A_R Team at CLEF 2026 SimpleText Task 2.1: Source-Aware Grounding Verification with Scientific Transformer Ensembles

Mhd Adnan Albuni[1], Riad Sonbol[2]

[1]Independent Researcher
[2]Higher Institute for Applied Sciences and Technology

Abstract

Large Language Models (LLMs) have substantially improved the quality of scientific text simplification, enabling the generation of more accessible versions of complex scientific content. However, these systems often introduce information that is not supported by the original source text, a phenomenon commonly referred to as overgeneration. Task 2.1 of the CLEF 2026 SimpleText Track focuses on detecting such unsupported content in simplified scientific texts.

We approach this task as a source-aware grounding verification problem, where generated sentences are assessed against their corresponding source abstracts. Our method employs three complementary transformer encoders—ModernBERT, SciBERT, and SPECTER2—which are independently fine-tuned and combined through a Logistic Regression stacking ensemble. To determine whether generated content is grounded in the source document, each model jointly processes the full source abstract and the generated sentence. We also investigate retrieval-based context selection but find that using the complete source abstract consistently yields stronger performance, suggesting that evidence required for verification is often distributed throughout the document rather than concentrated in a small set of sentences.

The stacking ensemble achieved an official sentence-level F1-score of 0.8527 and a document-level F1-score of 0.7852, outperforming the strongest individual model by approximately 4 percentage points on our internal evaluation. These results demonstrate that combining general-purpose and scientific-domain representations provides a robust solution for detecting overgenerated content. Overall, our findings highlight the importance of full-document grounding and the effectiveness of scientific transformer ensembles for reliable verification in scientific text simplification.

Keywords

overgeneration detection, hallucination detection, scientific text simplification, transformer ensembles, CLEF 2026 SimpleText

1. Introduction

Recent advances in Large Language Models (LLMs) have significantly improved automatic text simplification. Modern systems can transform complex scientific texts into more accessible versions while maintaining fluency and readability. Despite these improvements, generated outputs are not always faithful to the source material. Simplification models may introduce unsupported information, generate explanatory comments, leak prompts or instructions, or produce other forms of content that are not grounded in the original document. Such behavior, commonly referred to as overgeneration, presents a significant challenge for trustworthy simplification systems.

The impact of overgeneration is particularly important in scientific domains. Unlike general-purpose text generation, scientific simplification aims to preserve factual accuracy while making content easier to understand. Unsupported additions may distort research findings, misrepresent evidence, or introduce misleading interpretations. Consequently, reliable methods for detecting overgenerated content are essential for evaluating and improving simplification systems.

Task 2.1 of the CLEF 2026 SimpleText Track [1] focuses on identifying overgenerated content in simplified scientific texts [2]. Given a source abstract and a generated simplified sentence, participants

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
Mhd.adnan.404@gmail.com (M. A. Albuni); riad.sonbol@hiast.edu.sy (R. Sonbol)

must determine whether the generated sentence is fully supported by the source document. The task is evaluated at both sentence and document levels, making it necessary to identify local overgeneration instances while maintaining consistent document-level predictions.

Several characteristics make this task challenging. First, the dataset is highly imbalanced, with positive overgeneration examples representing only a small proportion of all training instances. Second, overgeneration appears in multiple forms, including unsupported factual additions, leaked instructions, repetitive content, and generation failures. Finally, the automatic annotation process introduces a degree of label noise, making robust generalization more difficult.

To address these challenges, we propose a source-aware grounding verification framework based on transformer ensembles. Our approach uses the complete source abstract as evidence and evaluates generated sentences through three independently fine-tuned transformer encoders: ModernBERT [3], SciBERT [4], and SPECTER2 [5]. These models provide complementary perspectives on scientific language understanding through their differing pretraining objectives and training corpora. To leverage their complementary strengths, we combine their outputs using a Logistic Regression stacking classifier [6].

During development, we additionally explored retrieval-based context selection, where only the most semantically relevant source sentences were provided to the classifier. While this strategy reduced input length, empirical evaluation showed that preserving the entire source abstract consistently produced stronger results. This observation suggests that overgeneration detection often requires information distributed across multiple parts of the source document rather than isolated local evidence.

The contributions of this work are threefold:

- We investigate source-aware grounding verification for overgeneration detection in scientific text simplification.
- We fine-tune and evaluate three complementary transformer encoders—ModernBERT, SciBERT, and SPECTER2—and combine their predictions using a stacking ensemble.
- We show that full-document grounding is more effective than restricting the model to locally retrieved evidence, highlighting the importance of broader source context for overgeneration detection.

2. Related Work

Hallucination detection and factuality verification have become increasingly important research areas with the rapid adoption of LLMs. Prior work has explored methods for assessing whether generated content is supported by source evidence, particularly in summarization, question answering, and dialogue generation. Many of these approaches frame the problem as a verification task that compares generated content against supporting documents.

