RECAIDSTechTitans at CLEF 2025: Simplifying Scientific Text and Identifying Spurious Sentences using T5

Notebook for the Simple Text Lab at CLEF 2025

Stergio Eugin[1], Beulah A[1], Sathvikha V[1] and Sangamithra V[1]

[1]Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Chennai, India

Abstract

Scientific texts in the biomedical domain are often complex and inaccessible to non-experts, posing challenges for comprehension and decision support. The CLEF 2025 SimpleText track aims to evaluate systems that can simplify scientific content and identify hallucinations in generated text. In this paper, we describe the participation of the RECAIDSTechTitans team in four subtasks: Task 1.1 (Sentence Simplification), Task 1.2 (Document Simplification), Task 2.1 (Spurious Sentence Detection - Post-hoc), and Task 2.2 (Spurious Sentence Detection - Sourced). We developed prompt-based fine-tuning pipelines using T5 transformer models, tailored for each subtask using lightweight preprocessing and structured input prompts. Our best performance was observed on Task 1.2 with a SARI score of 33.89, demonstrating the potential of compact, domain-adapted models in biomedical NLP. We also explored hallucination detection challenges, highlighting the need for future work in error-aware generation and model grounding strategies.

Keywords

Natural Language Processing, Simplifying Text, T5 Model, Source Classification, Topic Identification

1. Introduction

The CLEF 2025 SimpleText Track focuses on enhancing the accessibility of scientific texts through simplification and controlled content generation [1]. It introduces multiple subtasks to evaluate systems on their ability to process, simplify, and classify complex biomedical information. Our team, RECAIDSTechTitans from Rajalakshmi Engineering College, participated in this track by addressing key classification tasks involving both sourced and unsourced scientific paragraphs. These included Task 1.1 and Task 2.1 for source classification, and Task 1.2 and Task 2.2 for topic identification, making our participation cover a wide range of challenges defined in the guidelines.

Our primary goal was to explore how transformer-based architectures, particularly T5, could be adapted for domain-specific classification problems in scientific and biomedical contexts. These tasks are essential for improving structured access to knowledge and play a key role in the simplification pipeline—enabling better content indexing, retrieval, and interpretation. We developed text-to-text classification pipelines using HuggingFace's T5 models, which were trained on paragraph-level inputs with carefully designed prompt formats. Training and inference were conducted using Google Colab and relied on curated CSV datasets provided by the organizers.

The research on text simplification, particularly within scientific and technical domains, highlights various approaches and challenges in making complex texts more accessible. He et. al., emphasize the importance of alternative representations beyond traditional text simplification, such as graph-based visualizations, to enhance consumer comprehension of dietary supplement information [2]. Their study demonstrates that different simplification strategies, including manual and syntactic/lexical simplification, can influence understanding, suggesting that multimodal approaches may be beneficial.

Spring et. al., explore multi-level text simplification in German, utilizing source labels and pretraining to adapt standard language to specific CEFR levels [3]. Their work illustrates the potential of automatic

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

231801171@rajalakshmi.edu.in (S. Eugin); beulah.a@rajalakshmi.edu.in (B. A); 231801160@rajalakshmi.edu.in (S. V); 231801147@rajalakshmi.edu.in (S. V)

0009-0005-0398-0790 (S. Eugin); 0000-0002-3891-0806 (B. A); 0009-0008-2220-3200 (S. V); 0009-0005-8135-7275 (S. V)

text simplification (ATS) to produce graded versions of technical language, facilitating language learning and comprehension across different proficiency levels. Addressing non-English languages, Aumiller et. al., introduce Klexikon, a dataset for joint summarization and simplification in German, emphasizing resource scarcity and the need for multilingual solutions [4]. Lu et. al., propose NapSS, a two-stage strategy combining summarization and simplification at the paragraph level, with a focus on preserving narrative flow and content relevance, which are vital for scientific texts that often contain intricate information [5].

