Towards Understanding LLM Generated Biomedical Lay Summaries

Rohan Charudatt Salvi[1], Swapnil Panigrahi[2], Dhruv Jain[2], Md. Shad Akhtar[2], Shweta Yadav[1]

[1]University of Illinois, Chicago, [2]Indraprastha Institute of Information Technology, Delhi
{rcsalvi2, shwetay}@uic.edu, shad.akhtar@iiitd.ac.in

Abstract

In this paper, we investigate the effectiveness of large language models in generating accessible lay summaries of medical abstracts, targeting non-expert audiences. We assess the ability of models like GPT-4, Biomistral, and LLaMA 3-8B-Instruct to simplify complex medical information, focusing on layness, comprehensiveness, and factual accuracy. Utilizing both automated and human evaluations, we discover that automatic metrics do not always align with human judgments. Our analysis highlights the potential benefits of developing clear guidelines for consistent evaluations conducted by non-expert reviewers. It also points to areas for improvement in the evaluation process and the creation of lay summaries for future research.

1 Introduction

In the dynamic field of medical research, rapid and clear dissemination of knowledge is essential. Automatic text summarization of medical abstracts serves as an efficient method for providing access to crucial information to both medical professionals and researchers, facilitating quicker and clearer information exchange (Luo et al., 2022). The need to communicate complex medical findings also extends to non-expert audiences such as caregivers, journalists, and the general public, who often struggle with the complex sentence structures and specialized terminology of medical literature (Guo et al., 2021; Goldsack et al., 2023b; Friedman et al., 2002; Korsch et al., 1968). Lay summaries of these abstracts are crucial in making scientific discoveries accessible to these groups by avoiding medical jargon and using clear, straightforward language (Guo et al., 2021; Chandrasekaran et al., 2020; Goldsack et al., 2023b). Initially, the generation of biomedical lay summaries utilized the fine-tuning of transformer-based models (Guo et al., 2021; Goldsack et al., 2022). However, recent progress has shown that large language models (LLMs) are especially effective in this area, with LLM-generated summaries not only surpassing traditional references in news datasets (Zhang et al., 2024a) but also demonstrating robust capabilities in generating comprehensible summaries in biomedical contexts, using techniques like retrieval-augmented generation (Guo et al., 2024) and zero-shot approaches (Jahan et al., 2024). Additionally, employing methods such as few-shot learning and the use of context-specific prompts has been shown to enhance the accuracy and relevance of the generated summaries (Pakull et al., 2024). Moreover, the use of few-shot learning with pre-trained LLMs has proven to be a robust approach at the BioLaySumm shared task 2023 (Turbitt et al., 2023).

In this paper, we investigate the effectiveness of LLMs in generating lay summaries from biomedical abstracts. Using a few-shot prompting strategy, we evaluate the performance of four distinct LLMs: GPT-4 (Achiam et al., 2023), Mistral-large-Instruct-2407(AI), LLaMA 3-8B-Instruct (Meta-Llama), and BioMistral (Labrak et al., 2024). We assess the ability of models to generate lay summaries, focusing on comprehensiveness, layness, and factual accuracy. Three research questions guide our evaluation:

1. How comprehensive are lay summaries generated by various LLMs?
2. How readable are biomedical summaries for lay audiences?
3. How faithful are lay summaries to their original abstracts?

Our evaluation methodology incorporates both automated and human assessments of the generated summaries on the publicly available PLABA (Attal et al., 2023) and the PLOS dataset (Goldsack et al.,

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 260–268

2022). We also conducted an in-depth analysis of various evaluation metrics that are widely used for lay summarization tasks.

In this study, we introduce detailed guidelines for the manual evaluation of lay summaries, designed as a comprehensive rubric that enables non-expert audiences to effectively assess lay summaries. By integrating human evaluations alongside automated metrics, we indicate the crucial role of human judgment in assessing summary quality, highlighting the inconsistencies that may emerge by relying solely on automatic metrics and discussing future directions for this research area.

2 Background

The BioLaySumm shared task was first introduced at the BioNLP Workshop during ACL 2023 (Goldsack et al., 2023a). This task focuses on abstractive summarization of biomedical articles, with the goal of creating lay summaries accessible to general audiences. It makes use of the PLOS and eLife corpus for this task and assesses summaries according to three criteria: Relevance, Readability, and Factuality. Each of these criteria is measured using one or more automatic metrics. Initial research on lay summarization primarily employed fine-tuned transformers such as BART (Guo et al., 2021), which were prominently featured at the BioLaySumm shared task in 2023. However, strong performance on the task was demonstrated by employing zero-shot and few-shot prompts with pretrained LLMs (Turbitt et al., 2023).

