University of Amsterdam at the CLEF 2025 SimpleText Track

Taiki Papandreou, Jan Bakker and Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper reports on the University of Amsterdam's participation in the CLEF 2025 SimpleText track. We participated in Task 1 both for sentence-level and document-level text simplification. We explored scientific text simplification using BART fine-tuning and jargon-aware prompting with LLaMA 3.1. Our plan-guided BART model achieved the highest SARI score at the sentence level, while long input document-level text simplification approaches scored close. LLaMA performed competitively without domain-specific training, highlighting the promise of large language models for zero-shot simplification. More generally, document-level coherence and handling of domain-specific terms remain key challenges for future work.

Keywords

Information Storage and Retrieval, Natural Language Processing, Text Simplification, Jargon Detection, Overgeneration Detection

CEUR-WS.org/Vol-4038/paper_359.pdf

1. Introduction

The rise of the internet and social media has granted us access to an extraordinary amount of information, but it also brings significant risks, particularly the rapid spread of misinformation and disinformation. Scientific knowledge has long been regarded as the most effective counter to such falsehoods, and the importance of scientific literacy is widely acknowledged. However, in reality, many non-experts shy away from scientific sources, often perceiving them as too complex. Therefore, it is crucial to eliminate barriers that prevent the general public from engaging with scientific texts.

The CLEF 2025 SimpleText track investigates the barriers ordinary citizens face when accessing scientific literature head-on, by making available corpora and tasks to address different aspects of the problem. For details on the exact track setup, we refer to the Track Overview paper CLEF 2025 LNCS proceedings [1] as well as the detailed task overviews in the CEUR proceedings [2, 3].

We conduct an extensive analysis of the three tasks of the track: Task 1 on *Text Simplification*; Task 2 on *Controlled Creativity*; and Task 3 on *SimpleText 2024 Revisited.* We submitted multiple runs for Task 1, focusing on both sentence- and document-level simplification approaches. We also submitted runs for Task 2, although they are closely related to Task 1. No runs were submitted for Task 3.

The rest of this paper is structured as follows. Next, in Section 2, we discuss our experimental setup and the specific runs submitted. Section 3 discusses the results of our runs and provides a detailed analysis of the corpus and results for each task. We end in Section 4 by discussing our results and outlining the lessons learned.

2. Experimental Setup

In this section, we will detail our approach for the CLEF 2025 SimpleText track tasks.

CLEF 2025 Working Notes, 9–12 September 2025, Madrid, Spain

j.bakker@uva.nl (J. Bakker); kamps@uva.nl (J. Kamps)
0009-0002-9085-8491 (J. Bakker); 0000-0002-6614-0087 (J. Kamps)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Table 1
CLEF 2025 SimpleText Track Submissions, official run names start with UvA_task11_... or UvA_task12_...

Task	Run	Description
1.1	bartsent-cochraneauto	Sentence-level BART model trained on Cochrane-auto
1.1	o-bartsent-cochraneauto	Plan-guided sentence-level BART model trained on Cochrane-auto
1.1	llama31	Jargon aware prompt with llama3.1-8B-instruct model (sentence level)
1.2	bartpara-cochraneauto	Paragraph-level BART model trained on Cochrane-auto
1.2	bartdoc-cochraneauto	Document-level BART model trained on Cochrane-auto
1.2	baseline-cochrane	Document-level BART baseline trained on Cochrane corpus
1.2	llama31	Jargon aware prompt with llama3.1-8B-instruct model (document level)

2.1. Experimental Data

For details of the exact task setup and results, we refer the reader to the detailed overview of the track in [1]. Our focus is on text simplification (Tasks 1.1, 1.2, and 2.3), and the basic ingredients of the track are:

Corpus The new CLEF 2025 SimpleText corpus is based on biomedical literature abstracts and lay summaries from Cochrane systematic reviews, and is called Cochrane-auto [4].

Train data The specific train data for Task 1 consists of 1,085 documents, 4,171 paragraphs, and 14,719 sentences, with paired content from the abstract and the plain language summary.

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

References There are two sets of references for the new Cochrane data in the test set. First, a subset of 37 abstracts and 587 sentences, paired with 37 plain language summaries with 388 sentences, aligned and filtered as in Cochrane-auto [4]. Second, the full set of 217 abstracts with 4,293 source sentences, paired with 217 plain language summaries with 3,641 sentences, contained document level pairs of the *results and conclusions* sections [5].

We used a number of additional resources for our jargon aware simplification approaches.

Additional Sources For Task 1, we used the MedReadMe training set [6] for jargon detection, which was used as a part of our prompt to simplify text during inference.

