":1,
"snt_id":"G11.1_2892036907_2",
"simplified_snt":"Drones are increasingly used in the civilian and commercial domain
↪  and need to be autonomous."}
```

### 2.2. Evaluation metrics

To evaluate the simplification results, we used the EASSE implementation [7] of the following metrics:

- **FKGL**: The Flesch-Kincaid Grade Level [8] is a readability metric that provides an estimate of the education level required to understand the text. FKGL is based on two factors: average sentence length and average number of syllables per word. The resulting grade level indicates the U.S. school grade equivalent required to comprehend the text.
- **SARI** metric compares the system's output to multiple simplification references and the original sentence based on the words added, deleted, and kept by a system [9].
- **BLEU** (Bilingual Evaluation Understudy) is a metric commonly used in machine translation to assess the quality of a translated text by comparing it to one or more reference translations [10]. It operates by comparing n-grams (contiguous sequences of words) in the candidate translation to the n-grams in the reference translations.
- **Compression ratio** is calculated by comparing the size of the original text to the size of the simplified version.
- **Sentence splits** refer to the division of a source sentence into multiple sentences.
- **Levenshtein similarity** measures the number of edits (insertions, deletions, or substitutions) needed to transform one sentence into another.
- **Exact copies** refer to the number of unchanged original (source) sentences during the simplification process.
- **Additions proportion** calculates the ratio of added content introduced in the simplified text compared to the original text.
- **Deletions proportion** calculates the ratio of content deleted from the original text during the simplification process.
- **Lexical complexity score** computed by taking the log-ranks of each word in the frequency table [7].

## 3. Scientific Text Simplification Approaches

In this section, we discuss a range of text simplification approaches that have been applied to scientific text as provided by the track.

**Chaoyang University of Technology (CYUT)** [11] submitted four runs for Task 3, experimenting with the GPT-4 API provided by OpenAI. They experimented with three different prompts, even using GPT-4 to suggest better prompts for the task.

**National Polytechnic Institute of Mexico** (NLPalma) [12] submitted a single run for Task 3. They experimented with BLOOMZ with different prompts to generate text simplifications.

**University of Amsterdam** [13] submitted two runs (*UAms_**) for Task 3, using the zero-shot application of GPT-2 based text simplification model. Their approach aimed to address one of the main issues in text generation approaches, which are prone to 'hallucinate' and generate spurious content unwarranted by the input. Specifically, they did this by post-processing the generated output to ensure grounding on input sentences, spurious generated output was identified and removed.

**University of Applied Sciences, Cologne** [14] submitted four runs (*irgc_**) for Task 3, with two runs using T5, one run using PEGASUS, and the final run exploiting ChatGPT. They performed a detailed analysis

**University of Cadiz/Split** (Smroltra) [15] submitted a single run for Task 3. They experimented with a SimpleT5 model for text simplification.

**University of Kiel** [16] submitted a single run (*TeamCAU_**) for Task 3, based on the SimpleT5 pre-trained language model.

**University of Kiel/Cadiz/Gdansk** [17] submitted two runs for Task 3 (as *Pun Detective*). They used SimpleT5 and GPT-3 models under resource-constrained conditions such as the limited task-specific train data, and showed the SimpleT5 model outperforming GPT-3 in key metrics.

**University of Kiel/Split/Malta** (MicroGerk) [18] submitted a total of 3 runs for Task 3. They experimented with BLOOMZ, GPT-3, and SimpleT5 models for text simplification.

**University of Southern Maine** (AIIR Lab) [19] submitted a total of 2 runs for Task 3. They experimented with two models, a GPT-2 based model and an OpenAI DaVinci model for generating text simplifications.

**University of Zurich (Andermatt)** [20] submitted 6 runs (*Pandas_**) for Task 3, experimenting with four large pretrained language models: T5, Alpaca 5B, and Alpaca LoRA. They exploited Task 2 data as additional train data, and experimented with prompt engineering.

**University of Zurich (Hou)** [21] submitted three runs (*QH_**) for Task 3, adapting the Multilingual Unsupervised Sentence Simplification (MUSS) model to HuggingFace's BART, and using a T5-Large model. They experimented with a template consisting of 5 control tokens and also added the original request.

**University of Kiel/Gdansk/Cadiz** (TheLangVerse) submitted a single run for Task 3. They experimented with a fine-tuned OpenAI Curie model for text simplification.

**University of Western Brittany** (UBO) [22] submitted a single run for Task 3. They experimented with a SimpleT5 model to generate simplifications.

Another team from the

**University of Western Brittany** (not in the Table) [23] experimented with ChatGPT for scientific text simplification, conducting a qualitative experiment with various analyses of the prompts and generated output.

## 4. Results

In this section, we discuss the results of the track based on the evaluation data.

### 4.1. Evaluation on the test data

A total of 14 teams submitted 32 runs for Task 3, mainly LLMs. Table 1 presents the results of participants' runs according to the automatic evaluation listed in Section 2.2. Surprisingly, all systems modified the original sentences (Exact copies = 0). While many participants applied the same LLMs, such as GPT-3 and T5, their results differ a great deal.

According to the evaluation results, all runs in the track demonstrated improvements in FKGL readability score compared to the identity baseline (i.e., the source sentences). This suggests that the systems were able to generate shorter sentences with shorter words on average. However, it is important to note that shorter words are not necessarily synonymous with simplicity, as they may include numerous abbreviations or specialized terms. The original sentences had an FKGL score of approximately 14, which corresponds to a university-level text. However, the majority of the submitted runs achieved lower FKGL scores ranging from 11 to 12, indicating a level of complexity comparable to that of texts encountered at the completion of compulsory education.

In terms of the SARI (System Output Against References and Input) score, all runs exhibited significant improvements compared to the original sentences. However, it is worth noting that the source sentences had the highest vocabulary overlap with the reference sentences according to the BLEU score on the test data, suggesting a closer resemblance in terms of the words used.

Overall, the results indicate that the track participants were successful in generating simplified sentences with improved readability, as evidenced by the FKGL and SARI scores.

### 4.2. Evaluation on the train data

The test data included all the input sentences corresponding to the reference sentences used as train data in 2023. Therefore, we can provide an additional evaluation on 2022 text simplification data [3, 24].

