UBOnlp Report at the SimpleText lab of CLEF 2025

Benjamin Vendeville[1,2,3], Liana Ermakova[2], Pierre De Loor[3] and Jaap Kamps[4]

[1]Université de Bretagne Occidentale, Brest, France
[2]HCTI, Brest, France
[3]Lab-STICC (UMR CNRS 6285), Brest, France
[4]University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper presents the UBOnlp team's participation in the SimpleText lab at CLEF 2025, focusing on scientific text simplification and controlled creativity tasks. We evaluate the performance of GPT-4o using simple prompt-based approaches across multiple subtasks without specialized training or fine-tuning. For Task 1 (Text Simplification), we applied GPT-4o to both sentence-level and document-level simplification of scientific abstracts from the Cochrane-Auto corpus. Our system achieved competitive SARI scores (42.20 for sentence-level, 43.37 for document-level) while maintaining low complexity metrics, demonstrating effective simplification through content reduction rather than lexical substitution. For Task 2 (Controlled Creativity), we addressed spurious generation detection and error classification in simplified texts. Our approach showed strong performance in fluency error detection (F1 = 0.322, ranking first) and alignment error detection (F1 = 0.381, ranking third), but struggled with general spurious content detection, particularly in post-hoc scenarios without source documents. These results highlight both the potential and limitations of large language models for specialized text simplification tasks. While GPT-4o demonstrates capabilities in linguistic quality assessment, task-specific architectures remain superior for comprehensive error detection and generation control. Our findings contribute to understanding the practical applicability of general-purpose language models in scientific text processing workflows.

Keywords

Automatic text simplification, Science popularization, Large Language Models

1. Introduction

With Internet becoming ever more common, people start relying more and more on it for communicating and learning. That includes learning about science and communicating on scientific research. However, science education has failed to keep up with these advances, leading people to misunderstand the science they are reading about online. With the advance of AI, there is a growing potential to build tools that address this issue. This is what the SimpleText lab at CLEF [1, 2, 3] aims to study. The 2025 edition of the SimpleText lab [4] is divided into 3 tasks:

- *Task 1*: **Text Simplification** simplify scientific text [5].
  - **Subtask 1.1**: Simplify sentences.
  - **Subtask 1.2**: Simplify abstracts.
- *Task 2*: **Controlled Creativity** identify and avoid hallucination [6].
  - **Subtask 2.1**: Identifying creative generation.
  - **Subtask 2.2**: Classifying information distortion.
  - **Subtask 2.3**: Avoiding creative generation.
- *Task 3*: **SimpleText 2024 Revisited** selected tasks by popular request.
  - **Subtask 3.1**: *Content Selection*: Retrieving passages to include in a simplified summary
  - **Subtask 3.2**: *Complexity Spotting*: Identifying and explaining difficult concepts

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
benjamin.vendeville@univ-brest.fr (B. Vendeville); liana.ermakova@univ-brest.fr (L. Ermakova); deloor@enib.fr (P. D. Loor); kamps@uva.nl (J. Kamps)
0009-0003-5298-147X (B. Vendeville); 0000-0002-7598-7474 (L. Ermakova); 0000-0002-5415-5505 (P. D. Loor); 0000-0002-6614-0087 (J. Kamps)

- **Subtask 3.3**: *Text Simplification*: Simplify Scientific Text

This paper will detail the participation of team UBOnlp for tasks 1 and 2, where we used GPT-4o [7] to generate predictions. We will present the task and data provided, as well as the prompts we used for prediction.

2. Task 1: Text Simplification

2.1. Task Description

The goal of this task was to generate simplifications of scientific texts. It was divided into two subtasks: sentence-level simplification (Subtask 1.1) and document-level simplification (Subtask 1.2). This task used the Cochrane-Auto corpus, built from the Cochrane systematic reviews and their associated lay summaries. Cochrane-Auto consists of professionally written abstract-summary pairs aligned at sentence, paragraph, and document levels. The dataset was constructed by realigning biomedical abstracts and lay summaries at different levels of granularity-sentence, paragraph, and full document. The alignment is restricted to ensure accurate correspondences, enabling meaningful evaluation at each level. The dataset was split into training and test datasets:

- *train*: 4,171 sentences (Task 1.1) and 4,171 paragraphs (Task 1.2)
- *test*: 4,293 sentences (Task 1.1) and 217 abstracts (Task 1.2)

Participants were welcomed to use training data to train models, but we decided to use an untrained, prompt-based approach.

