LIS at SimpleText 2025: Enhancing Scientific Text Accessibility with LLMs and Retrieval-Augmented Generation

Notebook for the SimpleText Lab at CLEF 2025

Anya Amel Nait djoudi*[1], Sarah Nouali*[1], Mohsine Aabid[1], Ismail Badache[1], Adrian-Gabriel Chifu[1] and Patrice Bellot[1]

[1]*Aix Marseille Université, CNRS, LIS, Marseille, France*

Abstract

To improve public access to scientific knowledge, this work introduces a scientific text simplification model that combines Large Language Models (LLMs) with a Retrieval-Augmented Generation (RAG) framework. This paper presents the contribution of the R2I[1] team from the LIS Laboratory[2] to the SimpleText 2025 Lab, specifically Task 1.2: Document-level Scientific Text Simplification. Our main contribution is MedSimplify, a glossary of over 3,000 simplified definitions complied from multiple public medical sources. These definitions are integrated into a prompt-based simplification pipeline. We evaluated several LLM configurations for text simplification. Our results show that Mistral 7B with zero-shot prompting and MedSimplify definitions (`Mistral_DASP_0`) was the best-performing system, achieving a SARI score of 43.51 the highest among all our submissions. This result placed our team 5th in the CLEF 2025 SimpleText Task 1.2. Our findings show that grounding simplification in curated domain specific definitions improves readability while maintaining factual accuracy.

Keywords

Text simplification, Scientific text simplification, Biomedical text simplification, Large Language Models (LLM), Retrieval-Augmented Generation (RAG)

1. Introduction

Scientific publications serve as a primary medium for communicating research findings across a wide spectrum of disciplines, including but not limited to biomedical science. While scientific publications such as those in the biomedical domain are essential for advancing knowledge and informing public understanding, their heavy use of technical terminology and limited background context often makes them inaccessible to non-expert audiences [1, 2]. Understanding language spoken or written requires constructing a situation model that integrates both the explicit content and inferred meanings drawn from relevant background knowledge. Without this domain-specific knowledge, even fluent readers may struggle to comprehend scientific texts, regardless of their ability to decode the words (e.g., when reading a technical article on astrophysics) [3]. This helps explain why jargon not only impedes comprehension but also cognitive processing, leading to increased resistance to persuasion, heightened risk perceptions, and reduced support for scientific advancements [4]. This accessibility gap has sparked growing interest in the field of scientific text simplification, which seeks to bridge the divide between complex academic writing and broader public understanding.

In response to this challenge, initiatives such as the CLEF 2025 SimpleText Track [5] have emerged to promote the development and evaluation of automated systems capable of simplifying scientific

[1]Recherche d'Information et Interactions
[2]Laboratoire d'Informatique et des Systèmes
CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*These authors contributed equally to this work
anya.nait-djoudi@lis-lab.fr (A. A. Nait djoudi*); sarah.nouali@lis-lab.fr (S. Nouali*); mohsine.aabid@lis-lab.fr (M. Aabid); ismail.badache@lis-lab.fr (I. Badache); adrian.chifu@lis-lab.fr (A. Chifu); patrice.bellot@lis-lab.fr (P. Bellot)
https://github.com/Anyantd (A. A. Nait djoudi*); https://pageperso.lis-lab.fr/ismail.badache/ (I. Badache)
0009-0008-7037-5154 (A. A. Nait djoudi*); 0009-0003-3792-7678 (M. Aabid); 0000-0003-1868-1185 (I. Badache); 0000-0003-4680-5528 (A. Chifu); 0000-0001-8698-5055 (P. Bellot)

literature for non-expert audiences. In particular, Task 1.2 [6] focuses on identifying and explaining scientific jargon in a simplified and accessible manner.

Recent work [7, 8, 9] has demonstrated the potential of large language models (LLMs) in text simplification. Furthermore, Retrieval-Augmented Generation (RAG) [10] has been explored as a way to enrich outputs with external knowledge. For instance, [11] applied RAG to incorporate definitions and explanatory content from sources like Wikipedia [1], yielding slight improvements in relevance and readability.