Scientific and biomedical NLP has also benefited from the development of domain-specific transformer models. SciBERT [4] demonstrated that pretraining on scientific literature substantially improves performance across a variety of scientific NLP tasks. Similarly, SPECTER [7] introduced citation-informed pretraining objectives that capture semantic relationships between scholarly documents. More recently, ModernBERT [3] has shown strong performance across a broad range of classification tasks through improvements in encoder architecture and training methodology. BERT [8] remains the foundational architecture underlying all these models.

Ensemble learning has long been recognized as an effective strategy for improving robustness and generalization [6]. By combining models with different inductive biases and representations, ensembles often outperform individual classifiers. Motivated by these findings, we investigate whether combining general-purpose and scientific transformer encoders can improve source-aware overgeneration detection.

3. Methodology

3.1. Task Formulation

Task 2.1 is formulated as a binary sentence-level classification problem. Given a source abstract and a generated simplified sentence, the objective is to predict one of two labels: None or Overgeneration. All positive categories provided in the original dataset are merged into the Overgeneration class.

Following the official evaluation protocol, document-level labels are derived from sentence-level predictions. A document is labeled as Overgeneration whenever at least one sentence receives a positive prediction.

3.2. Input Representation

The final system uses the complete source abstract as grounding evidence. For every generated sentence, we construct an input pair consisting of the original source abstract and the generated sentence. The pair is encoded using the following format:

```
[CLS] Source Abstract [SEP] Generated Sentence [SEP]
```

This representation enables the model to jointly process the source evidence and generated content, allowing it to determine whether the generated sentence is supported by the original document.

3.3. Context Selection Experiments

During development, we explored whether reducing source context could improve grounding verification. Source abstracts were segmented into individual sentences, and sentence embeddings [9] were used to compute semantic similarity between source sentences and generated sentences. For each generated sentence, the five most similar source sentences were selected and provided as evidence.

Although this strategy reduced input length and focused the model on locally relevant content, validation experiments consistently indicated that using the complete source abstract yielded stronger performance. Consequently, all submitted systems and the final ensemble were trained using full-document context.

3.4. Transformer-Based Classifiers

We investigate three transformer encoders with complementary characteristics.

ModernBERT. ModernBERT [3] provides strong general-purpose contextual representations and supports sequences of up to 8,192 tokens, allowing it to process the full source abstract without truncation. It serves as the primary encoder in our ensemble.

SciBERT. SciBERT [4] is pretrained on scientific publications and offers domain-specific representations tailored to scientific language. It uses a scientific-domain vocabulary constructed from scratch, which benefits tokenization of technical terminology.

SPECTER2. SPECTER2 [5] captures semantic relationships characteristic of scientific literature and scholarly communication. Its pretraining on citation graphs makes it sensitive to document-level semantic similarity, complementing the token-level representations of the other encoders.

Each model is independently fine-tuned on the official training data using a binary classification head. Focal loss [10] is used during training to address class imbalance.

3.5. Stacking Ensemble

The three transformer models differ substantially in their pretraining objectives and learned representations, and may therefore capture complementary signals associated with overgeneration. To exploit these differences, we employ a stacking ensemble [6]. After fine-tuning, each model produces a probability estimate for the Overgeneration class. The six probability values (two per model) are concatenated and provided as input to a Logistic Regression meta-classifier, which learns the optimal combination and produces the final sentence-level prediction.

3.6. Document-Level Aggregation

Document-level predictions follow the official task definition. If at least one sentence within a document is classified as Overgeneration, the document is assigned a positive label. Otherwise, the document is labeled None.

4. Experimental Setup

4.1. Dataset

Experiments were conducted using the official CLEF 2026 SimpleText Task 2.1 dataset [**?**]. The dataset consists of scientific publications sourced from arXiv, a widely-used open-access repository covering a broad range of academic disciplines including physics, computer science, mathematics, and economics. The collected articles are annotated for overgeneration detection, with sentence-level labels identifying content that is not supported by the original source document. The dataset exhibits substantial class imbalance, with overgeneration examples representing less than 10% of all training sentences.

4.2. Training Configuration

Each transformer model was independently fine-tuned with the following configuration:

- **Learning rate:** 2×10^{-5}
- **Batch size:** 32 per device (ModernBERT: 16 per device with 2-GPU DistributedDataParallel)
- **Epochs:** up to 4 with early stopping (patience = 3, metric: macro F1 on validation set)
- **Maximum sequence length:** 2048 tokens (ModernBERT); 512 tokens (SciBERT, SPECTER2)
- **Optimizer:** AdamW with weight decay 0.01 and warmup ratio 0.1
- **Loss:** Focal loss ($\gamma = 2.0$) with class weights
- **Hardware:** NVIDIA RTX PRO 6000 Blackwell (96 GB VRAM) for SciBERT and SPECTER2; 2× NVIDIA RTX PRO 6000 Blackwell with DistributedDataParallel for ModernBERT

The logistic regression meta-classifier was trained on the held-out validation set probabilities, keeping the individual model parameters frozen.

