WisPerMed at BioLaySumm: Adapting Autoregressive Large Language Models for Lay Summarization of Scientific Articles

Tabea M. G. Pakull[1,2],   Hendrik Damm[2,3],   Ahmad Idrissi-Yaghir[2,3],
Henning Schäfer[1,2],   Peter A. Horn[1], and Christoph M. Friedrich[2,3]

[1] Institute for Transfusion Medicine, University Hospital Essen,
Hufelandstraße 55, 45147 Essen, Germany

[2] Department of Computer Science, University of Applied Sciences and Arts Dortmund,
Emil-Figge-Straße 42, 44227 Dortmund, Germany

[3] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE),
University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
tabeamargaretagrace.pakull@uk-essen.de

Abstract

This paper details the efforts of the WisPerMed team in the BioLaySumm2024 Shared Task on automatic lay summarization in the biomedical domain, aimed at making scientific publications accessible to non-specialists. Large language models (LLMs), specifically the BioMistral and Llama3 models, were fine-tuned and employed to create lay summaries from complex scientific texts. The summarization performance was enhanced through various approaches, including instruction tuning, few-shot learning, and prompt variations tailored to incorporate specific context information. The experiments demonstrated that fine-tuning generally led to the best performance across most evaluated metrics. Few-shot learning notably improved the models' ability to generate relevant and factually accurate texts, particularly when using a well-crafted prompt. Additionally, a Dynamic Expert Selection (DES) mechanism to optimize the selection of text outputs based on readability and factuality metrics was developed. Out of 54 participants, the WisPerMed team reached the 4th place, measured by readability, factuality, and relevance. Determined by the overall score, our approach improved upon the baseline by ≈ 5.5 percentage points and was only ≈ 1.5 percentage points behind the first place.

1 Introduction

In the biomedical domain, scientific publications and research play a central role in communicating research findings and results. However, these documents are usually written in complex language and use terminology and technical jargon that can be challenging for lay readers or researchers from different fields to understand (Goldsack et al., 2022). In this context, lay summarization can be utilized to extract the most relevant information from the original article or publication while also providing supplementary explanations. This often entails incorporating background information that may not be contained within the article itself.

In this context, this paper presents the participation of the team WisPerMed in the BioLaySumm2024 Shared Task (Goldsack et al., 2024) on automatic lay summarization and describes the employed approaches to tackle this challenge.

Summaries generated by LLMs, as demonstrated by Zhang et al. (2024), can be of equivalent or superior quality to original references. Additionally, instruction tuning is an effective approach for enhancing performance. However, LLMs face limitations when applied to domain-specific abstractive summarization. Key challenges include the quadratic complexity of transformer-based models (Vaswani et al., 2017) concerning input text length, model hallucination, where factually incorrect text is generated, and domain shift from training to test data (Afzal et al., 2023). Similarly, studies on text simplification (Amin et al., 2023) indicate that although general-purpose LLMs are capable of effectively simplifying clinical reports, they sometimes generate factual inaccuracies and omit crucial information.

To adapt LLMs to a specific domain or task (Ling et al., 2024), it is possible to fine-tune the models, leverage few-shot learning or further pre-train the models on domain data. Examples of domain-adapted LLMs for the biomedical domain include the BioMistral (Labrak et al., 2024) and OpenBioLLM (Pal and Sankarasubbu, 2024) model series. The BioMistral models are based on the Mistral 7B Instruct v0.1 (Jiang et al., 2023) model. They are further pre-trained on the PMC Open Access Subset[1]. OpenBioLLM models are based on the Llama3 (AI@Meta, 2024) models and were

---

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ Accessed: 2024-05-17

adapted to the biomedical domain through fine-tuning.