By the following year, the majority of proposed approaches by participating teams involved the use of LLMs (Goldsack et al., 2024). At BioLaySumm 2024, models such as GPT-3.5, GPT-4, and LLAMA3 were used in few-shot settings to generate lay summaries (Chizhikova et al., 2024). Another approach highlighted that fine-tuning LLMs like Biomistral with few-shot learning significantly enhances the accuracy of these summaries (Pakull et al., 2024). Additionally, recent research has explored retrieval-augmented generation (RAG), which utilizes LLMs and external knowledge sources such as Wikipedia to refine lay summarization (Guo et al., 2024). This RAG-based approach can be further enhanced by coupling it with reinforcement learning, optimizing the readability of the generated summaries (Ji et al., 2024).

3 Analysis on LLM generated plain language summaries

3.1 Lay Summary and Evaluation Guidelines

Based on our three research questions, we decided to evaluate the summaries on comprehensiveness, layness, and factuality. To ensure a consistent and robust assessment, we assume that our target audience has a limited background in biology (high school level) and intends to understand the article on a high level. Therefore, we aim for a lay summary that uses minimal medical jargon and effectively employs definitions or analogies to explain challenging biological concepts. Furthermore, it should be complete, explaining the topic, implementation, and findings of the study so that our intended readers can grasp the study (King et al., 2017).

Guided by previous research (Goldsack et al., 2022; Zhang et al., 2024b), our assessment methodology employs a 1-5 Likert scale for each defined metric. We sampled 15 abstracts from the PLABA and PLOS test set for lay summary generation by the models. Two undergraduates evaluated each generated summary using the guidelines. For both datasets, evaluators first read each abstract independently, and then the corresponding lay summaries. The evaluators were computer science majors who studied biology only until high school (10th grade).

We developed explicit scoring criteria, which was used for assessing summaries from both datasets, aiming to standardize evaluations and ensure reliability across different evaluators.

Comprehensiveness

Through comprehensiveness, we assess the extent to which the model-generated summaries encapsulate all the essential information necessary for a non-expert to grasp the high-level topic and significance of the research. The specifics of each score are as follows:

Score 1: The summary is incomplete; an evaluator cannot understand the topic or the significance of the research.

Score 2: The summary is partially complete; an evaluator gains a vague idea of the topic but cannot grasp the significance due to missing key details.

Score 3: The summary allows an evaluator to understand the topic but lacks important details that

convey the research's significance.

Score 4: The summary enables an evaluator to understand both the topic and significance, missing only minor details that could enhance understanding.

Score 5: The summary thoroughly covers all necessary information, allowing an evaluator to fully understand the topic and the significance of the research.

Layness

Layness measures the extent to which the model-generated text reduces medical jargon, enhances understanding of the summary by adding definitions and background context for the study's topic, and employs simpler sentence structures or analogies, making the content accessible to a general audience. The specifics of each score are as follows:

Score 1: There is not much difference between the plain text summary and the abstract.

Score 2: The plain text summary omits a few sentences that include jargon or omits a few words in sentences. It becomes easier to read but does not truly simplify the content.

Score 3: The summary is a mix of jargon and simple terms, as well as simple and complex sentences, along with some definitions. Laypersons may understand the main points but could find specific terms or sentences confusing.

Score 4: The summary is overall easy to understand, with the occasional presence of a complex sentence or medical terms that are not explained to the reader.

Score 5: The summary removes jargon or uses simple synonyms for them. If it cannot do either, it adds context for the evaluator to grasp the complex term. It uses simple, straightforward sentences or makes use of examples, making it easy for anyone to understand.

Factuality

Factuality measures the degree to which the information in the model-generated summaries remains true to the original abstracts. The specifics of each score are as follows:

Score 1: The study alters the findings or methodology, misrepresenting the study. The misrepresentation might be intentional or due to a misunderstanding of the original data.

Score 2: The study alters part of the study that can lead to misinterpretation of sections such as method or results, but not the entire study. These alterations could potentially skew the reader's understanding.

Score 3: The summary contains accurate information about the study but with frequent minor inconsistencies such as typos, incorrect figures, or omitting key details in findings. These inconsistencies do not majorly affect the overall integrity of the summary.

Score 4: The study contains accurate information about the study but with one or two minor exceptions. These exceptions are usually not critical to the study's main conclusions.

Score 5: The summary is fully factual and aligns completely with the study. It provides a detailed and accurate depiction of the original research without any significant omissions or errors.

3.2 Data

We evaluated our approach using the publicly available PLABA dataset (Attal et al., 2023) and the PLOS dataset (Goldsack et al., 2022). In the case of the PLOS dataset, we noted that associated author-written lay summaries presented readability challenges for a layman. Consequently, we used these summaries as the baseline for evaluating the effectiveness of our approach with the PLOS abstracts. We would like to point out that in Table 2, we keep the factuality score for them as 'N/A' since they were written by humans and not generated by a language model. For the PLABA dataset, we used the summaries generated by the fine-tuned Biomistral model as the baseline.

3.3 Evaluation Metrics

We evaluated the generated summaries for PLABA using several metrics. To measure comprehensiveness, we used: ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004), and SARI (Xu et al., 2016).

Method	Model	ROUGE-1	ROUGE-2	ROUGE-L	SARI	FKGL	DCRS	CLI	LENS	SummaC	AlignScore
Fine-tuned	Biomistral	**0.634**	**0.369**	**0.514**	**48.611**	12.278	9.876	13.812	56.433	54.5	79.9
Prompt	Biomistral	0.443	0.264	0.373	36.307	14.376	11.967	15.959	40.816	**82.3**	**87.7**
Prompt	GPT-4	0.548	0.213	0.351	40.973	9.692	**8.753**	**11.094**	**74.958**	34.0	75.7
Prompt	Mistral	0.528	0.206	0.335	40.722	**9.514**	9.001	12.063	72.021	32.3	70.7
Prompt	Llama3	0.547	0.258	0.385	41.680	11.764	9.474	13.462	67.407	47.9	81.4

Table 1: Model performance measured by automatic metrics on the PLABA dataset

For readability, we used the Coleman-Liau Index (CLI), Dale-Chall Readability Score (DCRS), Flesch-Kincaid Grade Level (FKGL), and LENS (Maddela et al., 2023). Additionally, we used AlignScore (Zha et al., 2023) and SummaC (Conv) (Laban et al., 2022) to assess the factuality of the summaries.

4 Our Analysis

RQ1: How comprehensive are lay summaries generated by various LLMs?

We observed in Table 1, in terms of automatic metrics, none of the prompt-based models were able to outperform the baseline. The fine-tuned Biomistral achieved scores of 0.634 and 48.611 on ROUGE-1 and SARI, respectively. This highlights that LLMs using prompts have added more abstractiveness to the plain text summaries, resulting in less overlap. However, in human evaluation, we found GPT-4 and Mistral achieved better ratings, scoring 4.165 and 4.565, respectively, compared to the baseline at 4.065. These final scores were computed by taking the average of the sum of ratings for the comprehensive facet.

These scores suggest that if models are better at restructuring or emphasizing key points by leveraging simple sentence structures and omitting non-essential details, it could enhance human understanding of the article's content and significance, leading to higher comprehensiveness scores (as presented in Table 3 in the appendix). We also observed this with the PLOS dataset, where GPT-4 and Mistral achieved higher comprehensiveness than the reference lay summaries. Lastly, the gap between automatic ROUGE and SARI scores and human ratings for models like Biomistral reveals the shortcomings of current metrics in fully assessing how much information the summary conveys. This indicates a need for novel metrics that better evaluate sentence structure and highlight informational content, essential to measure comprehensiveness.

RQ2: How readable are biomedical summaries for lay audiences?

We observe from Table 1 that GPT-4 gets the lowest scores on DCRS (8.75) and CLI readability ratings (11.09) and the highest on the LENS metric (74.96), indicating high simplicity and readability of the text it generates. Mistral achieves the lowest score on the FKGL metric (9.5). In the human evaluation, GPT-4 and Mistral again showed strong performance, with layness scores of 4.735 and 4.770, respectively. The lower FKGL rating indicates that Mistral most likely generated shorter sentences with simple syllables, whereas GPT-4 relies on more common words and potentially longer sentences than Mistral.