We evaluate system outputs using a range of standard and simplification-specific metrics provided by EASSE [8]. Flesch-Kincaid Grade Level (FKGL) [9] estimates the reading difficulty of a text based on average sentence length and syllables per word, returning a U.S. school grade level; higher values indicate more complex texts, with a theoretical lower bound of -3.40 and no upper limit.

BLEU [10] assesses n-gram overlap between generated and reference texts. Although originally developed for machine translation, it is commonly applied in simplification by treating standard and simplified English as distinct languages. Scores range from 0 (no overlap) to 1 (perfect match).

SARI [11] is specifically designed for text simplification, comparing the system output not only to references but also to the input. It evaluates the quality of additions, deletions, and words retained, with scores ranging from 0 to 100, where higher indicates better simplification.

To characterize structural transformations, we compute the compression ratio, which compares the token count of the output to that of the reference; higher values reflect more compressed outputs. Sentence splits count the number of input sentences divided into multiple ones in the output, with higher counts indicating more frequent segmentation.

We also use Levenshtein similarity to quantify the edit distance between the input and the output, where higher values denote greater surface similarity. The exact copy rate measures the proportion of output sentences that are identical to sentences in the input.

In addition, we track the proportion of additions and deletions, indicating the extent of lexical changes between input and output. Finally, lexical complexity is computed following Alva-Manchego et al. [8], by aggregating the third quartile of the log-frequency ranks of words, capturing the relative rarity of the vocabulary used.

For sentence-level simplification (Task 1.1) sentences were concatenated into abstract and evaluated as such. Furthermore, two different sets of references were used. One was based on the plain language summary (PLS) from the original Cochrane references and contained references for 217 abstracts, while the 2nd was made from Cochrane-auto and contained references for 37 abstracts.

2.2. Test Data

The provided test data for Task 1.1 was of the form:

```
{
    "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."
},
```


While test data for Task 1.2 was of the form:


```
{
    "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. Interventions in all studies
        included [...]"
 },
```


2.3. Submission Description

Our goal for this task was to see the performance of state of the art models used in a simple way. Therefore, we decided to use GPT-4o to generate simplifications based only on a simple prompt and the source sentence. The decoding was made with a temperature of 0, and we used the following prompt:

```
prompt = f"""You are a classification expert for simplification errors. You need to
    simplify the following scientific text for the general public.
The goal is to make the provided text more easily understandable.
It is important to keep un easy vocabulary, a simple semantic structure, and to not have
     too much information density.
You also need to be informative and make the user understand important facts in the
    source.
----------
Source: "{source}" """
```

The same prompt is used for both subtasks 1.1 and 1.2.

2.4. Results

2.4.1. Task 1.1

The evaluation of our run, along with scores of other participants, are presented in Table 1 and Table 2. We see our system being one of the best on SARI on sentence-level simplification while keeping one of the lowest FKGL and lexical complexity scores. Looking at the addition and deletion proportions, our model removed more content than other models, while adding less.

This suggests that our system adopts a more conservative rewriting strategy, favoring deletion over lexical addition. While this may help reduce complexity, it could also risk omitting important information.

On Cochrane-Auto aligned data, however, we observe a notable drop in our model's performance, especially on SARI and BLEU, while other systems such as *DSGT plan_guided_lla* remain closer to the PLS references. Interestingly, this drop coincides with a mismatch in sentence splitting behavior: while our model tends to preserve the original sentence boundaries, the PLS references in Cochrane-Auto may restructure content more, with significantly more sentence splits compared to those in the manually aligned references. This difference may have penalized our system, which performs better for sentence-level rewriting and performs well when reference simplifications follow similar segmentation. Despite this, our model maintains competitive scores on FKGL and lexical complexity, suggesting that it still

produces fluent and accessible output, albeit less aligned with the structural edits present in the PLS references.

Table 1
Scores of participants runs on Task 1.1 with references from the original plain language summaries from Cochrane.