2 Dataset

The dataset (Goldsack et al., 2022) of the Shared Task (Goldsack et al., 2024) contains two collections of scientific journal articles and the corresponding lay summaries, namely PLOS and eLife. PLOS and eLife also include the section headings and keywords of the article. The PLOS dataset has 24,773 examples in the training split and 1,376 examples in the validation split, whereas the eLife dataset is smaller with 4,346 examples in the training split and 241 examples in the validation split. The test split consists of 142 examples for both datasets. Lay summaries of the PLOS dataset were written by the authors of the articles and are approximately 150-200 words long, while eLife lay summaries were written by expert editors in correspondence with the authors and are about twice as long.

For the remainder of this paper, any reference to the validation or test set will include eLife and PLOS unless otherwise specified.

3 Evaluation Metrics

The generated summaries were evaluated across ten metrics that fall into the following categories: relevance, readability, and factuality. Relevance was assessed through Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004) (ROUGE-1, ROUGE-2, ROUGE-L) and BERTScore (Zhang et al., 2020). ROUGE counts the overlapping n-grams in the generated texts and target lay summaries, whereas BERTScore uses contextual word embeddings to compare the semantic similarity of the two texts. Readability was evaluated using the Flesch-Kincaid Grade Level (FKGL) (Kincaid, 1975), Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and Learnable Evaluation Metric for Text Simplification (LENS) (Maddela et al., 2023). The FKGL uses sentence lengths and syllable count per word to estimate readability. The DCRS uses a word list to compute the occurrences of words unknown to most 4-th graders and the CLI estimates the grade level necessary to comprehend the text. The LENS metric is a learnable evaluation metric trained on datasets containing human ratings of simplifications. In this setting, LENS measures the simplification of the

abstract by the generated text using the target lay summary as a reference. Factuality was assessed with AlignScore (Zha et al., 2023) and Summary Consistency (SummaC) (Laban et al., 2022). The AlignScore quantifies the degree of alignment between the facts in the summary and the scientific article, while SummaC also includes consistency.

4 Methods and Experiments

This section outlines the methodology employed in the experiments conducted on the specified dataset.

4.1 Fine-tuned Models

In this study, instruction tuning (Wei et al., 2022) was utilized to fine-tune various models. Instruction tuning refers to the process of fine-tuning language models on a collection of datasets described via instructions. BioMistral-7B-DARE (BioM) and Llama3-70B-Instruct (Llama3) were fine-tuned for one epoch utilizing Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) on the eLife and PLOS dataset individually. BioM was trained on the abstracts + lay summaries, whereas Llama3 was trained on the entire articles + lay summaries. The texts were structured using the Mistral and Llama3 instruction templates prior to the fine-tuning process. Please refer to the Appendix A, B, and C for details on prompts, parameters and licenses, respectively.

After evaluating the checkpoints of BioM on the validation set, the checkpoints with the best scores were selected for inference. For Llama3, the final checkpoints were selected. The models were given the same prompt as during fine-tuning but without the target.

4.2 Prompt Variations

Prompts can guide the LLM's content generation process without the need for fine-tuning. In the zero- and few-shot settings, different prompt variations and their effect on the evaluation metrics were examined. In the few-shot setting, example lay summaries from the training and validation set were included in the prompt when performing inference on the validation and test set, respectively. The format of these few-shot prompts is designed to emulate a preceding conversation with the model, with the included examples serving as the model's previous responses.