Building on these insights, our approach in Task 1.2 combines LLMs and RAG to enhance the simplification of scientific jargon. We employed models such as Mistral [12] and LLaMA [13], experimenting with diverse prompting strategies to generate clear and concise definitions tailored to non-expert users. To enrich the context and improve definition quality, we integrated retrieval from curated knowledge sources, enabling the models to ground their responses in relevant background information.

Our approach aims to improve the interpretability of domain-specific terminology while minimizing the risk of hallucinations or factual errors, thereby contributing to broader efforts in making scientific content more transparent and accessible.

Our main contributions are:

- MedSimplify: a glossary of over 3,000 simplified definitions for medical terms.
- An LLM-based pipeline for automatically extracting domain-specific terminology.
- Integration of simplified definitions into prompting strategies (zero-shot, one-shot, and iterative refinement) to support effective scientific text simplification.

The rest of this paper is organized as follows. In Section 2, we present our experimental setup and describe the specific runs submitted. Section 3 discusses the results obtained from these runs and includes additional post-competition experiments. Finally, in Section 4, we conclude the paper by summarizing the key lessons learned and outlining future perspectives.

2. Experimental Setup

In this section, we detail the approach we propose for Task 1.2 of the CLEF 2025 SimpleText Track.

2.1. Dataset

For Task 1.2 of the CLEF 2025 SimpleText Track, we relied on datasets provided by the organizers, primarily the **Cochrane-auto** corpus [14] which consists of aligned pairs of biomedical abstracts and corresponding lay summaries, originally sourced from the Cochrane Database of Systematic Reviews (CDSR). Cochrane-auto provides alignment at the document, paragraph, and sentence levels. It includes: 1,085 document pairs, 4,171 paragraph pairs, 14,719 sentence pairs. Unlike the training data, which was limited to Cochrane-auto, the test set as shown in Figure 1 comprised examples from a broader set of biomedical sources, including: 217 texts from Cochrane, 236 texts from Cochrane-auto (119 in the validation and 117 in the test splits, respectively), 110 texts from Medline, 103 texts from SimpleText2024 [15]. This diverse composition introduced variability in writing style, structure, and vocabulary, increasing the complexity of the task. In particular, Medline and SimpleText 2024 included content not aligned in the same way as Cochrane-auto, which required our models to generalize beyond the distribution of the training data.

2.2. Terminology Simplification Glossary

To address lexical complexity in biomedical texts, we developed **MedSimplify** [2], a specialized glossary of simplified definitions for complex biomedical terms. As discussed in [16], complex words,