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 GPT-4o	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
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
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
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
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
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
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
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
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

2.4.2. Task 1.2

The evaluation of our system, UBOnlp GPT-4o, alongside those of other participants, is presented in Table 3 and Table 4. Our system demonstrates competitive performance, particularly on SARI, indicating effective simplification strategies. It produces longer outputs and performs frequent sentence splitting, reflecting a consistent approach focused on decomposing and elaborating complex information rather than merely shortening the text. This is further supported by the high compression ratio and addition proportion, suggesting that the model often introduces explanatory content-such as definitions-to enhance clarity. Despite these strengths, the lower BLEU scores point to a greater divergence from reference phrasing, potentially impacting perceived fluency and alignment. The system also performs well on FKGL and lexical complexity metrics, confirming its ability to adapt the vocabulary and structure to a simpler register.

These tendencies are confirmed in the evaluation against the Cochrane-Auto references, where the results remain broadly consistent-SARI scores decrease slightly, while BLEU improves marginally, highlighting the model's stable behavior across reference sets.

Table 2
Scores of participants runs on Task 1.1 with references from Cochrane-Auto.

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
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
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 GPT-4o	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
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
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
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
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
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
flash.json 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-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

3. Task 2: Controlled Creativity

3.1. Task Description

In practice, when generating simplifications, organizers have found a high proportion and variety of spurious generation. The goal of this task is therefore to detect, classify and avoid *spurious generation*. In particular, we participated in the following subtasks 2.1 and 2.2.

3.1.1. Subtask 2.1

The goal of this subtask is to detect spurious generation. Participants were presented a system generated simplification, and had to classify it as *spurious* or not. in particular, two cases were studied: one (sourced) where participants had access to the source document of the simplification and one (posthoc) where they did not. The dataset was constructed from system simplifications retrieved from last year's submissions to the SimpleText lab and were automatically annotated based on token alignment, where if over 10% of the tokens in the generations were not aligned with the source, the generation was considered spurious. This created a high prevalence of the *spurious* label (90%). The train dataset was split into 13,341 sentences (posthoc) and 13,514 sentences (sourced) while the test dataset was split into 3,336 sentences (posthoc) and 3,379 sentences (sourced). Results are evaluated using Accuracy, Precision,

Table 3
Scores of participants runs on Task 1.2 with references from the original plain language summaries from Cochrane.

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
SINAI PRMZSTASK12V1	217	43.63	8.07	10.73	0.81	1.03	0.52	0.00	0.37	0.54	8.41
ASM MistralMinFKGL	217	43.51	8.26	11.85	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 GPT-4o	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
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
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
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	217	38.88	10.00	8.97	0.67	1.07	0.59	0.18	0.20	0.52	8.53
UvA bartpara-cochran	217	34.97	12.70	12.13	0.55	0.70	0.68	0.00	0.01	0.49	8.86
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

Recall, F1 score, AUROC and AUPRC.

3.1.2. Subtask 2.2

The goal of this subtask is to detect and classify hallucinations with regards to a taxonomy of [12]. The taxonomy classifies errors in text simplifications into one or multiple of 14 different *errors* classes, grouped into 4 error groups:

- **A. Fluency** Is the answer provided in a correct form that a fluent speaker would speak?
- **B. Alignment** Is the format of the answer correct?
- **C. Information** Is the information provided accurate and relevant to the input?
- **D. Simplification** Does the response focus on simplification?

In addition, a *"No Error"* class is also considered. The training data is constructed from 42,392 synthetically generated simplifications containing targeted errors generated from past submissions to the SimpleText lab. The test data was constructed from 2,659 manual annotations of past submissions to the SimpleText lab. Results are evaluated on the four aggregated error categories, using both F1 score and AUC.

3.2. Test data

3.2.1. Subtask 2.1

The provided test datasets were provided in a json format as such:

- *Subtask 2.1 Posthoc:*

Table 4
Scores of participants runs on Task 1.2 with references from Cochrane-Auto.