To choose the best few-shot examples, all examples were ranked based on their average nor-
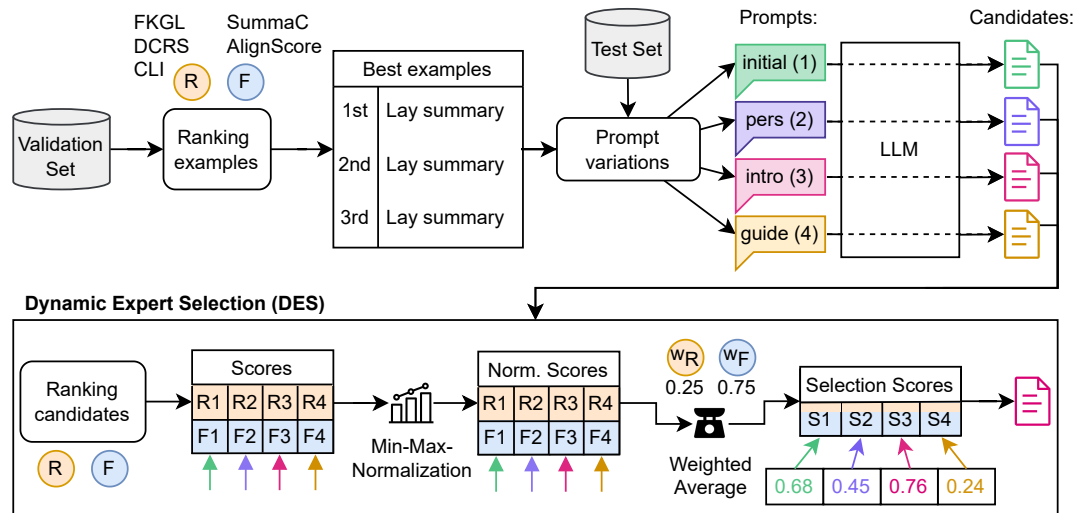
Figure 1: Workflow of the Dynamic Expert Selection (DES) mechanism in the few-shot setting using an example from the PLOS dataset. The process involves ranking examples, generating multiple summaries through various prompt variations, applying a large language model (LLM), and then normalizing and weighing the readability (R) and factuality (F) scores to rank and select the best summary based on the selection scores (S).

malized readability and factuality. The two and three highest-ranked examples were selected for the eLife and PLOS datasets, respectively.

An initial prompt was created by replicating the prompt used for inference with the fine-tuned BioM model (see Appendix A). This prompt was then tested with BioM and OpenBioLLM-70B (Open-Bio) on the validation set.

Additionally, three prompt variations were created, which provide the model with different kinds of context information. It was decided that BioM would be utilized for all experiments involving these variations due to its superior performance on the validation set in the few-shot setting (see Table 3 in Appendix D). LLMs can assume different roles and adapt their vocabulary accordingly (Salewski et al., 2023), resulting in enhanced performance in tasks related to the specified role. Accordingly, the first prompt variation comprises a persona description of a science communicator ($BioM_{pers}$), instructing the model to utilize the expertise of this persona to create the lay summary based on the abstract. The model is then instructed to channel the expertise of the described persona to craft the lay summary based on the abstract. The second prompt variation is a modification of the initial prompt, incorporating the introduction to provide additional background information because associated context can improve LLM performance (Karmaker and Feng, 2023). The second prompt variation is a modification of the initial prompt, but it includes the introduction as further context for background information ($BioM_{intro}$). The third prompt variation includes the abstract and a guide on how to write a lay summary ($BioM_{guide}$), accompanied by instructions concerning the content and style of the requested summary. This method leverages the importance of clear and detailed task directives. The selection of these prompts was based on a few preliminary experiments with the model and an initial assessment of the responses. However, no comprehensive optimization was performed. The wording of all prompts can be found in Appendix A.

Due to the efficacy of few-shot learning with the initial prompt, the prompt variations were implemented in a few-shot setting on the test set.

4.3 Dynamic Expert Selection (DES)

The success of an LLM depends on factors such as the properties of the dataset, the complexity of the domain, and the design of the prompt (Ling et al., 2024). Consequently, a model may yield a more suitable lay summary when prompted in a different manner. In addition, the output quality depends upon the selection of the inference parameters (Minaee et al., 2024). In consideration of this assumption, a Dynamic Expert Selection (DES) was developed. It selects the most appropriate text from a set of candidate texts based on metrics that do not require a reference lay summary.