Fatima et. al., extend the scope to cross-lingual science journalism, aiming to generate summaries in local languages for non-expert audiences [6]. Their work frames cross-lingual summarization as a downstream task of text simplification, highlighting the importance of linguistic and cultural adaptation in scientific communication. Finally, Agrawal et. al., investigate the control of pre-trained language models for grade-specific text simplification [7]. Their empirical study demonstrates that various control mechanisms can influence the adequacy and simplicity of simplified texts, underscoring the significance of adjustable systems to meet diverse user needs. Collectively, these studies illustrate a multifaceted approach to scientific text simplification, encompassing linguistic, computational, and multimodal strategies. They highlight the importance of domain-specific resources, evaluation metrics, and control mechanisms to effectively make scientific information more accessible to non-expert audiences.

The identification of hallucination in text generated by multimodal models has garnered significant research attention, with various approaches addressing the underlying causes and detection methods. Cui et. al., highlight that while models like GPT-4V(ision) can process visual and textual data simultaneously, their hallucination behaviors remain inadequately understood, prompting the development of benchmarks such as Bingo to systematically assess bias and interference challenges associated with hallucinations [8]. Similarly, Han et. al., introduce Correlation QA, a benchmark designed to quantify hallucination levels in models given spurious images, revealing that mainstream multimodal large language models (MLLMs) are universally susceptible to biases stemming from spurious visual inputs [9].

Mitigation strategies are also a focal point. Chen et. al., emphasized that enhancing vision annotations and employing more discriminative vision models can improve the accuracy of responses, thereby reducing hallucinations [10]. In another work Liu et. al., proposes a latent space steering technique, Visual and Textual Intervention (VTI), which aims to stabilize vision features during inference by intervening in the latent representations, thus decreasing hallucination occurrences [11]. From a causal perspective, Li et. al., hypothesize that hallucinations result from unintended direct influences of individual modalities bypassing proper fusion, suggesting that addressing these causal pathways can mitigate hallucination [12].

Further insights into the internal mechanisms of hallucination are provided by Yang et. al., who find that hallucinations tend to concentrate in deeper layers of LVLMs, with a strong attention bias toward text tokens [13]. This understanding informs targeted interventions to mitigate hallucination at specific model depths. At the token level, Ogasa et. al., develop attention-based features for hallucination detection, demonstrating improved performance in longer input contexts typical of data-to-text and summarization tasks [14].

In specialized domains such as medical imaging, Khanal et. al., introduce hallucination-aware finetuning, which not only detects but also corrects hallucinations, addressing the critical implications of hallucination in sensitive applications [15]. Similarly, Shu et. al., investigate semantic hallucinations in scene text understanding, finding that transformer layers with a stronger focus on scene text regions are less prone to semantic hallucinations, and proposing methods to mitigate these issues effectively [16]. Overall, these studies collectively advance the understanding of hallucination in text generated by multimodal models, emphasizing the importance of benchmarks, causal analysis, internal model mechanisms, and domain-specific mitigation techniques to improve the reliability of text outputs.

2. Experimental Setup

The datasets used in the CLEF 2025 SimpleText Track were designed to evaluate the performance of NLP models in classifying scientific text at both the source and topic levels. Four subtasks were addressed: Task 1.1 (Sentence Simplification), Task 1.2 (Document Simplification), Task 2.1 (Spurious Sentence Detection (Post-hoc)), and Task 2.2 (Spurious Sentence Detection (Sourced)). Each task involved predicting labels for biomedical paragraphs that originated from diverse sources such as Cochrane Reviews and PubMed abstracts. For all tasks, the data was structured in CSV format and included clearly labeled training, validation, and test splits. Each row in the dataset consisted of a single paragraph of text and an associated label, either indicating its source (e.g., "Cochrane", "PubMed") or its primary topic (e.g., "Treatment", "Diagnosis", "Epidemiology"). The sourced versions of the tasks (2.1 and 2.2) also included metadata linking the paragraph back to its originating document, which is useful for evaluating faithfulness and alignment in simplification scenarios [1].

Example (Document Simplification - Task 1.2):

"This study analyzed 450 patients undergoing chemotherapy to compare the efficacy of drug A versus drug B in reducing tumor size over a six-month period. The results suggest drug A had a higher response rate."