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
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 MistralMinFKGL	37	43.24	12.27	11.63	0.73	0.93	0.53	0.00	0.27	0.56	8.75
ASM MistralV7CleanLi	37	42.93	11.38	13.77	0.78	0.84	0.51	0.00	0.29	0.56	8.80
UvA baseline-cochran	37	42.10	24.27	11.71	0.57	0.71	0.61	0.00	0.06	0.49	8.74
UBOnlp GPT-4o	37	41.56	5.45	7.22	1.14	2.08	0.50	0.00	0.58	0.43	8.25
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
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
UvA bartpara-cochran	37	37.89	27.43	12.22	0.62	0.77	0.74	0.00	0.01	0.41	8.78
UvA bartdoc-ca	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
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


```
{
    "sentence": "I explained the complex terms directly within the simplified sentence:\
        n\n* 'Next-generation model' means a new and improved plan.",
    "anon_gen_id": "74704850//66348262//3"
 },
```


- *Subtask 2.1 Sourced:*


```
{
    "abs_id": "G01.1_1570837852",
    "sentence": "In this paper, we share our findings on how evolutionary algorithms and
         multi-agent systems can be used to understand a user's preferences while they
        interact with a digital assistant.",
    "gen_id": "11102757//G01.1_1570837852//1"
 },
```


The Sourced data could be merged with the abstract data of the following format:


```
{
    "query_id": "G11.1",
    "query_text": "drones",
    "doc_id": 2892036907,
```

```
    "abs_id": "G11.1_2892036907",
    "abs_source": "In the modern era of automation and robotics, autonomous vehicles are
         currently the focus of academic and industrial research. With the ever
        increasing number of unmanned aerial vehicles getting involved in activities in
        the civilian and commercial domain, there is [...]"
 },
```


3.2.2. Subtask 2.2

The test data was provided as a json file as such:


```
{
    "source sentence": "Partners In Health (PIH) and its sister organization in Lima,
        Peru, Socios En Salud (SES), treat a majority of multidrug-resistant
        tuberculosis (MDR-TB) patients in Peru, in conjunction with the Peruvian
        National TB Program (NTP).",
    "simplified sentence": "Socios En Salud (SES) and its sister organization in Lima,
        Peru, treat a majority of multidrug-resistant tuberculosis (MDR-TB) patients in
        Peru, in conjunction",
    "snt_id": "G01.1_147704292_1",
    "simp_id": 783909.0,
    "No error": "",
    "A1. Random generation": "",
    "A2. Syntax error": "",
    "A3. Contradiction": "",
    "A4. Simple punctuation / grammar errors": "",
    "A5. Redundancy": "",
    "B1. Format misalignement": "",
    "B2. Prompt misalignement": "",
    "C1. Factuality hallucination": "",
    "C2. Faithfulness hallucination": "",
    "C3. Topic shift": "",
    "D1.1. Overgeneralization": "",
    "D1.2 Overspecification of Concepts": "",
    "D2.1. Loss of Informative Content": "",
    "D2.2. Out-of-Scope Generation": ""
 },
```


3.3. Submission Description

In both subtasks, our goal was to try to measure performance of state of the art models used in a naive, simple way. In both subtasks, we relied on an untrained GPT-4o model using only a prompt wiht test data as input. The decoding was made with a temperature of 0.

3.3.1. Subtask 2.1

For this subtask, we used two slightly different prompts for the *sourced* and *posthoc* variations. For posthoc we used the following prompt:

```
prompt = f"""
You are an expert in detecting hallucinations in simplified scientific texts.

Hallucinations include:
- Information distortion: misrepresenting or oversimplifying facts in a misleading
    way.
- Spurious generation: adding information not supported by scientific content.

Your task: Analyze the simplified text and respond only with:
```

```
- True -> if the text likely contains a hallucination.
- False -> if the text seems accurate and faithful.

Respond with **only** `True` or `False`.

----------
Simplified Text:
{simplified}
"""
```

For the *sourced* variation, we used:

```
prompt = f"""
You are an expert in detecting hallucinations in simplified scientific texts.

Hallucinations include:
- **Information distortion**: when the simplified text misrepresents or alters the
    meaning of the source.
- **Spurious generation**: when the simplified text includes new information not
    present or supported in the source.

Your task is to compare the simplified text with the source and respond with:
- True -> if the simplified text contains hallucinations (of either type).
- False -> if the simplified text is faithful to the source.

Respond with **only** True or False.

----------
Source Text:
{source}

Simplified Text:
{simplified}
"""
```