Additional Train References For Task 1, the training set of MedReadMe [6] was used for jargon detection. This dataset contains 2,587 annotated sentences with a total of 5,207 jargon terms, sourced from 15 established medical simplification resources such as Cochrane Plain Language Summaries, NIH MedlinePlus articles, and clinical guideline adaptations from professional associations. Annotations were created by undergraduate students without medical training, simulating layperson comprehension challenges. Each sentence is labeled using a hierarchical classification scheme distinguishing between:

- Binary: jargon vs. non-jargon
- 3-class: medical jargon, general/multisense terms, abbreviations
- 7-class: including Google-Easy/Hard distinctions

We used the Roberta-large model trained on binary labeled training data since the detection-rate was the highest.

2.2. Official Submissions

We created runs for both tasks of the track, which we will discuss in order.

Table 2
Prompt used for runs UvA_task11_llama31 and UvA_task12_llama31

STRICT_INSTRUCTIONS
Remember: Simplify language but keep ALL medical details accurate.
- Keep exact numbers
- Keep medical patterns (like 'myoclonic' if no exact simple equivalent exists)
Replace these terms ONLY if you can keep their exact medical meaning:
Detected jargon terms listed here
Text to simplify: *Original text*
Write one simplified sentence:

Task 1 *This task asks simplify scientific text.* We submitted seven runs in total, for both the sentence-level (1.1) and document level (1.2) tasks, as shown in Table 1.

Five of our runs were created using the trained BART models that we introduced in the Cochrane-auto paper [4]. The baseline was trained on the document pairs in the original Cochrane corpus [5], while the other models were trained on the sentence, paragraph, and document pairs in Cochrane-auto. Our plan-guided system is inspired by the work of Cripwell et al. [7]. It consists of a classifier that specifies how each sentence should be simplified—should it be copied, rephrased, split, merged, or deleted?—and a BART model that simplifies each sentence conditioned on the predicted simplification action.

For the submissions UvA_task11_llama31 and UvA_task12_llama31, we used the trained Roberta-large model to detect jargon terms present in every sentence of each abstract. A Llama-3.1-8B-Instruct[1] model was then prompted to simplify each abstract either sentence by sentence or completely, replacing the detected jargon terms where possible. The prompt was designed to preserve numerical values and essential terminology while allowing lexical simplifications where possible to prevent hallucination. Hyperparameters included a temperature of 0.3, top-p sampling of 0.95, repetition penalty of 1.3, and max new tokens set to 512. For document-level processing, we used NLTK to split abstracts into sentences, simplified each independently, and reassembled the output. Post-processing of noisy outputs played a key role in improving clarity and factual consistency.

For both sentence level and document level text simplification, we used the prompt in Table 2. We first detected jargon terms as described above, and then we provided the detected terms to the prompt. The defined instructions helped us to keep the correct information.

Task 2 *This task asks identify and avoid hallucination.* Indirectly, our submissions to Task 1 above can also be evaluated in terms of the evaluation measures of Task 2.3. Hence, in a sense, we submitted the same runs as already shown in Table 1.

Task 3 *This task asks selected tasks by popular request.* We did not do specific experiments for this Task, but the Task 1 test sentences and abstracts include the sources of the CLEF 2024 Simplify Scientific Text Task. Hence, the Task 1 submissions as shown in Table 1 can also be evaluated in terms of their out-of-domain effectiveness against the CLEF 2024 reference simplifications.

3. Experimental Results

In this section, we will present the results of our experiments in three self-contained subsections following the CLEF 2025 SimpleText Track tasks.

3.1. Task 1: Text Simplification

We discuss our results for Task 1, asking to simplify scientific text. They are shown in Table 3.

---

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Table 3
Results for CLEF 2025 SimpleText

Run			SARI (auto)	SARI (original)
UvA	task11	bartsent-cochraneauto	41.3	38.7
UvA	task11	o-bartsent-cochraneauto	42.3	38.5
UvA	task11	llama31	40.9	38.7
UvA	task12	bartpara-cochraneauto	35.0	37.9
UvA	task12	bartdoc-cochraneauto	37.3	37.1
UvA	task12	baseline-cochrane	42.1	41.8
UvA	task12	llama31	38.3	35.6

Complex paragraph
There is some evidence that thrombolysis can probably improve outcomes after aneurysmal subarachnoid haemorrhage, without increasing the risk of haemorrhagic complications. Thrombolysis likely reduces the risk of poor functional outcome and cerebral artery vasospasm, and may reduce the risk of delayed cerebral ischaemia, but it likely makes little to no difference to case fatality or hydrocephalus, and may make little to no difference to the risk of cerebral infarction. However, the current evidence is still uncertain. The uncertainty is primarily due to the small total number of participants and outcome events. Data from further studies are required to confirm the efficacy of thrombolysis for improving outcomes after aneurysmal subarachnoid haemorrhage.