Table 2 shows the evaluation over these 648 sentences. A few observations can be made. First, we broadly see a similar effectiveness pattern as on the new test data in Table 1 above, with larger language models outperforming smaller ones. Second, the difference in performance between the very large models (e.g., OpenAI) and trainable models (e.g., GPT-2) is smaller on the train evaluation. Third, some models score exceedingly well on the train evaluation (e.g., SARI and BLEU above 50%) but less well on the new evaluation. This signals both the ability

**Table 1**
Results for task 3 (task number removed from the run_id) on the test set

| run_id | count | FKGL | SARI | BLEU | Compression ratio | Sentence splits | Levenshtein similarity | Exact copies | Additions proportion | Deletions proportion | Lexical complexity score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Identity_baseline | 245 | 13.64 | 15.09 | 26.22 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 0.00 | 8.64 |
| Reference | 245 | 12.03 | 100.00 | 100.00 | 0.95 | 1.10 | 0.66 | 0.00 | 0.33 | 0.40 | 8.64 |
| AiirLab_davinci | 243 | 11.17 | 47.10 | 18.68 | 0.75 | 1.00 | 0.68 | 0.0 | 0.20 | 0.45 | 8.59 |
| AiirLab_run1 | 245 | 9.86 | 30.07 | 15.93 | 1.26 | 1.67 | 0.80 | 0.0 | 0.30 | 0.17 | 8.47 |
| CYUT_run1 | 245 | 9.63 | 47.98 | 14.81 | 0.87 | 1.14 | 0.56 | 0.0 | 0.47 | 0.55 | 8.35 |
| CYUT_run2 | 245 | 8.43 | 44.93 | 12.09 | 0.76 | 1.06 | 0.56 | 0.0 | 0.46 | 0.62 | 8.31 |
| CYUT_run3 | 245 | 10.00 | 46.81 | 14.70 | 0.81 | 1.02 | 0.59 | 0.0 | 0.44 | 0.57 | 8.36 |
| CYUT_run4 | 245 | 9.24 | 47.69 | 15.41 | 0.78 | 1.03 | 0.58 | 0.0 | 0.41 | 0.58 | 8.32 |
| MiCroGerk_BLOOMZ | 245 | 12.54 | 32.01 | 22.24 | 0.92 | 0.99 | 0.89 | 0.0 | 0.13 | 0.21 | 8.54 |
| MiCroGerk_GPT-3 | 245 | 10.74 | 46.90 | 16.98 | 0.72 | 1.01 | 0.67 | 0.0 | 0.19 | 0.47 | 8.67 |
| MiCroGerk_simpleT5 | 245 | 12.96 | 25.43 | 21.26 | 0.91 | 0.99 | 0.92 | 0.0 | 0.09 | 0.18 | 8.52 |
| NLPalma_BLOOMZ | 245 | 9.61 | 35.66 | 5.76 | 0.68 | 1.00 | 0.51 | 0.0 | 0.35 | 0.66 | 8.26 |
| Pandas_alpaca-lora-alpaca-simplifier-alpaca-simplifier | 245 | 10.96 | 38.31 | 17.88 | 0.74 | 1.00 | 0.77 | 0.0 | 0.10 | 0.36 | 8.51 |
| Pandas_alpaca-lora-both-alpaca-normal-tripple | 245 | 12.02 | 36.10 | 20.89 | 0.89 | 1.05 | 0.82 | 0.0 | 0.16 | 0.29 | 8.57 |
| Pandas_alpaca-lora-both-alpaca-simplifier-tripple_10 | 244 | 11.71 | 36.38 | 19.62 | 0.89 | 1.07 | 0.78 | 0.0 | 0.16 | 0.31 | 8.55 |
| Pandas_alpaca-lora-simplifier-alpaca-short | 245 | 12.90 | 31.88 | 24.08 | 0.93 | 1.02 | 0.89 | 0.0 | 0.13 | 0.20 | 8.58 |
| Pandas_clean-alpaca-lora-simplifier-alpaca-short | 245 | 12.90 | 31.88 | 24.08 | 0.93 | 1.02 | 0.89 | 0.0 | 0.13 | 0.20 | 8.58 |
| Pandas_submission_ensemble | 245 | 10.51 | 40.25 | 17.40 | 0.77 | 1.09 | 0.73 | 0.0 | 0.15 | 0.40 | 8.52 |
| QH_run1 | 245 | 12.45 | 26.46 | 21.23 | 0.94 | 1.07 | 0.92 | 0.0 | 0.11 | 0.17 | 8.50 |
| QH_run2 | 245 | 13.05 | 24.40 | 21.33 | 0.96 | 1.03 | 0.92 | 0.0 | 0.12 | 0.15 | 8.48 |
| QH_run3 | 245 | 12.74 | 27.56 | 20.24 | 0.90 | 1.01 | 0.91 | 0.0 | 0.09 | 0.19 | 8.50 |
| Smroltra_SimpleT5 | 245 | 12.88 | 26.25 | 21.43 | 0.90 | 1.00 | 0.91 | 0.0 | 0.09 | 0.19 | 8.54 |
| TeamCAU_ST5 | 245 | 12.77 | 27.19 | 21.06 | 0.90 | 1.00 | 0.91 | 0.0 | 0.10 | 0.20 | 8.52 |
| TheLangVerse_openai-curie-finetuned | 245 | 12.21 | 30.78 | 18.92 | 0.86 | 1.00 | 0.86 | 0.0 | 0.11 | 0.24 | 8.49 |
| ThePunDetectives_GPT-3 | 245 | 7.52 | 41.56 | 6.10 | 0.46 | 0.97 | 0.50 | 0.0 | 0.16 | 0.68 | 8.46 |
| ThePunDetectives_SimpleT5 | 245 | 12.92 | 25.87 | 21.79 | 0.91 | 0.99 | 0.92 | 0.0 | 0.09 | 0.18 | 8.53 |
| UAms_Large_KIS150 | 245 | 10.50 | 33.02 | 14.59 | 1.26 | 1.48 | 0.76 | 0.0 | 0.34 | 0.20 | 8.45 |
| UAms_Large_KIS150_Clip | 245 | 11.12 | 33.47 | 16.59 | 1.01 | 1.23 | 0.82 | 0.0 | 0.24 | 0.23 | 8.48 |
| UBO_SimpleT5 | 245 | 12.33 | 30.89 | 21.08 | 0.88 | 1.05 | 0.89 | 0.0 | 0.10 | 0.22 | 8.51 |
| irgc_ChatGPT_2stepTurbo | 245 | 12.31 | 46.98 | 16.86 | 0.94 | 1.04 | 0.63 | 0.0 | 0.37 | 0.46 | 8.46 |
| irgc_pegasusTuner007plus_plus | 245 | 12.74 | 23.28 | 17.42 | 1.23 | 1.28 | 0.83 | 0.0 | 0.22 | 0.15 | 8.55 |
| irgc_t5 | 245 | 9.56 | 37.83 | 15.85 | 0.76 | 1.35 | 0.73 | 0.0 | 0.15 | 0.38 | 8.49 |
| irgc_t5_noaron | 245 | 9.55 | 37.84 | 15.84 | 0.76 | 1.35 | 0.73 | 0.0 | 0.15 | 0.38 | 8.49 |

to effectively train text simplification models for the domain, but also the risk of overfitting models and the need for independent evaluation data.