3.3.2. Subtask 2.2

For the subtask 2.2, we used a prompt describing the taxonomy, as well as the format required, and included examples. The taxonomy is the definition of the errors as provided in [12] while *possible codes* are the codes corresponding to the error.

```
prompt = f"""
    You are a classification expert for simplification errors.  Whenever you answer, you
        MUST call the function classify_simplification_errors and pass exactly one JSON
        object of the form:
    {{ "labels": [<list of codes>] }} and nothing else-no prose, no extra keys.
    The only valid key inside that object is "labels" (an array of strings).
    TAXONOMY:
    {taxonomy}
    AVAILABLE CODES:
    {POSSIBLE_CODES}
    IMPORTANT RULES:
    1.  The function call must look exactly like:
           classify_simplification_errors({{ "labels": [ ... ] }})
        with no extra text before or after.
    2.  If **no error** applies, return exactly:
           classify_simplification_errors({{ "labels": ["No"] }})
    3.  If you choose any code *other than* "No," you **MUST NOT** include "No" in your
        array.
```

```
    4. The list inside "labels" must be in **lexicographic order** (e.g. ["A1","B2","
       D22"], not ["B2","A1","D22"]).
    5. If multiple codes apply, include all of them (still in lexicographic order).  If
        only one applies, return only that one (e.g. ["A4"]).
    EXAMPLES (you see INPUT -> function call OUTPUT):
      - Example 1 (no error):
        Input:
          Source: "The cat meowed softly."
          Simplification: "The cat meowed softly."
        Output:
          classify_simplification_errors({{ "labels": ["No"] }})
      - Example 2 (single-label):
        Input:
          Source: "The cat chattered."
          Simplification: "The cat meowed."
        Explanation: "chattered" -> "meowed" fixes a word-choice/typo error -> A2 Syntax
            error.
        Output:
          classify_simplification_errors({{ "labels": ["A2"] }})
      - Example 3 (multi-label):
        Input:
          Source: "Bob went to the market. He loves fruit."
          Simplification: "Bob went market; he adores apples and wrote a poem about them
              ."
        Explanation:
          - "went market;" has a grammar/syntax issue -> A2 Syntax error
          - "he adores apples and wrote a poem about them" adds extra content (not part
              of simplification) -> D22 Out-of-Scope Generation
        Output:
          classify_simplification_errors({{ "labels": ["A2","D22"] }})
    NOW CLASSIFY ONLY the pair below.  Do not write any additional text-just invoke
        classify_simplification_errors(...) once.
    ----------
    Source: "{source}"
    Simplification: "{simplification}"
"""
```

3.4. Results

3.4.1. Subtask 2.1

Results for this subtask are presented in table 5 and table 6 In the posthoc detection scenario, our GPT-4o approach ranked last among the participating teams. The results reveal a characteristic pattern: while our method achieved high precision (0.92), indicating that when it predicted spurious generation it was usually correct, it suffered from extremely low recall (0.21). This suggests our GPT-4o approach was overly conservative in identifying spurious content when operating without access to source documents. The low accuracy (0.27) and near-random AUROC (0.52) indicate that our approach struggled significantly with the posthoc detection task. Given that the dataset has a 90% prevalence of spurious examples, our low recall particularly hurt overall performance.

When source documents were available, our GPT-4o approach showed improved but still limited performance. The recall increased from 0.21 to 0.71, and accuracy improved from 0.27 to 0.70. This suggests that GPT-4o benefits significantly from having reference material to compare against when detecting spurious generation. However, our approach still ranked in the lower tier of submissions, with several teams achieving accuracy scores above 0.90 and F1 scores above 0.95.

The performance difference between our approach and top-performing methods (which achieved F1-scores above 0.95) suggests that task-specific model architectures, such as BERT-based classifiers

and ensemble methods, may still be more suitable for this type of detection task than general-purpose language models used in a zero-shot or few-shot manner.

