Saama Technologies at BioLaySumm: Abstract based fine-tuned models with LoRA

Hwanmun Kim, Kamal Raj Kanakarajan, Malaikannan Sankarasubbu
Saama Technologies
{hwan.kim, kamal.raj, malaikannan.sankarasubbu}@saama.com

Abstract

Lay summarization of biomedical research articles is a challenging problem due to their use of technical terms and background knowledge requirements, despite the potential benefits of these research articles to the public. We worked on this problem as participating in BioLaySumm 2024. We experimented with various fine-tuning approaches to generate better lay summaries for biomedical research articles. After several experiments, we built a LoRA model with unsupervised fine-tuning based on the abstracts of the given articles, followed by a post-processing unit to take off repeated sentences. Our model was ranked 3rd overall in the BioLaySumm 2024 leaderboard. We analyzed the different approaches we experimented with and suggested several ideas to improve our model further.

1 Introduction

While many academic publications in the biomedical field can potentially benefit a wide readership including many non-experts, their accessibility is often limited by their use of technical terms and relatively sophisticated expressions. Therefore the summarization of biomedical research articles is an interesting and important task that can benefit the general public, and BioLaySumm 2024 (Goldsack et al., 2024) aims to solve this question by adopting techniques of NLP. BioLaySumm asks participants to suggest models that summarize the biomedical articles based on the PLOS and eLife datasets (Goldsack et al., 2022) composed of original research articles and lay summaries written by experts.

In this paper, we explain our approaches to the BioLaySumm 2024 in detail. To generate better lay summaries, we experimented with multiple fine-tuning approaches with LoRA based on the abstract part of the biomedical research papers. As a result of a series of experimentations, we concluded that our best-performing model is the unsupervised fine-tuned model with LoRA followed by a post-processing unit that chops off repeated sentences in the raw predictions. At the end of the competition, our model was ranked 3rd overall in BioLaySumm 2024 leaderboard.

2 Background

2.1 Task description

In BioLaySumm 2024, participants are expected to generate lay summaries for the research articles in the test set made from PLOS and eLife journals. For the development of summarization systems, PLOS (eLife) dataset provides 24773 (4346) articles for the train split and 1376 (241) articles for the validation split. For both PLOS and eLife datasets, the test split is composed of 142 articles. For each data point, the whole article including the abstract is provided along with the keywords and article id. For the train splits and the validation splits, ground-truth lay summaries targeted for non-experts are provided. These summaries are written by authors (PLOS) or expert editors (eLife). Participants can submit summaries generated from either individual models for each dataset or a unified model for both datasets. The qualities of submitted summaries are evaluated in three criteria: relevance, readability, and factuality. Each criterion is composed of multiple automatic metrics:

- **Relevance**: ROUGE (1,2, and L) (Lin, 2004), BERTScore (Zhang et al., 2020)
- **Readability**: Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948), Coleman-Liau Index (CLI) (Coleman and Liau, 1975), LENS (Maddela et al., 2023)
- **Factuality**: AlignScore (Zha et al., 2023), SummaC (Laban et al., 2022)

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 786–792
August 16, 2024. ©2024 Association for Computational Linguistics

These metrics are calculated through the BioLaySumm 2024 evaluation script[1]. For each metric, the average score over the entire prediction is reported. The goal of competition is to minimize FKGL, DCRS, and CLI and maximize all other metrics.

2.2 Related works

While automatic text summarization has long been the subject of interest for its wide applicability in various domains (El-Kassas et al., 2021; Allahyari et al., 2017), the advent of large language models (LLMs) has innovated the field drastically (Chang et al., 2024; Zhang et al., 2024; G et al., 2024).

As a subfield of text summarization, automatic lay summarization of biomedical literature obtained further attention for its close relationship with health literacy (Guo et al., 2021). Since most biomedical research articles assume readers are familiar with the scientific concepts and domain-specific languages of the field, it is important to measure and evaluate the readability of the generated summaries as well (Guo et al., 2021; Goldsack et al., 2022). On the other hand, fact-checking the lay summaries has been important as the use of LLMs becomes popular since LLMs are known to often experience hallucinations that generate misinformed texts (Zhang et al., 2023).

In this context, BioLaySumm provides a meaningful challenge where both the readability and factuality of summaries are evaluated (Goldsack et al., 2023, 2024). While various approaches were used for last year's competition (Goldsack et al., 2023), the most successful approaches include few-shot prompting on GPT models (Turbitt et al., 2023), fine-tuning on FLAN-T5 models (Sim et al., 2023), and factorized energy-based model trained on BioBart model (Phan et al., 2023).