**Table 2**
Results for task 3 (task number removed from the run_id) on the train set

| run_id | count | FKGL | SARI | BLEU | Compression ratio | Sentence splits | Levenshtein similarity | Exact copies | Additions proportion | Deletions proportion | Lexical complexity score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Identity_baseline | 648 | 14.54 | 20.50 | 43.24 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 8.74 |
| Reference | 648 | 11.58 | 100.00 | 100.00 | 0.80 | 1.05 | 0.74 | 0.05 | 0.16 | 0.35 | 8.63 |
| AiirLab_davinci | 469 | 12.40 | 42.01 | 24.52 | 0.74 | 1.00 | 0.68 | 0.0 | 0.18 | 0.44 | 8.73 |
| AiirLab_run1 | 469 | 10.62 | 33.46 | 27.16 | 1.21 | 1.68 | 0.83 | 0.0 | 0.26 | 0.14 | 8.62 |
| Croland_T3_SimpleT5 | 130 | 14.62 | 31.68 | 42.47 | 0.91 | 0.98 | 0.93 | 0.0 | 0.06 | 0.15 | 8.72 |
| Croland_GPT3 | 50 | 8.77 | 39.72 | 10.40 | 0.43 | 1.0 | 0.50 | 0.0 | 0.12 | 0.70 | 8.68 |
| MiCroGerk_BLOOMZ | 469 | 13.67 | 34.70 | 37.74 | 0.91 | 0.99 | 0.91 | 0.0 | 0.09 | 0.17 | 8.68 |
| MiCroGerk_GPT-3 | 469 | 12.23 | 40.49 | 21.53 | 0.69 | 0.99 | 0.66 | 0.0 | 0.15 | 0.47 | 8.79 |
| MiCroGerk_simpleT5 | 469 | 13.77 | 35.97 | 41.28 | 0.91 | 0.99 | 0.92 | 0.0 | 0.07 | 0.15 | 8.70 |
| NLPalma_BLOOMZ | 469 | 10.00 | 33.63 | 10.86 | 0.64 | 0.99 | 0.50 | 0.0 | 0.37 | 0.69 | 8.43 |
| Pandas_alpaca-lora-alpaca-simplifier-alpaca-simplifier | 648 | 11.70 | 40.25 | 31.64 | 0.71 | 0.99 | 0.75 | 0.0 | 0.09 | 0.38 | 8.66 |
| Pandas_alpaca-lora-both-alpaca-normal-tripple | 648 | 12.91 | 37.19 | 33.96 | 0.87 | 1.05 | 0.81 | 0.0 | 0.12 | 0.27 | 8.72 |
| Pandas_alpaca-lora-both-alpaca-simplifier-tripple_10 | 647 | 11.97 | 38.04 | 32.46 | 0.83 | 1.12 | 0.78 | 0.0 | 0.14 | 0.32 | 8.68 |
| Pandas_alpaca-lora-simplifier-alpaca-short | 648 | 14.12 | 32.26 | 37.60 | 0.93 | 1.00 | 0.91 | 0.0 | 0.09 | 0.16 | 8.71 |
| Pandas_clean-alpaca-lora-simplifier-alpaca-short | 648 | 14.12 | 32.26 | 37.60 | 0.93 | 1.00 | 0.91 | 0.0 | 0.09 | 0.16 | 8.71 |
| Pandas_submission_ensemble | 648 | 11.36 | 40.44 | 28.47 | 0.69 | 1.00 | 0.71 | 0.0 | 0.12 | 0.42 | 8.66 |
| QH_run1 | 648 | 12.09 | 79.56 | 68.63 | 0.85 | 1.09 | 0.79 | 0.0 | 0.17 | 0.31 | 8.62 |
| QH_run2 | 648 | 12.55 | 77.68 | 65.83 | 0.89 | 1.06 | 0.80 | 0.0 | 0.18 | 0.28 | 8.61 |
| QH_run3 | 648 | 12.34 | 75.78 | 67.93 | 0.83 | 1.02 | 0.79 | 0.0 | 0.15 | 0.31 | 8.62 |
| Smroltra_BLOOM | 100 | 9.69 | 36.27 | 21.60 | 0.68 | 1.23 | 0.70 | 0.0 | 0.11 | 0.43 | 8.67 |
| Smroltra_GPT | 100 | 12.14 | 44.04 | 22.79 | 0.70 | 0.99 | 0.68 | 0.0 | 0.14 | 0.44 | 8.78 |
| Smroltra_SimpleT5 | 648 | 13.49 | 40.73 | 44.58 | 0.88 | 0.99 | 0.91 | 0.0 | 0.07 | 0.18 | 8.68 |
| TeamCAU_task_2.2_AI21 | 100 | 13.03 | 37.23 | 24.89 | 0.80 | 1.00 | 0.81 | 0.0 | 0.11 | 0.31 | 8.78 |
| TeamCAU_task_2.2_BLOOM | 100 | 9.70 | 34.38 | 18.22 | 0.65 | 1.19 | 0.72 | 0.0 | 0.10 | 0.43 | 8.57 |
| TeamCAU_task_2.2_ST5 | 648 | 12.30 | 64.99 | 59.61 | 0.81 | 1.01 | 0.83 | 0.0 | 0.10 | 0.28 | 8.63 |
| TheLangVerse_openai-curie-finetuned | 648 | 11.80 | 89.31 | 79.34 | 0.80 | 1.03 | 0.75 | 0.0 | 0.17 | 0.36 | 8.57 |
| ThePunDetectives_GPT-3 | 648 | 8.08 | 34.59 | 6.90 | 0.43 | 0.98 | 0.48 | 0.0 | 0.15 | 0.71 | 8.61 |
| ThePunDetectives_SimpleT5 | 648 | 13.39 | 41.40 | 45.18 | 0.89 | 0.99 | 0.91 | 0.0 | 0.07 | 0.17 | 8.68 |
| UAms_Large_KIS150 | 648 | 11.40 | 36.38 | 25.81 | 1.16 | 1.41 | 0.79 | 0.0 | 0.28 | 0.19 | 8.56 |
| UAms_Large_KIS150_Clip | 648 | 11.92 | 36.65 | 28.68 | 0.98 | 1.22 | 0.84 | 0.0 | 0.21 | 0.21 | 8.58 |
| UBO_BLOOM | 618 | 12.91 | 37.29 | 39.13 | 0.80 | 0.99 | 0.83 | 0.0 | 0.07 | 0.26 | 8.68 |
| UBO_SimpleT5 | 469 | 11.87 | 89.94 | 77.98 | 0.79 | 1.06 | 0.75 | 0.0 | 0.16 | 0.36 | 8.62 |
| irgc_ChatGPT_2stepTurbo | 648 | 12.67 | 37.53 | 14.75 | 0.90 | 1.04 | 0.61 | 0.0 | 0.38 | 0.49 | 8.55 |
| irgc_pegasusTuner007plus_plus | 648 | 13.97 | 27.53 | 32.25 | 1.23 | 1.26 | 0.86 | 0.0 | 0.18 | 0.10 | 8.71 |
| irgc_t5 | 648 | 9.88 | 38.68 | 29.68 | 0.76 | 1.42 | 0.72 | 0.0 | 0.15 | 0.38 | 8.70 |
| irgc_t5_noaron | 648 | 9.87 | 38.70 | 29.68 | 0.75 | 1.42 | 0.72 | 0.0 | 0.15 | 0.38 | 8.70 |
| CYUT_run1 | 648 | 10.21 | 35.73 | 8.61 | 0.86 | 1.13 | 0.54 | 0.0 | 0.49 | 0.57 | 8.40 |
| CYUT_run2 | 469 | 9.17 | 34.71 | 7.31 | 0.78 | 1.05 | 0.56 | 0.0 | 0.49 | 0.61 | 8.26 |
| CYUT_run3 | 469 | 10.47 | 36.23 | 9.40 | 0.80 | 1.03 | 0.59 | 0.0 | 0.45 | 0.58 | 8.29 |
| CYUT_run4 | 469 | 10.29 | 36.58 | 9.65 | 0.77 | 1.01 | 0.58 | 0.0 | 0.43 | 0.59 | 8.31 |

**Table 3**
Comparison of manually simplified and source sentences in Task 3