Table 5
Results for CLEF 2025 SimpleText Task 2.1 Detecting Overgeneration: Test data, posthoc detection without sources, best five runs per team

Team/Method	count	Acc.	Prec	Rec	F1	AUROC	AUPRC
SINAI basic-prefilter-all-true	3,336	0.91	0.91	1.00	0.95	0.55	0.91
DSGT bertclassifier	3,336	0.91	0.93	0.97	0.95	0.64	0.93
DSGT bert_nli_llm_ensemble	3,336	0.90	0.93	0.97	0.95	0.64	0.93
DSGT bertnlillmensemble	3,336	0.90	0.93	0.97	0.95	0.64	0.93
DUTH Task21posthoc_et	3,336	0.90	0.92	0.97	0.95	0.62	0.92
DSGT llm	3,336	0.77	0.95	0.78	0.86	0.70	0.94
DSGT nli_entailment	3,336	0.45	0.95	0.41	0.57	0.61	0.92
SINAI improved-prefilter-confidence-95	3,336	0.35	0.95	0.29	0.44	0.57	0.91
UBOnlp GPT-4o	3,379	0.27	0.92	0.21	0.35	0.52	0.90

Table 6
Results for CLEF 2025 SimpleText Task 2.1 Detecting Overgeneration: Test data, detection with sources.

Team/Method	count	Acc.	Prec	Rec	F1	AUROC	AUPRC
AIIRLab CrossEncoder	3,379	0.98	0.99	0.99	0.99	0.95	0.99
Mtest bartfinetuned	3,379	0.97	0.99	0.97	0.98	0.96	0.99
SINAI improved-prefilter-all-true	3,379	0.96	1.00	0.95	0.98	0.98	0.99
SINAI improved-prefilter-confidence-99	3,379	0.93	1.00	0.93	0.96	0.96	0.99
DSGT bertclassifier	3,379	0.91	0.93	0.98	0.95	0.65	0.93
DSGT bertnlillmensemble	3,379	0.91	0.93	0.97	0.95	0.68	0.93
DUTH Task21sourced_et	3,379	0.91	0.93	0.97	0.95	0.66	0.93
SINAI improved-prefilter-confidence-95	3,379	0.81	1.00	0.79	0.88	0.89	0.98
DUTH Task21sourced_ridge	3,379	0.77	0.94	0.79	0.86	0.68	0.93
DSGT llm	3,379	0.74	0.94	0.76	0.84	0.68	0.93
UBOnlp GPT-4o	3,379	0.70	0.95	0.71	0.81	0.69	0.93
RECAIDS T5	3,379	0.49	0.89	0.49	0.63	0.47	0.89
DSGT nli_entailment	3,379	0.35	0.92	0.31	0.46	0.53	0.90
AIIRLab LLMs	3,379	0.10	0.00	0.00	0.00	0.50	0.90

3.4.2. Subtask 2.2

Our system achieved the best F1 score (0.322) for fluency error detection, outperforming all competing systems including specialized fine-tuned models. This demonstrates GPT-4o's capabilities for identifying grammatical errors and fluency issues. It also showed strong performance in alignment error detection (F1 = 0.381, 3rd place), showing effective identification of format and structural issues. However, our system showed lower performance in "No Error" classification (F1 = 0.680) suggesting tendency toward false positives. Information and simplification error detection showed moderate results, indicating challenges with task-specific requirements.

The results highlight GPT-4o's strength in linguistic tasks while revealing limitations in specialized error detection, showing the usefulness of building task-specific error detection models.

4. Conclusion

This paper evaluated GPT-4o's effectiveness for scientific text simplification and controlled creativity tasks at CLEF 2025 SimpleText using straightforward prompt-based approaches without specialized

Table 7
Model Performance by Error Categories (Best Scores in Bold) for No error, Fluency (A), Alignment(B), Information (C), and Simplification (D) categories, with F_1 and AUC-PR.