The mechanism uses the readability metrics FKGL, DCRS, and CLI, as well as the factuality

metrics AlignScore and SummaC. These metrics are computed for each candidate text. Readability scores are multiplied by -1 so that higher scores indicate better readability. All scores are normalized using min-max normalization to range between 0 and 1, where 1 is the best and 0 is the worst. For each candidate text, an overall score is calculated by multiplying the means with different weights. Given that the target lay summaries in eLife have a higher readability than those in PLOS (Goldsack et al., 2022), the overall scores are computed with different weights for the two aspects. For eLife summaries: Readability is weighted at 0.675 and factuality at 0.325. For PLOS summaries: Readability is weighted at 0.25 and factuality at 0.75. The candidate text with the highest overall score is selected as the most suitable lay summary. The selection of the weights is based on the assumptions about the target texts and comparisons of the overall scores on the validation dataset.

This approach was applied to BioM in the few-shot setting using all prompt variants (see Figure 1) and to the fine-tuned BioM using two distinct inference parameter settings (see Appendix B).

5 Results

The results of the experiments using BioM, Llama3, and OpenBio are presented in table 1. The experiments are categorized into zero-shot learning, few-shot learning, and fine-tuning.

BioM exhibits the highest LENS score in the zero-shot setting. However, its relevance and factuality performance are the lowest. Few-shot learning resulted in enhanced performance across all metrics except for LENS. The persona prompt ($BioM_{pers}$) led to an improvement in relevance. Including the introduction in the prompt ($BioM_{intro}$) resulted in a reduction in all aspects despite the fact that the model had access to more information from the article itself. In comparison, the prompt with the guide ($BioM_{guide}$) exhibits minimal enhancements. The optimal few-shot learning for BioM occurred with the initial prompt, which achieved the highest readability and factuality in the few-shot setting, excluding the DES approach. However, OpenBio slightly underperformed with this prompt in the few-shot setting, except for the LENS score, where it performed best in this setting.

The DES used all four prompts and outperformed the baseline with improvements in factuality and readability, achieving the best results in the few-shot setting.

Fine-tuning BioM improved relevance and factuality scores, though the LENS score decreased slightly, with other readability metrics similar to the few-shot setting. The fine-tuned BioM outperformed the baseline in terms of relevance and overall quality. The DES approach improved all metrics except for a slight drop in the LENS score. In contrast, Llama3 underperformed despite being larger. It was less effective at extracting relevant information from full articles and produced lower-quality text in terms of readability, even though its LENS score was higher than BioM's. Additionally, Llama3's factuality scores decreased, leading to an overall performance drop compared to the baseline.

6 Conclusion

This paper presents the WisPerMed team's approaches to automatic lay summarization within the biomedical domain, utilizing a combination of fine-tuning, prompt variations, and Dynamic Expert Selection.

Among these approaches, fine-tuning emerged as an effective method, leading to the best performance across most metrics. This underscores the importance of task-specific training in optimizing model output for complex summarization tasks. Additionally, BioM showed strong few-shot learning capabilities, illustrating its robustness and versatility in generating accurate and relevant summaries even without extensive training. As the model adjusts to the factuality and readability of given examples, providing better examples could lead to further enhancements in these aspects.

BioM reached high factuality, even when provided solely with abstracts as input, suggesting that BioM leveraged domain-specific knowledge acquired during pre-training. This indicates that domain adaptation remains an important factor when using LLMs for lay summarization of scientific articles, as BioM outperformed the larger general model Llama3.

