FutureMakers at the CLEF 2026 SimpleText Track: Retrieval-Augmented Evidence-Focused Overgeneration Detection for Task 2.2

Notebook for the SimpleText Lab at CLEF 2026

Wenlong Liu[1], Sixian Qi[2], Yusheng Yi[1,*] and Ruilan Lin[1]

[1]Foshan University, Foshan, China
[2]Guangdong University of Technology, Guangzhou, China

Abstract

When automatically processing scientific literature, text simplification models may generate content that cannot be attributed to the source document, a phenomenon known as overgeneration. Detecting such content in CLEF 2026 SimpleText Task 2.2 is challenging because source documents are often too long to be encoded directly by pre-trained models. To address this problem, we propose a retrieval-augmented, evidence-focused framework that uses TF-IDF based sparse retrieval at both training and inference stages to select source sentences most relevant to each candidate sentence, thereby reformulating long-document overgeneration detection as a short-text entailment problem between retrieved evidence and the candidate sentence. On the official test set, our method achieves a sentence-level F1 of 0.8673, a document-level F1 of 0.8069, and an Accuracy (Multiclass) score of 0.7795, demonstrating that retrieval augmentation and train–inference alignment can effectively improve overgeneration detection in long-document text simplification.

Keywords

overgeneration detection, text simplification, retrieval-augmented framework

1. Introduction

We participated in Task 2 (Identify and Avoid Hallucination) of the CLEF 2026 SimpleText track [1, 2]. This task aims to detect hallucinations in automatically simplified scientific texts. Current text simplification systems may introduce information that cannot be supported by the source document, a phenomenon referred to as overgeneration [2]. Such errors compromise the faithfulness of simplified texts and limit their applicability in science communication. CLEF 2026 SimpleText Task 2.2 formulates this challenge as a multi-class natural language inference problem: given a biomedical source abstract and a set of simplified sentences, systems must determine whether each sentence contains overgeneration and classify it into one of six categories.

Existing overgeneration detection approaches typically rely on pre-trained language models to perform textual entailment classification between source and generated texts [3]. However, source documents are often complete scientific abstracts that exceed the 512-token input limit of models such as BERT [4], making truncation a common issue in neural text classification pipelines. As a result, many methods truncate the source text, potentially discarding relevant evidence and reducing detection accuracy.

To address this limitation, we design a retrieval-augmented evidence-focused framework. During both training and inference, TF-IDF retrieval is used to select the source sentences most relevant to the sentence under examination. These retrieved sentences are concatenated into a compact evidence context and provided as input to PubMedBERT for multi-class classification. By applying the same

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
1967407393@qq.com (W. Liu); 2325515377@qq.com (S. Qi); yiys@fosu.edu.cn (Y. Yi); 2577877903@qq.com (R. Lin)
0009-0006-0850-597X (W. Liu); 0009-0001-5930-9120 (S. Qi); 0009-0006-7098-3681 (Y. Yi); 0009-0005-1273-0136 (R. Lin)

retrieval procedure during training and inference, we maintain consistent input construction and enable the model to make predictions based on focused evidence rather than truncated document contexts.

Our main contributions are as follows: (1) we design a retrieval-augmented evidence-focused framework that helps alleviate information loss caused by document truncation; (2) we maintain training–inference consistency through identical retrieval procedures; and (3) experimental results on CLEF 2026 SimpleText Task 2.2 demonstrate the effectiveness of the proposed method.

2. Related Work

2.1. Hallucination Detection in Text Simplification

The problem of hallucination in text simplification has been brought into focus by recent shared tasks such as CLEF 2026 SimpleText Task 2.2 [2], which targets the detection and classification of overgeneration errors in simplified biomedical abstracts. Early approaches to assessing the faithfulness of generated text relied on surface-level statistical features, notably n-gram overlap and bag-of-words similarity, to measure the consistency between a source text and its simplification [5]. Although straightforward, these methods struggle to capture deeper semantic relations and often fail to discriminate between legitimate paraphrasing and unsupported content insertion.