| Metric (Avg) | Source snt | Simplified snt |
|---|---|---|
| FKGL | 15.16 | 12.12 |
| # Abbreviations | 0.24 | 0.13 |
| # Difficult terms | 0.41 | 0.28 |

**Table 4**
Statistics on the levels of the difficulty of simplified sentences on the scale of 1-7

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| syntax complexity | 259 | 51 | 9 | | | | |
| lexical complexity | 93 | 119 | 62 | 26 | 19 | | |

# 5. Analysis of simplification quality

## 5.1. Lexical and syntax difficulty analysis

In order to evaluate the quality of our train data (648 manually simplified sentences), we compared simplified and source sentences according to the following metrics:

- FKGL readability score that relies on average sentence lengths and number of syllables per word [8];
- Average number of abbreviations per sentence. The list of abbreviations was taken from Task 2.1 [6, 5].
- Average number of difficult terms per sentence. The list of difficult terms was constructed from the data used for the evaluation of Task 2.1 [6, 5].

Table 3 reports the scores of manually simplified and source sentences used in Task 3 according to these three metrics. The table provides evidence that our manual simplifications reduce text difficulty not only in terms of readability score, but our simplified sentences have more than 50% less difficult terms and abbreviations. These results also show that our tasks are closely interconnected.

A master's student in translation and technical writing manually assessed the syntactic and lexical complexity of the simplification of 319 simplified sentences from the participants' runs, corresponding to 17 distinct source sentences evaluated using the same criteria. This evaluation is also a score ranging from 1 to 7 assigned to the simplification, with 1 representing *simple* and 7 representing *complex*. Table 4 provides evidence that automatic simplification is effective in terms of reducing syntax difficulty. However, lexical difficulty, i.e. the presence of difficult scientific terms, is much higher and thus remains the main barrier to understanding a scientific text.

We illustrate the syntactic and lexical complexity with a number of examples.

**Table 5**
Examples of perceived syntax and lexical complexity of systems' simplifications of the passage from Example 5.1

| System simplification | Syntax complexity | Lexical complexity |
|---|---|---|
| enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles. | 1 | 3 |
| "abstract: something that is not easily understood. novel: something that is new and different. technological advancements: improvements in technology. mobile devices: handheld devices such as smartphones and tablets. applications: programs that run on mobile devices can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles." | 2 | 3 |
| mobile devices and applications can help with wildfire confrontation. | 1 | 2 |

**Example 5.1.** *Abstract Novel technological advances in mobile devices and applications can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles.*

**Example 5.2.** *Four kinds of monitor units were specially designed for a wireless communication, including a control center, a local monitor unit, mobile devices (personal digital assistant; PDA), and a Web page (for both patient and doctor).*

The passage from Example 5.1 was scored 1 according to the complexity of its syntax and 3 according to the complexity of its vocabulary, while Example 5.2 was scored 2 according to the complexity of its syntax and 4 according to the complexity of its vocabulary.

Tables 5 and 6 provide syntax and lexical complexity annotations for the simplifications of the passages from Examples 5.1 and 5.2 respectively. Simplification complexity scores in Example 5.2 (Table 6) show that reducing syntax complexity may lead to an increase in lexical complexity as some difficult scientific terms may not be explained in the text, such as *personal digital assistant; PDA* in the example.

**Example 5.3.** *Several tasks can be accomplished from the AEGIS App, such as routing, spatial search for closest facilities and firefighting support infrastructures, access to weather data and visualization of fire management data (water sources, gas refill stations, evacuation sites etc.).*

The complexity of the source sentence from Example 5.3 was reduced drastically (syntax $2 \rightarrow 1$, lexical $4 \rightarrow 1$) by a system as shown in Table 7. However, the first simplification is a clear example of the hallucination problem in LLMs.

