BioSimplex at the CLEF 2026 SimpleText Task 1.1: A Two-Stage Operation-Aware Framework for Scientific Text Simplification*

CLEF 2026 SimpleText Task 1.1

Sejal Maurya[1,*], Ridhi Ladda[1,*,†], Gayatri Tupe[1,*,†] and Ritesh Kumar[1,*,†]

[1] *Department of Computer Science and Engineering, Indian Institute of Information Technology Surat, Gujarat, India*

Abstract

The growing volume of biomedical research literature presents a significant accessibility challenge, as technical terminology and complex sentence structures often limit understanding among non-expert readers. This paper describes our participation in the CLEF 2026 SimpleText Lab Task 1.1, which focuses on automatic sentence-level simplification of scientific text. We propose a label-conditioned two-stage framework in which a SciBERT-based classifier first predicts the most appropriate simplification operation (*ignore*, *rephrase*, or *split*), and a fine-tuned FLAN-T5 model subsequently generates simplified text conditioned on the predicted label. Both components are trained using the Cochrane-Auto biomedical simplification corpus.

Our system is evaluated on the official CLEF 2026 benchmark consisting of 48,809 scientific sentences. We compare FLAN-T5-Base and FLAN-T5-Large variants within the proposed framework. The best-performing system, based on FLAN-T5-Large, achieved a SARI score of **37.98**, a BLEU score of **8.31**, and an FKGL score of **9.81** on the official CodaBench evaluation. The results demonstrate that combining operation-aware classification with instruction-tuned sequence-to-sequence generation provides an effective approach for improving the accessibility of biomedical scientific text while maintaining content fidelity.

Keywords

text simplification, scientific language accessibility, biomedical NLP, sequence-to-sequence models, label-conditioned generation, CLEF 2026, SimpleText

1. Introduction

The rapid growth of scientific and biomedical literature has created significant challenges for non-expert readers seeking to access and understand research findings. Scientific publications often contain specialised terminology, complex sentence structures, and dense information that can limit accessibility for patients, students, journalists, policymakers, and the general public. Automatic text simplification aims to address this challenge by transforming complex text into simpler and more accessible language while preserving its original meaning.

The CLEF 2026 SimpleText Track provides a shared benchmark for evaluating automatic simplification of scientific text and related tasks [1]. The track consists of three tasks: scientific sentence simplification (Task 1), hallucination detection and avoidance (Task 2), and research area classification (Task 3). In this work, we participate in Task 1.1, which focuses on sentence-level scientific text simplification and is based on the benchmark introduced by the task organisers [2]. Given a complex scientific sentence, participating systems are required to generate a simplified version that improves readability while preserving the original meaning.

Recent advances in instruction-tuned large language models and domain-specific transformer architectures have significantly improved the state of the art in text simplification. However, different

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

* Submitted to the CLEF 2026 SimpleText Lab, Task 1: Sentence-level scientific text simplification.

* Corresponding author.

† These authors contributed equally.

sejalmaurya1aug@gmail.com (S. Maurya); ui23cs37@iiitsurat.ac.in (R. Ladda); ui23cs22@iiitsurat.ac.in (G. Tupe); riteshkumar@iiitsurat.ac.in (R. Kumar)

scientific sentences often require different simplification strategies. Some sentences require only minor lexical substitutions, whereas others benefit from sentence splitting or structural rephrasing. Motivated by this observation, we propose a two-stage, label-conditioned framework that explicitly models different simplification operations.

Our approach consists of two components:

- **Stage 1 — Sentence classifier**: A domain-adapted SciBERT classifier predicts one of three simplification operations: *ignore*, *rephrase*, or *split*.
- **Stage 2 — Label-conditioned simplifier**: A fine-tuned FLAN-T5 model generates simplified text conditioned on the predicted operation label using task-specific prompts.

Both components are trained using the Cochrane-Auto corpus [3], a large-scale biomedical text simplification dataset derived from Cochrane systematic reviews.

Our best official submission achieved a SARI score of **37.98** on the CLEF 2026 leaderboard using a SciBERT-guided FLAN-T5-Large model. We additionally compare FLAN-T5-Base and FLAN-T5-Large variants within the proposed label-conditioned framework to investigate the effect of model scale on biomedical text simplification.

