Team YXZ at BioLaySumm: Adapting Large Language Models for Biomedical Lay Summarization

Jieli Zhou[1*], Cheng Ye[2*], Pengcheng Xu[3], and Hongyi Xin[1]

[1]UM-SJTU Joint Institute, Shanghai Jiao Tong University
[2]Zhejiang Laboratory
[3]University of Illinois Urbana-Champaign
[*]These authors contribute equally to this work.
Correspondence: zhoujieli@sjtu.edu.cn, hongyi.xin@sjtu.edu.cn

Abstract

Biomedical literature are crucial for disseminating new scientific findings. However, the complexity of these research articles often leads to misinterpretations by the public. To address this urgent issue, we participated in the BioLaySumm task at the 2024 ACL BioNLP workshop, which focuses on automatically simplifying technical biomedical articles for non-technical audiences. We conduct a systematic evaluation of the SOTA large language models (LLMs) in 2024 and found that LLMs can generally achieve better readability scores than smaller models like Bart. Then we iteratively developed techniques of title infusing, K-shot prompting, LLM rewriting and instruction fine-tuning to further boost readability while balancing factuality and relevance. Notably, our submission achieved the first place in readability at the workshop, and among the top-3 teams with the highest readability scores, we have the best overall rank. Here, we present our experiments and findings on how to effectively adapt LLMs for automatic biomedical lay summarization. Our code is available at `https://github.com/zhoujieli/biolaysumm`.

1 Introduction

The biomedical literature is one of the most important information sources for researchers to share their latest discoveries. However, the increasing volume of information has become overwhelming, e.g. PubMed alone hosts over 36 million papers, with more than one million new articles added annually (Jin et al., 2024). This information deluge makes it challenging for even specialized biomedical experts to keep up with the latest research, let alone the general public. General public, although having a keen interest in biomedical research due to its relevance to everyday life, may be prohibited by the difficult biomedical terminology, experimental setups, or the metric abbreviations. Currently, media outlets play a crucial role in bridging the gap between scientific literature and public understanding (Peters, 2013). However, media often lack the necessary context when reporting new studies, which can lead to exaggerated claims. For instance, reports may claim that certain diets promote longevity[1], but closer inspection of the literature reveals that these claims are typically based on preliminary animal studies and lack robust support in human studies (Bruijnis et al., 2013; Janssen et al., 2019; Murphy et al., 2014).

In recent years, with the rapid advancements of Transformer-based natural language processing algorithms, many intelligent systems have been developed to automate the process of explaining and summarizing research papers to lay people. Dangovski et al. (2021) collected 100,000 webpages from Science Daily, a popular press release websites for research papers, and finetuned a Bert model to produce automatic science journalism; Cohan et al. (2018) developed a hierarchical disource-aware attention model to effectively summarize the long and hierarchical research paper. Interactive research paper summarization systems like SciSummary [2], Scholarcy[3] and SciSpace[4] can help researchers quickly get the gist from the literature and develop a map of connected research (Nahas, 2024).

In particular, biomedical literature contains many domain jargons which lack readable explanations; background stories are frequently omitted, and findings are hidden in obscure metrics names without direct discussions. In essense, the literature cannot "talk" by itself and answer general public's questions. To advance the research in automatic lay summary generation for biomedical literature, the ACL BioNLP task of BioLaySumm was started

[1] `https://www.hsph.harvard.edu/nutritionsource/media/`

[2] `https://scisummary.com/`

[3] `https://www.scholarcy.com/`

[4] `https://typeset.io/`

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 818–825
August 16, 2024. ©2024 Association for Computational Linguistics

in 2023 (Goldsack et al., 2023, 2024). Accompanying this task, a lay summary dataset was curated to facilitate model training and evaluation (Goldsack et al., 2022). The dataset contains 31,020 article-summary pairs[5] from two journal series, Public Library of Science (PLOS) and eLife. Each article-summary pair contains the paper full text and the corresponding lay summaries written by paper authors (PLOS) or journal editors (eLife). Due to the different sources of lay summaries, the characterities of these summaries vary, e.g. PLoS lay summaries are on average 175.6 words and eLife summaries are on average 347.6 words [6]. The task of BioLaySumm is then given the full article text, a system should automatically output a lay summary, which will be measured in 10 metrics of factuality, readability and relevance.