To address these limitations, entailment-based methods built on pre-trained language models [4] have become the dominant paradigm. By framing the source document as a premise and the generated sentence as a hypothesis, models such as BERT can be fine-tuned to predict whether an entailment relation exists. This idea has been successfully applied to factual consistency evaluation in summarization (e.g., FactCC [6]) and to correctness ranking of generated summaries. However, the adoption of these techniques for biomedical text simplification remains relatively limited. A further challenge is that biomedical abstracts frequently exceed the typical 512-token input limit of encoder-based models, making effective evidence selection a critical bottleneck for overgeneration detection.

2.2. Biomedical Pre-trained Language Models

General-purpose language models often perform sub-optimally in specialised domains due to the lack of domain-specific terminology and knowledge structures [7]. To overcome this, several models have been obtained by continued pre-training on large biomedical corpora. Among them, PubMedBERT—pre-trained from scratch on 14 million PubMed abstracts—has achieved state-of-the-art results on a range of biomedical NLP tasks, including text classification and named entity recognition [8]. Despite its strong domain adaptation, PubMedBERT inherits the same architectural constraint as BERT: a maximum input length of 512 tokens. When applied to long documents such as Cochrane systematic review abstracts, truncation is unavoidable and potentially discards essential evidential information [9].

2.3. Retrieval-Augmented Methods and Sparse Retrieval

Retrieval-augmented techniques enhance model inputs by first selecting relevant pieces of information from an external knowledge source and then feeding them together with the original query to a downstream model. This paradigm has proved particularly effective in knowledge-intensive tasks such as open-domain question answering and fact verification [10]. For evidence selection, the retrieval component can be implemented in a variety of ways, ranging from dense neural retrievers to classical sparse retrieval models.

Among sparse models, TF-IDF (Term Frequency–Inverse Document Frequency) [11] remains one of the most widely used techniques for information retrieval and text similarity computation. Its simplicity, efficiency, and strong performance on keyword-matching tasks make it an attractive choice for building lightweight retrieval components. In the context of hallucination detection, a TF-IDF-based retriever can quickly locate the source sentences that are most relevant to a given target sentence, providing a focused evidence context for the classifier.

However, most existing retrieval-augmented approaches for detection tasks apply retrieval only at inference time, while training continues to rely on the full source document. This misalignment between training and inference input distributions can significantly impair model generalisation. Motivated by this issue, we propose a retrieval-augmented framework in which a TF-IDF retriever is employed during both training and inference in a strictly aligned manner, ensuring that the model always reasons on compact, relevant evidence.

2.4. Summary and Motivation

In summary, current methods for overgeneration detection in long biomedical documents suffer from two main limitations: (1) the maximum input length of pre-trained models necessitates truncation, which can discard relevant evidence, and (2) when retrieval augmentation is applied only at inference, the mismatch between training and inference input distributions reduces performance. To address these issues, we propose a retrieval-augmented, evidence-focused approach that uses TF-IDF as a lightweight retriever and ensures strict train–inference alignment.

3. Method

3.1. Evidence Retrieval

For a given source document S and a target sentence t, we first split S into a sequence of individual sentences using regular expressions:

$$S = \{s_1, s_2, \ldots, s_n\}$$

The splitting rules explicitly account for common abbreviations in academic literature (e.g., et al., Fig.) to avoid incorrect sentence boundaries. A TF-IDF vectorizer is then fitted on all sentences of S. For the target sentence t, the cosine similarity with each source sentence s_i is computed as:

$$\mathrm{sim}(t, s_i) = \frac{v_t \cdot v_{s_i}}{\|v_t\| \|v_{s_i}\|}$$

where v_t and v_{s_i} denote the TF-IDF vector representations of t and s_i, respectively.