1.1. Contributions

The main contributions of this work are:

- A two-stage label-conditioned biomedical text simplification pipeline for CLEF 2026 SimpleText Task 1.1.
- A SciBERT-based classifier that predicts simplification operations (*ignore*, *rephrase*, and *split*) prior to text generation.
- A prompt-guided FLAN-T5 simplification model that conditions generation on the predicted operation label.
- An empirical comparison of FLAN-T5-Base and FLAN-T5-Large within a label-conditioned simplification framework for biomedical scientific text.

The remainder of this paper is organised as follows. Section 2 reviews related work on text simplification and biomedical NLP. Section 3 describes the datasets used. Section 4 presents the proposed methodology. Section 5 details the experimental setup. Section 6 defines the evaluation metrics. Section 7 presents the official evaluation results and comparison of submitted systems. Section 8 discusses the findings and limitations of the proposed approach. Finally, Section 9 concludes the paper and outlines directions for future work.

2. Related Work

Text simplification is a well-established task in NLP, with roots in lexical simplification [4], syntactic simplification [5], and more recently neural sequence-to-sequence approaches [6]. We review work most directly relevant to our contributions: CLEF SimpleText, biomedical text simplification, pretrained language models for simplification, and evaluation methodology.

2.1. CLEF SimpleText Shared Task

The CLEF SimpleText Lab has provided a community benchmark for automatic simplification of scientific passages since its first edition in 2022 [7]. Each annual iteration refines the task definition, data, and evaluation protocol. The 2025 edition [8] attracted a diverse set of participants employing methods ranging from instruction-tuned large language models to supervised fine-tuning of domain-specific transformers.

Among the most competitive CLEF 2025 systems, Largey et al. [9] (AIIRLab) employed instruction-tuned LLaMA [10] and Mistral [11] models with few-shot task demonstrations, achieving strong SARI scores without any task-specific gradient updates. Their work demonstrates that large instruction-following models can serve as competitive baselines for simplification even without fine-tuning, but also highlights that domain-specific fine-tuning consistently improves performance on technical biomedical text.

The NITK SCaLAR Lab (CLEF 2025) explored a complementary strategy, fine-tuning domain-adapted models including BioBERT [12], BioBART, and a modified GPT-2 on biomedical simplification data. Their results support the hypothesis that domain pretraining provides a meaningful head-start for downstream simplification, particularly for sentences containing rare medical terminology.

2.2. Biomedical Text Simplification

Biomedical text simplification has attracted dedicated research attention because of its high real-world impact and the unique linguistic properties of clinical and scientific writing. Early work focused on rule-based lexical substitution using medical thesauri such as UMLS [13].

The development of large aligned corpora substantially changed the landscape. Bakker and Kamps [3] introduced the Cochrane-Auto dataset, which automatically aligns sentences from the full text of Cochrane systematic reviews with corresponding sentences from their structured plain-language summaries. With hundreds of thousands of aligned sentence pairs, Cochrane-Auto enables supervised training of neural simplification models at scale, and has been adopted as a standard resource in the biomedical ATS community.

Van den Berckt et al. [14] benchmarked a range of neural approaches on biomedical corpora and found that models pre-trained on biomedical text (e.g., PubMedBERT, BioBERT) consistently outperformed general-domain counterparts, underscoring the importance of domain alignment. More recently, Guo et al. [15] demonstrated that controlled generation—conditioning output on readability constraints—can improve FKGL alongside SARI, motivating the label-conditioned design we adopt in the present work.

2.3. Pretrained Language Models for Simplification

The transformer architecture [16] has become the backbone of modern ATS systems. Among encoder–decoder models, BART [17] and T5 [18] have been extensively evaluated on text simplification benchmarks due to their ability to generate fluent and semantically faithful rewrites of complex text.

