SINAI at BioLaySumm: Self-Play Fine-Tuning of Large Language Models for Biomedical Lay Summarisation

Mariia Chizhikova, Manuel C. Díaz Galiano
L. Alfonso Ureña López and **M. Teresa Martín Valdivia**
Department of Computer Science, University of Jaén
Campus Las Lagunillas, s/n, 23071, Jaén, Spain
mchizhik@ujaen.es

Abstract

An effective disclosure of scientific knowledge and advancements to the general public is often hindered by the complexity of the technical language used in research which often results very difficult, if not impossible, for non-experts to understand. In this paper we present the approach developed by the SINAI team as the result of our participation in BioLaySumm shared task hosted by the BioNLP workshop at ACL 2024. Our approach stems from the experimentation we performed in order to test the ability of state-of-the-art pre-trained large language models, namely GPT 3.5, GPT 4 and Llama-3, to tackle this task in a few-shot manner. In order to improve this baseline, we opted for fine-tuning Llama-3 by applying parameter-efficient methodologies. The best performing system which resulted from applying self-play fine tuning method which allows the model to improve while learning to distinguish between its own generations from the previous step from the gold standard summaries. This approach achieved 0.4205 ROUGE-1 score and 0.8583 BERTScore.

1 Introduction

Science outreach and scientific advocacy are crucial for the development of the science itself, as most funding still comes from public sources and thus demands public's support. Furthermore, science is central to most of the grand challenges of today's society, such as climate change, economic productivity, health and new drug discovery. These factors highlight the relevance of making information about scientific advancements accessible for general public. Moreover, it also may help the public make sound and informed choices about issues like participating in a clinical trial or getting a vaccination (Varner, 2014).

Nevertheless, even with an increased online availability of scientific publications, accessing the information from these sources remains a challenging task for non-experts due to the technical language and specific terminology used to write scientific work. One viable solution for addressing the informational requirements of the general public or gatekeepers like journalists are plain language summaries or lay summaries - a format that presents scientific research in an easily understandable manner for non-experts (King et al., 2017). However, manual generation of lay summaries is a tedious and costly process that involves contracting expert writers specialised in science outreach. For this reason, the development of effective automatic lay summarisation systems is attracting an increasing amount of attention of the Natural Language Processing (NLP) researchers (Ermakova et al., 2022).

BioLaySumm shared task held on the BioNLP 2024 workshop at ACL brings the community effort to tackle the task of automatic abstractive summarisation of biomedical articles for non-technical audiences by leveraging the extensive work done by the creators of eLife (King et al., 2017) and the Public Library of Science (PLOS) database in manually composing lay summaries.

This paper presents the methodology developed by the SINAI team as a part of our participation in the BioLaySumm shared task. Our experimentation involved comparison between few-shot learning (FSL) of instruction-tuned pre-trained large language models (LLMs), parameter-efficient tuning of open-source pre-trained LLMs and Self-Play fine tuning (SPIN) methodology which allows the model to improve while learning to distinguish between its own generations from the previous step form the gold standard summaries. The latter mentioned approach resulted to be the highest-scoring submission among all made by our team achieving 0.4205 ROUGE-1 score and 0.8583 BERTScore.

The remainder of the paper is structured as follows: Section 2 provides a concise description of the data utilized for this task. Section 3 details

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 804–809
August 16, 2024. ©2024 Association for Computational Linguistics

the systems developed by our team for the official evaluation. The details of the evaluation process and the results are presented in Section 4. Finally, Section 5 concludes our work.

2 Data

The organisers of the BioLaySumm shared task put at the disposal of the participants two datasets, PLOS and eLife, each of which consisted of biomedical research articles (including their technical abstracts) and their expert-written lay summaries (Goldsack et al., 2022). As for the dimensions of the data, while PLOS may be considered large-scale dataset with 24,773 instances for training and 1,376 for validation, eLife is a medium-scale dataset comprised of 4,246 training instances and 241 validation instances.

One important difference between the two datasets lies in the process of its generation. The Public Library of Science (PLOS) is a publisher that hosts peer-reviewed journals several of which require authors to submit an 150-200 word long *author summary* alongside their work. In contrast with this, eLife is an open-access journal that started creating plain-language summaries of its research articles in 2021 (King et al., 2017). As a result, lay summaries from two datasets differ from each other according to several characteristics, such as length and the extent to which they are abstractive. As can be seen on Table 1, which presents the statistics of token counts[1], eLife lay summaries and abstracts are almost twice as long. Furthermore, the authors claim that lay summaries of eLife appear to be significantly more abstractive based on the fact that these consistently contain more novel *n*-grams than abstracts across both datasets (King et al., 2017).

