NUQLab at SimpleText 2026: Plan-Guided Biomedical Scientific Text Simplification

Shimaa Ibrahim[1], Wajdi Zaghouani[1]

[1]Northwestern University in Qatar, Education City, Doha, Qatar

Abstract

This paper describes the NUQLab submission to the CLEF 2026 SimpleText shared task on scientific text simplification, covering both Task 1.1 (sentence-level simplification) and Task 1.2 (document-level simplification) in the biomedical domain. Our system adopts a fully inference-based, plan-guided pipeline built on pretrained Cochrane-auto checkpoints without any additional fine-tuning. The core architecture consists of a RoBERTa-based operation classifier that predicts sentence-level simplification actions (*rephrase*, *split*, *merge*, *ignore*, *delete*) and a BART-based generator that produces simplifications conditioned on operation-specific control tokens. For document-level simplification, we extend this design with a biomedical-aware sentence splitter that protects statistical expressions, p-values, and abbreviations, followed by a document recomposition stage. We additionally include a tokenizer-based diagnostic that detects control-token failures at load time and falls back to unconstrained generation. A plain BARTsent configuration is retained as a secondary fallback. On the official CLEF 2026 evaluation, our plan-guided system obtains a SARI of **38.96** on Task 1.1 and **38.97** on Task 1.2. On our offline Cochrane-auto validation subset (non-delete rows), the same Task 1.1 system reaches a SARI of 51.96, above the plain BARTsent fallback (~30–32) and an earlier plan-guided run (~35.5). We further discuss the extent to which the predicted operation is realized by the generator and the practical case for a lightweight, non-fine-tuned pipeline relative to larger, more expensive LLMs.

Keywords

scientific text simplification, biomedical NLP, controllable generation, plan-guided simplification

1. Introduction

Scientific text simplification aims to rewrite specialized writing in a more accessible form while preserving meaning. In the biomedical domain, research abstracts typically contain dense technical terminology, statistical expressions, and highly compressed discourse that are difficult for non-expert readers to process. The CLEF SimpleText initiative addresses this challenge through shared tasks and benchmark datasets focused on improving access to scientific information [1, 2, 3, 4, 5, 6]. The 2026 edition requires simplification at both sentence level (Task 1.1) and document level (Task 1.2), with evaluation on the Cochrane-auto biomedical corpus [7, 8].

A central challenge in this setting is producing simplifications that are *controlled*: systems that generate freely tend to introduce unsupported or overly creative content, as highlighted by the CLEF 2025 hallucination track [9]. Our system responds to this challenge by adopting a plan-guided approach in which a sentence-level classifier first predicts a simplification operation and the generator is then conditioned on the predicted control token. This design makes the simplification process more structured and interpretable compared to unconstrained sequence-to-sequence generation.

The system is intentionally lightweight: we reuse released pretrained Cochrane-auto models without any additional fine-tuning, making the pipeline reproducible from public checkpoints. The main contributions of this paper are:

- a unified plan-guided inference pipeline for both sentence- and document-level biomedical simplification;
- adaptation of Cochrane-auto checkpoints to the CLEF 2026 SimpleText Task 1 without further training;

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

shimaa.ibrahim@northwestern.edu (S. Ibrahim); wajdi.zaghouani@northwestern.edu (W. Zaghouani)

0009-0000-2044-2306 (S. Ibrahim); 0000-0003-1521-5568 (W. Zaghouani)

Table 1
Pretrained models used in our system.

Role	Checkpoint	Architecture
Operation classifier	`janbakker/classifier-cochraneauto` [15]	RoBERTa
Plan-guided generator	`janbakker/o-bartsent-cochraneauto` [16]	BART
Fallback generator	`janbakker/bartsent-cochraneauto` [17]	BART

- practical inference-time mechanisms for robust generation, including delete handling, a tokenizer-based control-token diagnostic, and biomedical-aware document splitting.

2. Related Work

The CLEF SimpleText lab has progressively expanded its focus on scientific text simplification across several editions [1, 2, 3, 4]. The 2025 and 2026 tracks promote simplification to the primary task and explicitly require that systems preserve factual correctness without introducing unsupported content [5, 10, 6, 8].

The Cochrane-auto corpus, introduced by Bakker and Kamps [7], provides aligned biomedical abstract–plain-language-summary pairs at sentence, paragraph, and document level. Unlike general-domain Wikipedia-based simplification resources [11], Cochrane-auto reflects the terminology and discourse structure of biomedical communication and directly supports the SimpleText 2025/2026 evaluation. Subsequent work on section-level biomedical simplification further demonstrates the importance of going beyond isolated sentence rewriting [12].