FLAN-T5 [19], an instruction-tuned variant of T5, extends this capability through large-scale instruction tuning across a diverse collection of NLP tasks. Its strong instruction-following behaviour makes it particularly suitable for controlled text generation tasks, including simplification. Furthermore, supervised fine-tuning on domain-specific corpora can adapt FLAN-T5 to specialised domains such as biomedicine, where technical terminology and complex sentence structures present additional challenges.

For sentence classification, SciBERT [20]—a BERT model pretrained on 1.14 million scientific papers from Semantic Scholar—provides strong representations for scientific text. Its exposure to domain-specific vocabulary, including biomedical terminology, makes it a natural choice for predicting simplification operations on scientific sentences. The combination of SciBERT for operation prediction and FLAN-T5 for controlled generation forms the basis of the label-conditioned framework proposed in this work.

2.4. Evaluation of Text Simplification

Evaluating simplification quality is non-trivial because a good simplified sentence must simultaneously be fluent, meaning-preserving, and genuinely simpler than the source. SARI [21], BLEU [22], and FKGL [23] together provide complementary coverage of these dimensions and constitute the official metrics for CLEF 2026 Task 1.1. We formalise these metrics in Section 6.

3. Dataset

3.1. Cochrane-Auto Dataset

The Cochrane-Auto dataset [3] is constructed by automatically aligning sentences from full Cochrane systematic reviews with sentences from their corresponding plain-language summaries. Each record in the dataset contains: (i) a complex source sentence (`sentence_src`), (ii) one or more simplified target sentences (`sentence_tgt`), and (iii) an operation label indicating the simplification type applied (`ignore`, `rephrase`, or `split`).

We apply the following preprocessing steps: (1) flattening multi-target records so that each complex sentence is paired with a single simplified reference (`flatten_simple`); (2) removing records where either the source or target field is empty or contains only whitespace; and (3) integer-encoding operation labels for classifier training. After preprocessing, the training split contains approximately 7,082 sentence pairs, and the validation split approximately 1,041 pairs.

3.2. CLEF 2026 SimpleText Task 1.1 Test Set

The official test set consists of **48,809 complex scientific sentences** provided by the CLEF 2026 SimpleText organisers in the file `sentences_src.json`. Sentences are drawn from scientific abstracts across multiple disciplines, though the Cochrane-Auto training domain (systematic reviews) partially overlaps with the biomedical subset. No reference simplifications are released to participants during the evaluation phase; scoring is performed server-side on CodaBench.

4. Methodology

4.1. Overview

Our approach follows a two-stage architecture designed to explicitly model different simplification operations. First, a SciBERT classifier predicts the most suitable simplification action for an input sentence: *ignore*, *rephrase*, or *split*. The predicted label is then converted into a task-specific prompt that conditions a fine-tuned FLAN-T5 model. This design enables the generator to apply operation-aware simplification strategies rather than relying on a single generic prompt for all inputs.

Our system consists of two sequential components trained independently and composed at inference time. Figure 1 shows the overall architecture.

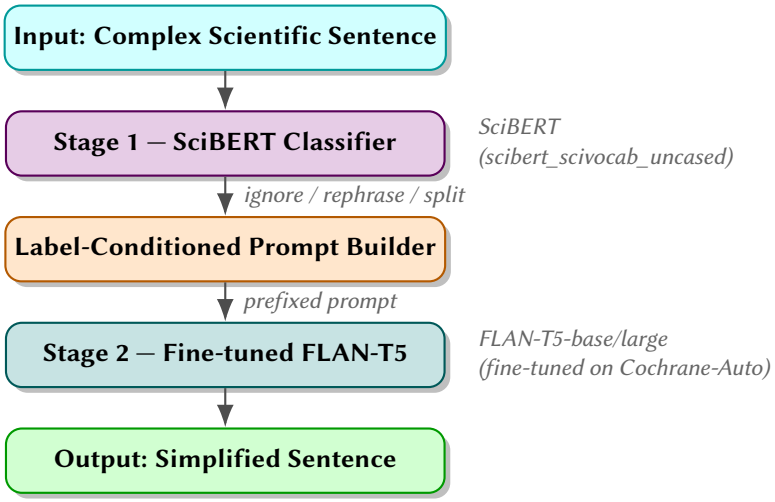

Figure 1: Architecture of the label-conditioned two-stage simplification pipeline. Stage 1 (SciBERT) classifies the required operation; Stage 2 (FLAN-T5) generates simplified text conditioned on the predicted label.