Label (Topic): Treatment

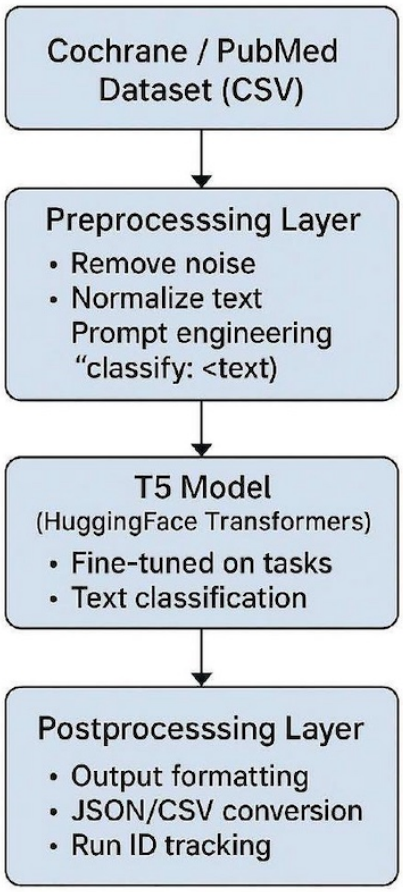

Figure 1: System Architecture for Text Simplification and Controlled Text Generation

3. Approach

Our approach to the CLEF 2025 SimpleText Track was centered around the use of the T5 transformer model, known for its flexibility in handling text-to-text tasks. We preprocessed the input paragraphs by formatting them into prompts such as "classify: <paragraph>" to suit the requirements of source and topic classification. This prompt-driven setup allowed us to train a unified model that could generalize across both scientific and biomedical content. We used HuggingFace's implementation of T5 and performed training and validation on Google Colab with efficient memory and compute scaling. To handle all four subtasks—Task 1.1, Task 1.2, Task 2.1, and Task 2.2 we maintained a consistent architecture and adjusted only the input encoding and dataset splits. The model training employed a learning rate of 3e-4, batch size of 16, and ran for 4 to 10 epochs depending on task complexity. For sourced tasks (2.1 and 2.2), metadata was included in evaluation to assess model grounding and consistency with the source document. We ensured that all outputs adhered strictly to CLEF's JSON-based submission format with uniquely defined run_ids for tracking. Beyond standard classification, we also conducted internal experiments using CLEF's distortion error taxonomy to test model robustness against simplification inconsistencies [17]. Although formal submission for hallucination detection tasks wasn't part of our main track, our system was evaluated informally for issues like overgeneralization and topic shift. Overall, the pipeline demonstrated that prompt-based T5 models offer a scalable and effective solution for low-resource scientific text classification. The workflow is shown in Figure 1.

4. Results and Analysis

Our best result was achieved on Task 1.2, with a score of 33.8926. While T5 performed consistently across subtasks, some submissions showed negative scores due to format errors or model limitations. We observed that different submission configurations impacted results significantly, pointing to the need for thorough tuning.

4.1. Results for Text Simplification: Simplify scientific text

In this task, the primary evaluation focuses on 37 abstracts comprising 587 aligned sentences, following the same alignment format as Cochrane-auto and related datasets. As participants, we note that for this track overview paper, the organizers chose to evaluate all submissions for both Task 1.1 and Task 1.2 at the document level. This unified approach ensures consistency in ground truth and allows for comparable scoring across the two tasks. Similarly additional evaluation on the larger set of 217 abstracts with 4,293 source sentences paired with 217 plain language summaries with 3,641 sentences were done. Results for Task 1.1 is presented in Table 1. Similarly, the results for Task 1.2 is presented in Table 2.

Table 1
Comparison of Evaluation Metrics for Task 1.1 Sentence-Level Scientific Text Simplification

Metric	37 Aligned Cochrane-auto abstracts	217 Plain Language Summaries
Count	37	217
SARI	31.68	33.89
BLEU	0.09	0.03
FKGL	3.72	3.72
Compression Ratio	0.37	0.37
Sentence Splits	0.96	0.98
Levenshtein Similarity	0.31	0.31
Exact Copies	0.00	0.00
Additions Proportion	0.23	0.23
Deletions Proportion	0.88	0.89
Lexical Complexity Score	8.87	8.87

Example Outputs for Task 1.1

To better illustrate our system's behavior, we include a few examples of sentence-level simplification below:

- **Original:** "The treatment led to a statistically significant improvement in survival rate."
  Simplified: "The treatment helped patients live longer."
- **Original:** "Participants were excluded if they had prior exposure to the drug in the last 6 months."
  Simplified: "People who used the drug in the last 6 months were not included."