3 System overview

To find the best-performing system for BioLaySumm 2024, we experimented with several different systems based on the abstracts of the research articles. In this section, we introduce the systems we experimented including the system we submitted to the leaderboard of BioLaySumm 2024. Throughout all experiments, we used eLife (PLOS) training data only for model training or prompting to generate summaries for eLife (PLOS) validation/test data.

3.1 Submitted system: Unsupervised fine-tuned LoRA model

The system we submitted for the competition is the unsupervised fine-tuned LoRA model. Due to the context-size limitation of most LLMs, it is nearly impossible to fit the entire articles into the inputs for the LLMs. Instead, inspired by the system (Turbitt et al., 2023) which took 1st place in the last year's competition (Goldsack et al., 2023), we only appended the abstract and the lay summary for the inputs to the model (Template 1). We used the entire input text for our training phase while we only used the input text just before the lay summary starts for the text generation. For parameter-efficient training, we adopted low-rank adaptation (LoRA) (Hu et al., 2021) for our training.

```
### Provide a lay summary of the following
    research abstract.

Abstract: In temperate climates , winter deaths
    exceed summer ones . However , there is
    limited information on the timing and the
    relative magnitudes of maximum and minimum
    mortality , (...)
Lay summary: In the USA , more deaths happen in
    the winter than the summer . But when deaths
     occur varies greatly by (...)
```

Template 1: Input text for unsupervised fine-tuning. The bold-faced text is the part used for the text generation as well.

While examining the generated summaries, we found that our fine-tuned model tends to repeat identical sentences rather than ending the summary. To regulate this, we post-processed our summary to chop off the redundant sentences. See appendix A For the details of the post-processing.

3.2 Other approaches

3.2.1 Baseline: zero-shot and few-shot prompting

While we use some form of fine-tuning in all the other approaches, we set a few-shot prompting system as our baseline following the best-performing system from the previous year's competition (Turbitt et al., 2023). While we adopted this abstract-based few-shot approach from the last year's competition, we randomly sampled 6 examples from the train set instead of hand-picked 3 examples used in the last year. We listed 6 abstract-summary pairs out of these sampled examples. See appendix B

[1] https://github.com/TGoldsack1/BioLaySumm2024-evaluation_scripts

for the sample prompt we used. Also, to provide a baseline that indicates the bare ability of the LLM we use, we tested zero-shot prompts where the same template was used as the few-shot prompts but with no examples listed.

3.2.2 Supervised fine-tuning with LoRA

Since the input text used in unsupervised fine-tuning in Section 3.1 trains not only the styles of lay summaries but also the styles of the original abstracts to the model, the quality of generated summaries may be affected by these abstracts in unwanted ways. To prevent this, we experimented with supervised fine-tuning. In particular, we treated the content of the lay summary as the label and the rest of the input text as the context by excluding input text tokens from the calculation of the loss function. To make this 'label' to be automatically detected after tokenization, we slightly changed the format of input text from Template 1 (see Appendix C).

3.2.3 Direct preference optimization on the fine-tuned model

Since our fine-tuning approaches only use abstract-summary pairs, it does not see the full contents of the research article during the training. Therefore the generated summaries may struggle with the factuality criterion. To mitigate this problem, we experimented with direct preference optimization (DPO) (Rafailov et al., 2024). DPO trains the human preference on a language model by providing pairs of similar samples where the relative preference within each pair is labeled (preferred sample vs rejected sample). To provide these relative preference labels, we generated summaries on randomly sampled 1000 articles in the train set using the unsupervised fine-tuned model and calculated factuality metrics (AlignScore, SummaC) on both the ground-truth lay summary and the generated summary. After comparing the average of the calculated factuality metrics within each pair, we label the summary with the higher score as the preferred sample and the summary with the lower score as the rejected sample. This DPO training is performed on top of the unsupervised fine-tuned model in Section 3.1.

4 Experimental setup

4.1 Hardware

All our experiments performed on a 4× Quadra RTX 8000 (48GB VRAM) card.

4.2 Text generation

We used `mistral-7B-instruct-v0.2` throughout all experiments. For both the few-shot approach and the fine-tuned approach, text generation is performed through vLLM[2] (Kwon et al., 2023) for faster experimentation. We set the temperature to 0 for all text generation.

4.3 Fine-tuning experiments with LoRA

For both unsupervised and supervised fine-tuning experiments, we utilized libraries from Huggingface (Transformers, PEFT[3], TRL[4]). We used AdamW optimizer (Loshchilov and Hutter, 2017) to optimize cross-entropy loss with label smoothing (Pereyra et al., 2017). Experimented hyperparameters are available in Appendix D.

4.4 Direct preference optimization experiments

For DPO experiments, we utilized Axolotl library[5]. We used the sequence size of 4096, the batch size 8, and the learning rate 1.0×10^{-5} with a linear scheduler over 3 epochs.

5 Results

5.1 Experiment results

We report the results of all our experiments in Table 1. Averages of result 7 and result 8 are the scores submitted to the leaderboard of BioLaySumm2024, and our model is ranked 2nd in relevance, 16th in readability, 18th in factuality, and 3rd in average scores of all categories out of 55 participants (Goldsack et al., 2024). Overall, our model delivered decent summaries in all 3 evaluation criteria while particularly successful in the relevance criterion.

5.2 Analysis on approaches

5.2.1 Baseline approaches: zero-shot and few-shot prompting

We set the zero-shot and few-shot prompting system as our baseline following the most successful approach last year (Turbitt et al., 2023). When comparing the baseline results from others in Table 1 (result 1, 3 vs. result 5, 9∼13 and result 2, 4 vs. result 6), fine-tuning approaches outperform zero-shot or few-shot prompting in relevance. For readability, fine-tuning is superior for the eLife

[2] https://github.com/vllm-project/vllm

[3] https://github.com/huggingface/peft

[4] https://github.com/huggingface/trl

[5] https://github.com/OpenAccess-AI-Collective/axolotl

#	Apporach	Dataset	Relevance				Readability				Factuality	
			R-1	R-2	R-L	BS	FKGL	DCRS	CLI	LENS	AS	SC
1	Baseline: Zero-shot	eLife, V	0.335	0.089	0.308	0.843	13.34	10.44	14.90	**74.90**	0.680	0.503
2	Baseline: Zero-shot	PLOS, V	0.442	0.128	0.400	0.861	13.50	**10.46**	14.90	**75.27**	0.680	0.527
3	Baseline: Few-shot	eLife, V	0.466	0.128	0.437	0.859	11.63	9.33	12.80	69.60	**0.711**	0.506
4	Baseline: Few-shot	PLOS, V	0.465	0.150	0.427	0.867	**12.86**	11.00	**13.97**	65.59	0.838	0.684
5	Unsup. FT	eLife, V	**0.497**	**0.150**	**0.477**	**0.865**	8.70	7.46	10.41	64.24	0.623	0.531
6	Unsup. FT	PLOS, V	**0.500**	**0.191**	**0.464**	**0.873**	14.16	10.67	15.52	45.25	**0.941**	**0.873**
7	Unsup. FT	eLife, T	0.477	0.133	0.456	0.863	8.52	7.36	10.42	62.31	0.601	**0.553**
8	Unsup. FT	PLOS, T	0.480	0.176	0.443	0.871	14.20	10.84	15.89	41.48	0.956	0.901
9	Sup. FT	eLife, V	0.488	0.143	0.467	0.863	10.86	7.90	10.13	63.58	0.607	0.510
10	Unsup. FT + DPO	eLife, V	0.487	0.144	0.467	0.863	8.43	7.34	10.40	63.40	0.630	0.537
11	Unsup. FT, no PP	eLife, V	0.493	0.149	0.473	**0.865**	8.72	7.40	10.40	63.97	0.621	0.531
12	Sup. FT, no PP	eLife, V	0.478	0.141	0.457	0.862	10.89	7.69	**10.10**	62.68	0.602	0.509
13	Unsup. FT + DPO, no PP	eLife, V	0.473	0.141	0.453	0.863	**8.41**	**7.10**	10.38	62.29	0.624	0.536

Table 1: All experiment results. The # column indicates the experiment result number. The approach column describes the components of the approach used for that experiment, such as zero-shot, few-shot, unsupervised fine-tuning (unsup. FT), supervised fine-tuning (sup. FT), direct preference optimization (DPO), or post-processing (PP). The dataset column indicates the dataset and the split (T for test, V for validation). For further clarification, we highlighted the results for the PLOS dataset with blue shades. Here we report all the 10 metrics used for BioLaySumm 2024: ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), BERTScore (BS), Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), LENS, AlignScore (AS), and SummaC (SC). Bold-faced numbers indicate the best scores we obtained on the validation split of each dataset.

dataset (except for the LENS score) while the opposite is true for the PLOS dataset. This might be related to the worse readability of PLOS summaries that the authors write themselves. On the other hand, fine-tuning approaches yield higher factuality scores for the PLOS dataset while giving worse AlignScore and better SummaC scores for the eLife dataset. These contrastive patterns in readability and factuality among different datasets might indicate that readability and factuality are in a trade-off relationship, as simplified summaries may deliver less accurate information.

5.2.2 Unsupervised vs. supervised fine-tuning

By comparing the unsupervised fine-tuning experiments (results 5, 11) with the supervised fine-tuning experiments (results 9, 12) in Table 1, we find that unsupervised fine-tuning outperforms supervised fine-tuning in all metrics except CLI. Despite our expectation of supervised fine-tuning performing better in the readability scores from not learning the patterns in the abstracts, the supervised fine-tuning was not superior in the readability neither. Detailed investigations on the reasons for this difference between the supervised and the unsupervised fine-tuning would be a good subject for the future research.

5.2.3 Direct preference optimization

When comparing the results of DPO experiments (results 10, 13) with the results of their fine-tuned model before DPO training (results 5, 11) in Table 1, we observe that DPO training gives better factuality scores as expected, as well as improved readability scores except for LENS. Yet, DPO training makes relevance scores worse at the same time, as its training process suggests some ground truth summaries as rejected samples.

5.2.4 Post-processing

To investigate the effect of the post-processing unit, we evaluated predictions with no post-processing (results 11, 12, 13 in Table 1). The comparison with the results of post-processed summaries (results 5, 9, 10) shows that post-processed summaries are superior to non-processed summaries in both relevance and factuality. Regarding the readability, the effect of the post-processing unit is mixed, where the post-processing improves LENS while it worsens DCRS and CLI. For FKGL, the effect is not consistent over different experiments.

6 Conclusion

As we participated in BioLaySumm 2024, we experimented with different fine-tuning approaches with LoRA to generate summaries based on the given abstract of a biomedical research article. In particular, we explored unsupervised fine-tuning, supervised fine-tuning, and direct preference optimization, and we concluded that our best-performing model is the unsupervised fine-tuned model with post-processing to chop off repeated sentences. Our model achieved 3rd place overall in the leaderboard of BioLaySumm 2024. While our model was successful, it would be interesting to extend our approach to a variety of larger LLMs or to adopt other schemes to utilize the full article of the research paper instead of the abstracts. Potential future researches on analysis on different fine-tuning methodologies and benchmarking on evaluation criteria beyond the current challenge may deepen the understanding of our approach.

7 Limitations

Due to our limited resources, we only experimented with a single type of relatively small open-sourced model. Due to the limited context size of the model we used, our exploration of methods to utilize full research articles was limited to DPO which interacts with the full articles only through the factuality scores.

It is also worthwhile to mention that our approach was more successful in the relevance than other than two other evaluation criteria. This might be related with the fact that summaries more readable than the suggested golden summary might score less in the BERTScore. It would be interesting subject for the future researches to see how our approach performs in other summary evaluation criteria beyond the current challenge.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *International Journal of Advanced Computer Science and Applications*, 8(10).

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

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

Edgar Dale and Jeanne Sternlicht Chall. 1948. *A formula for predicting readablility*. Bureau of Educational Research, Ohio State University.

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

Bharathi Mohan G, Prasanna Kumar R, Vifert Jenuben Daniel V, Archanaa. N, Mohammed Faheem, Suwin Kumar. J.D. T, and Kousihik. K. 2024. Comparative evaluation of large language models for abstractive summarization. In *2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 59–64.

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*, 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 *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.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

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*.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Phuc Phan, Tri Tran, and Hai-Long Trieu. 2023. VBD-NLP at BioLaySumm task 1: Explicit and implicit key information selection for lay summarization on biomedical long documents. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 574–578, Toronto, Canada. Association for Computational Linguistics.

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.

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, Toronto, Canada. 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, Toronto, Canada. Association for Computational Linguistics.

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. *Preprint*, arXiv:1904.09675.