The four prompt variations exhibited differing effects on the evaluation metrics. BioM is adept in fulfilling the role of a science communicator ($BioM_{pers}$), as evidenced by the enhanced relevance. $BioM_{intro}$ and $BioM_{guide}$ did not significantly enhance the metrics, indicating that the increase in context was not beneficial for all texts. Without DES, a shorter prompt ($BioM_{initial}$) yielded the optimal results, suggesting that the model effectively

Expt.	R-1	R-2	R-L	BERT	FKGL	DCRS	CLI	LENS	Align	SC
Baseline	0.470	0.140	0.436	0.862	12.036	10.148	13.485	48.096	<u>0.779</u>	0.703
Zero-shot Learning										
BioM	0.329	0.071	0.298	0.845	12.404	10.093	13.974	<u>80.396</u>	0.541	0.458
Few-shot Learning										
BioM	0.440	0.124	0.409	**0.857**	11.287	**8.954**	<u>12.755</u>	75.744	0.728	0.604
BioM$_{pers}$	**0.442**	0.125	**0.412**	0.856	11.318	9.066	13.031	63.766	0.721	0.607
BioM$_{intro}$	0.391	0.106	0.359	0.851	12.233	9.618	13.693	76.638	0.669	0.529
BioM$_{guide}$	0.434	0.117	0.403	0.856	11.773	9.553	13.662	76.912	0.692	0.557
BioM$_{DES}$	0.439	**0.128**	0.409	0.855	<u>10.969</u>	8.993	12.819	74.025	**0.767**	**0.673**
OpenBio	0.415	0.104	0.382	0.855	11.657	9.848	13.711	**79.519**	0.731	0.558
Fine-tuning										
BioM	0.470	<u>0.152</u>	0.442	<u>0.865</u>	11.338	8.872	13.064	51.058	0.775	0.705
BioM$_{DES}$	<u>0.471</u>	<u>0.152</u>	<u>0.443</u>	<u>0.865</u>	**11.072**	**8.862**	**12.871**	51.028	**0.782**	<u>0.722</u>
Llama3	0.418	0.108	0.391	0.856	11.622	10.628	15.080	**72.860**	0.602	0.592

Table 1: Performance metrics of experiments on the test set. The models include BioMistral-7B (BioM), Llama3-70B (Llama3), and OpenBioLLM-70B (OpenBio). The experiments are categorized into fine-tuned, zero-shot, and few-shot settings. The metrics reported are ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), BERTScore (BERT), FKGL, DCRS, CLI, LENS, AlignScore (Align), and SummaC (SC). Bolded values indicate the best in each section, and underlined values the best overall performance.

comprehends the task from the provided examples. The DES mechanism further refined readability and, in particular, factuality by retrospectively selecting the best text outputs based on evaluation metrics. This highlights the potential of metric-driven selection to improve the quality of lay summaries further.

In conclusion, our study demonstrates that fine-tuning, the use of informed prompt variations, and selection mechanisms can enhance the capability of autoregressive LLMs to produce lay summaries that are factually accurate, relevant, and readily accessible to non-specialist audiences. This approach fosters broader public engagement with scientific findings, advancing the goal of making biomedical research comprehensible and accessible.

Limitations

Only four discrete prompts in combination were tested with DES, and only two sets of inference parameters were explored. This limited scope means that the findings may not fully capture the potential variability and performance of the the various models under different conditions. The weights for the Dynamic Expert Selection method were chosen based on heuristics without any formal op-timization, which could impact the robustness and generalizability of the results. Another limitation is the possibility that BioM may have been previously exposed to the gold standard summaries. If this is the case, it could skew the results by artificially inflating the model's performance. These limitations indicate potential avenues for future research, including the necessity for more comprehensive prompt engineering, optimization of DES weights, and a wider range of tasks to ensure the robustness of the approach. Another potential future direction is adapting these methods for other complex domains or languages and exploring additional metrics.

References

Anum Afzal, Juraj Vladika, Daniel Braun, and Florian Matthes. 2023. Challenges in Domain-Specific

Abstractive Summarization and How to Overcome Them. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 682–689, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

AI@Meta. 2024. Llama 3 Model Card.

Kanhai Amin, Pavan Khosla, Rushabh Doshi, Sophie Chheang, and Howard P. Forman. 2023. Artificial Intelligence to Improve Patient Understanding of Radiology Reports. *The Yale Journal of Biology and Medicine*, 96(3):407–417.