**Table 6**
Examples of perceived syntax and lexical complexity of systems' simplifications of the passage from Example 5.2

| System simplification | Syntax complexity | Lexical complexity |
|---|---|---|
| four kinds of monitor units were specially designed for a wireless communication, including a control center, a local monitor unit, mobile devices (personal digital assistant; pda), and a web page. | 2 | 4 |
| four monitor units were specially designed for wireless communication: a control center, a local monitor unit, a pda, and a web page for both patient and doctor. | 1 | 5 |
| four types of monitor units were designed for wireless communication: a control center, a local monitor unit, personal digital assistants (pdas), and a web page for both patient and doctor. | 1 | 4 |

**Table 7**
Examples of hallucinations and perceived syntax and lexical complexity of systems' simplifications of the passage from Example 5.3

| System simplification | Syntax complexity | Lexical complexity |
|---|---|---|
| the aegis app is available for download on the google play store | 1 | 1 |
| these tasks can be accomplished from the aegis app, such as searching for facilities closest to a player's position and firefighting support, access to weather data and visualization of firefighting support, as well as access to training courses for fire fighters. | 1 | 3 |

**Table 8**
Example of annotation of errors, source sentence from Example 5.1

| System simplification | Incorrect syntax | Unresolved anaphora due to simplification | Unnecessary repetition/iteration | Spelling, typographic or punctuational errors |
|---|---|---|---|---|
| enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles. | YES | YES | NO | YES |

**Table 9**
Error type statistics

| Error type | Instances | |
|---|---|---|
| | # | % |
| Incorrect syntax | 10 | 3 |
| Unresolved anaphora due to simplification | 34 | 11 |
| Unnecessary repetition/iteration | 16 | 5 |
| Spelling, typographic or punctuational errors | 94 | 30 |

## 5.2. Errors & Information distortion

For these 319 simplified sentences, we also verified other parameters: the syntax, the presence of unresolved references due to simplification, unnecessary repetition, and any spelling, typographic, or punctuation errors. The example of the annotation is given in Table 8. Table 9 provides statistics on the error types found in generated simplifications. The most common errors (30%) are spelling, typographic, and punctuational. These are followed by unresolved anaphora due to simplification (11%), and unnecessary repetion (5%).

In order to analyze information distortion [3], a master student in translation and technical writing and a university translator manually annotated 425 pairs of source sentences and simplifications submitted by the participants, corresponding to 22 distinct source sentences. Sentences were assigned with binary labels corresponding to the occurrence of the information distortion types. The objective was to determine whether there had been any information distortion during the simplification process, and if so, what type of distortion and what level of information loss severity there has been, on a scale from 1 to 7. We considered the following types of information distortions: *contresens (misinterpretation), topic shift, ambiguity, omission of essential details with regard to a query, overgeneralization, oversimplification, wrong synonym, insertion of false or unsupported information, insertion of unnecessary details with regard to a query, redundancy, repetition/iteration, style, nonsense*. If we considered there to be a distortion, we entered "YES" in the *information distortion* column, and also entered "YES" in the column corresponding to the specific error found. In cases of distortion, we also indicated the *information loss severity*, as we considered it, ranging from 1 to 7. For many sentence pairs, multiple types of information distortion were assigned.

Table 10 provides an example of annotation of information distortion. In 31% of cases we have not identify information distortion. Statistics on the information distortion severity on the scale of 1-7 is given in Table 11. Table 12 provides statistics on the information distortion types identified in the participants' runs.

Most information distortions are *Omission of essential details with regard to a query* and *Oversimplification* especially in sentences with a few difficult scientific terms. It might be explained by the fact that typically, text simplification involves the removal of complex or difficult-to-understand elements and the reduction of text length. However, significant information loss due to omission of essential details or unresolved anaphora might lead to complete loss of the sense of the text as shown in Table 13. An example of contresens (misinterpretation) due to

**Table 10**
Example of annotation of information distortion

| Source sentences | System simplification | Information distortion | Omission of essential details | Oversimplification | Information severity loss |
|---|---|---|---|---|---|
| This work describes an innovative mobile application for wildfire information management that operates on Windows Phone devices and acts as a complementary tool to the web-based version of the AEGIS platform for wildfire prevention and management. | the application is based on the same data and functionalities as the web-based version of the platform, | YES | YES | YES | 7 |

**Table 11**
Statistics on the information distortion severity on the scale of 1-7

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| information loss severity | 34 | 25 | 29 | 15 | 16 | 4 | 9 |
| information loss severity % | 8 | 6 | 7 | 4 | 4 | 1 | 2 |

**Table 12**
Information distortion type statistics

| **Information distortion type** | **Instances** | |
|---|---|---|
| | # | % |
| Contresens | 20 | 4.7 |
| Topic shift | 10 | 2.36 |
| Omission of essential details with regard to a query | 62 | 14.59 |
| Oversimplification | 41 | 9.65 |
| Insertion of false or unsupported information | 25 | 5.88 |
| Insertion of unnecessary details with regard to a query | 7 | 1.65 |
| Redundancy | 4 | 0,94 |
| Style | 5 | 1.17 |
| Nonsense | 2 | 0.47 |

extreme suppression of words is given in Table 14.

We also noticed that many simplifications made by AI systems, such as ChatGPT, are not exactly simplifications [23], e.g. ChatGPT's simplification involved rearranging the word order in the sentence and replacing some words with synonyms. In sentences with high lexical complexity (6 or 7), the AI does not simplify the technical terms, only the syntax.