In the previous iteration of BioLaySumm, we observe that the best performing teams used advanced large language models (LLMs) like GPT-3.5 to produce zero-shot lay summaries (Turbitt et al., 2023) and augment training data (Sim et al., 2023). In addition, we have frequently used tools like GPT4, Kimi.ai, and ChatDoc.com to facilitate rapid understanding of research papers. In this work, we make several key contributions. Firstly, we conduct a comprehensive set of experiments demonstrating that while directly prompting Large Language Models (LLMs) can enhance the readability of lay summaries, it often compromises their factuality and relevance. Secondly, we have developed a suite of adaptation techniques, including **title infusion**, **K-shot prompting**, **LLM rewriting**, and **instruction fine-tuning**, which enable LLMs to generate lay summaries that are well-balanced in terms of factuality, relevance, and readability. Notably, our approach achieved first place in readability in the 2024 BioLaySumm competition, while also maintaining a strong balance across the other evaluated metrics.

2 Dataset and Evaluation Metrics

The dataset (Goldsack et al., 2022) of 31,020 article-summary pairs is divided into three parts, train, validation and test, where the train and validation sets are given for model development, and the lay summaries of the test set are hidden. The distribution of the data is shown in Table 1. Model outputs are submitted to a competition website and the scores are computed in around 1.5 hours.

Table 1: **Distribution of the BioLaySumm data**

Dataset	Train	Val	Test	# words
PLOS	24,773	1,376	142	175.6
eLife	4,346	241	142	347.6

Model-generated lay summaries are evaluated against 10 metrics of three categories, relevance, readability and factuality.

Relevance: ROUGE or Recall-Oriented Understudy for Gisting Evaluation (1, 2, and L) (Lin, 2004) and **BERTScore** (Zhang et al.) are the relevance metrics which uses lexical or embedding based methods to measure the overlap between the generated summaries to the reference ones. The higher the scores the more relevant the summaries.

Readability: Flesch-Kincaid Grade Level (**FKGL**)(Kincaid et al., 1975), Dale-Chall Readability Score (**DCRS**)(Chall and Dale, 1995), Coleman-Liau Index (**CLI**)(Coleman and Liau, 1975) are lightweight readability metrics that predict the US grade level of education needed to understand the generated summaries, so the lower the scores, the more readable the summaries. Whereas **LENS** (Learnable Evaluation Metric for Simplification) (Maddela et al., 2022), a readability metric trained on human judgment data, aligns more closely with human preferences, the goal is to achieve a higher LENS score.

Factuality: Alignscore (Zha et al., 2023) is an automatic factual consistency metric for checking whether all information in the summary is contained in the reference. Similarly, **SummaC** (Laban et al., 2022) or Summary Consistency is a natural language inference (NLI) method that measures the factual consistency between generated summary and reference sentence-wise. The goal is to maximize these two factuality metrics.

3 Method and Results

Following previous work (Turbitt et al., 2023), we mainly focus on using LLMs in this work. In our settings, the input is the article text and relevant metadata, and different LLMs are prompted along with a system prompt to generate the lay summary. In this section, we explain our methods and the results in detail.

[5] https://biolaysumm.org/

[6] https://aclanthology.org/2023.bionlp-1.44.mp4

Table 2: **Large Language Models' Performances on BioLaySum test set**. 5 SOTA LLMs are prompted to generate lay summarization for the test dataset, the title for each article is retrieved and added to the prompt. Overall, LLM models, although lacking in relevance and factuality, generally outperform the baseline model in readability metrics. Notice that the Llama3 models beat the baseline by a large margin in the LENS metric.

Method	Relevance				Readability				Factuality	
	R1 ↑	R2 ↑	RL ↑	BS ↑	FKGL↓	DCRS↓	CLI↓	LENS↑	AS ↑	SC ↑
Baseline (BART)	**0.4696**	**0.1395**	**0.4358**	**0.8623**	12.0359	10.1475	13.4852	48.0963	**0.7788**	**0.7026**
Claude-3-Opus	0.4426	0.1194	0.3966	0.8506	13.2162	10.1755	15.2807	73.093	0.5794	0.4912
Gemini-1.5-pro	0.4405	0.11	0.4076	0.8554	13.6968	10.4408	16.0192	73.5596	0.6823	0.5004
GPT-4	0.4299	0.0983	0.3837	0.8524	14.362	10.7441	15.9209	72.2514	0.5818	0.452
Llama3-8B-Instruction	0.4152	0.1065	0.3847	0.854	11.6099	**9.2043**	**12.8627**	80.1454	0.6539	0.5172
OpenBioLLM-Llama3-70B	0.4104	0.0993	0.3801	0.855	**11.0162**	9.369	12.9965	**81.2052**	0.7018	0.5463