4.2. Stage 1: SciBERT Sentence Classifier

We fine-tune `allenai/scibert_scivocab_uncased` [20] as a three-class sequence classifier over the Cochrane-Auto operation labels. The classification head is a linear layer applied to the `[CLS]` token representation.

The three labels correspond to distinct simplification strategies: *Ignore* (sentence is already simple or is a heading/title), *Rephrase* (sentence should be rewritten with simpler vocabulary and shorter clauses while preserving its syntactic structure), and *Split* (sentence should be decomposed into two or more shorter independent sentences).

4.3. Stage 2: Label-Conditioned FLAN-T5 Simplifier

We fine-tune `flan-t5-base/large` [19] as a sequence-to-sequence simplifier. Each training example is formed by prepending a label-specific prefix to the source sentence. The prefix mapping is:

```
LABEL_PROMPTS = {
  "ignore":   "Keep this sentence unchanged: ",
  "rephrase": "Simplify this scientific sentence "
              "using easier words: ",
  "split":    "Split this sentence into two or more "
              "simpler sentences: ",
}
```

4.4. Inference Pipeline

At inference time, the SciBERT classifier first predicts an operation label for each of the 48,809 test sentences. The label selects the corresponding prefix from LABEL_PROMPTS, which is prepended to the source sentence and fed to the fine-tuned FLAN-T5 decoder. Generation uses beam search (`num_beams` $= 2$) with a batch size of 256 and `fp16` precision. Model checkpoints are saved via `save_pretrained`, compressed, and loaded from cloud storage for batch inference, enabling processing of the full test set without re-training.

5. Experimental Setup

5.1. Hardware and Software Environment

All experiments were conducted using Google Colab with NVIDIA A100 GPUs. Models were implemented using the HuggingFace Transformers library and trained using PyTorch. Mixed-precision (`fp16`) training was employed where applicable to reduce memory usage and accelerate training.

5.2. SciBERT Training Configuration

Table 1
SciBERT classifier hyperparameters.

Parameter	Value
Epochs	3
Learning Rate	2×10^{-5}
Batch Size	32
Max Sequence Length	128
Optimizer	AdamW
Loss Function	Cross Entropy

5.3. FLAN-T5 Training Configuration

Table 2
FLAN-T5 fine-tuning hyperparameters.

Parameter	Value
Epochs	3
Learning Rate	5×10^{-4}
Batch Size	16
Precision	fp16
Collator	DataCollatorForSeq2Seq
Optimizer	AdamW

Preliminary experiments indicated performance degradation when training beyond three epochs, suggesting overfitting to the Cochrane-Auto training distribution. Consequently, training was intentionally limited to three epochs to improve generalisation on unseen scientific text.

6. Evaluation Metrics

We formalise the three official evaluation metrics used in CLEF 2026 SimpleText Task 1.1.

6.1. SARI

SARI (System output Against References and against the Input) [21] measures simplification quality by simultaneously rewarding the system for: (1) adding words present in the reference but not in the input; (2) keeping words present in both the reference and the input; and (3) deleting words present in the input but not in the reference. SARI is computed for each n-gram order $n \in \{1, 2, 3, 4\}$ and averaged.

Let $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{H}$ denote the multisets of n-grams in the source (input), reference, and hypothesis (system output) respectively. For each operation $o \in \{\text{add}, \text{keep}, \text{del}\}$, a precision p_o and recall r_o are computed, and an F1-like score F_o is derived. SARI is then:

$$\text{SARI} = \frac{1}{3}\big(F_{\text{add}} + F_{\text{keep}} + F_{\text{del}}\big) \tag{1}$$

where each component is defined as:

$$F_{\text{add}} = \frac{|\mathcal{H} \cap \mathcal{R}| - |\mathcal{H} \cap \mathcal{S}|}{|\mathcal{H}| - |\mathcal{H} \cap \mathcal{S}|}, \quad F_{\text{keep}} = \frac{|\mathcal{H} \cap \mathcal{S} \cap \mathcal{R}|}{|\mathcal{H} \cap \mathcal{S}|}, \quad F_{\text{del}} = \frac{|\mathcal{S}| - |\mathcal{H} \cap \mathcal{S}|}{|\mathcal{S}|} \tag{2}$$

Higher SARI indicates better simplification quality.