Several information distortion types were observed in the simplifications of the Example 5.3

**Table 13**
Examples of severe information loss (Omission of essential details / Unresolved anaphora)

| Source sentence | System simplification |
|---|---|
| Four kinds of monitor units were specially designed for a wireless communication, including a control center , a local monitor unit, mobile devices (personal digital assistant; PDA), and a Web page (for both patient and doctor). | four types of monitor units were designed for wireless communication. |
| This work describes an innovative mobile application for wildfire information management that operates on Windows Phone devices and acts as a complementary tool to the web-based version of the AEGIS platform for wildfire prevention and management. | the application is based on the same data and functionalities as the web-based version of the platform, |
| This paper proposes an electrophysiological wireless patient monitoring system which integrates a Wireless ECG signal transmitter, GPS device and a mobile phone to acquire physiological signals and transmit them to a local server via Bluetooth wireless technology. | it also integrates a wireless ecg signal transmitter, gps device and a mobile phone to acquire physiological signals and transmit them to a local server. |
| This application was developed to promote the mobility and flexibility for the patients and also for the medical personnel, which further will improve both the quality of health care and lifestyle of the patient. | this application was developed to improve health care and lifestyle. |
| This application was developed to promote the mobility and flexibility for the patients and also for the medical personnel, which further will improve both the quality of health care and lifestyle of the patient. | this application was developed to improve the quality of health care and the lifestyle of patients. |
| Abstract Novel technological advances in mobile devices and applications can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks, such as access to data and information, sharing of intelligence and coordination of personnel and vehicles. | abstract novel technological advances in mobile devices and applications can be exploited in wildfire confrontation, enabling end-users to easily conduct several everyday tasks. |
| In view of the application of mobile technology in education, some previous studies have addressed that mobile learning is a meaningful learning that can improve the interaction between students and the situations and reach the purposes of learning. | mobile learning can improve learning. |
| In view of the application of mobile technology in education, some previous studies have addressed that mobile learning is a meaningful learning that can improve the interaction between students and the situations and reach the purposes of learning. | mobile technology can improve student learning. |

**Table 14**
Examples of contresens

| Source sentence | System simplification |
|---|---|
| In the main stream of mobile learning, using mobile carriers with suitable learning methods or strategies in mobile learning activities for different students to enhance learning have gradually become a important and concern issue. | mobile learning strategies for different students is an important issue. |

(see Table 7 before), but the most important is *Insertion of false or unsupported information* which can be caused by hallucination of a LLM.

## 5.3. Analysis of hallucination

The hallucination problem refers to the tendency of LLMs to generate plausible but incorrect or misleading information that is not grounded in reality. In terms of information distortion types we introduced in [3], LLMs' hallucinations refer mainly to *Topic shift, Insertion of false or unsupported information, Insertion of unnecessary details with regard to a query.* Note that *Non-sense* was attributed to incomplete sentences as in Examples 5.4 and 5.5. *Contresens* as in Table 14 seems to be caused by extreme deletion of terms and not by LLMs' hallucination.

**Example 5.4.** *the purpose of this study is to investigate the effect of mobile learning on the interaction between students and the*

**Example 5.5.** *such as routing, search for closest facilities and firefighting support infrastructures, access to weather data and visualization of fire management data (water sources, gas refill stations, evacuation sites, etc.).*

Table 15 provides some examples of hallucination in generated simplifications.

We conduct a deeper analysis of some of the sources of hallucination and it's prevalence in the submitted generated text simplifications.

We first show the human text simplifications in the ground truth data. Table 16 shows an example of a human reference simplification, combining the input sentences belonging to the abstract of the document $id = 130055196$ retrieved for query G01.1. Here, we show the deletions and insertions relative to the source input sentences (in this case on the first 4 sentences). Table 17 shows an example output simplification of one of the participating teams, for the same input sentences as in Table 16 above. Most simplifications are revisions of the input, but we also observe that sometimes an entire sentence is inserted (shown as xxx in Table 17).

Modern text simplification approaches are based on sequence-to-sequence models and generate simplified text by exploiting the superior text generation capacity of these models. However, these models are trained to complete any textual prompt and in text generation mode may generate additional output. A typical case is an additional sentence at the end of the input, which is often a very plausible continuation. This makes such spurious generation very difficult for human readers to spot, yet at the same time unwarranted by the input text from the scientific

**Table 15**
Hallucination examples in generated simplifications

| Source sentence | System simplification |
|---|---|
| The purpose of this study is to investigate the learning achievement and learning attitude of elementary school students on a campus plant learning activity when they use mobile carriers and competitive learning strategies. | in this paper, the writers show that we can use wireless technologies in smart homes and get the most out of powerful pdas. |
| Several tasks can be accomplished from the AEGIS App, such as routing, spatial search for closest facilities and firefighting support infrastructures, access to weather data and visualization of fire management data (water sources, gas refill stations, evacuation sites etc.). | the aegis app is available for download on the google play store |
| In recent years, mobile technologies have developed and applied in education fields, and some mobile emerging carriers with mobile technologies include a personal digital assistant (PDA), smart phone, and e-book. | this study aims to investigate the effects of mobile emerging carriers on the learning performance of students |
| The purpose of this study is to investigate the learning achievement and learning attitude of elementary school students on a campus plant learning activity when they use mobile carriers and competitive learning strategies. | in this paper, the writers show that we can use wireless technologies in smart homes and get the most out of powerful pdas. |
| In the main stream of mobile learning, using mobile carriers with suitable learning methods or strategies in mobile learning activities for different students to enhance learning have gradually become a important and concern issue. | this paper presents a mobile learning system based on mobile carriers and learning methods |
| This application was developed to promote the mobility and flexibility for the patients and also for the medical personnel, which further will improve both the quality of health care and lifestyle of the patient. | the application is developed for the android platform |
| In the main stream of mobile learning, using mobile carriers with suitable learning methods or strategies in mobile learning activities for different students to enhance learning have gradually become a important and concern issue. | in the main stream of mobile learning , using mobile carriers with suitable learning methods or strategies for different students to help them master mobile skills has become a key concern for teachers . |

abstract. Hence, this is a form of 'hallucination' that is a serious risk for the use-case of the track.