Controllable simplification is a central thread in this literature. The CLEF 2025 Task 2 overview motivates methods that constrain generation to avoid hallucination [9]. Our plan-guided approach connects directly to this line: rather than training new models, we direct a pretrained generator through explicit operation prediction, restricting its output space without any additional supervision.

3. System Overview

3.1. Architecture

Our system addresses both subtasks using a two-stage plan-guided pipeline. In the first stage, a RoBERTa-based [13] operation classifier predicts a simplification action for each input sentence. In the second stage, the predicted label is converted into a control token that is prepended to the source sentence and passed to a conditional BART [14] generator. All models are used in inference mode only; no weights are updated. Table 1 lists the pretrained checkpoints.

The classifier exposes generic output labels (LABEL_0–LABEL_4), which we map manually to the underlying operation space: LABEL_0→*ignore*, LABEL_1→*rephrase*, LABEL_2→*split*, LABEL_3→*merge*, LABEL_4→*delete*. The mapping is verified at load time using validation-set label-distribution inspection.

3.2. Plan-Guided Generation

For each non-delete sentence, the predicted operation is converted to a control token and prepended to the source:

`<operation_token> + source sentence.`

Control tokens are detected automatically from the generator tokenizer by checking both plain forms (e.g., `rephrase`) and bracketed forms (e.g., `<rephrase>`). A diagnostic check at load time verifies that distinct operations resolve to distinct tokens; if all operations collapse to the same token, plan-guided mode is disabled and the system falls back to the plain BARTsent generator. Non-English inputs are copied unchanged, since the released checkpoints are English biomedical models. The control

token is used as a soft conditioning signal rather than as a hard decoding constraint; therefore, the generated output may sometimes simplify the sentence without fully realising the predicted structural operation. Delete-predicted sentences are assigned an empty string and excluded from the generation step. Generated outputs of three words or fewer are treated as degenerate and replaced by the original source sentence.

3.3. Document-Level Extension (Task 1.2)

For Task 1.2, each source document is first split into sentences using a biomedical-aware splitter built on NLTK Punkt [18]. Before tokenization, the splitter protects patterns that are frequently mis-segmented in biomedical text: p-values, confidence intervals, statistical abbreviations, common biomedical abbreviations, and decimal numbers. After tokenization, placeholders are restored. Very short fragments are merged into the preceding sentence, and segments shorter than two words are discarded. If NLTK fails, the splitter falls back to a regex-based routine.

Each resulting sentence is then processed by the same classify–generate pipeline as Task 1.1, and the simplified outputs are concatenated with a single space to form the final document. Delete-predicted sentences are omitted from the recomposed document. A configurable `DELETE_OUTPUT` switch governs the degenerate case in which every sentence of a document is predicted as delete: the submitted Task 1.2 run uses the `source` setting, which falls back to the first source sentence and thereby avoids empty document-level predictions, whereas the `empty` setting (used for the Task 1.1 submission) emits an empty string for such cases.

3.4. Offline Validation

SARI [19] is used for offline validation following the task setup [8]. For Task 1.1, SARI is computed on the Cochrane-auto validation rows with non-empty reference simplifications (the non-delete rows of the primary-source subset); degenerate generations of three words or fewer are replaced by the source sentence to preserve the keep component. For Task 1.2, SARI is computed over all 119 validation documents, with the source document substituted for any all-delete document.

4. Experimental Setup

Data. Task 1.1 uses a test file of 48,809 sentences from 1,734 documents. Of these, 20,259 (41.5%) are English and 28,550 (58.5%) are non-English. The primary evaluation subset is the English `Cochrane-auto 2026` portion (3,679 sentences). Task 1.2 uses a test file of 8,069 documents; the primary evaluation subset contains 175 English Cochrane-auto documents. The validation split for Task 1.2 contains 119 documents.

Implementation. All experiments run on a single NVIDIA T4 GPU in Google Colab using PyTorch and Hugging Face Transformers. The random seed is set to 42. Table 2 summarizes the inference settings.

5. Results and Discussion

5.1. Task 1.1: Sentence-Level Simplification

Table 3 summarizes the validation SARI scores across configurations.

On the official CLEF 2026 evaluation, the submitted plan-guided system achieves a SARI of **38.96**. On our offline Cochrane-auto validation subset, the same system reaches a SARI of 51.96 on the non-delete rows, above both the plain BARTsent fallback and an earlier plan-guided run without the full control-token diagnostic; the gap between the offline and official figures reflects the broader, multilingual

Table 2
Inference settings for both subtasks.