6.2. BLEU

BLEU (Bilingual Evaluation Understudy) [22] measures n-gram precision between the hypothesis and one or more references. For N-gram orders up to $N = 4$:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

where p_n is the modified precision for n-grams of order n, $w_n = \frac{1}{N}$ is a uniform weight, and BP is the brevity penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{4}$$

with c the total length of the hypothesis corpus and r the effective reference corpus length. Higher BLEU indicates greater n-gram overlap with the reference.

6.3. FKGL

The Flesch-Kincaid Grade Level [23] estimates the US school grade level needed to comprehend a text based on sentence length and syllable count:

$$\text{FKGL} = 0.39 \cdot \frac{\text{total words}}{\text{total sentences}} + 11.8 \cdot \frac{\text{total syllables}}{\text{total words}} - 15.59 \tag{5}$$

Lower FKGL values indicate simpler text; scientific abstracts typically score between 14 and 18.

7. Results

7.1. Official Submission Results

Table 3 presents the official results of our submitted systems on the CLEF 2026 SimpleText Task 1.1 benchmark.

Table 3
Official CLEF 2026 SimpleText Task 1.1 sentence-level results obtained from the CodaBench evaluation platform.

System	SARI	BLEU	FKGL	Compr.
SciBERT + FLAN-T5-Base	37.25	14.74	11.21	0.68
SciBERT + FLAN-T5-Large	37.98	8.31	9.81	0.51

The FLAN-T5-Large submission achieved the highest overall performance, obtaining a SARI score of 37.98. It also achieved the lowest FKGL score of 9.81, indicating improved readability and accessibility of the generated simplifications.

The FLAN-T5-Base submission achieved a slightly lower SARI score of 37.25 but obtained substantially higher BLEU (14.74), suggesting greater lexical overlap with the reference simplifications.

7.2. Comparison of Submitted Systems

The results reveal a trade-off between simplification quality and lexical preservation.

Increasing the model size from FLAN-T5-Base to FLAN-T5-Large improves both SARI and FKGL, suggesting that the larger model is more effective at generating readable simplifications while preserving the meaning of the original sentence.

In contrast, the FLAN-T5-Base model achieved considerably higher BLEU, indicating that its outputs remained closer to the reference texts at the lexical level. This behaviour suggests a more conservative simplification strategy that prioritises preserving original wording rather than performing stronger paraphrasing operations.

Overall, the FLAN-T5-Large model produced the strongest simplification performance among the submitted systems and was selected as our best official submission for CLEF 2026 SimpleText Task 1.1.

8. Discussion

8.1. Effect of Model Scale

The comparison between FLAN-T5-Base and FLAN-T5-Large highlights the impact of model scale on scientific text simplification. The larger model achieved higher SARI (37.98) and lower FKGL (9.81), indicating that it generated outputs that were both simpler and more readable. These results suggest that increased model capacity enables more effective simplification while maintaining the meaning of the original sentence.

However, the FLAN-T5-Base model achieved substantially higher BLEU (14.74), indicating greater lexical overlap with the reference simplifications. This behaviour suggests that the base model adopts a more conservative generation strategy, preserving a larger proportion of the original wording and sentence structure.

8.2. Role of Label-Conditioned Generation

Our proposed framework incorporates a SciBERT classifier that predicts the most appropriate simplification operation before generation. Rather than relying on a single generic prompt for all inputs, the model receives operation-specific instructions corresponding to *ignore*, *rephrase*, and *split* actions.

This design allows the simplification model to adapt its generation strategy according to the characteristics of the input sentence. Sentences requiring minimal modification can be preserved, whereas more complex sentences can be simplified through rephrasing or sentence splitting. Such operation-aware generation provides a more structured approach to simplification than applying a single transformation strategy to all sentences.