Table 18 quantifies how often such spurious generation occurs. We re-aligned the generated output with the original source sentences, and flag here only entire output sentences that do not

**Table 16**
Example of SimpleText Task 3 reference versus input: ~~deletions~~ and insertions

| Topic | Document | Output |
|---|---|---|
| G01.1 | 130055196 | ~~As various kinds~~ The rise of output devices ~~emerged , such as highresolution~~ like high-resolution printers ~~or a display of~~ and PDA ~~( Personal Digital Assistant ) ,~~ displays has increased the ~~importance of~~ need for high-quality resolution conversion ~~has been increasing~~ . \|~~This~~ The paper proposes a new method ~~for enlarging image with~~ to make images bigger while maintaining high quality . \|~~One of the largest problems on image enlargement~~ The main issue with enlarging images is ~~the exaggeration of the jaggy~~ that jagged edges can become exaggerated . \|To ~~remedy~~ solve this problem , we ~~propose~~ suggest a new ~~interpolation~~ method ~~, which uses artificial~~ that helps us to estimate the value of the newly generated pixels using a neural network ~~to determine the optimal values of interpolated pixels~~ . \|The ~~experimental~~ experiment 's results are ~~shown~~ presented and ~~evaluated~~ analyzed . \|~~The~~ We evaluate the effectiveness of our methods ~~is discussed~~ by comparing ~~with the conventional methods~~ them to traditional approaches . \| |

**Table 17**
Example of SimpleText Task 3 output versus input: ~~deletions~~, insertions, and whole sentence insertions

| Topic | Document | Output |
|---|---|---|
| G01.1 | 130055196 | As various kinds of output devices emerged , such as highresolution printers or a display of PDA ( Personal Digital Assistant ) ~~, the~~ . The importance of high-quality resolution conversion has been increasing . \|This paper proposes a new method for enlarging an image with high quality . It will involve using a combination of high-speed imaging and high-resolution video . \|One of the ~~largest~~ biggest problems on image enlargement is the exaggeration of the jaggy edges . This is especially true when the image is enlarged , as in this case . \|To remedy this problem , we propose a new interpolation method ~~, which~~ . This method uses artificial neural network to determine the optimal values of interpolated pixels . \|The experimental results are shown and evaluated . The results are compared to other studies and found to be inconclusive . \|The effectiveness of our methods is discussed by comparing with the conventional methods . Our methods are designed to help people with mental health problems , not just as a way to cure them . \| |

share a single token with the input.[1] We can make several observations. First, the number and fraction of spurious sentences vary considerably, with the majority of submissions having zero or very few cases (0-1%). Second, several submissions have notable fractions of spurious sentences in the output, with four runs over 20% and even one over 50% of input sentences. Third, standard test simplification evaluation measures (as shown in Tables 2 and 1) are almost completely insensitive to spurious sentences. Standard measures such as BLEU or SARI are based on (positive) text overlap with the reference simplifications. Hence, there is only a very marginal

[1]This approach is indicative but imperfect. For example, significant reordering of content may lead to false positives.

**Table 18**
Results for SimpleText Task 3: Spurious generation

| Run | # Input Sentences | Spurious Content | |
|---|---|---|---|
| | | Number | Fraction |
| AiirLab_davinci | 245 | 3 | 0.01 |
| AiirLab_run1 | 245 | 58 | 0.24 |
| CYUT_run1 | 757 | 12 | 0.02 |
| CYUT_run2 | 245 | 3 | 0.01 |
| CYUT_run3 | 245 | 1 | 0.00 |
| CYUT_run4 | 245 | 1 | 0.00 |
| MiCroGerk_BLOOMZ | 245 | 0 | 0.00 |
| MiCroGerk_GPT-3 | 245 | 1 | 0.00 |
| MiCroGerk_simpleT5 | 245 | 0 | 0.00 |
| NLPalma_BLOOMZ | 245 | 135 | 0.55 |
| Pandas_alpaca-lora-alpaca-simplifier-alpaca-simplifier | 245 | 1 | 0.00 |
| Pandas_alpaca-lora-both-alpaca-normal-tripple | 245 | 0 | 0.00 |
| Pandas_alpaca-lora-both-alpaca-simplifier-tripple_10 | 245 | 3 | 0.01 |
| Pandas_alpaca-lora-simplifier-alpaca-short | 245 | 0 | 0.00 |
| Pandas_clean-alpaca-lora-simplifier-alpaca-short | 245 | 0 | 0.00 |
| Pandas_submission_ensemble | 245 | 2 | 0.01 |
| QH_run1 | 245 | 3 | 0.01 |
| QH_run2 | 245 | 3 | 0.01 |
| QH_run3 | 245 | 1 | 0.00 |
| Smroltra_SimpleT5 | 245 | 0 | 0.00 |
| TeamCAU_task_2.2_ST5 | 245 | 0 | 0.00 |
| TheLangVerse_openai-curie-finetuned | 245 | 1 | 0.00 |
| ThePunDetectives_GPT-3 | 245 | 0 | 0.00 |
| ThePunDetectives_SimpleT5 | 245 | 0 | 0.00 |
| UAms_Large_KIS150 | 757 | 213 | 0.28 |
| UAms_Large_KIS150_Clip | 757 | 0 | 0.00 |
| UBO_SimpleT5 | 245 | 0 | 0.00 |
| irgc_ChatGPT_2stepTurbo | 245 | 0 | 0.00 |
| irgc_pegasusTuner007plus_plus | 245 | 57 | 0.23 |
| irgc_t5 | 245 | 11 | 0.04 |
| irgc_t5_noaron | 245 | 11 | 0.04 |

decrease in the longer output. In fact, some of the quality indicators (e.g., sentence splits, sentence length, and hence FKGL, etc.) tend to be increased by the adding of spurious sentences. As a general conclusion, our analysis clearly prompts the need to address hallucination as a central aspect of text simplification approaches and text simplification evaluation.

## 6. Conclusion

The paper provides an overview of the CLEF 2023 SimpleText Task 3, which focuses on the simplification of sentences found in scientific abstracts. The objective of the task is to simplify

these sentences to enhance their accessibility and comprehensibility for a general audience. The paper highlights the key aspects and goals of the task within the broader context of the CLEF 2023 SimpleText track. This task is closely connected to Task 2 [5] of the SimpleText track on difficult term identification and explanation.

A parallel corpus of sentences was constructed by extracting sentences from scientific publication abstracts, and these sentences were subsequently manually simplified. These manual simplifications reduce text difficulty in terms of readability score and a significant decrease (50%) in difficult terms and abbreviations from Task 2 [5] compared to the source sentences. This parallel corpus serves as a valuable resource for training and developing models for text simplification in the context of the task. The availability of a substantial amount of manually simplified sentences facilitates the exploration and development of various approaches and techniques in the field.

Regarding the different approaches to the tasks and their effectiveness, a few notable observations were made. Firstly, the results of similar methods varied considerably depending on factors such as implementation, fine-tuning, and the use of prompts. Secondly, apart from effectiveness, the efficiency of the approaches is crucial, as limitations on tokens or time often made participants submit incomplete runs based on language models.