[1] https://fr.wikipedia.org/
[2] https://github.com/Anyantd/MedSimplify/

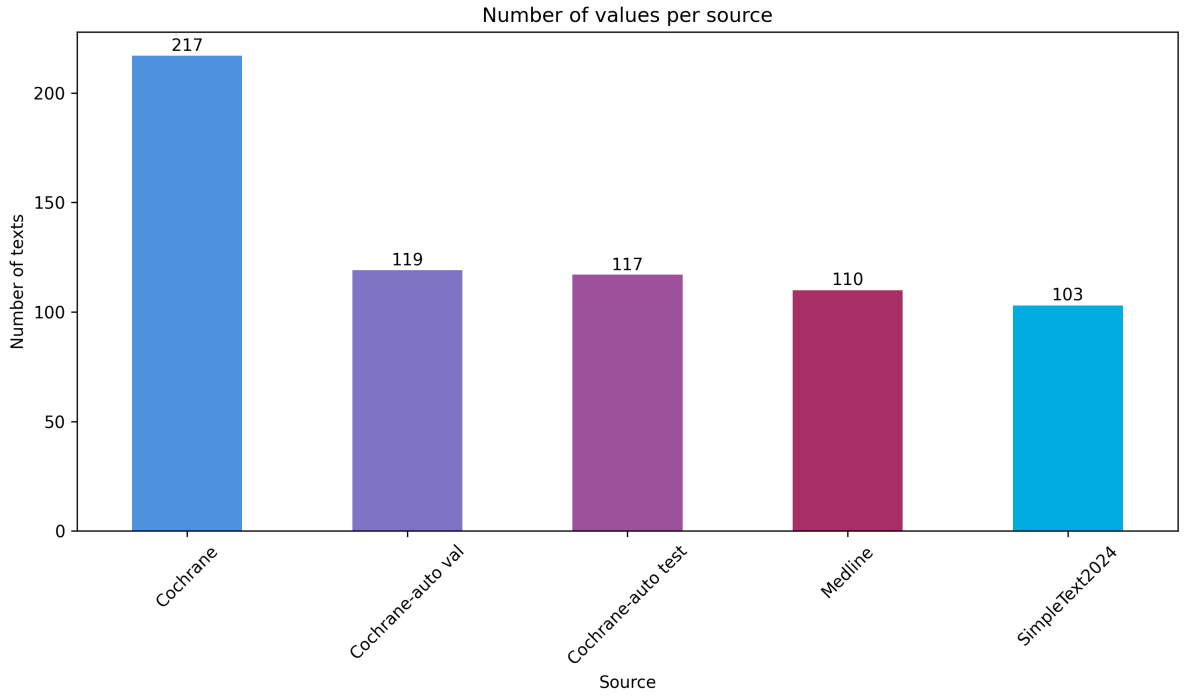

Figure 1: Test set Task 1.2 of the CLEF 2025 SimpleText Track

particularly those that are rare, domain specific or used in unusual contexts, can hinder understanding and disturb reader focus. These challenges are especially pronounced in biomedical communication, where technical terminology may not align with readers' background knowledge. MedSimplify was constructed by aggregating and unifying content from multiple publicly available medical glossaries that were contructed by researchers and experts in the medical field: the *Glossary of Lay Terminology for Consent Forms*[3], *Plain Language Thesaurus for Health Communications*[4], *CLAD-Thesaurus*[5], *Plain English Health Dictionary*[6], *Glossary (in Lay Terms)*[7]. They were systematically extracted, cleaned, and de-duplicated to produce a unified dataset, which was then stored in CSV format. Each entry within the MedSimplify glossary comprises a medical term and its corresponding layman-friendly definition. In instances where multiple definitions were available for a given term, the shortest one was consistently selected to accommodate token limitations inherent in the Retrieval-Augmented Generation (RAG) prompting pipeline. The final glossary encompasses over 3,000 entries and occupies approximately 178 kB. While this dataset was primarily designed to enhance contextual grounding within our RAG framework, it is important to acknowledge that each term is associated with only a single definition, which may not fully capture nuanced, context-specific meanings.

2.3. Preprocessing

Given the domain-specific complexity of biomedical texts, we employed the Mistral 7B language model to identify potentially challenging terms for non-expert readers. For each test instance, the full input text was provided to the model, with prompts crafted to elicit a list of domain specific keywords likely to hinder comprehension. We did not impose a fixed number of keywords, allowing the model to determine

[3]https://feinstein.northwell.edu/sites/northwell.edu/files/2019-07/Glossary-of-Lay-Terminology-for-Consent-Forms-07-19.pdf
[4]https://stacks.cdc.gov/view/cdc/11500
[5]https://clad.tccld.org/wp-content/uploads/2014/12/CLAD-Thesaurus.pdf
[6]https://nt.gov.au/__data/assets/pdf_file/0006/1257567/aid-plain-english-health-dictionary-spread.pdf
[7]https://rcm1.rcm.upr.edu/institutionalreview/wp-content/uploads/sites/16/2020/04/layterms.pdf

term salience dynamically (see Table 1 for prompt details). The output was post-processed to remove extraneous content such as formatting artifacts, or irrelevant introductory phrases. Extracted terms were then matched against entries in our terminology simplification glossary (Section 2.2) using exact string matching. Only terms with valid matches were retained, ensuring that subsequent simplification steps were grounded in high confidence, human-authored definitions. We choose for this first version of our work to use exact matching, to ensure high precision by only retrieving intended terms, avoids false positives from similar-looking but unrelated words and provides clean, reliable inputs for the model. These term-definition pairs were later embedded in our best performing prompt 1 to provide explicit contextual support during generation.

2.4. Architecture

Our pipeline for text simplification follows a Retrieval-Augmented Generation (RAG) paradigm, comprising the following key stages:

2.4.1. Retrieval via Exact Matching

For each keyword of the input text, we retrieve a corresponding definition from our custom-built dictionary using an exact string match strategy. This ensures deterministic retrieval of precise, human-authored definitions written in plain English (layman's terms) and avoids ambiguity introduced by semantic or fuzzy retrieval.

2.4.2. Prompt Design

We designed and evaluated multiple prompts to guide the language model in producing accurate and accessible simplifications. To explore different strategies for leveraging contextual definitions and examples, we designed four distinct prompt types:

- *Keyword-Guided Retrieval Prompt (KGR)*: Presents only extracted keywords from the input text to guide targeted keyword extraction;
- *Definition-Augmented Simplification (Zero-Shot) (DASP_0)*: Includes the complex input text and associated definitions, with no examples provided;
- *Definition-Augmented Simplification (One-Shot) (DASP_1)*: Adds a single example of input–output simplification alongside definitions to support in-context learning;
- *Iterative Refinement Prompt (IRP)*: Takes an initial simplified version and the original text, prompting the model to improve clarity and accuracy without distorting the meaning, inspired by [17].

2.4.3. Generation Models

The enriched prompt (original input + definitions) is passed to an LLM model to generate the simplified text. We utilized different models with the prompts DASP-0 and DASP-1 described in Table 1:

- The Mistral 7B model[8] was used in `Mistral_DASP_0` and `Mistral_DASP_1`;
- the Gemma 2-9B model[9] [18] was used in `Gemma2_DASP_1`;
- and finally the Med42-v2 model[10][19] in `Med42_DASP_0`.

The different solutions are summarized in Table 2.

[8]Mistral 7B: https://huggingface.co/mistralai/Mistral-7B-v0.3
[9]Gemma 2-9B: https://huggingface.co/google/gemma-2-9b-it
[10]Med42-v2: https://huggingface.co/m42-health/Llama3-Med42-8B

Table 1
Prompt templates

Prompt Name	Prompt Text
KGR	I have the following document: {Complex text} Please give me the keywords that are present in this document and separate them with commas. Make sure you to only return the keywords and say nothing else. For example, don't say: "Here are the keywords present in the document"
DASP_0	Using these definitions, please simplify the following scientific text for a general audience. Use plain language and explain any complex terms or acronyms. Ensure that all numbers, results, and facts remain exactly the same. Do not paraphrase numerical data or alter the meaning of findings. **DEFINITIONS:** {List of definitions} **TEXT:** {Complex text}
DASP_1	You are a helpful assistant that simplifies biomedical or scientific texts. **Task:** Using these definitions, simplify the following scientific text for a general audience. Use plain language and explain any complex terms or acronyms. Ensure that all numbers, results, and facts remain exactly the same. Do not paraphrase numerical data or alter the meaning of findings. **Example:** (Example of a pair of complex text and its simplified version) *Definitions: {List of definitions}* *Text:{example of complex text}* *Simplified:{example of simplified text}* Now do the same for the following: **Definitions:** {definition} **Text:**{text} **Simplified:**
IRP	Improve the simplified version of the scientific text below to make it clearer and easier for a general audience. Your goal is to maximize the SARI score by simplifying language and structure, while keeping all facts, numbers, and findings exactly the same. Do this step by step. **ORIGINAL TEXT:** {Complex text} **FIRST SIMPLIFIED VERSION:** {Generated simple text} **REFINED VERSION:**

2.4.4. Post Competition Approaches

Following the competition, we conducted a series of additional experiments to further enhance model performance and evaluate output quality using the SARI score, employing alternative strategies as outlined bellow :

1. **Gemma3_DASP_0**: we used in this solution the model[11] [20] with the prompt DASP_0.
2. **Mistral_IRP**: In one of our post-competition experiments, we aimed to improve the performance of `Mistral_DASP_0` through an iterative refinement strategy. Specifically, we reintroduced the initially generated simplified outputs from `Mistral_DASP_0` back into the model as input, using the prompt *Iterative Refinement Prompt (IRP)* described in Section 1. The objective was to encourage the model to produce a second, potentially more refined simplification by leveraging its own prior output as an intermediate step.
3. **Mistral_FKGL**: We assessed the readability of the simplified texts generated using the models from the evaluation phase (see Table 2) by computing their Flesch-Kincaid Grade Level (FKGL) [21] scores. FKGL scores were calculated using the Python library Textstat[12]. FKGL is a standard

[11]Gemma 3: https://huggingface.co/google/gemma-3-4b-it
[12]Textstat:https://pypi.org/project/textstat/

readability metric that estimates the U.S. school grade level required to understand a given text. Since we aim to produce simplified texts that are more accessible to the general public, we selected the version with the lowest FKGL score, indicating higher readability. The post-competition results are shown in Table 2.

3. Experimental Results

In this section, we present the results of our experiments, submitted on the platform Codabench[13] that calculated the SARI scores which can be found in Table 2. Our runs are referred to as ASM in the CLEF overview paper [5].

Table 2
Overview of best submitted runs and corresponding SARI Scores

Phase	Run	Description	Score
Evaluation	`Mistral_DASP_0`	Definitions used, zero-shot prompting	**43.5051**
	`Med42_DASP_0`	Definitions used, zero-shot prompting	41.9708
	`Gemma2_DASP_1`	Definitions used, one-shot prompting	42.5695
	`Mistral_DASP_1`	Mistral 7b with one-shot prompting	42.4718
Post-Competition	`Gemma3_DASP_0`	Definitions used, zero-shot prompting	41.6236
	`Mistral_baseline`	No definitions, zero-shot prompting	42.9018
	`Mistral_IRP`	Mistral 7B with iterative refinement prompt (IRP)	43.1044
	`Mistral_FKGL`	FKGL-based output selection	**43.5120**

We extracted a total of 848 definitions from our glossary that matches the complex texts. These were matched to 386 complex texts, with an average of 1.2 definitions per text.

We observe that augmenting the prompt with plain English definitions of complex keywords, as done in `Mistral_DASP_0` compared to `Mistral_baseline`, leads to a slight improvement in the SARI score.

Among the models using the zero-shot prompt `DASP_0` namely `Mistral_DASP_0`, `Med42_DASP_0`, and `Gemma3_DASP_0`. Mistral consistently delivers the best performance.

Additionally, the one-shot prompt variant `DASP1` does not yield better results than `Mistral_DASP_0`, indicating that oneshot prompting did not provide a significant advantage in this context. This result indicates that the example used in our one-shot prompt might have disturbed the model.

During the postcompetition phase, we explored further refinements using the Iterative Refinement Prompt (IRP) with Mistral. While `Mistral_IRP` showed a slight decrease compared to `Mistral_DASP_0`, it still achieved a strong performance. Notably, the FKGL-based approach (`Mistral_FKGL`) matched and even slightly surpassed the highest SARI score, demonstrating the potential of readability-based ranking in producing simplified outputs.

4. Conclusion

This work explored the simplification of biomedical scientific texts using a Retrieval-Augmented Generation (RAG) framework. The main contributions include: (1) the development of a glossary aggregating over 3,000 simplified definitions from public medical sources, (2) a pipeline for extracting expert domain terms using an LLM-based keyword extractor, and (3) the integration of these definitions into diverse prompting strategies including zero-shot, one-shot, and iterative refinement to support simplified texts generation of scientific texts.

While the test corpus primarily consisted of biomedical documents, it also included other scientific texts, requiring models to generalize beyond the biomedical domain.

[13]Codabench : https://www.codabench.org/

The best-performing configuration was the Mistral 7B model with zero-shot prompting `Mistral_DASP_0` , which achieved a SARI score of 43.50. A small post-competition enhancement using `Mistral_FKGL` further improved the score to 43.51 the highest observed performance. These results confirm that grounding simplification in curated, domain specific definitions can enhance readability without compromising factual integrity.

Despite these promising outcomes, limitations remain. Relying on a single static definition per term fails to capture contextual nuance, and the glossary though extensive may lack coverage of rare or emerging terms. We aimed with this first version of our work, to test whether adding definitions to the context would be beneficial. In future versions of our work, we plan to include synonyms and explore additional matching approaches. Additionally, one-shot prompting yielded limited gains, indicating potential for improved prompt design or few-shot learning strategies.

Future work will focus on expanding the glossary with context-sensitive or multi-definition entries, leveraging semantic retrieval to better align definitions with context, and fine-tuning LLMs specifically for biomedical simplification. These improvements aim to further enhance the accessibility, relevance, and trustworthiness of simplified scientific content.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 for grammar and spelling checks, as well as for paraphrasing and rewording. After using these tools, the author(s) reviewed and edited the content as needed, and take full responsibility for the publication's content.

References

[1] S. Bao, R. Zhao, S. Zhang, J. Zhang, W. Wang, Y. Ru, Ctyun ai at biolaysumm: Enhancing lay summaries of biomedical articles through large language models and data augmentation, in: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, 2024, pp. 837–844.

[2] Y. Guo, W. Qiu, Y. Wang, T. Cohen, Automated lay language summarization of biomedical scientific reviews, in: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI Press, 2021, pp. 160–168. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16089. doi:10.1609/aaai.v35i1.16089.

[3] D. T. Willingham, How knowledge helps, American Educator 30 (2006) 30–37.

[4] O. M. Bullock, D. Colón Amill, H. C. Shulman, G. N. Dixon, Jargon as a barrier to effective science communication: Evidence from metacognition, Public Understanding of Science 28 (2019) 845–853.

[5] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText track: Simplify scientific texts (and nothing more), in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), European Conference on Information Retrieval, Lecture Notes in Computer Science, Springer, Springer, 2025, pp. 425–433.

[6] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3147–3162.

[7] M. Färber, P. Aghdam, K. Im, M. Tawfelis, H. Ghoshal, Simplifymytext: An llm-based system for inclusive plain language text simplification, in: European Conference on Information Retrieval, 2025, pp. 418–424.

[8] S. Agrawal, M. Carpuat, Controlling pre-trained language models for grade-specific text simplification, arXiv preprint arXiv:2305.14993 (2023).

[9] N. Colic, J.-D. Kim, F. Rinaldi, Pre-gamus: Reducing complexity of scientific literature as a support against misinformation, in: Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024, 2024, pp. 196–201.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[11] Z. You, S. Radhakrishna, S. Ming, H. Kilicoglu, Uiuc_bionlp at biolaysumm: an extract-then-summarize approach augmented with wikipedia knowledge for biomedical lay summarization, in: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, 2024, pp. 132–143.
[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.
[13] 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).
[14] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), 2024, pp. 41–51.
[15] L. Ermakova, V. Laimé, H. McCombie, J. Kamps, Overview of the clef 2024 simpletext task 3: Simplify scientific text, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), 2024, pp. 3147–3162.
[16] A. Kelious, M. Constant, C. Coeur, Complex word identification: A comparative study between chatgpt and a dedicated model for this task, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 3645–3653.
[17] T. Guidroz, D. Ardila, J. Li, A. Mansour, P. Jhun, N. Gonzalez, X. Ji, M. Sanchez, S. Kakarmath, M. M. Bellaiche, et al., Llm-based text simplification and its effect on user comprehension and cognitive load, arXiv preprint arXiv:2505.01980 (2025).
[18] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024).
[19] C. Christophe, P. K. Kanithi, T. Raha, S. Khan, M. A. Pimentel, Med42-v2: A suite of clinical llms, arXiv preprint arXiv:2408.06142 (2024).
[20] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
[21] J. P. Kincaid, R. P. Fishburne Jr, 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, Institute for Simulation and Training, University of Central Florida, 1975. URL: https://api.semanticscholar.org/CorpusID:61131325.