Setting	Task 1.1	Task 1.2
Classifier batch size	64	64
Classifier max length	128	128
Generator batch size	32	32
Generator max input length	256	256
Generator max new tokens	150	128
Beam width	8	5
Length penalty	1.5	1.0
Degenerate fallback	source if $\leq$3 words	

Table 3
SARI scores for Task 1.1. Offline validation is computed on the non-delete Cochrane-auto validation rows; the official score is the CLEF 2026 leaderboard result on the full multilingual test set.

System	Evaluation setting	SARI
Plain BARTsent (no control tokens)	Offline validation	≈30–32
Plan-guided (earlier run)	Offline validation	≈35.5
Plan-guided (this system)	Offline validation	51.96
Plan-guided (this system)	Official leaderboard	**38.96**

composition of the official test set. The submission file contains all 48,809 test rows; 13.9% are empty (delete-predicted) outputs, and the `source` and `language` fields are echoed back correctly.

Three design choices account for the good result. First, explicit operation prediction before generation makes simplification more structured than unconstrained sequence-to-sequence generation. Second, the control-token diagnostic prevents silent plan-guided failures by reverting to plain BART when the tokenizer cannot distinguish operations. Third, the English-only restriction prevents the English biomedical model from being applied to non-English inputs, avoiding noisy cross-lingual predictions.

A notable characteristic of the output is that delete predictions are common: the final submission contains thousands of empty outputs (13.9%). While this increases brevity, overuse of deletion can reduce informativeness. The plain BARTsent fallback configuration is provided to investigate this trade-off in future runs.

5.2. Task 1.2: Document-Level Simplification

On the official CLEF 2026 evaluation, the document-level system achieves a SARI of **38.97**, essentially on par with the sentence-level Task 1.1 result. This is obtained with the same plan-guided classify–generate backbone, extended with the biomedical-aware sentence splitter and single-space recomposition described in Section 3. The submitted run uses the `source` delete-handling setting, so fully-deleted documents fall back to their first source sentence and the submission contains no empty document-level predictions. Offline validation over the 119 Cochrane-auto validation documents was used during development to confirm the pipeline before submission.

5.3. Discussion

Across both subtasks, our system demonstrates that pretrained plan-guided biomedical simplification checkpoints can be transferred effectively to the CLEF 2026 benchmark without any fine-tuning. The key enablers are careful control-token handling, a tokenizer-based failure diagnostic, and biomedical-aware sentence segmentation for the document task.

The 2026 benchmark is substantially more challenging than earlier English-only editions because 58.5% of the test sentences are non-English. Our conservative decision to copy non-English inputs

unchanged is appropriate given that the released checkpoints are English-only models; it reduces the risk of unsupported cross-lingual generation at the cost of leaving non-English simplification to the evaluation protocol.

For Task 1.2, the custom biomedical-aware splitter is a critical preprocessing component: protecting p-values, confidence intervals, abbreviations, and decimal numbers prevents many spurious boundaries that would otherwise propagate errors into the classification and generation stages.

5.4. Plan Adherence and Comparison with Larger LLMs

Two related issues are important for interpreting the behaviour of the proposed pipeline: whether the predicted plan is actually followed by the generator, and whether a pretrained, non-fine-tuned pipeline of this kind remains a useful alternative to larger, more expensive LLMs.On the first point, the control token conditions the generator rather than constraining its decoding (Section 3), so plan adherence is not architecturally guaranteed: the generator can produce a fluent simplification while only partially realizing the requested operation, particularly for the structurally more demanding *split* and *merge* labels, since these require the model to reorganize sentence boundaries rather than substitute or drop content. We did not run a systematic manual audit of plan adherence for this submission, so we report this as an expected consequence of the soft-conditioning design rather than as a measured error rate; a labeled audit of a sample of Task 1.1 and Task 1.2 outputs, checking whether the realized operation matches the predicted one, is a natural next step and would let us report concrete adherence figures rather than a qualitative expectation.

On the second point, the case for a lightweight, non-fine-tuned pipeline over a modern LLM does not rest on output fluency, where larger models are likely to have an advantage, but on reproducibility and cost: the system runs entirely from public checkpoints on a single T4 GPU, requires no additional training or external API calls, and its behavior (including the delete rate and control-token resolution) is fully inspectable at inference time. This makes it a useful, low-resource baseline for settings where reproducibility and local inference matter, even if a larger LLM might improve fluency or close part of the remaining evaluation gap.