In terms of automatic simplification, it proved effective in reducing syntax difficulty and optimizing the FKGL (Flesch-Kincaid Grade Level) score. However, the presence of difficult scientific terms (lexical difficulty) remains a major obstacle to understanding scientific texts. By focusing on reducing syntax complexity in text simplification, there is a possibility of encountering an increase in lexical complexity. This is because some challenging scientific terms may not be adequately explained or simplified in the simplified text. These findings highlight the interdependence of the SimpleText tasks and the importance of investigating their key dependencies.

The most common errors introduced during simplification were spelling, typographic, and punctuational mistakes (30%), followed by unresolved anaphora introduced by simplification (11%). In 15% of cases the information was distorted by omission of essential details, while oversimplification occurs in 10% of analyzed instances. Although, text simplification aims to make the content more understandable and accessible, often by removing complex or difficult-to-understand elements, it is crucial to strike a balance and avoid removing essential information that is necessary for the overall meaning and coherence of the text and avoid severe information distortion. In the process of simplification, there is a risk of significant information loss when essential details are omitted or when anaphora (referencing previous words or phrases) is left unresolved. This can result in a complete loss of the overall meaning or sense of the text. In some cases, the excessive suppression of words may lead to the introduction of contresens (contradictory or nonsensical statements) which further affects the accuracy and coherence of the simplified text. Another issue noted was when there were no errors or distortion but also no simplification: the model produced a sentence nearly identical to the original with no improvement.

LLMs tend to generate responses that might sound plausible and coherent but are not necessarily accurate or factual. Improving the factuality, accuracy, and overall reliability of LLM outputs is a crucial goal to ensure that these models provide reliable and trustworthy information to users. We observed the following information types which might be caused by LLMs'

hallucinations: *Topic shift, Insertion of false or unsupported information, Insertion of unnecessary details with regard to a query* representing around 10% of analyzed simplifications.

The overall conclusion drawn from the CLEF 2023 SimpleText track is that significant progress has been made by the state-of-the-art models, but there is still ample room for improvement. In future work, we will aim to quantify this and, through manual annotation, score 'good' simplifications as well as poor transformations. It would be interesting to look for any association of such 'adventurousness' of simplification back to the particular models under test, as it would be to know if there are links between models and distortion type. Exploring the information distortion introduced by simplification is an area of focus for further research.

### Acknowledgments

*This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01. We would like to thank Quentin Dubreuil, and all other colleagues and participants who helped run this track.*

## References

[1] M. Maddela, F. Alva-Manchego, W. Xu, Controllable Text Simplification with Explicit Paraphrasing (2021). URL: http://arxiv.org/abs/2010.11004.

[2] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Overview of simpletext 2021 - CLEF workshop on text simplification for scientific information access, in: CLEF'21: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 432–449. URL: https://doi.org/10.1007/978-3-030-85251-1_27.

[3] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, É. Mathurin, P. Bellot, Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts, in: CLEF'22: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://doi.org/10.1007/978-3-031-13643-6_28.

[4] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2023 SimpleText Task 1: Passage selection for a simplified summary, in: [25], 2023.

[5] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation, in: [25], 2023.

[6] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the CLEF 2023 SimpleText Lab: Automatic simplification of scientific texts, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2023.

[7] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, EASSE: Easier automatic sentence simplification evaluation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language

Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 49–54. URL: https://aclanthology.org/D19-3009. doi:10.18653/v1/D19-3009.
[8] R. Flesch, A new readability yardstick., Journal of Applied Psychology 32 (1948) p221 – 233.
[9] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the ACL 4 (2016) 401–415.
[10] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proc. of the 40th annual meeting on ACL, ACL, 2002, pp. 311–318.
[11] S.-H. Wu, H.-Y. Huang, A Prompt Engineering Approach to Scientific text simplification: CYUT at SimpleText2023 Task3, in: [25], 2023.
[12] V. M. Palma, C. P. Preciado, G. Sidorov, NLPalma @ CLEF 2023 SimpleText: BLOOMZ and BERT for complexity and simplification task, in: [25], 2023.
[13] R. Hutter, J. Sutmuller, M. Adib, D. Rau, J. Kamps, University of Amsterdam at the CLEF 2023 SimpleText Track, in: [25], 2023.
[14] B. Engelmann, F. Haak, C. K. Kreutz, N. Nikzad-Khasmakhi, P. Schaer, Text Simplification of Scientific Texts for Non-Expert Readers, in: [25], 2023.
[15] P. Dadić, O. Popova, CLEF 2023 SimpletText Tasks 2 and 3: Enhancing Language Comprehension: Addressing Difficult Concepts and Simplifying Scientific Texts Using GPT, BLOOM, KeyBert, Simple T5 and More, in: [25], 2023.
[16] A. Anjum, N. Lieberum, Automatic Simplification of Scientific Texts using Pre-trained Language Models: A Comparative Study at CLEF Symposium 2023, in: [25], 2023.
[17] F. Ohnesorge, M. A. Gutierrez, J. Plichta, Scientific Text Simplification and General Audience, in: [25], 2023.
[18] D. R. Davari, A. Prnjak, K. Schmitt, CLEF2023 SimpleText Task 2, 3: Identification and Simplification of Difficult Terms, in: [25], 2023.
[19] B. Mansouri, S. Durgin, S. Franklin, S. Fletcher, R. Campos, AIIR and LIAAD Labs Systems for CLEF 2023 SimpleText, in: [25], 2023.
[20] P. S. Andermatt, T. Fankhauser, UZH_Pandas at SimpleTextCLEF-2023: Alpaca LoRA 7B and LENS Model Selection for Scientific Literature Simplification, in: [25], 2023.
[21] R. Hou, X. Qin, An Evaluation of MUSS and T5 Models in Scientific Sentence Simplification: A Comparative Study, in: [25], 2023.
[22] Q. Dubreuil, UBO Team @ CLEF SimpleText 2023 Track for Task 2 and 3 - Using IA models to simplify Scientific Texts, in: [25], 2023.
[23] S. Bertin, Scientific simplification, the limits of ChatGPT, in: [25], 2023.
[24] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the CLEF 2022 SimpleText Task 3: Query biased simplification of scientific texts, volume 3180 of *CEUR Workshop Proceedings*, 2022. URL: https://ceur-ws.org/Vol-3180/.
[25] M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2023.