The readability metrics depend on sentence lengths (FKGL), word familiarity (DCRS), characters per word (CLI), and LENS evaluates simplification on a sentence level and not a paragraph(Xu et al., 2016). Thus, the scores potentially may look a bit aligned because we prompt models to generate simple sentences. What is not currently captured is a measure of how many complex words were omitted by the model, how many were simplified, and how many contexts or definitions were added since these are other characteristics apart from simple sentences on which humans evaluated the summaries for layness. This may potentially be the reason Mistral gets a higher rating on Layness than GPT-4 in the human evaluation. Additionally, this could also be a great metric to reflect on the nature of model-generated summaries, whether the LLM prefers to simplify sentences, omit jargon, replace terms, or add more context. For instance, in the PLOS evaluation, Mistral and GPT-4 received higher layness scores as they added definitions and used simpler terms, in contrast to the baseline summary that, despite its simple sentence structure, included medical jargon that reduced its layness.

Model	PLABA Dataset			PLOS Dataset		
	Comprehensiveness	Layness	Factuality	Comprehensiveness	Layness	Factuality
Baseline	4.065	4.165	4.230	4.1	2.233	N/A
GPT-4	4.165	4.735	4.150	4.767	**4.667**	4.7
Mistral	**4.565**	**4.77**	**4.835**	**4.767**	4.567	**4.8**
Llama3	4.000	4.099	4.395	4.533	3.2	4.667
Biomistral	3.520	3.105	3.935	3.933	2.2	3.967

Table 2: Human Evaluation Results on PLABA and PLOS Datasets

RQ3: How faithful are lay summaries to their original abstracts?

SummaC and AlignScores evaluated the factual alignment of generated summaries with original abstracts. High scores of 87.7 (AlignScore) and 82.3 (SummaC) for Biomistral in a prompt setting indicate these metrics favor text similar to the abstracts, despite occasionally favoring incomplete summaries. SummaC scores showed inconsistency with human evaluations, while AlignScores performed slightly better with respect to alignment with human evaluations. In the human evaluation, we observe Mistral achieving the highest rating for factuality, followed by Llama3, scoring 4.835 and 4.395, respectively. In the PLOS, along with GPT-4 and Mistral, Biomistral and Llama-3 achieved high factuality by maintaining sentence structures similar to the abstracts. However, we would like to highlight that the factuality score on both human and automatic metrics reflects solely intrinsic factuality. The LLMs also incorporate additional definitions and context to enhance user understanding, which may sometimes be inaccurate, leading to extrinsic hallucinations (Ramprasad et al., 2024). However, in this scenario, it is unreasonable to expect non-experts to identify and assess these inaccuracies.

5 Discussion

Our research examined how various language models generate lay summaries to simplify scientific findings. The Biomistral fine-tuned model effectively reflected the reference summaries and occasionally added definitions for complex terms. However, its prompt-based version often generated the same abstract or missed crucial information. LLama3 did simplify sentences, but it did not add necessary definitions and contexts, impacting its layness. Both GPT-4 and Mistral models excelled in creating understandable summaries, though they sometimes omitted detailed information. This underscores the trade-off between simplicity and factual accuracy in lay summaries.

Our results indicate that for prompt-based approaches, model size correlates with performance, with larger models like GPT-4 and Mistral showing superior adherence to guidelines and creativity in using analogies. Lastly, clear guidelines enhance the consistency of lay summary evaluations (as seen in Table 4 in appendix) by standardizing assessment criteria for non-expert reviewers.

6 Conclusion

In this study, we investigated how LLMs generate lay summaries for non-experts. Our findings show that while LLMs can simplify complex medical information effectively, there's a significant gap between automated metrics and human evaluations of the summary quality. This gap reveals the limitations of current evaluation methods and the need for metrics that align more closely with human perceptions of comprehensiveness, layness, and factuality. In the future, we plan to analyze other summarization methods and develop an effective human evaluation design that includes extrinsic factuality, on a larger dataset to refine our understanding of evaluation metrics perform across broader contexts.

7 Limitations

Our work has a few limitations. Firstly, LLMs exhibit an indeterministic nature, as they generate different lay summaries for the same input. Secondly, the format of the generated text often deviates from the example provided in the prompt, particularly in the cases of the Llama3 and Mistral models. Therefore, post-processing with regular expressions might be necessary to achieve the most effective results from these prompts. Additionally, we used a fixed prompt, which may not work equally well across all models, potentially leading to poorer-quality lay summaries. Lastly, there is a potential limitation concerning the training data of the LLMs. It is possible that the models were unintentionally trained on or exposed to the reference summaries used in our evaluations, which

could boost their performance on the lay summarization task.

8 Ethical Considerations

Although the LLMs perform well, they occasionally add additional definitions and context that could be incorrect. Moreover, in their efforts to simplify complex medical information, LLMs sometimes oversimplify, potentially leading to misinterpretations of the results. Therefore, non-experts should exercise caution when using LLM-generated lay summaries to ensure they are not misled by inaccuracies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mistral AI. Large Enough — mistral.ai. https://mistral.ai/news/mistral-large-2407/. [Accessed 16-10-2024].

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224.

Mariia Chizhikova, Manuel Carlos Díaz-Galiano, L Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2024. Sinai at biolaysumm: Self-play finetuning of large language models for biomedical lay summarisation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 804–809.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35(4):222–235.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023a. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomsa Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023b. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. *arXiv preprint arXiv:2309.17332*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171:108189.

Yuelyu Ji, Zhuochun Li, Rui Meng, Sonish Sivarajkumar, Yanshan Wang, Zeshui Yu, Hui Ji, Yushui Han, Hanyu Zeng, and Daqing He. 2024. Rag-rlrc-laysum at biolaysumm: Integrating retrieval-augmented generation and readability control for layman summarization of biomedical texts. *arXiv preprint arXiv:2405.13179*.

Stuart RF King, Emma Pewsey, and Sarah Shailes. 2017. An inside guide to elife digests. *Elife*, 6:e25410.

Barbara M Korsch, Ethel K Gozzi, and Vida Francis. 1968. Gaps in doctor-patient communication: I. doctor-patient interaction and patient satisfaction. *Pediatrics*, 42(5):855–871.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. *arXiv preprint arXiv:2210.04705*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Meta-Llama. Llama3/model$_{c}ard.mdatmainmeta-llama/llama3$.

Tabea MG Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A Horn, and Christoph M Friedrich. 2024. Wispermed at biolaysumm: Adapting autoregressive large language models for lay summarization of scientific articles. *arXiv preprint arXiv:2405.11950*.

Sanjana Ramprasad, Kundan Krishna, Zachary Lipton, and Byron Wallace. 2024. Evaluating the factuality of zero-shot summarizers across varied domains. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 50–59, St. Julian's, Malta. Association for Computational Linguistics.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. Mdc at biolaysumm task 1: Evaluating gpt models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024a. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhihao Zhang, Tomas Goldsack, Carolina Scarton, and Chenghua Lin. 2024b. ATLAS: Improving lay summarisation with attribute-based control. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 337–345, Bangkok, Thailand. Association for Computational Linguistics.

A Prompt

This appendix outlines the prompt employed for generating lay summaries as described in this paper. This prompt was used across all four models with a minor change. The symbol '#' was included in the prompt for the Mistral and LLaMA3 models.

Prompt:

You are a biology teacher in a high school and want to teach students in 10th grade about a research study. Your goal is to convey the information in the abstract in plain and easy to understand language that students can follow.

You decide to generate a plain text for the same abstract keeping in mind what makes a text simple and easy to understand.

1. It attempts to avoid as much scientific jargon as possible. If it cannot avoid it, then it replaces it with easy to understand synonyms.

2. It has an explanation and definition for complex biological terms and can include simple real-life examples to make it easier to understand.

3. The sentence structure is simple, and the text has a good coherent flow.

4. The word count cannot exceed 300 words.

5. The text should have all the important points. And if words are replaced by simpler terms, it is good to connect them to original words by referencing them using brackets.