Team/Method	No Error		A		B		C		D	
	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
DSGT DebertaLlmensemble	0.763	0.561	0.283	0.133	0.354	0.173	0.301	0.156	0.374	0.224
AIIRLab paraphrase_mpnet	0.755	0.567	0.255	0.154	0.258	0.113	0.136	0.084	0.147	0.168
DSGT roberta	0.694	0.491	0.233	0.121	0.249	0.101	0.114	0.089	0.128	0.164
UBOnlp GPT-4o	0.680	0.505	0.322	0.150	0.381	0.192	0.250	0.122	0.292	0.136
DSGT llama	0.680	0.483	0.282	0.132	0.324	0.182	0.269	0.147	0.306	0.196
AIIRLab OpenChat	0.640	0.421	0.154	0.070	0.141	0.061	0.144	0.080	0.222	0.156
AIIRLab MajorityVoting	0.633	0.415	0.156	0.071	0.110	0.045	0.170	0.088	0.239	0.160
DSGT BERT	0.515	0.330	0.214	0.133	0.208	0.103	0.167	0.095	0.129	0.161
DUTH scibert	0.436	0.321	0.088	0.045	0.035	0.025	0.100	0.066	0.145	0.135
Mtest bartfinetuned	0.404	0.322	0.270	0.143	0.472	0.265	0.078	0.074	0.128	0.167
DSGT bert_llama_ensemble	0.404	0.322	0.231	0.137	0.253	0.107	0.116	0.088	0.128	0.163
RECAIDSTechTitans T5	0.404	0.322	0.022	0.046	0.000	0.026	0.004	0.065	0.000	0.136
DUTH logreg	0.404	0.322	0.000	0.044	0.000	0.026	0.000	0.064	0.000	0.136

training. Our results demonstrate both strengths and limitations of general-purpose language models for specialized text processing tasks. In text simplification, GPT-4o achieved competitive SARI scores (42.20 sentence-level, 43.37 document-level) through a conservative strategy that prioritized content reduction over lexical substitution. For controlled creativity, the model excelled in fluency error detection (highest F1 score among participants) and alignment error detection, but struggled with spurious generation detection, particularly in post-hoc scenarios without source documents. These findings highlight that while GPT-4o demonstrates strong linguistic capabilities for quality assessment tasks, task-specific architectures remain superior for comprehensive error detection and generation control. The substantial performance gap between our approach and specialized systems indicates that domain-specific fine-tuning or architectural modifications are necessary for optimal performance in critical applications. Future work should explore hybrid approaches combining the linguistic sophistication of large language models with the precision of specialized architectures. Our results underscore the importance of careful evaluation when deploying general-purpose language models in specialized domains where accuracy and reliability are essential.

Acknowledgments

This research was funded by the French National Research Agency (ANR) under the projects ANR-22-CE23-0019-01 and ANR-19-GURE-0001 (program *Investissements d'avenir* integrated into France 2030).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Claude in order to: Grammar and spelling check, Paraphrase and reword, and Drafting content.

References

[1] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the clef 2024 simpletext track: Improving access to scientific texts for everyone, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September

9–12, 2024, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2024, p. 283–307. URL: https://doi.org/10.1007/978-3-031-71908-0_13. doi:10.1007/978-3-031-71908-0_13.

[2] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, S. Auer, J. Kamps, CLEF 2024 SimpleText Track, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 28–35. doi:10.1007/978-3-031-56072-9_4.

[3] L. Ermakova, S. Bertin, H. McCombie, J. Kamps, Overview of the clef 2023 simpletext task 3: Simplification of scientific texts, Overview of the CLEF 2023 SimpleText Task 3 (2023).

[4] L. Ermakova, et al., Overview of CLEF 2025 SimpleText Track: Simplify Scientific Texts (and Nothing More), in: J. C. de Albornoz, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer-Verlag, 2025.

[5] J. Bakker, et al., Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, in: G. Faggioli, et al. (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[6] B. Vendeville, et al., Overview of the CLEF 2025 SimpleText Task 2: Identify and Avoid Hallucination, in: G. Faggioli, et al. (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[7] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, Perelman, et al., GPT-4o System Card, 2024. doi:10.48550/arXiv.2410.21276. arXiv:2410.21276.

[8] 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. doi:10.18653/v1/D19-3009.

[9] J. P. Kincaid, Jr. Fishburne, R. Robert P., C. Richard L., Brad S., Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:, Technical Report, Defense Technical Information Center, Fort Belvoir, VA, 1975. doi:10.21236/ADA006655.

[10] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: A Method for Automatic Evaluation of Machine Translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

[11] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing Statistical Machine Translation for Text Simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. doi:10.1162/tacl_a_00107.

[12] B. Vendeville, L. Ermakova, P. D. Loor, Resource for Error Analysis in Text Simplification: New Taxonomy and Test Collection, 2025. doi:10.1145/3726302.3730304. arXiv:2505.16392.