3 Methods

This section details the implementation of systems presented by our team for the official evaluation. We tested three approaches to lay summary generation: FSL of instruction-tuned models, parameter efficient fine-tuning of text generation pre-trained models and a novel method of fine-tuning of LLMs called SPIN.

3.1 Few-shot learning

FSL of pre-trained LLMs like GPT-3.5 proved to be a robust approach to the task of lay summarisation during the previous editions of the BioLaySumm shared task (Turbitt et al., 2023). For this reasons we decided to perform experiments with both closed, such as GPT-3.5 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023) and open models, namely Llama-3 (AI@Meta, 2024). The GPT models were accessed via OpenAI API[2], while the Llama-3-8B model was run on 1 NVIDIA Tesla V100-PCIE-32GB by making use of `transformers` Python library (Wolf et al., 2019).

Our team adopted 2-shot-prompting approach that introduced one randomly selected example from each of the datasets and injected a mention of the dataset to which each text abstract belonged. Thus, our method creates one system for the two datasets. The details regarding the prompt can be found in Appendix A.

While performing the experiments, we empirically found out that outputs from Llama-3 occasionally contained definition of what a lay summary is, repetition of prompt's content and `LAY SUMMARY` prefix. In order to remove this noise from the generation, we designed a post-processing procedure based on regular expressions.

3.2 Fine-tuning

While FSL allows adaptation of the model to a task without the need of further training of an LLM, fine-tuning involves training the model using additional, task-specific data.

We selected Llama-3-8B[3], the newest open LLM at the time of system's development process, for fine-tuning employing the QLoRA method (Dettmers et al., 2023), which minimizes memory usage and the number of trainable parameters by backpropagating gradients through a frozen, 4-bit quantized pre-trained LLM into Low Rank Adapters (LORA).

In order to obtain a single model capable of generating lay summaries for instances of both datasets, we trained Llama-3-8B on the data obtained by merging both training dataset into one.

As for the hyperparameters set for training, the LoRA alpha was set to 16, LoRA dropout was

[1] We used the tokenizer of BioMistral-7B model to split the texts into tokens

[2] https://openai.com/api/

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

		Lay summaries			Abstracts		
Dataset	**Subset**	**Avg(STD)**	**Min**	**Max**	**Avg(STD)**	**Min**	**Max**
eLife	Train	479.53 (84.69)	226	875	255.87 (45.67)	95	798
	Val	486.97 (93.62)	285	894	255.55 (43.24)	120	524
PLOS	Train	269.11 (58.15)	18	675	400.85 (111.11)	18	1198
	Val	269.53 (57.46)	72	530	404.1 (109.78)	112	877

Table 1: Token count statistics across two datasets.

equal to 0.1, LoRa rank - to 64 and the batch size was set to 1. The number of training epochs was initially set to 10, but an early stopping mechanism was implemented to prevent overfitting by stopping the training when validation loss does not decrease for 3 consecutive epochs. This allowed us to determine that one epoch was optimal for this kind of training.

3.3 Self-play fine-tuning

A significant advancement in LLMs performance is often achieved by applying post-pretraining alignment with mode desirable behaviour by using such techniques as Direct Preference Optimization (DPO) (Rafailov et al., 2024). Nevertheless, most alignment methods require a large volume of high-quality human-annotated data, which was not available for this challenge. For this reason, we opted for experimenting with SPIN (Chen et al., 2024), a novel fine-tuning method which begins from a supervised fine-tuning model, Llama-3-8B-instruct [4] denoted by $\mathbf{p}_{\theta_\mathbf{t}}$ which is employed to generate responses $\mathbf{y}'$ to the prompt $\mathbf{x}$ in the gold standard dataset, $\mathbf{y}$. The objective is to find a new LLM $\mathbf{p}_{\theta_\mathbf{t}+\mathbf{1}}$ capable of distinguishing $\mathbf{y}'$ from $\mathbf{y}$.

We employed this method to train a QLoRA adapter in order to be able to perform training on a single NVIDIA Ampere A100 50Gb GPU. We applied this method to fine-tune the model on each dataset separately, which resulted in two different adapters for eLife and PLOS datasets respectively.

4 Evaluation

This section presents the results of the official evaluation campaign that was carried out by the organizers by assessing the predictions made by our system on 142 articles for each of the two datasets.