4.3. Evaluation Metrics

Performance is evaluated using Precision, Recall, and F1-score at both sentence and document levels. Sentence-level evaluation measures the ability to identify individual overgenerated sentences, while document-level evaluation measures the ability to identify documents containing at least one overgenerated sentence.

Table 1
Results for individual models and the stacking ensemble. Prec./Recall refer to the Overgeneration class. Ctx: Full = full source abstract, Top-5 = top-5 retrieved source sentences. Level: Internal = internal held-out test set; Official = CLEF 2026 official evaluation.

Model	Acc.	Macro F1	Prec.	Recall	Ctx	Level
ModernBERT-large	0.9714	0.9129	0.8500	0.9799	Full	Internal
SciBERT	0.9703	0.9064	0.7600	0.9310	Top-5	Internal
SPECTER2	0.9709	0.9074	0.7700	0.9215	Top-5	Internal
Ensemble (LR)	**0.9851**	**0.9488**	**0.8900**	**0.9240**	—	Internal
Ensemble (LR) – Doc F1	—	0.7852	0.8257	0.7484	—	Official
Ensemble (LR) – Sent F1	—	0.8527	—	—	—	Official

5. Results and Discussion

5.1. Main Results

Table 1 reports the performance of the individual transformer models and the stacking ensemble on our internal held-out test set and the official CLEF 2026 test set.

Among the individual models, ModernBERT achieved the strongest performance with a macro F1 of 0.9129 and an Overgeneration recall of 0.9799 on our internal test set, benefiting from its 2048-token context window which allows full-source processing without truncation. SciBERT achieved a macro F1 of 0.9064 with recall of 0.9310, while SPECTER2 achieved 0.9074 with recall of 0.9215. On the official CLEF 2026 test set, the logistic regression ensemble achieved a sentence-level F1 of 0.8527 and a document-level F1 of 0.7852 (Precision: 0.8257, Recall: 0.7484).

The stacking ensemble improved macro F1 to 0.9488 on the internal test set, with Overgeneration precision of 0.89 and recall of 0.9240, outperforming each individual model by approximately 4 macro F1 points.

5.2. The Importance of Full-Document Grounding

One of the most important observations of this work concerns the role of source context. Initial experiments explored retrieval-based context selection, where only the most semantically similar source sentences were provided to the classifier. While this strategy appeared promising, empirical evaluation consistently favored full-document grounding.

This finding suggests that overgeneration detection differs from traditional retrieval-based verification tasks. Determining whether a generated sentence is supported by the source often requires information distributed across multiple parts of the abstract. Restricting the model to a small subset of retrieved sentences may therefore remove evidence necessary for accurate verification.

5.3. Ensemble Effectiveness

The three transformer encoders capture different aspects of scientific language understanding. ModernBERT provides strong contextual reasoning capabilities across long sequences, SciBERT contributes domain-specific scientific knowledge through its scientific vocabulary and pretraining corpus, and SPECTER2 introduces representations learned from scholarly citation structure. By combining these complementary signals, the stacking framework improves robustness of the final prediction process.

6. Conclusion

This paper presented the A_R Team submission to CLEF 2026 SimpleText Task 2.1. We proposed a source-aware grounding verification framework for detecting overgenerated content in simplified

scientific texts. Our approach fine-tunes three complementary transformer encoders—ModernBERT, SciBERT, and SPECTER2—using complete source abstracts and generated sentences. Their outputs are subsequently combined through a Logistic Regression stacking ensemble.

Our experiments demonstrate the importance of full-document grounding and show that broader source context is more informative than locally retrieved evidence for overgeneration detection. The ensemble achieved a sentence-level F1 of 0.8527 and a document-level F1 of 0.7852 on the official CLEF 2026 test set. Future work will investigate stronger ensemble strategies, cross-encoder architectures, and more advanced grounding verification methods for trustworthy scientific text simplification.

Acknowledgments

We thank the CLEF 2026 SimpleText Track organizers for organizing the shared task and providing the dataset and evaluation framework.

Declaration on Generative AI

During the preparation of this work, the authors used *Claude (Anthropic)* in order to: **Improve writing style**, **Grammar and spelling check**, and **Code writing and formatting**. After using these tool(s)/service(s), 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. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), 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.

[2] B. Chakar, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 2: Identify and Avoid Hallucination, in: E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[3] B. Warner, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024).

[4] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2019.

[5] K. Lo, et al., SPECTER2: Adapting scientific document representation with heterogeneous graph neural networks, arXiv preprint arXiv:2404.04178 (2024).

[6] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.

[7] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level representation learning using citation-informed transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT 2019, Association for Computational Linguistics, 2019.

[9] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2019.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.