8.3. Limitations

Despite promising results, several limitations remain.

- **Domain dependence**: Both the classifier and generator were trained on the Cochrane-Auto corpus, which primarily consists of biomedical systematic reviews. Performance may vary when applied to scientific domains with substantially different writing styles or terminology.
- **Classifier error propagation**: Incorrect operation predictions produced by the SciBERT classifier may affect the quality of the generated simplifications. Future work should investigate the impact of classifier accuracy on overall system performance.
- **Readability ceiling**: Although the FLAN-T5-Large system reduced FKGL to 9.81, the generated text may still be challenging for some non-expert readers. Additional lexical simplification techniques could further improve accessibility.
- **Limited model exploration**: Due to computational constraints, experimentation was restricted to FLAN-T5 variants. Investigating larger domain-adapted language models may provide further improvements in simplification quality.

9. Conclusion and Future Work

We presented a label-conditioned two-stage framework combining SciBERT-based operation classification with fine-tuned FLAN-T5 generation for CLEF 2026 SimpleText Task 1.1. The FLAN-T5-Large variant achieved the best performance with a SARI score of 37.98, BLEU of 8.31, and FKGL of 9.81, demonstrating that operation-aware generation improves accessibility of biomedical scientific text.

Several directions may further improve the framework, including exploring larger biomedical language models such as BioMistral, improving classifier robustness across scientific domains, integrating lexical simplification resources, and investigating joint learning approaches that unify operation prediction and text generation.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Claude for: language refinement, grammar and spelling correction, formatting assistance, and improving the clarity and structure of the manuscript.

After using these tools, the authors reviewed and edited all content as needed. All research ideas, experimental design, data processing, model development, evaluation, interpretation of results, and conclusions were carried out and verified by the authors. The authors take full responsibility for the content of this publication.

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] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text, 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] M. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for biomedical text simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR), 2024.

[4] L. Specia, S. K. Jauhar, R. Mihalcea, SemEval-2012 task 1: English lexical simplification, in: Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), 2012, pp. 347–355.

[5] Z. Zhu, D. Bernhard, I. Gurevych, A monolingual tree-based translation model for sentence simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING), 2010, pp. 1353–1361.

[6] S. Nisioi, S. Stajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 85–91.

[7] L. Ermakova, P. Bellot, J. Brasquet, J. Kamps, E. SanJuan, S. Sheremetyeva, Overview of the CLEF 2022 SimpleText track: Automatic simplification of scientific texts, in: Proceedings of the CLEF 2022 Working Notes, CEUR-WS, 2022.

[8] L. Ermakova, H. Azarbonyad, M. Bertin, E. Marty, J. Kamps, Overview of the CLEF 2025 SimpleText track: Automatic simplification of scientific texts, in: Proceedings of the CLEF 2025 Working Notes, CEUR-WS, 2025.

[9] A. Largey, P. Mulhem, G. Quenot, AIIRLab systems for CLEF 2025 SimpleText: Using LLaMA and Mistral for scientific text simplification, in: Proceedings of the CLEF 2025 Working Notes, CEUR-WS, 2025.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[11] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).

[12] 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.

[13] D. M. McDonald, H. Chen, Using sentence-selection heuristics to rank text segments in TXTRACTOR, Information Processing and Management 40 (2004) 227–243.

[14] N. Van den Berckt, A. Tack, K. Sirts, W. Daelemans, Med-EASY: Benchmarking automated simplification of medical texts, arXiv preprint arXiv:2309.12029 (2023).

[15] H. Guo, R. Pasunuru, M. Bansal, Automated rewriting of scientific papers for lay audiences, arXiv preprint arXiv:2109.08776 (2021).

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.

[17] 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 (ACL), 2020, pp. 7871–7880.
[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.
[19] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[20] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[21] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, in: Transactions of the Association for Computational Linguistics, volume 4, 2016, pp. 401–415. doi:10.1162/tacl_a_00107.
[22] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318. doi:10.3115/1073083.1073135.
[23] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, B. S. Chissom, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Technical Report RBR 8-75, Naval Technical Training Command, 1975.