To force the model to learn semantic reasoning rather than relying on simple string matching, if t is identical to some s_i (i.e., $t = s_i$), its similarity score is set to -1.0. The adjusted similarity is defined as:

$$\mathrm{sim}'(t, s_i) = \begin{cases} -1.0, & \text{if } t = s_i \\ \mathrm{sim}(t, s_i), & \text{otherwise} \end{cases}$$

Based on the adjusted similarities, the top-K most similar source sentences are selected and sorted by their original positions in the document. They are then concatenated into a compact evidence context $C(t)$:

$$C(t) = \bigoplus_{i \in I_K} s_i, \quad I_K = \underset{i \in \{1,\ldots,n\}}{\mathrm{topK}} \ \mathrm{sim}'(t, s_i)$$

where topK denotes the operation of taking the indices of the K highest similarity scores in descending order, and $\bigoplus$ denotes text concatenation in the original document order.

3.2. Biomedical Entailment Classifier

We adopt PubMedBERT-base as the backbone network. The model input is formatted as a text pair: the evidence context $C(t)$ serves as the first segment, and the target sentence t as the second segment. The tokenizer automatically inserts the special delimiter token and assigns token type ids. The final hidden state of the `[CLS]` token is fed into a linear classification head, which outputs a probability distribution over the six categories.

3.3. Train–Inference Alignment

The core design principle of our method is to maintain strictly identical retrieval logic during training and inference. In the training phase, for each training sample (S, t, y), we generate the evidence context $C(t)$ using exactly the same retrieval procedure described in Section 3.1, and feed $(C(t), t)$ into the model. Consequently, the input distribution seen by the model at training time is perfectly aligned with that at inference time, eliminating the train–inference discrepancy. This alignment ensures that the model learns entailment relationships from precise evidence rather than from truncated full texts.

4. Experiments

4.1. Dataset

Our experiments are conducted on the official training, development, and test sets provided by the CLEF 2026 SimpleText Task 2. The training set contains 21,057 documents with 350,583 sentences in total, while the development set contains 1,120 documents with 18,638 sentences. The source abstracts are often complete scientific texts whose length substantially exceeds the maximum input length of PubMedBERT. The label distribution is highly imbalanced: `None` accounts for 92.3% of all sentences, while GROUNDED_OVERGENERATION has only 100 instances (0.03%), and REPETITIVE_CONTENT has 1,477 instances (0.42%).

4.2. Experimental Setup

We adopt `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext` as the backbone model with a maximum input length of 512 tokens. The retrieval parameter K is set to 4. We train the classifier using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and 4 epochs. A warmup ratio of 0.1 is applied, together with mixed-precision training. LayerNorm parameters are forced to remain in FP32 to avoid gradient overflow. All experiments are run with a fixed random seed of 42. The evaluation metrics include sentence-level and document-level precision, recall, and F1-score, as well as the macro-averaged F1-score for multi-class classification.

4.3. Baselines

To investigate the contribution of each module, we compare the following configurations:

- **Full-text Truncation Baseline**: PubMedBERT is trained and evaluated using the full source abstract (truncated to 512 tokens), without any retrieval component.
- **Train–Inference Misaligned**: The model trained on full texts is evaluated with retrieval-augmented inference, creating a distribution mismatch.
- **Our Method (Retrieval-Augmented Aligned)**: Both training and inference employ the same retrieval logic to construct the evidence context.

4.4. Main Results

Table 1 reports the performance of each method on the official test set.

Table 1
Performance comparison on the test set.

Method	2.2 Acc (Multiclass)	2.1 Document-level			2.1 Sentence-level		
		F1	Prec	Rec	F1	Prec	Rec
Full-text Baseline	0.7702	0.7732	0.7895	0.7576	0.8459	0.8485	0.8434
Train–Inference Misaligned	0.7768	0.7432	0.7154	0.7732	0.8012	0.7596	0.8478
Aligned (Ours)	**0.7795**	**0.8069**	**0.8393**	**0.7769**	**0.8673**	**0.8722**	**0.8624**

As shown in Table 1, the aligned retrieval-augmented method improves the sentence-level F1-score by approximately 2.2 percentage points and the document-level F1-score by about 3.4 percentage points over the full-text baseline. Notably, the misaligned configuration underperforms the full-text baseline, which indicates that introducing retrieval solely at inference time without aligning the training phase actually harms performance due to the input distribution mismatch. This confirms the critical role of our train–inference alignment strategy.