4.1 Evaluation metrics

The generated summaries were evaluated across three aspects: Relevance, Readability and Factuality. To assess the relevance, n-gram based metrics (ROUGE 1, 2 and L) and semantic similarity metrics were calculated (BERTScore). In order to evaluate the readability Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), and LENS were used. Finally, to assess the factuality, the organisers calculated AlignScore and SummaC (Goldsack et al., 2024).

4.2 Results

Table 2 presents the details of the relevance metrics scored by each of the presented systems. Among the FSL experiments, Llama-8B-Instruct demonstrated the lowest performance among the models evaluated. Nonetheless given the unknown number of parameters of GPT-3.5-turbo and GPT-4-turbo, as well as the lack of information about whether these two closed-source systems add any post-processing to their outputs, it is difficult to draw conclusions about the performance of the models themselves. Nevertheless, we could empirically observe that the generations of both GPT-3.5 and GPT-4 are always complete and concluded texts, while Llama-3-8B-Instruct often outputted truncated or, on the contrary, noisy at the end of the sequence text. For this reason, as we noted previously, we introduced a rule-based post-processing procedure, which resulted in achieving the highest relevance scores for the eLife dataset.

As for GPT-3.5 and GPT-4, we were not able to find a substantial difference in overall performance between those two systems in the FSL setting. However, it is noticeable that GPT-3.5 showcased one of the best performances in terms of BERTScore for the eLife dataset and outperformed GPT-4 in all relevance metrics for PLOS dataset.

Comparing the results from the two fine-tuning methods employed, we can see that SPIN

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

fine-tuning of Llama3-8B-Instruct outperformed QLoRA in generating lay summaries for PLOS dataset. This can result from a larger amount of data available for this dataset, which makes SPIN to produce a more robust model. Nevertheless, for a much smaller dataset such as eLife, training a separate adapter with SPIN did not yield a performance improvement, while merging eLife with PLOS for training a universal QLoRa adapter for Llama-3-8B resulted to be a better solution in terms of relevance metrics.

As for readability and factuality, Table 3 presents the values these metrics. Overall, most of the systems produced less complex text that the reference lay summaries for PLOS datasets were reported to be (14.76, 10.91 and 15.90 for FKGL, DCRS and CLI, respectively) (Goldsack et al., 2022), while the reported lack of complexity for eLife's lay summaries (10.92, 8.83 and 12.51 for FKGL, DCRS and CLI, respectively) was more difficult to achieve even with our best system in terms of the readability metrics, namely FSL with Llama-3-8B-Instruct and rule-based post-processing.

Notably, the GPT-4 model, among all the presented systems, was the one that produced generally more complex text than others, while scoring one of the lowest values for factuality metrics. With regard to that, there can be perceived a trade-off between factuality and readability, with higher ranked models in terms of readability criteria achieving lesser factuality scores and vice-versa. The best performing model in terms of factuality resulted from fine-tuning of Llama-3-8B with QLoRA.

5 Conclusions

In this study, we explored various methodologies to generate lay summaries of biomedical articles, an important task for improving public accessibility to scientific information. Our participation in the BioLaySumm shared task at the BioNLP2024 workshop involved experimenting with FSL, parameter-efficient tuning and SPIN methods. Among these, SPIN fine-tuning demonstrated the highest performance in terms of relevance metrics, achieving a 0.4205 ROUGE-1 score and 0.8583 BERTScore.

The evaluation of readability and factuality revealed a trade-off between these two aspects. Models that generated more readable texts tended to have lower factuality scores, with the GPT-4 based FSL systems exemplifying this trend. Conversely, the fine-tuned Llama-3-8B with QLoRA achieved the best factuality scores while getting fairly good readability scores as well, indicating its potential for producing accurate and readable summaries.

6 Limitations

The limitations of the presented approaches stem from the inherent characteristics and potential biases of the pre-trained models they are based on. Specifically, models like Llama-3-8B-Instruct, GPT-3.5, and GPT-4 were pre-trained on extensive text datasets, which were not thoroughly evaluated for existing biases. Consequently, these models may generate inappropriate content or replicate biases present in the underlying data. Therefore, it is crucial to conduct comprehensive evaluations of safety and fairness concerns before deploying these systems in any practical applications.

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*.

AI@Meta. 2024. Llama 3 model card.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Elise Mathurin, and Patrice Bellot. 2022. Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 470–494. Springer.