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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

A Details of post-processing

In the raw predictions of fine-tuned models, we observed that identical sentences are repeated without completing the paragraph in a small fraction of the generated summaries. To mitigate this, we introduced the post-processing unit to chop off the repeated sentences from the prediction. To do this, we first split the prediction into a sequence of sentences. Then we examine these sentences from the beginning of the sequence and drop the rest of the sequence when the given sentence has appeared before during the examination.

We split the prediction into sentences based on the appearance of sentence-ending punctuation marks like period (".") or question mark ("?"). Yet, there are some exceptions we had to handle in this process:

- If punctuation is in the middle of parentheses, does not end the sentence there.
- If a period is part of a URL address, which is specified by the beginning sequences ("www" or "http") and the ending sequences ("com", "edu", "gov", or "org"), then do not end the sentence at that period.
- If a period is part of commonly used abbreviations in academic writing ("et al .", "vs .", and "e . g ."), do not end the sentence at that period.
- If the previous word of a period is a single letter English alphabet, do not end the sentence there, since it is likely a part of a phrase for a subsection or abbreviation of names (ex: "a.1", "c. elegans", "George R. R. Martin").
- If a period is surrounded by Arabic numerals, do not end the sentence since it is likely a part of a floating number.

B Few-shot prompt for the baseline system

```
### Provide a lay summary of the following
    research abstract.

Abstract: The role of the cellular
    microenvironment in enabling metazoan tissue
     genesis remains obscure . Ctenophora has
    recently emerged as (...)
Lay summary: The emergence of the diversity of
    multicellular animals involved cells joining
     together to form tissues and organs . The
    glue that (...)

Abstract: To evolve and to be maintained ,
    seasonal migration , despite its risks , has
     to yield fitness benefits compared with
    year-round residency . Empirical data
    supporting this (...)
Lay summary: Winter is one of the most
    challenging seasons for many animals . Cold
    temperatures , bad weather , short days ,
    long nights and a shortage of food can
    impose (...)

Abstract: The adaptive prokaryotic immune system
     CRISPR-Cas provides RNA-mediated protection
     from invading genetic elements . The
    fundamental basis of the system is (...)
Lay summary: In most animals , the adaptive
    immune system creates specialized cells that
     adapt to efficiently fight off any viruses
    or other pathogens that have invaded . (...)

Abstract: Adipose tissue is crucial for the
    maintenance of energy and metabolic
    homeostasis and its deregulation can lead to
     obesity and type II diabetes ( T2D ) .
    (...)
Lay summary: Obesity is a growing public health
    concern around the world , and can lead to
    the development of type 2 diabetes , heart
    disease and cancer . (...)

Abstract: The roles played by cortical
    inhibitory neurons in experience-dependent
    plasticity are not well understood . Here we
     evaluate (...)
Lay summary: What we see or fail to see through
    our eyes leaves a lasting impression by
    changing the strength of connections between
     (...)

Abstract: Numerous studies have established
    important roles for microRNAs ( miRNAs ) in
    regulating gene expression . Here , we
    report that miRNAs also serve as (...)
Lay summary: To produce a protein from a gene ,
    the sequence of the gene must be transcribed
     to produce a molecule of messenger RNA (
    mRNA ) . (...)

Abstract: Midbrain dopamine neurons have been
    proposed to signal reward prediction errors
    as defined in temporal difference ( TD )
    learning algorithms. (...)
Lay summary:
```

Template 2: Sample few-shot prompt used for our baseline system. The 6 examples listed here are the actual examples we used for the eLife articles.

C Input text for supervised fine-tuning

```
### Provide a lay summary of the following
    research abstract.

### Abstract: In temperate climates , winter
    deaths exceed summer ones . However , there
    is limited information on the timing and the
     relative magnitudes of maximum and minimum
    mortality , (...)

### Lay summary: In the USA , more deaths happen
     in the winter than the summer . But when
    deaths occur varies greatly by (...)
```

Template 3: Input text for supervised fine-tuning. The bold-faced text is the context and the rest of the text is the label.

D Fine-tuning hyperparameters

Hyperparameter	Values
Epochs	**3**, 5
Batch size	**8**
Sequence size	2048, **4096**
Learning rate (LR)	1.0E-5, **2.0E-5**
LR scheduler	**Linear**
LoRA r	**8**
LoRA α	**16**

Table 2: Hyperparameters we investigated in the fine-tuning experiments. Hyperparameters in bold are what we used for the submitted model.