4.5. Ablation Study

Effect of Retrieval Augmentation. The comparison between the full-text baseline and our aligned retrieval method directly quantifies the contribution of the evidence focusing module. Retrieval augmentation brings consistent improvements on both sentence-level and document-level metrics, validating the importance of extracting precise evidence from long documents for entailment judgment.

Effect of Train–Inference Alignment. The misaligned configuration, which uses a full-text trained model with retrieval at inference, achieves lower scores than the full-text baseline. This proves that input distribution consistency is crucial for model generalization. Applying retrieval only at inference does not bring any gain; on the contrary, the train–inference mismatch degrades performance.

Impact of K. To investigate the influence of the number of retrieved sentences, we train four aligned models with $K \in \{3, 4, 5, 6\}$ and evaluate them on the test set. The results are shown in Table 2.

Table 2
Effect of Retrieval Size K.

K	2.1 F1 (doc)	2.1 F1 (snt)	2.2 Acc (Multiclass)
3	0.8002	0.8615	0.7762
4	**0.8069**	**0.8673**	**0.7795**
5	0.8030	0.8641	0.7778
6	0.7960	0.8589	0.7743

As K increases from 3 to 4, the document-level F1 improves, demonstrating that more relevant sentences provide useful supplementary evidence. However, when K continues to grow to 5 and 6, performance declines, especially at $K = 6$ where the document-level F1 drops to 0.7960. This indicates that including too many low-relevance sentences introduces noise and disturbs the classifier. Therefore, $K = 4$ strikes the best balance between information gain and noise suppression, and is selected as the final retrieval parameter.

4.6. Error Analysis

Despite strong overall performance, the GROUNDED_OVERGENERATION category remains extremely challenging. A likely contributing factor is the combination of scarce training data—only 100 instances—and the noise in its boundary with UNGROUNDED_INJECTION, which the organizers themselves identified as particularly difficult. These conditions make it hard for the model to learn a reliable decision pattern for this class. Future work could explore contrastive learning or other debiasing strategies to improve separation between such subtle categories.

5. Conclusion

This paper proposes a retrieval-augmented, evidence-focused method to address the challenge of overgeneration detection in long biomedical documents, as defined in CLEF 2026 SimpleText Task 2.2. By applying the same TF-IDF retrieval procedure during both training and inference, our method selects a compact evidence context for each candidate sentence and reformulates long-document overgeneration detection as a short-text entailment judgment between the retrieved evidence and the candidate sentence. Experimental results demonstrate that the proposed method outperforms the full-text truncation baseline on both sentence-level and document-level metrics. Ablation studies

further confirm the respective contributions of the retrieval augmentation module and the train–inference alignment strategy. Overall, our method provides an efficient and reproducible solution for overgeneration detection in long-document text simplification.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62276064). We thank the organizers of the CLEF 2026 SimpleText Track for providing the datasets, evaluation platform, and task guidelines.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for text polishing and grammar checking. After using this tool, 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] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, I. Gurevych, Ranking generated summaries by correctness: An interesting but challenging application for natural language inference, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2214–2220. URL: https://aclanthology.org/P19-1213/. doi:10.18653/v1/P19-1213.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[5] W. Kryscinski, N. S. Keskar, B. McCann, C. Xiong, R. Socher, Neural text summarization: A critical evaluation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 540–551. URL: https://aclanthology.org/D19-1051/. doi:10.18653/v1/D19-1051.

[6] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346. URL: https://aclanthology.org/2020.emnlp-main.750/. doi:10.18653/v1/2020.emnlp-main.750.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. doi:10.1093/bioinformatics/btz682.

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare 3 (2021). URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.

[9] K. Huang, S. S. Garapati, A. Rich, An interpretable end-to-end fine-tuning approach for long clinical text, ArXiv abs/2011.06504 (2020). URL: https://api.semanticscholar.org/CorpusID:226306829.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[11] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manage. 24 (1988) 513–523. URL: https://doi.org/10.1016/0306-4573(88)90021-0. doi:10.1016/0306-4573(88)90021-0.