3.1 Title Infusing

We observe that article titles are missing in the test data. However, we think titles are essential for summarization since it encapsulates the high-level description of the article by the authors. To retrieve the title for the articles, we used **BeautifulSoup4**[7] to retrieve the article titles from the eLife and PLOS website based on the DOI urls. The titles are then infused into the prompt for LLMs to better position its lay summarization. All models in this work have the information of the title at inference time.

3.2 Large Language Models

Since the last iteration of BioLaySumm, more advanced LLMs emerged which showcased better reasoning and text processing skills. We benchmarked five SOTA LLMs against the official baseline method based on BART: Anthropic's Claude-3-Opus[8], Google's Gemini-1.5-Pro[9], OpenAI's GPT-4 [10] and Meta's Llama-3 [11]. These LLMs have over billions or even hundreds of billions of parameters and are very good at instruction following. We also included, OpenBioLLM-LLama3-70B (Ankit Pal, 2024), which is a Llama-3-70B model finetuned on biomedical domain and is reported to specialize in various BioNLP tasks. Due to the prohibitive costs, we limit the input tokens to only the abstract part of the test set. The result of these 5 LLMs on test data is shown in Table 2. It is surprising that even the most advanced LLMs cannot surpass the Bart-baseline model in terms of relevance or factuality. However, it is evident that LLMs hold a distinct advantage in readability, particularly as measured by the LENS score. Next, building on these findings, we aimed to enhance readability while also improving the metrics for relevance and factuality.

3.3 Finetuning for Relevance and Factuality

LLMs are autoregressive transformer models which are trained to predict the next token[12]. Without further fine-tuning, they lack capabilities in producing factual and relevant summaries, as shown in table 2. Inspired by the success of techniques like Supervised Fine-Tuning (SFT) (Alt et al., 2019) and instruction tuning (IT) (Wu et al., 2024) on LLMs in solving downstream NLP tasks, we adopt a parameter efficient finetuning technique called Low-Rank Adaption (LoRA) (Hu et al., 2021) to further fine-tune two models, one for PLOS and another for eLife, due to the different characteristics of these two journal lay summaries as shown in table 1. We first construct instruction tuning data from the article-summary pair by including an instruction prompt, **"Write lay summary for the given input (a summary that is suitable for non-experts). Here is the article:..."**. We use the article full text (up to 8K as of the Llama3's context window) as input, and the corresponding lay summary as output. We report the result on the validation dataset in Table 3, and for the test data submission, we perform instruction fine-tuning on the entire train-val dataset and predict on the test dataset. The experiments were done on a GPU server with 8 NVIDIA RTX 4090Ti 48GB GPUs. We used Llama-3-8B-Instruct [13] as our base model,

[7] https://beautiful-soup-4.readthedocs.io/en/latest/

[8] https://www.anthropic.com/api

[9] https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

[10] https://openai.com/index/gpt-4/

[11] https://llama.meta.com/llama3/

[12] https://huggingface.co/docs/transformers/llm_tutorial

[13] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 3: Performance of Llama-3-8B model on the validation dataset with and without Instruction-finetuning. With finetuning the metrics of 8B model compare favorably to larger un-finetuned 70B model, and is consistently better than the baseline Llama3-8B model in relevance and factuality.

Method	Relevance				Readability				Factuality	
	R1 ↑	R2 ↑	RL ↑	BS ↑	FKGL↓	DCRS↓	CLI↓	LENS↑	AS ↑	SC ↑
Llama3-8B	0.4185	0.1049	0.3855	0.8563	11.7285	9.2582	**12.9303**	80.1033	0.6316	0.5282
Llama3-8B-FT	**0.4297**	**0.111**	**0.3977**	0.8513	11.8257	**8.6854**	13.1627	**80.7012**	**0.70628**	**0.5816**
OpenBioLLM-70B	0.4180	0.1071	0.3821	**0.8583**	**11.6109**	9.8948	13.3855	80.0218	0.6966	0.5763

and training goes on for 3 epochs, with learning rate of 5e-5. More details are available in our code repo.