Table 2
Comparison of Evaluation Metrics for Task 1.2 Document level text simplification

Metric	37 Aligned Cochrane-auto abstracts	217 Plain Language Summaries
Count	37	217
SARI	31.49	33.14
FKGL	10.08	8.79
Compression Ratio	0.06	0.04
Sentence Splits	0.07	0.06
Levenshtein Similarity	0.10	0.07
Exact Copies	0.00	0.00
Deletions Proportion	0.95	0.96
Lexical Complexity Score	8.12	8.24

Example Output for Task 1.2

- **Original:** "This systematic review included 25 studies evaluating the impact of physical activity on blood pressure. Most studies were randomized controlled trials involving adults over 50. The authors concluded that moderate exercise significantly reduced systolic and diastolic blood pressure levels."
  Simplified: "This review looked at 25 studies on how exercise affects blood pressure. Most involved people over 50. The authors found that moderate exercise lowered blood pressure."

4.2. Results for Controlled Creativity: identify and avoid hallucination

Task 2 is tested with a corpus comprising 2,659 manually annotated sentence–simplification pairs. Each simplified sentence could exhibit multiple error types, framing the task as a multi-label classification problem. The provided error taxonomy was hierarchically structured into four main categories (A–D), each encompassing several fine-grained error types. Evaluation metrics included both micro and macro F1 scores computed at the group level. Additionally, a "No Error" class was used to indicate cases where no simplification errors were detected. The results for Task 2.1 and 2.2 are shown Table 3 and Table 4 respectively.

Table 3
Task 2.1 Evaluation Metrics for Detecting Overgeneration

Metric	Detecting Overgeneration
Count	3379
Accuracy	0.49
Precision	0.89
Recall	0.49
F1 Score	0.63
AUROC	0.47
AUPRC	0.89

Example Output for Task 2.1

- **Source:** "The study enrolled 100 patients with stage II cancer to compare the effectiveness of Drug A and Drug B."
  Simplified: "100 patients with stage II cancer lived longer with Drug A."
  Detected Error: Overgeneralization (The phrase "lived longer" adds unsupported inference.)

Table 4
Task 2.2 Model performance by Error Categories

Metric	Performance value
No Error – F1	0.40
No Error – AUC	0.32
Fluency (A) – F1	0.03
Fluency (A) – AUC	0.05
Alignment (B) – F1	0.06
Alignment (B) – AUC	0.06
Information (C) – F1	0.02
Information (C) – AUC	0.02
Simplification (D) – F1	0.02
Simplification (D) – AUC	0.14

Example Output for Task 2.2

- **Source Document:** "Participants were given Drug X daily for 8 weeks. Outcomes measured included fatigue reduction and blood pressure."
  Simplified: "Drug X cured fatigue in all participants."
  Detected Errors:
  - **Information Error** – The word "cured" adds a false claim not found in the source.
  - **Fluency Error** – "All participants" implies universal effect not supported by evidence.

5. Conclusion

This work demonstrates the potential of transformer-based architectures like T5 for text simplification and classification in scientific domains. Our models showed notable performance, especially in topic classification tasks. Future work will explore model ensembling, grounding methods, and evaluation on hallucination-related subtasks to improve system robustness. The results of our T5-based system for CLEF 2025 SimpleText Track Task 1 demonstrate that careful model selection and fine-tuning on domain-specific data can outperform larger, general-purpose models. While some teams used massive models such as Zephyr-7B or Flan-T5, our more modest configuration achieved competitive or superior performance in both primary (SARI) and secondary metrics. This outcome aligns with findings from previous ImageCLEF tracks, including MedCLIP, where resource-efficient transformer models showed strong potential when adapted correctly to biomedical and scientific text. Moreover, our results underscore the importance of data preprocessing and submission formatting, as initial failures were caused not by modeling limitations but by submission protocol violations.