Simplification by jargon-aware Llama model
Some studies suggest that a treatment called thrombolytic therapy may help patients recover better from bleeding in the brain caused by a ruptured blood vessel, without causing further bleeding problems. This treatment might lower chances of: * Poor recovery * Blood vessel narrowing in the brain (after a stroke) * Brain damage from lack of blood flow (later on) It probably won't affect how many people die after getting this condition. It also doesn't seem to greatly change the chance of another blockage happening in the same area of the brain. The facts aren't clear yet. There aren't enough people in the study for a reliable answer. More research is needed to see if a treatment called thrombolysis can help people recover better from bleeding in their brain caused by a weak blood vessel bursting.

Table 4
The last paragraph of an abstract in the CLEF 2025 SimpleText test set, along with the corresponding simplification from UvA_task11_llama31.

3.1.1. Task 1.1 Sentence-level Simplification

When computed against the references in the newly created Cochrane-auto test set, our plan-guided system achieves the highest SARI score, but it is only very slightly higher than that of the baseline. This indicates that training on Cochrane-auto does not offer a substantial advantage over training on the Cochrane corpus for document-level simplification.

As for the jargon-aware prompt method, despite not being trained on the text simplification domain dataset, it achieves a similar SARI score compared to the ones trained on the Cochrane-auto corpus.

3.1.2. Task 1.2 Document-level Simplification

The initial results for the BART models trained at the paragraph and document levels further demonstrate that training on Cochrane-auto rather than the Cochrane corpus does not improve performance and may even harm the SARI score. The Llama model that simplifies entire abstracts also achieved a lower SARI score than the Llama model operating at the sentence level.

3.1.3. Analysis

Table 4 displays the last paragraph of an abstract in the CLEF 2025 SimpleText test set, which is dense with jargon. It also shows the output of the LLaMa-3.1 model that simplified this abstract sentence-

Table 5
Analysis of SimpleText Task 2.3: Spurious generation at the sentence (top) and document (bottom) level

Run	SARI	Source	Spurious Content	
	(217)	Number	Number	Fraction
UvA_Task11_bartsent-cochraneauto	38.7	9,160	6	0.00
UvA_Task11_o-bartsent-cochraneauto	38.5	9,160	1	0.00
UvA1_llama31	38.7	9,160	4226	0.46
UvA_Task12_baseline-cochrane	41.8	666	182	0.27
UvA_Task12_bartdoc-cochraneauto	37.1	666	103	0.15
UvA_Task12_bartpara-cochraneauto	37.9	666	44	0.07
taiki_task12_llama31	35.6	666	603	0.91

by-sentence, conditioned on the jargon terms as detected by the trained RoBERTa classifier. It can be seen that the model successfully simplifies various jargon terms such as *aneurysmal subarachnoid haemorrhage* into easier-to-read alternatives like *bleeding in the brain caused by a ruptured blood vessel.* Thus, the meaning of the original paragraph is preserved, while the text is made more accessible to a general audience. However, *thrombolysis* is substituted with *a treatment called thrombolysis* twice, while it would be sufficient to mention that thrombolysis is a treatment once. This is a result of simplifying each sentence in isolation.

3.2. Task 2: Controlled Creativity

3.2.1. Task 2.3 Sentence-level Simplification

We continue with Task 2, asking to identify and avoid hallucination. While we did not submit a special pair of runs with and without particular grounding by design components, our Task 1 submission did take special care to avoid overgeneration or other information distortion.

First, our Cochrane-auto trained models are conservative and avoid gratuitous changes. This may not optimize readability as much as some other approaches, but it leads to an accurate rendition of the content without risk of information distortion. We feel that such a conservative approach is important in the context of scientific text simplification.

Second, our jargon-aware runs attempt to address impenetrable terminology by actively promoting the deletion, rephrasing, or explanation of jargon in the text. This can result in significant content insertion. However, the used models and prompt were prone to ”noise” resulting in potential overgeneration.

3.2.2. Results

We employ a simple alignment of source and prediction sentences, specifically examining overgeneration or noise at the end of the prediction. Suppose the alignment reaches the end of the source tokens, while the prediction still has another sentence (or additional content after the last sentence). In that case, this is flagged as “overgeneraton.” This approach is more reliable at the sentence level, as the alignment and spurious content can be detected with relative ease.