6. Conclusion

We presented the NUQLab system for CLEF 2026 SimpleText Task 1, covering both sentence-level and document-level biomedical simplification. Our approach reuses pretrained Cochrane-auto plan-guided models without fine-tuning, combining a RoBERTa operation classifier with a BART generator conditioned on predicted control tokens. Practical inference-time mechanisms – control-token diagnostics, delete handling, source-sentence fallback, and biomedical-aware sentence splitting – make the system robust and reproducible from public checkpoints.

On the official CLEF 2026 evaluation the system scores 38.96 SARI on Task 1.1 and 38.97 on Task 1.2, with a higher 51.96 SARI on the non-delete Cochrane-auto validation subset used during development. These results suggest that structured operation prediction can provide a useful and reproducible way to guide biomedical simplification without additional training, although the control signal is a soft conditioning mechanism rather than a hard constraint, so plan adherence is not guaranteed and delete overuse remains a risk. Future work will explore stronger multilingual support, a systematic audit of plan adherence, and more faithful document-level control mechanisms beyond sentence-wise recomposition.

Acknowledgments

This work was funded by the NPRP-C Grant No. 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings and conclusions expressed in this paper are solely those of the authors.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) and ChatGPT (OpenAI) to assist with grammar and style editing, paraphrasing and rewording, and reviewer-response editing. The authors reviewed and edited all content and take full responsibility for the publication's content.

References

[1] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 432–449. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_27. doi:10.1007/978-3-030-85251-1_27.

[2] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, E. Mathurin, P. Bellot, Overview of the CLEF 2022 SimpleText lab: Automatic simplification of scientific texts, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://link.springer.com/chapter/10.1007/978-3-031-13643-6_28. doi:10.1007/978-3-031-13643-6_28.

[3] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 482–506. URL: https://link.springer.com/chapter/10.1007/978-3-031-42448-9_30. doi:10.1007/978-3-031-42448-9_30.

[4] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. D. 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, volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 283–307. URL: https://link.springer.com/chapter/10.1007/978-3-031-71908-0_13. doi:10.1007/978-3-031-71908-0_13.

[5] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText track: Simplify scientific text (and nothing more), in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 436–463. URL: https://doi.org/10.1007/978-3-032-04354-2_23. doi:10.1007/978-3-032-04354-2_23.

[6] L. Ermakova, H. Azarbonyad, J. Bakker, B. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026. To appear.

[7] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: 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.

[8] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText task 1: Simplify scientific text, in: Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026. To appear.

[9] B. Vendeville, J. Bakker, H. Azarbonyad, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText task 2: Identify and avoid hallucination, in: Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 4186–4204. URL: https://ceur-ws.org/Vol-4038/paper_345.pdf.

[10] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText task 1: Simplify scientific text, in: Working Notes of CLEF 2025: Conference and Labs of the Evaluation

Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 4167–4185. URL: https://ceur-ws.org/Vol-4038/paper_344.pdf.

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

[12] J. Bakker, J. Kamps, Section-level simplification of biomedical abstracts, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13819–13833. URL: https://aclanthology.org/2025.emnlp-main.697/. doi:10.18653/v1/2025.emnlp-main.697.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703/. doi:10.18653/v1/2020.acl-main.703.

[15] J. Bakker, janbakker/classifier-cochraneauto, https://huggingface.co/janbakker/classifier-cochraneauto, 2024. Hugging Face model card. Classifier pretrained on Cochrane-auto. Accessed 2026-05-10.

[16] J. Bakker, janbakker/o-bartsent-cochraneauto, https://huggingface.co/janbakker/o-bartsent-cochraneauto, 2024. Hugging Face model card. Plan-guided sentence-level BART model pretrained on Cochrane-auto. Accessed 2026-05-10.

[17] J. Bakker, janbakker/bartsent-cochraneauto, https://huggingface.co/janbakker/bartsent-cochraneauto, 2024. Hugging Face model card. Sentence-level BART model pretrained on Cochrane-auto. Accessed 2026-05-10.

[18] T. Kiss, J. Strunk, Unsupervised multilingual sentence boundary detection, Computational Linguistics 32 (2006) 485–525. URL: https://aclanthology.org/J06-4003/. doi:10.1162/coli.2006.32.4.485.

[19] 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. URL: https://aclanthology.org/Q16-1029/. doi:10.1162/tacl_a_00107.