J. S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*, volume 1. Brookline Books.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, New Orleans, LA, USA.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825.

Shubhra Kanti Karmaker and Dongji Feng. 2023. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203, Singapore. Association for Computational Linguistics.

J.P. Kincaid. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis, Millington, Tennessee.

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 *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, pages 74–81, Barcelona, Spain.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *arXiv preprint*. ArXiv:2305.18703.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations*, New Orleans, LA, USA. OpenReview.net.

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.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *arXiv preprint*. ArXiv:2402.06196.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. In *Advances in Neural Information Processing Systems*, volume 36, pages 72044–72057. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language

Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Prompts

The prompts used in the experiments are shown in Figures 2, 3, 4, 5, and 6.

Fine-tuning/Initial Prompt with Abstract for BioM

You will be provided with the abstract of a scientific article. Your task is to write a lay summary that accurately conveys the key findings and significance of the research in non-technical language understandable to a general audience.:

Abstract of a scientific article:

[Abstract]

Lay summary for this article:

Figure 2: The prompt used for fine-tuning BioM and as the initial prompt in the zero- and few-shot settings. For fine-tuning the prompt also includes the target lay summary.

Fine-tuning/Inference Prompt with Article for LLama3

You will be provided with a scientific article. Your task is to write a lay summary that accurately conveys the key findings and significance of the research in non-technical language understandable to a general audience.:

Scientific article:

[Abstract]

Lay summary for this article:

Figure 3: The prompt used for fine-tuning Llama3. For fine-tuning the prompt also includes the target lay summary.

Persona Prompt

Meet Layla, your fantastic science communicator committed to breaking down complex research for everyone! Layla's mission is to create summaries that make scientific literature easy to understand for the general public. Before writing, Layla thoroughly reads the abstract to grasp the research goals and findings accurately. Precision is crucial for Layla; she makes sure her summaries align with the abstract's research while expanding on key points and methods. Layla ensures each summary gives a complete understanding of the findings and their importance. She offers detailed explanations and backgound information as context to aid comprehension. She highlights the main discoveries and their real-world implications, explaining study mechanisms and methods in reader-friendly language. Layla brings research to life with vivid descriptions and relatable examples, showing its impact on society. Her tone is informative yet engaging, avoiding jargon to be inclusive.

Now, let's channel Layla's expertise to craft a comprehensive lay summary for a scientific article.

Abstract of the scientific article:

[Abstract]

Layla:

Figure 4: The Persona-Prompt used in zero- and few-shot setting with BioM.

Intro Prompt

You will be provided with the abstract of a scientific article and the introduction section for background information. Your task is to write a lay summary that accurately conveys the key findings and significance of the research in non-technical language understandable to a general audience. Please ensure that your summary is mainly based on the information provided in the abstract. You may also use information from the introduction for additional context if necessary.

Introduction of the scientific article:

[Introduction]

Abstract of the scientific article:

[Abstract]

Lay summary for this article:

Figure 5: The Intro-Prompt used in zero- and few-shot setting with BioM.

Guide Prompt

You will be provided with the abstract of a scientific article. Your task is to write a lay summary that accurately conveys the key findings and significance of the research in non-technical language understandable to a general audience.

Abstract of the scientific article:

[Abstract]

Guidelines for crafting a lay summary:

- Craft a detailed summary that explains the research findings and their implications, providing thorough explanations where necessary.

- Ensure factual accuracy and alignment with the research presented in the abstract, elaborating on key points and methodologies.

- Highlight the main findings and their implications for real-world scenarios, delving into specific mechanisms or methodologies used in the study and their broader significance.

- Incorporate descriptive language to explain complex concepts.

- Maintain a balanced tone that is informative and engaging, avoiding technical jargon or overly formal language.

- Ensure the summary provides sufficient depth and context to guide the reader through the research journey and address potential questions or areas of confusion.