3.4 K-Shot Prompting for Factuality

In constructing the prompts for LLMs, we used in-context learning (Dong et al., 2022) or few-shot learning [14], in which we provide some examples to the LLMs. For test data, we proposed to use the top K semantically similar articles in the train and validation dataset or K-shot prompting. To compute the semantic similarities, we used a recent state-of-the-art embedding model BGE M3 (Chen et al., 2024) from BAAI[15] to compute and pick the K most similar abstracts from the train-val dataset to the given test article. Together with the input article, the K-Shot example pairs are sent to the LLMs. Due to time limit, we set K=1. We experimented this K-shot Prompting technique with the smaller Llama-3-8B model, and observe better performance metrics in factuality as shown in Table 4. This implies LLMs can be prompted with semantically similar lay summary example to be more grounded in the origincal text, boosting factuality.

Table 4: Factuality Scores for Llama3-8B-Instruction Models with and without K-shot Prompting on the test dataset, K=1

Model	AlignScore↑	SummaC↑
Llama3-8B (kshot)	**0.7523**	**0.5582**
Llama3-8B	0.6539	0.5172

3.5 LLM-rewrite for Readability

In observing the results from the five LLMs as in table 2, we hypothesized that Bart model finetuned on the given dataset does better in relevance and factuality while LLMs may be better in readability. We aim to boost the readability of Bart, thus confirming our hypothesis that LLMs' lay summaries are better in readability, so we developed a strategy of LLM-rewrite, where we first finetuned a Bart model, then used a specialized biomedical LLM, OpenBioLLM-LLama3-70B to rewrite the summary. It is observed that readability of the LLM-rewrote Bart summary is improved and especially in the LENS metric as shown in Table 5. However, Bart+rewrite still have consistently lower readability compared to LLM (OpenBioLLM-70B).

Table 5: Readability Metrics for BART-Finetuned Models with and without rewrite with OpenBioLLM-LLama3-80B.

Model	FKGL↓	DCRS↓	CL↓	LENS↑
BART	12.5053	9.948	13.5215	60.1214
BART+rewrite	11.1444	9.9033	13.4803	80.3856
LLM-70B	**11.0162**	**9.369**	**12.9965**	**81.2052**

4 Results and Conclusion

We constructed our final submission based a combination of three techniques: title infusing, instruction finetuning and K-Shot prompting. Overall, our submission achieved the 1st place in the readability catorgy, and had the better overall score compared with the top-3 team in the readability category (29th vs. 36th and 40th)

Team	**Relevance**	**Readability**	**Factuality**	**Overall**
YXZ	0.6845	0.8395	0.3190	29th
NLPSucks	0.3870	0.8297	0.5299	36th
jimmyapples	0.7008	0.8270	0.1875	40th

In this work, through systematic experiments on the capabilities and limitatons of the SOTA Large Language Models, we developed strategies to adapt LLMs for the task of BioLaySumm, and achieved the best result in readability while balancing other metrics. We provide experiment details and findings for researchers to keep advancing in this field.

[14] https://www.promptingguide.ai/techniques/fewshot

[15] https://www.baai.ac.cn/english.html

Limitations

Our study demonstrates effective strategies for adapting large language models (LLMs) to biomedical lay summarization, achieving good performances in readability while balancing factuality and relevance. However, several limitations warrant attention. First, while our techniques, title infusion, K-shot prompting, instruction tuning, and LLM rewriting showed promising results on the BioLaySumm datasets, their applicability to other types of biomedical papers, such as systematic reviews, remains untested. These methods may require further domain-specific adaptations for them to work well on other datasets. Second, the computational demands of current state-of-the-art LLMs are high, restricting their use in resource-limited settings. We will explore techniques to reduce model sizes and enable them for low-resource scenarios. Third, our efforts to balance readability with factuality and relevance reveal inherent trade-offs, that is enhancing readability may sometimes oversimplify complex biomedical concepts, risking factual accuracy and detail omission. We plan to develop more balanced summarization strategies in future studies. Lastly, the ethical implications of using LLMs for generating lay summaries in highly sensitive biomedical fields are significant, especially given the risk of misinformation due to LLMs' hallucination issues. We will develop methods to automatically detect the toxic contents in the LLM outputs, and develop effective methods to correct them. In conclusion, while our study advances the field of LLM-based biomedical summarization, ongoing efforts are necessary to address these limitations and enhance the reliability and scope of our methodologies.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv preprint arXiv:1906.08646*.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

MRN Bruijnis, FLB Meijboom, and EN Stassen. 2013. Longevity as an animal welfare issue applied to the case of foot disorders in dairy cattle. *Journal of Agricultural and Environmental Ethics*, 26:191–205.

Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited: The new dale-chall readability formula. *(No Title)*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakov, and Marin Soljačić. 2021. We can explain your research in layman's terms: Towards automating science journalism at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12728–12737.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. 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*, 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 *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lieneke K Janssen, Nadine Herzog, Maria Waltmann, Nora Breuer, Kathleen Wiencke, Franziska Rausch, Hendrik Hartmann, Maria Poessel, and Annette Horstmann. 2019. Lost in translation? on the need for convergence in animal and human studies on the role of dopamine in diet-induced obesity. *Current addiction reports*, 6:229–257.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2024. Pubmed and beyond: biomedical literature search in the age of artificial intelligence. *Ebiomedicine*, 100.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.

Tytus Murphy, Gisele Pereira Dias, Sandrine Thuret, et al. 2014. Effects of diet on brain plasticity in animal and human studies: mind the gap. *Neural plasticity*, 2014.

Kamal Nahas. 2024. Is ai ready to mass-produce lay summaries of research articles? *Nature*.

Hans Peter Peters. 2013. Gap between science and media revisited: Scientists as public communicators. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14102–14109.

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and Sarvnaz Karimi. 2023. Csiro data61 team at biolaysumm task 1: Lay summarisation of biomedical research articles using generative models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 629–635.

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.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2024. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 LLMs

In total, we spent around 200 USD on prompting various LLMs to generate lay summaries. For Claude-3-opus, we used API from Antropic, and for GPT-4, we used API from OpenAI. For Gemini-1.5-pro we used API from Google. Due to the prohibitive costs, we only tested the LLMs' performances on the test dataset.

For the local LLM models, BART, Llama3-8B-Instruction and OpenBioLLM-Llama3-70B, we downloaded the weights and checkpoints from Huggingface[16] and Modelscope[17]. All experiments were done on a single GPU-server with 8 NVIDIA RTX 4090 GPUs. We also used the Google Colab platform [18] for ideation and prototyping.

A.2 Prompts for LLMs

Our system prompt is set to be

System Prompt: "Please write a corresponding lay summary based on the content of the article provided. Requirements:
1. Lay summary needs to be easy to understand so that it can be quickly understood by non-specialists;
2. Lay summary needs to grasp the point of the article and be concise and to the point;
3. Just output a lay summary, no other content is required."

Then, the computed K-Shot example pairs are added along with the input (abstract/full text). An example of K-shot prompt is shown as below.

```
messages = [
{"role": "system",
"content": system_prompt},
{"role": "user",
"content": "Abstract: xxx"},
{"role": "assistant",
"content": "Lay summary: xxx"},
{"role": "user",
"content": "Abstract: xxx"},
```

[16] https://huggingface.co/

[17] https://www.modelscope.cn/

[18] https://colab.research.google.com/

```
{"role": "assistant",
 "content": "Lay summary: xxx"},
{"role": "user",
 "content": "Abstract: xxx"},
{"role": "assistant",
 "content": "Lay summary: xxx"},
{"role": "user",
 "content": "Abstract: {xxx}"}
 ]
```

A.3 Llama-3 Instruction Tuning

We construct instruction tuning dataset based on the given train and validation dataset, an example of the instruction tuning data is shown as below.

```
{
  "instruction": "Write lay summary
  for the given input (a summary that
  is suitable for non-experts).
  Here is the article.",
  "input":  article,
  "output": lay summary
}
```

Then we used unsloth [19] and LoRA [20] to conduct instruction finetuning. Code implementation is made available in our code repo https://github.com/zhoujieli/biolaysumm.

A.4 Analysis on Readability

To further corroborate our findings on readability, we evaluated local LLMs on the validation dataset and analyze their readability performances over the two journal datasets. The results are shown in Figure 1 and Figure 2. This is in line with our findings, and highlights a novel observation: lay summaries written by journal editors in eLife have better and harder-to-beat readability whereas in PLOS, authors' own lay summaries are generally worse in readabilities. This observation provides a unique opportunity for lay summarization systems like ours to help in generating lay summaries for the authors, and make their research more readable.