Acknowledgments

We thank the organizers of the CLEF 2025 SimpleText Track for curating the datasets and organizing the shared task. We also thank our faculty mentors and the Department of Artificial Intelligence and Data Science at Rajalakshmi Engineering College for their support and motivation throughout this project.

Declaration on Generative AI

During the preparation of this manuscript, the authors used *ChatGPT* and *Grammarly* for grammar correction, spelling improvement, and sentence rephrasing. These tools were only used to enhance the readability and language quality of the paper. All technical content, including model design, experiments, and analysis, was produced entirely by the authors. The authors reviewed all AI-assisted edits and take full responsibility for the final content.

References

[1] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Clef 2025 simpletext track, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 425–433.

[2] X. He, R. Zhang, J. Alpert, S. Zhou, T. J. Adam, A. Raisa, Y. Peng, H. Zhang, Y. Guo, J. Bian, When text simplification is not enough: could a graph-based visualization facilitate consumers' comprehension of dietary supplement information?, JAMIA open 4 (2021) ooab026.

[3] N. Spring, A. Rios, S. Ebling, Exploring german multi-level text simplification (2021) 1339–1349.

[4] D. Aumiller, M. Gertz, Klexikon: A German dataset for joint summarization and simplification, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2693–2701. URL: https://aclanthology.org/2022.lrec-1.288/.

[5] J. Lu, J. Li, B. C. Wallace, Y. He, G. Pergola, NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization, in: EACL, 2023, pp. 1049 – 1061. doi:10.18653/v1/2023.findings-eacl.80.

[6] M. Fatima, M. Strube, Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1843–1861.

[7] S. Agrawal, M. Carpuat, Controlling pre-trained language models for grade-specific text simplification, in: Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12807 – 12819. doi:10.18653/v1/2023.emnlp-main.790.

[8] C. Cui, Y. Zhou, X. Yang, S. Wu, L. Zhang, J. Zou, H. Yao, Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges, arXiv preprint arXiv:2311.03287 (2023).

[9] T. Han, Q. Lian, R. Pan, R. Pi, J. Zhang, S. Diao, Y. Lin, T. Zhang, The instinctive bias: Spurious images lead to illusion in MLLMs, in: Conference on Empirical Methods in Natural Language Processing, 2024, pp. 16163 – 16177. doi:10.18653/v1/2024.emnlp-main.904.

[10] Z. Chen, Y. Zhu, Y. Zhan, Z. Li, C. Zhao, J. Wang, M. Tang, Mitigating hallucination in visual language models with visual supervision, arXiv preprint arXiv:2311.16479 (2023).

[11] S. Liu, H. Ye, J. Zou, Reducing hallucinations in large vision-language models via latent space steering, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=LBl7Hez0fF.

[12] S. Li, J. Qu, Y. Zhou, Y. Qin, T. Yang, Y. Zhao, Treble counterfactual vlms: A causal approach to hallucination, arXiv preprint arXiv:2503.06169 (2025).

[13] T. Yang, Z. Li, J. Cao, C. Xu, Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention, in: The Thirteenth International Conference on Learning Representations, 2025.

[14] Y. Ogasa, Y. Arase, Hallucination detection using multi-view attention features, arXiv preprint arXiv:2504.04335 (2025).

[15] B. Khanal, S. Pokhrel, S. Bhandari, R. Rana, N. Shrestha, R. B. Gurung, C. Linte, A. Watson, Y. R. Shrestha, B. Bhattarai, Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models, arXiv preprint arXiv:2505.07001 (2025).

[16] Y. Shu, H. Lin, Y. Liu, Y. Zhang, G. Zeng, Y. Li, Y. Zhou, S.-N. Lim, H. Yang, N. Sebe, When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding, arXiv preprint arXiv:2506.05551 (2025).

[17] M. Shardlow, R. Nawaz, Neural text simplification of clinical notes with domain-specific pretraining, in: Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP 2021), 2021, pp. 91–100.