

SIB Text-Mining at TREC PLABA 2024

Luc Mottin¹, Anaïs Mottaz^{1,2}, Julien Knafou^{1,2}, Alexandre Flament², Julien Gobeill^{1,2}, Patrick Ruch^{1,2}

¹ SIB, Swiss Institute of Bioinformatics, Geneva, Switzerland

² HES-SO, University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland

Abstract

The comprehension of health information by patients has a real influence on the efficacy of their treatment. However, while more health resources are increasingly available to the public, the use of medical jargon and complex syntax makes them difficult to understand. Recent advances in machine translation and text simplification may help to make these resources more accessible by adapting biomedical text into plain language. In this context, the TREC 2024 Plain Language Adaptation of Biomedical Abstracts track sought to develop specialized algorithms able to adapt biomedical abstracts into plain language for the general public. The SIB Text Mining group participated in the “Complete Abstract Adaptation” subtask. Our first approach examines how a specific prompting using a state-of-the-art Large Language Model performs in the global adaptation of biomedical documents, with the intention of proposing a baseline with no technical improvements to compare more advanced strategies. The second approach investigates how the fine tuning of the transformer handles the task, and the third approach integrates a Retrieval Augmented Generation function to help generate a new document based on information from relevant sources.

1 Introduction

The rapid advancements in information retrieval (IR) technologies have significantly enhanced the ability to process vast volumes of data and respond to complex queries. Benchmark competitions, such as the Text REtrieval Conference (TREC) organized by the National Institute of Standards and Technology (NIST), provide an essential platform for evaluating and advancing these IR systems across specialized domains. The objective of the TREC 2024 Plain Language Adaptation of Biomedical Abstracts (PLABA) track is to enhance health literacy by adapting documents into plain language, addressing the need for precise yet simplified biomedical information that can be readily understood by non-expert audiences [1]. This track is particularly relevant for addressing practical needs within the biomedical field. Our group is engaged in R&D projects such as Polypharmacy, which aim, among other objectives, to create individualized side-effect profiles based on medications recorded within an application. Such projects require the development of effective “laymanization” strategies (aka. the process of translating complex or technical information into layman language) and stand to benefit directly from the standardized evaluation methods and findings generated by the different groups participating in TREC PLABA.

This paper presents the participation of the SIB Text Mining group, part of the Swiss Institute of Bioinformatics (SIB), in the PLABA track. The SIB Text Mining group brings a wealth of experience from past TREC challenges, including participation in the TREC Genomics [2], TREC Medical Records [3], TREC Chemical IR [4], TREC Clinical Decision Support [5], TREC Precision Medicine [6,7] and TREC Deep Learning [8] tracks. In undertaking the PLABA 2024 task 2 “Complete Abstract Adaptation”, our approach was informed by the fact that only three runs could be subjected to manual evaluation by NIST experts. The first run served to establish a baseline for comparison using the LLAMA3 large language model

(LLM), and the remaining two were used to assess *ad hoc* adaptation of typical strategies for LLM improvement.

2 Methodology

2.1 Training and test data

The dataset for the 2024 PLABA track is derived from the PubMed database, an extensive collection of biomedical abstracts. A set of 75 topics representing biomedical interrogations (*e.g.* “what medication has a drug interaction with high blood pressure?”) was provided as a training set [9]. Each question is associated with 10 relevant MEDLINE abstracts, yielding a total of 750 documents. Then, each abstract was manually adapted to plain language by at least one annotator, resulting in 921 new documents (7,643 sentence pairs). The test set mirrors the structure of the training data. It includes 40 questions, each of which is also linked to 10 PubMed abstracts for a total of 400 documents with an average of 10.3 sentences per document (4,120 sentences).

2.2 Plain language adaptation strategies

This year, task 2 of TREC PLABA focuses on complete abstract adaptation. It involves simplifying complex scientific texts while ensuring that each source sentence is either transformed into one or more plain language sentences or omitted, if necessary, without merging information from different sentences. The outputs were manually evaluated by human experts across four dimensions: simplicity, accuracy, completeness, and brevity.

Run 1 - Improved prompting

To establish a baseline for adapting biomedical documents into plain language, we employed LLAMA3, a state-of-the-art LLM known for its high accuracy in generating coherent text [10]. LLAMA3 was selected for its ability to simplify complex language while preserving the core meaning of specialized content. To implement this strategy, the LLAMA 3.0 model with 8 billion parameters was downloaded and run locally using the Ollama framework with no further retraining. A temperature of 0.1 and a top-p of 0.9 were applied to produce deterministic, coherent and safe outputs, while allowing for some degree of variability.