[19] https://www.unsloth.ai/blog/llama3

[20] https://huggingface.co/docs/diffusers/training/lora

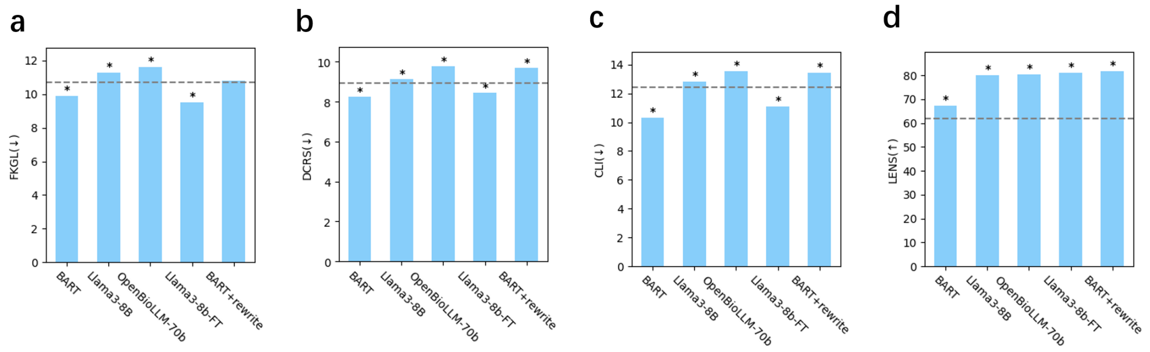

Figure 1: A comparison of readability performance of lay summaries generated by different methods on the eLife validation dataset. (a) Median FKGL score, (b) Median DCRS score, (c) Median CLI score, (d) Median LENS score. In each figure, the gray horizontal dashed line represents the median readability score of the reference lay summary in the validation set. The (*) symbol indicates that the Wilcoxon signed-rank test was passed (p-value less than 0.05), meaning that the readability score of the generated lay summary is significantly better or worse than that of the reference lay summary. From (d), it can be observed that leveraging the powerful language expression capabilities of LLM significantly enhances the LENS scores of the generated lay summary.

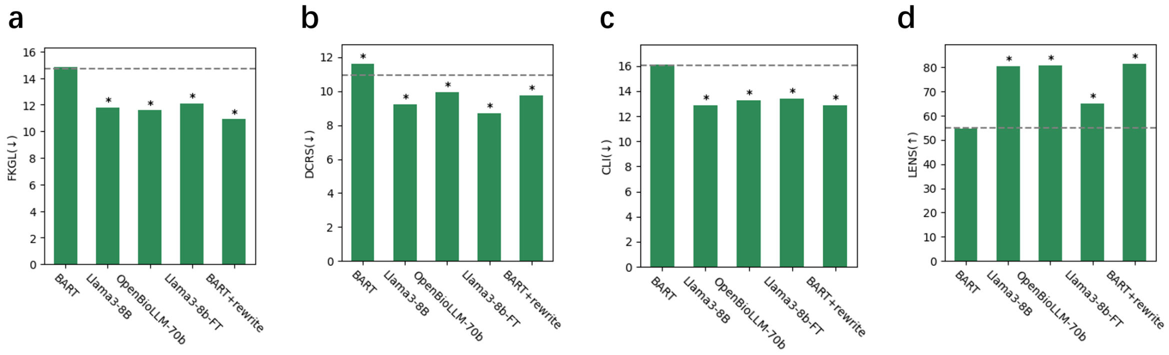

Figure 2: Analysis on models' readability performance on the PLOS validation dataset. (a) Median FKGL score, (b) Median DCRS score, (c) Median CLI score, (d) Median LENS score. gray horizontal dashed line represents the median readability score of the reference lay summary in the validation set. The (*) symbol indicates that the Wilcoxon signed-rank test was passed (p-value less than 0.05), meaning that the readability score of the generated lay summary is significantly better or worse than that of the reference lay summary. We found that the lay summaries written by authors in the PLOS journal generally have significantly lower scores in (a), (b), (c), and (d) readability compared to those generated with the help of LLM.