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*

System	Dataset	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	BERTScore↑
Llama-3 + QLoRa on merged datasets	PLOS	0.4038	0.1419	0.3766	0.8495
	eLife	0.418	0.1007	0.3969	0.836
	Overall	0.4109	0.1213	**0.3867**	0.8428
Chat-GPT-3.5-turbo FSL	PLOS	0.4216	0.1076	0.3805	0.8603
	eLife	0.3719	0.0987	0.3418	0.8491
	Overall	0.3969	0.1032	0.3612	0.8547
Chat-GPT-4-turbo FSL	PLOS	0.4016	0.0834	0.3637	0.8548
	eLife	0.4139	0.0941	0.3771	0.849
	Overall	0.4077	0.0888	0.3704	0.8519
Llama-3-8B-Instruct FSL	PLOS	0.2958	0.067	0.2757	0.8045
	eLife	0.4118	0.0933	0.3892	0.8118
	Overall	0.3537	0.0802	0.3324	0.8082
Llama-3-8B-Instruct FSL + post-processing	PLOS	0.3904	0.0896	0.3609	0.8536
	eLife	**0.4262**	**0.1091**	**0.3997**	**0.8493**
	Overall	0.4083	0.0994	0.3803	0.8514
Llama-3-8B SPIN	PLOS	**0.4591**	**0.1485**	**0.418**	**0.8692**
	eLife	0.3819	0.1013	0.3527	0.8474
	Overall	**0.4205**	**0.1249**	0.3853	**0.8583**

Table 2: Detailed scores of the relevance metrics' values obtained by the systems presented by the SINAI team

System	Dataset	FKGL↓	DCRS↓	CLI↓	LENS↑	AlignScore↑	SummaC↑
Llama-3 + QLoRa on merged datasets	PLOS	12.5437	9.2242	14.5609	53.2788	**0.7663**	**0.7752**
	eLife	11.7704	8.5377	13.5612	58.2338	**0.747**	**0.7216**
	Overall	12.157	8.881	14.0611	55.7563	**0.7567**	**0.7484**
Chat-GPT-3.5-turbo FSL	PLOS	12.5183	10.015	13.9546	78.4641	0.7311	0.5396
	eLife	12.9662	9.8737	14.2676	77.5932	0.737	0.531
	Overall	12.7423	9.9441	14.111	78.0286	0.734	0.5353
Chat-GPT-4-turbo FSL	PLOS	14.4599	11.0353	15.8399	72.3903	0.6255	0.4635
	eLife	14.6803	10.923	15.893	71.65	0.6598	0.4692
	Overall	14.5701	10.9791	15.8664	72.0301	0.6427	0.4663
Llama-3-8B-Instruct FSL	PLOS	12.2824	**8.2744**	12.0751	46.8115	0.4857	0.4943
	eLife	12.3472	**8.3412**	12.8586	50.1217	0.5071	0.5057
	Overall	12.315	**8.3078**	12.4668	48.4662	0.4964	0.4999
Llama-3-8B-Instruct FSL + post-processing	PLOS	**10.8711**	8.6886	**12.0567**	**81.3139**	0.6274	0.5167
	eLife	**11.0275**	8.5286	**12.4404**	**81.1903**	0.6603	0.5341
	Overall	**10.9493**	8.6086	**12.2486**	**81.2521**	0.6439	0.5254
Llama3 SPIN	PLOS	12.8408	10.667	14.8027	73.3912	0.7521	0.5505
	eLife	11.6155	9.0522	12.8268	80.5002	0.6713	0.5291
	Overall	12.2281	9.8609	13.8148	76.9457	0.7117	0.5398

Table 3: Detailed scores of the readability and factuality metrics' values obtained by the systems presented by the SINAI team

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.

Stuart RF King, Emma Pewsey, and Sarah Shailes. 2017. Plain-language summaries of research: An inside guide to elife digests. *eLife*, 6:e25410.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

Johanna Varner. 2014. Scientific Outreach: Toward Effective Public Engagement with Biological Science. *BioScience*, 64(4):333–340.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Prompt engineering

This appendix section details the prompts used for each of the lay summary generation methods described in this paper.

A.1 Few-shot learning

For the FSL approach we randomly selected 2 examples from the training sets of each database. The example A.1 presents the prompt template used for FSL generation.

Example A.1. You are an expert in generating of lay summaries - more readable summaries of scientific papers that are accessible to the general public. You will be given abstracts of scientific paper either from PLOS of eLife databases and will return only lay summaries like in the following examples:

This document is from {source} database, create the lay summary from abstract.

ABSTRACT: {Example abstract 1}

LAY SUMMARY: {Example lay summary 1}

This document is from {source} database, create the lay summary from abstract.

ABSTRACT: {Example abstract 2}

LAY SUMMARY: {Example lay summary 2}

This document is from {source} database, create the lay summary from abstract.

ABSTRACT: {Test set abstract}

LAY SUMMARY: