Pre-Gamus

Reducing Complexity of Scientific Literature as a Support against Misinformation

Nico Colic[1], Jin-Dong Kim[2], Fabio Rinaldi[1]
[1]Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland
[2]Database Center for Life Science, ROIS-DS, Chiba, Japan
[1]{fabio.rinaldi, nicola.colic}@idsia.ch
[2]jdkim@dbcls.rois.ac.jp

Abstract

Scientific literature encodes a wealth of knowledge relevant to various users. However, the complexity of scientific jargon makes it inaccessible to all but domain specialists. It would be helpful for different types of people to be able to get at least a gist of a paper. Biomedical practitioners often find it difficult to keep up with the information load; but even lay people would benefit from scientific information, for example to dispel medical misconceptions. Besides, in many countries, familiarity with English is limited, let alone scientific English, even among professionals. All this points to the need for simplified access to the scientific literature. We thus present an application aimed at solving this problem, which is capable of summarising scientific text in a way that is tailored to specific types of users, and in their native language. For this objective, we used an LLM that our system queries using user-selected parameters. We conducted an informal evaluation of this prototype using a questionnaire in 3 different languages.

Keywords: LLM, text summarisation, text simplification

1. Introduction

Today's age of information abundance presents the challenge of navigating complex scientific literature, particularly biomedical. Even professional practitioners struggle to keep up with most recent literature. For example, a clinical doctor might be confronted with an unusual disease, and might need to get information which is only available in the literature, yet might not have the time and disposition to read a scientific paper which might or might not answer that particular information need (Cohen and Hersh, 2005). More so, the gap between scientific publications and lay peoples' understanding hinders dissemination of biomedical insights, fuelling misinformation and making informed decision-making difficult (Kandula et al., 2010). The COVID-19 pandemic, in particular, has underscored this vital role of accurate biomedical information in public health (Bin Naeem and Kamel Boulos, 2021). However, traditional scientific literature often presents dense, technical content, inaccessible to non-experts. This disparity, coupled with the rapid pace of scientific research, exacerbates the challenge of making the knowledge of biomedical literature usable by experts and lay people. This challenge is particularly pronounced among language groups with limited proficiency in English, as the majority of medical research is published in English (Frayne et al., 1996). This problem is not just limited to lay people, but to biomedical professionals, as well.

Addressing this challenge, we present an application that facilitates the understanding of biomedical texts, generating concise, easily comprehensible summaries and simplifications in the users' chosen language. This application, thus, is faced with three different tasks:

- Text simplification
- Text summarisation
- Machine translation

We begin with an overview of current research in these fields (section 2); explain the implementation of our application (sections 3 and 4) and our first evaluation (section 5). We close with a brief discussion of the results and future work (section 6).

2. Background

Text simplification is the process of making a text easier to understand by rewriting it in simpler language, while retaining the original meaning and key information. This often involves replacing complex words (lexical simplification) and structure (grammatical simplification) with simpler alternatives, rephrasing sentences to be more concise, and breaking down complex ideas into more manageable portions (Al-Thanyyan and Azmi, 2021).

Text summarisation, on the other hand, refers to the task of condensing a longer piece of text into a

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 196–201

shorter version while preserving its essential information. Traditionally, this process involves identifying the most important sentences or paragraphs and presenting them in a cohesive and concise manner, thereby providing a condensed overview of the original content (El-Kassas et al., 2021).

The rise of Large Language Models (LLMs) marks a paradigm shift in natural language processing, revolutionizing the way text generated and manipulated (Min et al., 2023). LLMs, such as OpenAI's GPT series and Google's BERT, have demonstrated unprecedented capabilities in capturing and generating human-like text across various domains, including biomedical literature. Essentially, these models are pre-trained on vast amounts of text data, learning to predict the next word in a sequence based on the context provided by the preceding words. This pre-training process allows LLMs to capture complex linguistic patterns and semantic relationships within the data. In particular, they have also proven to outperform previous approaches in the above tasks by far (Van Veen et al., 2023; Al-Thanyyan and Azmi, 2021; Kocmi and Federmann, 2023).

In the context of biomedical text summarisation, simplification, and translation tasks, leveraging LLMs offers several advantages. Rather than treating these tasks as independent processes, performing them simultaneously in one integrated workflow makes intuitive sense. LLMs possess the capability to understand complex biomedical texts, extract salient information, paraphrase content into simpler language, and translate it into multiple languages in a unified manner.

Mainstream use of LLMs such as ChatGPT, however, seems to be mostly business-focused (Wenxue Zou and Tang, 2023), and we presume that especially among more elderly professionals, adoption is hindered by the somewhat more modern chat-based interaction (Sarcar et al., 2023).

One danger of using LLMs, however, is their well-known tendency to *hallucinate*, that is, to invent facts (Zhang et al., 2023). For our application, this is particularly detrimental, as they may pose a health risk and can be a source of the very misinformation we're trying to combat with our application.

3. Implementation

3.1. Design Decisions

By integrating the tasks of text simplification, summarisation and translation into a single workflow, we harness the full potential of LLMs to streamline the process and hope to produce more coherent and linguistically accurate outputs. Because of this, we generally use the term *summary* when referring to the model's output in this paper, and mean it to also imply simplification and translation. As it turns out, however, some "summaries" can be longer than the original text, as the simplification aspect causes additional sentences to be included that explain complicated concepts.

We designed our application with a simple, easy-to-use interface to ease adoption by elderly professionals, in particular. In our application, we have decided to use *personas* to allow users to select their preferred level of text simplification. These personas cater to different user groups, allowing individuals to choose a simplification level that aligns with their comprehension needs and background knowledge. The personas available for selection include:

- *Teenagers*, who need simpler language and lexical simplification.
- *Adult Laypeople*, who need explanation of medical concepts.
- *Professional Clinician*, who mostly need summarisation.

The application is thus implemented as a web service with a simple-to-use interface that allows our users to obtain simplified summaries of biomedical texts in various languages. It is available here[1]; and its code can be found there[2]. The application is composed of 3 sections:

- Text selection
- Parameter selection
- Output

The application is written in `python` using `streamlit`, which facilitates the development of web applications.

3.2. Text Selection

The text selection section allows the user to either enter their own text, select from 10 pre-selected demonstration papers, or to enter a PubMed ID to automatically fetch the corresponding abstract. In the latter case, the application downloads the abstract for the given PubMed ID from the PubMed repository using the Entrez library, and displays it to the user in case the text needs editing.

3.3. Parameter Selection

The parameter selection allows the user to enter all the necessary information to generate the query that is sent to the ChatGPT API. Here, the **maximum number of tokens** indicates only a hard

[1] `pre-gamos.streamlit.app/`
[2] `github.com/Aequivinius/pre-gamos.ai`

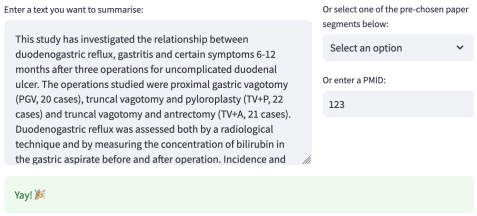

Figure 1: The user submitted a PubMed ID to download the corresponding abstract. It is displayed in the input field to the left, where they can make changes to the abstract, or enter a new text.

cut-off of the response in order to keeping costs incurred by using the API in check. However, it does not affect the length of the summary, as the model does not have a mechanism to control its output length. However, the choosing between different **personas** allows the user to direct the linguistic complexity of the summary and degree of simplification. Currently, the application supports *teenager*, *adult layperson* and *professional clinician* as possible personas. **Temperature** indicates to the model how much determinism is required, with lower temperature making its responses more deterministic (Ouyang et al., 2023). Finally, the user can select the target **language** of the response.

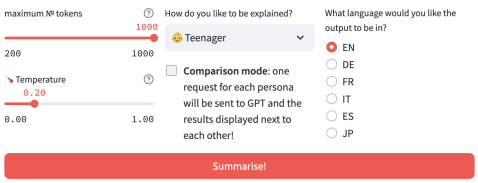

Figure 2: All values are left to their defaults. A simplification of the text selected above will be produced, suitable for a teenager in English, with moderate non-determinism.

There is an additional **Comparison mode** checkbox, which will generate simplifications for all personas simultaneously and display them side-by-side. This feature was added to allow for easier evaluation.

3.4. Output

Once the response from the model has been received, it is presented to the user. In addition, the Flesch readability score (Farr et al., 1951) is computed, and download buttons for different export formats displayed.

Since ChatGPT (3.5) only takes into account 5000 tokens for generating its responses (Floridi and Chiriatti, 2020), longer texts submitted through our applications are chunked and submitted individually for simplification.

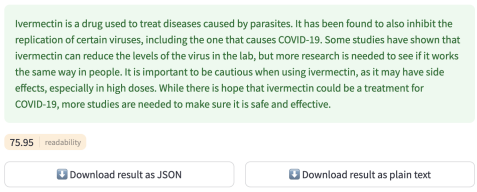

Figure 3: Resulting summary displayed along with its reading score, and different export options.

If the comparison mode is activated, this section will also display a pair-wise comparison of the responses generated for each persona.

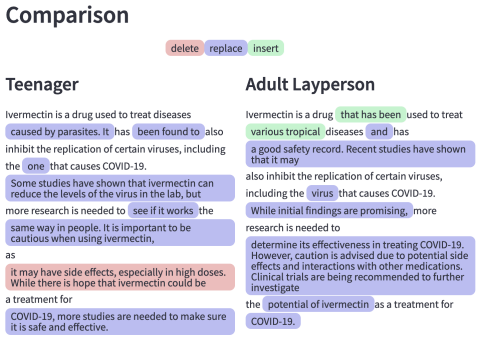

Figure 4: Outputs generated for teenagers and adult layperson displayed side-by-side, with differences highlightened.

For both the reading score and comparison, the text is naively tokenised by splitting on the empty space. In order to tokenise Japanese, where this approach is not viable at all, the dedicated tokeniser `fugashi` is employed (McCann, 2020).

The requests send to the model and its responses are cached; so for repeat queries, the application serves previous responses instantaneously.

4. PA-LLM

As an auxiliary tool, we developed a similar application using the same technology called PA-LLM[3]. However, here we allow the user to select *which* LLM the request is sent to, allowing them to compare the quality of the different models. Currently, only *GPT* and *BARD* are supported; but more APIs can be easily added.

PA-LLM also allows the user to upload the summaries and simplifications obtained through it to PubAnnotation, a repository for storing and displaying annotations of PubMed articles (Kim and

[3] `pa-llm.streamlit.app/`

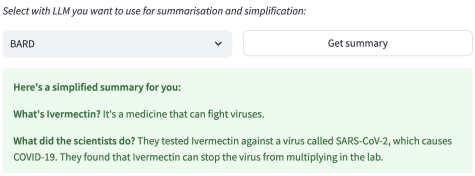

Figure 5: Showing part of the PA-LLM application, where the user has selected BARD to obtain their response.

Wang, 2012). The text simplifications are added as paragraph-level annotations. This allows the user to save their results, and superimpose different simplifications and more traditional annotations such as named entities.

While this is only an adjunct to the project, we realised that the explosive development of and progress in LLMs necessitates the need for researchers to be able to easily compare them.

5. Evaluation

Evaluating text summarisation automatically is notoriously difficult (Bhandari et al., 2020), and to evaluate it manually costly (Steinberger and Ježek, 2009). While we are preparing a formal evaluation, in the scope of this short-term project we can only present the results of an informal evaluation.

For this, we used in-depth questionnaires with 5 participants from 3 different language groups (1 for English, 3 for Spanish, 1 for Japanese). These questionnaires presented 9 summaries to the participants, generated for 3 different input texts and for 3 different personas, and asked participants to rate the summaries according to how appropriate they were for each of the personas, and how coherent (logical order of facts) and consistent (lack of contradiction).

For the input texts, we selected a balanced paper about the use of ivermectin during the Covid pandemic, a retracted paper about hydroxychloroquine, and finally a paper univocally endorsing the use of masks. We picked these different types of publications in order to see if it affects the model's performance.

The summaries were shown to the participants without any information about the persona they had been generated for; and participants were asked to rate their response on a scale of 1 to 5, with 5 being the highest.

In a second part, participants were shown a pair of summaries (as generated by the comparison mode described in 3.3), and were asked to indicate specificities that made one summary more appropriate for one persona than the other.

For English, the summaries generated for professional clinicians were always rated maximally appropriate for them; while for adult laypeople and teenagers they were only rated 4.3 and 4 out of 5, respectively.

For Spanish, summaries aimed at professional clinicians were rated only 4 out of 5 in average for appropriateness, somewhat higher than the 2.7 and 3.8 for adult laypersons and teenagers. In fact, the former was rated much more appropriate for professional clinicians (3.88 out of 5). For the Spanish-speaking participants of the questionnaire, however, we note difference in response profiles, with one participant having a very high variance of 2.2 across the questions; and the other two participants having a low variance of 0.3, but different averages of 2.7 and 4.7. In fact, the k-alpha score for the responses was -0.103, which indicates that there was a slight, but systematic disagreement between participants.

For Japanese, conversely, the summaries generated for professional clinicians were deemed most inappropriate, with only 3.7 out of 5, as opposed to the 4.3 rating both adult layperson and teenager texts received.

For coherence and consistency, all responses were rated 5 in all languages; with the exception for the responses from one Spanish participant who consistently rated responses either 2 or 3, for all questions.

We also computed Flesch reading ease scores, which gives a measure of how easily understandable a piece of text is. It is computed based on the average sentence length and the average number of syllables per word in the text; and higher scores point to simpler language (Farr et al., 1951).

The readability scores vary across the languages, but for all clearly show a difference for the generated summaries depending on the target persona. For English, the readability scores averaged to 75.9, 45.6 and 24.9 for teenagers, adult laypeople and professional clinicians across 10 sample abstracts. While the actual averages differ across languages (for German, for example, they are 40.5, 30.0 and 19.3 for teenagers, laypeople and professional clinicians, respectively), in all languages did summaries for teenagers result in the highest readability scores, and those for clinicians in the lowest.

We also asked participants to point out specific words or structures that made one summary more suitable for a teenager or for a professional clinician. For all five participants, they noted simpler terms for the former, and more accurate and more complicated terms for the latter. For Japanese, however, the participant noted that some terms were incorrectly translated.

For English, interestingly, our participant pointed out that the summary contained an explanation about a drug mentioned in the original text. The explanation itself, however, was not part of the original

text.

The questionnaires used for our evaluation and results can be found here[4].

6. Discussion and Conclusion

The evaluation above clearly shows that a more rigid evaluation is necessary; but already gives some first insights.

Firstly, for the evaluation of appropriateness, the same summaries received vastly different ratings from different survey participants. This shows that future evaluations need to include examples of appropriate summaries so that participants can calibrate their ratings.

Secondly, it shows that our approach is promising. The summaries were generally deemed most appropriate for the target personas they were generated for, and also their readability scores seem to support this.

Thirdly, a more careful evaluation of language differences is needed. While all texts for all languages were rated highly in terms of coherence and consistency, we noted some irregularities for some languages. For Japanese, it was the mistranslation of terms; for English, it was the hallucination of background information.

This last point deserves special attention, because it shows that even for summarisation tasks, the model does use knowledge not provided in the input text to generate its response. While in our particular case this was, in fact, helpful to make the text more easily understood, it can be a source of misinformation and put at risk the trustworthiness of applications such as ours. This is indeed in line with similar research (Zaretsky et al., 2024), where LLMs introduced misinformation on a similar simplification task.

7. Acknowledgements

The work described in this paper has been partially funded by the grant “Platform for an Epidemic-Related Guard Against Misinformation that is Understandable and grounded in Science” (PREGAMUS, Hasler Foundation) to Fabio Rinaldi, by the grant “Brisk.AI” (RPG2120, Leading House for the Latin American Region - University of St. Gallen) to Fabio Rinaldi, and by a “Strategic Research Projects” grant from ROIS (Research Organization of Information and Systems) to Jin-Dong Kim.

Special thanks to Oscar Lithgow Serrano (IDSIA) for contributing to the design of the experiments described in this paper, and to Yalbi Balderas Martinez (INER, Mexico) for contributing to the evaluation in Spanish.

[4] drive.google.com/drive/folders/12sBQDW_h59BWq-6dXgLZ1l6g0nHiQdwg

8. Bibliographical References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Salman Bin Naeem and Maged N Kamel Boulos. 2021. Covid-19 misinformation online and health literacy: a brief overview. *International journal of environmental research and public health*, 18(15):8091.

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Susan M Frayne, Risa B Burns, Eric J Hardt, Amy K Rosen, and Mark A Moskowitz. 1996. The exclusion of non-english-speaking persons from research. *Journal of general internal medicine*, 11:39–43.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366. American Medical Informatics Association.

Jin-Dong Kim and Yue Wang. 2012. Pubannotation-a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Paul McCann. 2020. fugashi, a tool for tokenizing japanese in python. *arXiv preprint arXiv:2010.06858*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

Sayan Sarcar, Cosmin Munteanu, Jaisie Sin, Christina Wei, and Sergio Sayago. 2023. Designing conversational user interfaces for older adults. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–5.

Josef Steinberger and Karel Ježek. 2009. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.

Yunkang Yang Wenxue Zou, Jinxu Li and Lu Tang. 2023. Exploring the early adoption of open ai among laypeople and technical professionals: An analysis of twitter conversations on #chatgpt and #gpt3. *International Journal of Human–Computer Interaction*, 0(0):1–12.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B. Blecker, and Jonah Feldman. 2024. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Network Open*, 7(3):e240357–e240357.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.