Your lay summary for the article:

Figure 6: The Guide-Prompt used in zero- and few-shot setting with BioM.

B Setup and Hyperparameter

Training All trainings were executed on a single Nvidia H100 80GB using the unsloth[2] framework and QLoRA (Dettmers et al., 2023). The following modules were targeted with QLoRA: "q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", and "down_proj". The QLoRA rank and alpha were both set to 16. The QLoRA dropout was set to 0. The optimization of the models was conducted using the 8-bit Adam optimizer (Loshchilov and Hutter, 2019), which was configured with a maximum learning rate of 2×10^{-4} and a weight decay factor of 0.01. The learning rate schedule included a linear decay following an initial phase consisting of five warm-up steps. Maximum sequence length was set to 4,096.

Inference For the inference process, a greedy search algorithm was employed as the decoding strategy (Minaee et al., 2024), with a configuration that allowed for the generation of up to 1024 new tokens per inference iteration.

DES The DES with the fine-tuned model used the inference parameter as described above for one candidate, and a repetition penalty of 1.1 was chosen to generate another candidate.

C Licenses

In Table 2 the Licenses as given by the owners of the Framework/Model are displayed.

Framework/Model	License
unsloth[3]	Apache License Version 2.0
BioMistral-7B-DARE[4]	Apache License Version 2.0
Llama-3-70B-I[5]	Llama 3 Community License Agreement
OpenBioLLM-70B[6]	Llama 3 Community License Agreement

Table 2: Licenses of the dataset, Framework and Models used for this Shared Task.

---

[2] https://github.com/unslothai/unsloth Accessed: 2024-05-17
[3] https://github.com/unslothai/unsloth Accessed: 2024-05-17
[4] https://huggingface.co/BioMistral/BioMistral-7B-DARE Accessed: 2024-05-17
[5] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct Accessed: 2024-05-17
[6] https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B Accessed: 2024-05-17

D Results on the Validation Set

The results of experiments on the validation set and the reference scores of the target lay summaries and input abstracts are presented in Table 3.

Expt.	R-1	R-2	R-L	BERT	FKGL	DCRS	CLI	LENS	Align	SC
Targets	-	-	-	-	12.857	9.944	14.251	57.988	0.670	0.512
Abstracts	0.410	0.135	0.380	0.855	15.260	11.378	16.961	38.259	-	-
Zero-shot Learning										
BioM	0.332	0.070	0.301	0.844	**12.530**	10.156	**13.957**	**80.159**	0.521	0.465
BioM$_{pers}$	0.411	0.118	0.379	0.847	12.579	**10.074**	14.897	69.732	0.741	0.628
BioM$_{intro}$	0.397	0.118	0.364	0.849	13.735	10.478	14.990	68.530	0.743	0.580
BioM$_{guide}$	**0.422**	**0.123**	**0.389**	**0.851**	13.971	10.478	15.667	68.561	**0.747**	**0.593**
Few-shot Learning										
BioM	**0.440**	**0.122**	**0.411**	**0.855**	**10.875**	**8.733**	**12.359**	76.358	**0.701**	**0.596**
OpenBio	0.423	0.107	0.390	0.854	12.429	9.729	14.721	**77.961**	0.678	0.554
Fine-tuning										
BioM	**0.478**	**0.148**	**0.446**	**0.866**	11.743	9.899	13.886	56.888	0.724	**0.677**

Table 3: Performance metrics of experiments on the validation set. The models include BioMistral-7B (BioM), Llama3-70B (Llama3), and Llama3-OpenBioLLM-70B (OpenBio). The experiments are categorized into fine-tuned, zero-shot, and few-shot settings. The metrics reported are ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), BERTScore (BERT), FKGL, DCRS, CLI, LENS, AlignScore (Align), and SummaC (SC). 'Targets' and 'Abstracts' provides benchmark scores of the target lay summaries and abstracts, respectively. Bolded values indicate the best in each section, and underlined values the best overall performance.