In this first run, our intention was to make the guidelines of the task as explicit as possible through the specification of the prompt. This approach stemmed from prompt engineering techniques to improve the efficacy of LLMs [11,12], such as the Chain-of-thought prompting that guides the model through intermediate reasoning steps. We empirically evaluated the output of different prompts, assessing their consistency with the PLABA task guidelines given to the annotators. The resulting prompt is presented below:

You must adapt the following biomedical sentence into clear, concise language, focusing on key points and avoiding ambiguity.

The adaptation must be easy to understand and concise, contain accurate information, and it must seek to minimize information lost from the original text.

You must substitute medical jargon with common alternatives or explain terms with no substitutes.

Do not change words from common English.

Provide a direct answer with no chatting, and do not repeat the original in the answer.

Sentence is: {sentence}

Run 2 - Improved model LLAMA 3.1

In Run 2, we maintained a temperature that tended toward 0 to reduce creative variability in responses and a prompt that focused on clarity, brevity, and minimal rephrasing. We tried Llama 3.1¹ 8B instead of Llama 3 8B to see if it improved its ability to simplify complex information. Llama 3.1 uses a larger context (128K tokens instead of 8K) and a more refined architecture [13]. It was trained with a more extensive and diverse dataset, including multilingual and logical reasoning data. The curation of the training data included data cleaning, such as removing noise and irrelevant content, removing duplicates and quality filtering. The run was thus designed to take advantage of Llama 3.1 while emphasizing brevity and avoiding unnecessary emphasis.

Adapt this sentence for a general audience, adding short and simple information for general public integrated (without too much emphasis) inside it for specific scientific and medical jargon terms and with minimal rephrasing. Output only the adaption surrounded by brackets, without any introduction or explanation

Run 3 - RAG

Retrieval-Augmented Generation (RAG) is a technique aimed at enhancing language model performance on tasks that require specific or up-to-date knowledge [14]. By combining a text generation model with a document retrieval module, RAG allows the model to access an external database and integrate relevant information directly into its outputs (*e.g.* recent developments from domain-specific sources). Given that biomedical content frequently includes complex or specialized terms, combining a LLM with RAG might enhance the adaptation process by providing clear explanations for terms that lack simpler substitutes in layman language.

In this run, we reused the prompt presented in the baseline run and we reproduced the LLM parameters to ensure consistency in evaluation. A collection of 90 documents from Wikipedia and the Monash University biomedical glossary² was selected for their alignment with the thematic scope of the test-set questions. These documents were extracted directly from web sources and segmented into chunks of 300 tokens without overlap to facilitate retrieval. A linear Support Vector Machine (SVM) model was employed to select the 2 most relevant documents for each query. This combined approach leverages trusted public biomedical sources and specific document filtering.

2.3 Alternative evaluation

As none of the models was fine-tuned using the training dataset, we employed this dataset as a benchmark to evaluate the performance of our runs. We used two methods: SARI and the cosine distance with LLaMA3 embeddings. The SARI metric, specifically designed for text simplification, assesses the quality of the simplified output by comparing it to a reference. It rewards changes that preserve essential content while enhancing readability through the evaluation of word addition, deletion and retention [15]. We used the Hugging Face implementation of SARI, as provided in the evaluate module version 0.4.3 [16]. Additionally, we applied cosine distance to the embeddings generated by LLaMA3 to measure the semantic similarity between our output and the benchmark.

¹ <https://ai.meta.com/blog/meta-llama-3-1/>

² <https://www.monash.edu/discovery-institute/study/biomedical-sciences-glossary>

3 Results

At the time of writing this working note, official results have not been released by NIST contractors. Table 1 presents the results obtained using the alternative method described in section 2.3. This analysis indicated a slight advantage for the run using RAG. However, the comparison would have been more informative if the prompts had been more consistent.

	Run 1	Run 2	Run 3
SARI	35	33	39
Cosine similarity	0.79	0.79	0.75

Table 1: Evaluation of the strategies based on SARI and cosine similarity.

4 Discussion

We chose to adapt the documents one sentence at a time, without considering their contextualization within the abstract to focus on producing non-complex outputs. However, this has resulted in two guidelines being deliberately overlooked: 1) ignoring the sentences that are not relevant to consumer understanding of the document; and 2) resolving anaphora (*i.e.* replacing the pronouns as “it”, in the source sentence, that refer to something in the previous sentence). Furthermore, Llama 3.1 could have been more effective in this case by leveraging its increased context length compared to Llama 3.

Regarding the prompt, incorporating examples from the training dataset could have been beneficial (*i.e.* few-shot prompting). As well, fine-tuning the model on this dataset could have yielded a more tailored solution. Given its purely decoder-only architecture, Llama may not be the most suitable model for a text transformation task. In retrospect, an encoder-decoder model such as T5 might have been a better choice for this task and a comparison with such a model would provide valuable insights.

Finally, other *ad hoc* strategies could be considered. With more evaluated runs available, we would have tried to leverage an English corpus (*e.g.* Wall Street Journal) to identify complex terms, where complexity is defined in terms of word frequency falling below a predefined threshold. Another option is to integrate a specialized tool for personalized information retrieval in biomedical literature such as SIBiLS [17].

5 Conclusion

Adapting biomedical documents for a general audience meets a real need to make complex medical information accessible to all. The initial evaluation criteria were simplicity, accuracy, completeness, and brevity, each of which was to be assessed manually by experts following specific guidelines. However, the absence of official feedback makes it difficult to assess the effectiveness of our approaches. To address this limitation, we explored the use of the training dataset to test automated metrics, including readability scores with SARI and semantic similarity measures with embedding cosine similarity. Nevertheless, alternative frameworks such as Simplicity-DA could provide different insights [18], although some adaptation may be necessary to align with PLABA.

Bibliography

1. Ondov B, Attal K, Demner-Fushman D (2022) A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11), 1976–1988. DOI: 10.1093/jamia/ocac149.
2. Ruch P, Chichester C, Cohen G et al. (2004). Report on the TREC 2003 Experiment: Genomic Track. In *Proceedings of the 2003 Text REtrieval Conference (TREC)*.

3. Gobeill J, Gaudinat A, Pasche E et al. (2011) Bitem group report for TREC medical records track 2011. In the Proceedings of the Twentieth text retrieval conference (TREC 2011). NIST Special Publication.
4. Gobeill J, Gaudinat A, Pasche E et al. (2011) Bitem group report for TREC chemical IR track 2011. In the Proceedings of the Twentieth text retrieval conference (TREC 2011). NIST Special Publication.
5. Gobeill J, Gaudinat A et Ruch P (2015) Exploiting incoming and outgoing citations for improving Information Retrieval in the TREC 2015 Clinical Decision Support Track. In Proceedings of The 24th Text REtrieval Conference (TREC 2015). NIST Special Publication.
6. Pasche E, Gobeill J, Mottin L et al (2017) Customizing a variant annotation-support tool: an inquiry into probability ranking principles for TREC precision medicine. In Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017). NIST Special Publication.
7. Pasche E, Caucheteur D, Mottin L et al. (2020) SIB text mining at TREC precision medicine 2020. In Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020). Gaithersburg (Maryland). NIST Special Publication. NIST Special Publication.
8. Knafou J, Jeffreys M., Mottin L et al. (2019) SIB Text Mining at TREC 2019 Deep Learning Track: Working Note. In Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019). Gaithersburg (Maryland). NIST Special Publication.
9. Attal K, Ondov B, Demner-Fushman D (2023) A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(8). DOI: 10.1038/s41597-022-01920-3.
10. Abhimanyu D, Abhinav J, Abhinav P et al. (2024) The Llama 3 Herd of Models. *arXiv*. arxiv.org/abs/2407.21783.
11. Wang L, Chen X, Deng X et al. (2024) Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ Digital Medicine*, 7(1), 41. DOI: 10.1038/s41746-024-01029-4.
12. White J, Fu Q, Hays S et al. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arxiv.org/abs/2302.11382*.
13. (PDF) Llama 3.1: An In-Depth Analysis of the Next Generation Large Language Model. (n.d.). Retrieved November 10, 2024, from https://www.researchgate.net/publication/382494872_Llama_31_An_In-Depth_Analysis_of_the_Next_Generation_Large_Language_Model#fullTextFileContent.
14. Guo Y, Qiu W, Leroy G et al. (2024) Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149, 104580. DOI: 10.1016/j.jbi.2023.104580.
15. Xu W, Napoles C, Pavlick E et al. (2016) Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4, 401–415. DOI: 10.1162/tac1_a_00107.
16. SARI - a Hugging Face Space by evaluate-metric. (n.d.). Retrieved November 13, 2024, from <https://huggingface.co/spaces/evaluate-metric/sari>.
17. Gobeill J, Caucheteur D, Michel PA et al. (2020) SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts, *Nucleic Acids Research*, 48(1), 12–16. DOI: 10.1093/nar/gkaa328.
18. Alva-Manchego F, Scarton C, Specia L (2021) The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4), 861–889. DOI: 10.1162/coli_a_00418.