6. Ensure the text is factually correct, this includes definitions, synonyms, important numeric figures, and findings.

Here is an example of what you should generate:
Abstract: [Sample Abstract]
Here is the rationale to use:

1. Read the abstract thoroughly to grasp main ideas.

2. Avoid scientific jargon; use simple synonyms or explanations.

3. Define complex terms and provide real-life examples for clarity.

4. Use simple sentence structures for better understanding.

5. Keep the text under 300 words.

6. Include all important points, referencing original terms in brackets when replacing them.

7. Ensure all information is factually correct, including definitions and key data.

Generated Plain Text: [Sample Plain Summary]
Abstract: [Test set abstract]
Generated Plain Text:

B Implementation Details

We accessed the GPT-4 ("gpt-4") and Mistral Large models("mistral-large-2407") using the API provided by OpenAI and Mistral respectively. The Llama-3-8B model and Biomistral were accessed through Huggingface and were run on 1 NVIDIA Tesla T4-16GB and 1 NVIDIA Tesla A100 Ampere-40GB respectively. Biomistral used as the baseline was fine-tuned on the PLABA dataset through QLORA (Dettmers et al., 2024).

We generated plain text abstracts for all 148 abstracts in the PLABA test dataset using the one-shot prompt approach for each model. For Biomistral and Llama3 we observed that the summaries contained the prompt prefix, the test abstract. Moreover, for Llama3 there were cases where we found rationale or references being generated after the plain text. Thus, we employed a post-processor using regular expression for both models to extract the plain text summary.

Model	Text
Abstract	To reduce the damage to posteromedial knee capsule, we developed a direct extra-articular arthroscopic approach for excision of symptomatic popliteal cysts. This study aimed to demonstrate the surgical technique and present the 2-year follow-up results. Cystectomy is performed by extra-articular surgical approach through a high posteromedial portal. Twenty-one consecutive patients diagnosed of symptomatic popliteal cysts with failed initial conservative treatments were included. At a median follow-up of 29.4 months, all knees had improved clinical function assessed by Rauschning and Lindgren knee classification ($p<0.001$). The cysts were either disappeared (95.2%) or reduced in size (4.8%). Only one (4.8%) patient had recurrent cyst, which was solved after ultrasound-guided aspiration. This direct extra-articular arthroscopic technique could be a feasible alternative for treatment of symptomatic popliteal cysts.
Biomistral fine-tuned (baseline)	To reduce the damage to the knee capsule, we created a new way to remove cysts (growths) in the knee. This new way is called a direct extra-articular arthroscopic approach. This study aimed to show how the surgery is done and present the 2-year follow-up results. Cystectomy is performed by a high posteromedial portal. Twenty-one patients with cysts in the knee were included. At a median follow-up of 29.4 months, all knees had improved function. The cysts were either disappeared or reduced in size. Only one patient had a recurring cyst, which was solved by draining the cyst. This direct extra-articular arthroscopic technique could be a good way to treat cysts in the knee.
Mistral-large	A popliteal cyst is a fluid-filled swelling that forms at the back of the knee. It can cause pain and other symptoms. To treat these cysts without damaging the nearby knee structures, doctors developed a new surgical method using a tiny camera called an arthroscope. This study aims to explain this new method and show the results two years after the surgery. The surgery is done through a small opening on the inner back side of the knee, allowing doctors to remove the cyst without going into the knee joint itself. Twenty-one patients who had painful popliteal cysts and didn't get better with other treatments were included in the study. After an average follow-up of about 29 months, all the patients reported improved knee function based on a specific knee classification system. In almost all cases (95.2%), the cysts disappeared, and in a few cases (4.8%), the cysts became smaller. Only one patient (4.8%) had the cyst come back, but it was successfully treated with an ultrasound-guided needle to drain the fluid. This new surgical technique, which involves removing the cyst from outside the knee joint using an arthroscope, seems to be a good option for treating painful popliteal cysts.

Table 3: Summaries generated by Biomistral and Mistral-large for an abstract in PLABA. Colors in the text indicate: additional background information (brown), simpler terms used (blue), and simplified sentences (orange).

Model	Comprehensiveness	Layness	Factuality
Baseline	0.667	0.45	0.880
GPT-4	0.435	0.717	1.000
Mistral	0.690	0.755	0.755
Llama3	0.606	0.74	0.698
Biomistral	0.693	0.822	1.000

Table 4: Inter annotator agreement (Cohen's Kappa) on PLABA Dataset

Model	Comprehensiveness	Layness	Factuality
Baseline	0.688	0.688	N/A
GPT-4	0.318	0.700	0.300
Mistral	0.800	0.605	1.00
Llama3	0.744	0.615	0.412
Biomistral	0.783	0.455	0.503

Table 5: Inter annotator agreement (Cohen's Kappa) on PLOS Dataset