Table 5 shows the results. We observe here that, indeed, the Cochrane-auto runs have marginal overgeneration and are conservative in their edits. we also see that the jargon-aware LLaMA run has a significant fraction of spurious content. While part of this may be due to additional explanations of jargon and helpful, there is also a significant number of cases in which ”noise” or LLM commentary is added. At the document level, the alignment can be tricky due to the length of the abstracts and extensive sentence deletions. Partly due to the smaller number of cases, errors seem more pronounced for document-level text simplification. This may be partly due to the more complex alignment of documents, but also due to complexities in removing ”noise” or spurious content in very long predictions. The relative fraction still serves as a useful indicator of spurious content, and we observe almost twice as much spurious content in the baseline (unaligned) Cochrane train data. The aligned Cochrane-auto models fare much better. The LLaMA model again suffers from a relatively high number of cases with spurious content.

3.3. Task 3: SimpleText 2024 Revisited

We continue with Task 3, asking for selected tasks by popular request. As noted, the CLEF 2024 text simplification test data were included in the CLEF 2025 test corpus. Hence, the performance of the exact same models on a different domain can be evaluated.

We leave this evaluation and analysis for future research, as the track organizers have not yet released the references and evaluation for the additional abstracts and sentences in the test set.

For reference, similar Cochrane-trained BART models were submitted to the CLEF 2024 Simpletext Track last year [8]. On the scientific abstracts on technology and AI, these models obtained SARI scores of 26.7 (sentence level), 33.2 (document level), and 35.1 (paragraph level). These scores in another domain are notably lower than those in the biomedical domain this year, possibly also due to the sentence-level references.

For further reference, similar Cochrane-trained sentence-level BART and Contextual BART models were submitted to the TREC 2024 PLABA Track last year [9]. On the Medline abstracts, these models obtained SARI scores of 28.8 (sentence level BART), and 30.5 (sentence level BART with whole abstract as context). These scores are also notably lower than those observed this year, possibly due to the different nature of Medline abstracts and the choice to run planner-based models on a one-to-one sentence simplification task.

4. Discussion and Conclusions

This paper detailed the University of Amsterdam's participation in the CLEF 2025 SimpleText track. We conducted a range of experiments for the different tasks of the track.

Our primary focus was on the core Task 1 on *Text Simplification*, where we evaluated multiple approaches to scientific text simplification, including BART-based fine-tuning and jargon-aware prompting with LLaMA 3.1. Our plan-guided BART model achieved the highest SARI score on sentence-level simplification, indicating that structured simplification actions can slightly improve performance. However, training on the Cochrane-auto dataset did not significantly outperform the baseline trained on the original Cochrane corpus, especially at the document level. The LLaMA-based method performed competitively without domain-specific training, demonstrating the potential of large language models in zero-shot simplification when guided by jargon detection and structured prompts. These results suggest that while current models are effective at sentence-level simplification, maintaining coherence and factual accuracy across longer texts remains a challenge. Future work should focus on better discourse modeling and more robust handling of domain-specific terminology.

Acknowledgments

This research was conducted as part of the final research project for the Master's in Artificial Intelligence program at the University of Amsterdam.

We thank the track and task organizers for their amazing service and effort in making realistic benchmarks for scientific text simplification available.

Jan Bakker is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

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

References

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

[2] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, in: [10], 2025.

[3] B. Vendeville, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 2: Identify and Avoid Hallucination, in: [10], 2025.

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

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

[6] C. Jiang, W. Xu, MedReadMe: A systematic study for fine-grained sentence readability in medical domain, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17293–17319. URL: https://aclanthology.org/2024.emnlp-main.958/. doi:`10.18653/v1/2024.emnlp-main.958`.

[7] L. Cripwell, J. Legrand, C. Gardent, Document-level planning for text simplification, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 993–1006. URL: https://aclanthology.org/2023.eacl-main.70/. doi:`10.18653/v1/2023.eacl-main.70`.

[8] J. Bakker, G. Yüksel, J. Kamps, University of amsterdam at the CLEF 2024 simpletext track, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 3182–3194. URL: https://ceur-ws.org/Vol-3740/paper-310.pdf.

[9] J. Bakker, T. Papandreou-Lazos, J. Kamps, Biomedical text simplification models trained on aligned abstracts and lay summaries, in: I. Soboroff, G. Awad, H. T. Dang, A. Ellis (Eds.), The Thirty-third Text REtrieval Conference Proceedings (TREC 2024), National Institute for Standards and Technology. NIST Special Publication 1329, 2025.

[10] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.