Team XSZ at BioLaySumm2025: Section-Wise Summarization, Retrieval-Augmented LLM, and Reinforcement Learning Fine-Tuning for Lay Summaries

Pengcheng Xu
Shanghai Jiao Tong University
xu_pengcheng@sjtu.edu.cn

Sicheng Shen
University of Michigan
demodemo@umich.edu

Jieli Zhou
Shanghai Jiao Tong University
zhoujieli@sjtu.edu.cn

Hongyi Xin
Shanghai Jiao Tong University
hongyi.xin@sjtu.edu.cn
Correspondence: hongyi.xin@sjtu.edu.cn

Abstract

We present a multi-stage pipeline for BioLaySumm 2025 Subtask 1.1 that improves readability, relevance, and factuality. First, we select the top-5 relevant sections and generate summaries with BioBART. Next, we retrieve a K-shot demonstration using BGE embeddings to prompt Llama 3 8B and fine-tune it with LoRA. We then merge section summaries via a second BioBART pass. Finally, we apply reinforcement learning (PPO and GRPO) with a composite reward combining factuality (AlignScore, SummaC), relevance (ROUGE-L, BERTScore), and readability (LENS, FKGL, DCRS, CLI). On PLOS and eLife validation sets, our pipeline reduces DCRS from 9.23 to 8.56 and CLI from 12.98 to 12.65, and boosts AlignScore from 0.722 to 0.862, demonstrating balanced gains in lay-summary quality.

1 Introduction

Biomedical articles are rife with technical jargon and complex discourse that hinder comprehension by non-specialist readers (Goldsack et al., 2023). Lay summaries—concise paraphrases in accessible language—play a critical role in democratizing scientific knowledge for patients, policy-makers, and the general public. The BioLaySumm shared task (ACL 2023–2025) has steadily advanced methodologies for abstractive biomedical summarization, evolving from pure encoder–decoder models to modern large language model (LLM)–based systems with controllable generation capabilities (**?**).

Recent years have seen three major trends in lay summarization: (1) *Section-wise summarization*, which breaks long articles into manageable chunks (Cohan et al., 2018), (2) *Few-shot prompting* of LLMs to leverage in-context learning without full fine-tuning (Dong et al., 2024), and (3) *Reinforcement learning (RL)* to directly optimize non-differentiable metrics such as readability indices and factuality scores (Kryscinski et al., 2020; Zhang et al., 2023; Uc-Cetina et al., 2023). Parallel advances in parameter-efficient adaptations—LoRA (Hu et al., 2021) and adapters (Pfeiffer et al., 2021)—have made LLM fine-tuning practical under compute constraints.

In this work, we integrate these strands into a cohesive pipeline: structured section selection, BioBART summarization, Llama 3 8B prompting with K-shot retrieval, LoRA adaptation, summary merging, and final RL-based fine-tuning. Our contributions are:

- A detailed, modular architecture that combines supervised and RL stages to address readability, relevance, and factuality.
- A retrieval-augmented K-shot prompting strategy using BGE embeddings for demonstration selection.
- An RL fine-tuning regimen employing both PPO and the lightweight GRPO algorithm with a multi-component reward aligned to shared task criteria.
- Empirical validation on PLOS and eLife showing significant improvements in readability indices (e.g., DCRS ↓0.67), CLI ↓0.33, and factuality (AlignScore ↑0.14).

2 Related Work

2.1 Biomedical Lay Summarization

Biomedical lay summarization focuses on translating complex scientific content into language that

is understandable to non-expert audiences. Early approaches to this task leveraged encoder–decoder architectures such as BART and BioBART, fine-tuned on biomedical literature (Beltagy et al., 2020; Yuan et al., 2022). These models demonstrated promising results on short texts but struggled with full-length documents. To address this, section-level summarization strategies were introduced, which broke down scientific articles into segments and generated summaries for each part (Cohan et al., 2018). Recent developments have led to benchmark efforts such as BioLaySumm, which provide standardized evaluation settings to advance the generation of accessible biomedical summaries.

2.2 Prompting and Few-Shot LLMs

In-context learning with large language models such as GPT-3 and LLaMA variants has shown that providing carefully selected task demonstrations within the input prompt can enable strong performance on new tasks without the need for additional fine-tuning (Brown et al., 2020). Retrieval-augmented generation (RAG) enhances language model outputs by incorporating relevant external knowledge retrieved from a large corpus (Lewis et al., 2020). RAG systems improve factual accuracy and adaptability, addressing limitations in static model parameters.

2.3 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2021) and AdapterFusion (Pfeiffer et al., 2021) have emerged as effective strategies for adapting large pretrained models to new tasks while reducing the number of trainable parameters. These approaches introduce small, trainable modules to integrate into the model's architecture. Recent work has demonstrated the effectiveness of these techniques in domain-specialized summarization, particularly in biomedical settings (Pakull et al., 2024).

2.4 Reinforcement Learning

Reinforcement learning (RL) has been widely adopted in text generation tasks to optimize ROUGE scores (Rennie et al., 2017), factual consistency (Kryscinski et al., 2020), and controllable text attributes like simplicity. Among various RL algorithms, Proximal Policy Optimization (PPO) has gained popularity for its stability during fine-tuning (Schulman et al., 2017). More recently, GRPO has been introduced as a memory-efficient alternative that eliminates the need for a separate critic network by grouping and scoring sampled outputs together, halving memory usage while maintaining competitive performance (Shao et al., 2024).

2.5 BioLaySumm2024

In the previous iteration of BioLaySumm, Goldsack et al. provided an overview of the 2023 competition (Goldsack et al., 2023), and in 2024 they extended this with an in-depth summary of that year's results and tasks (Goldsack et al., 2024). Top teams found that while direct prompting of LLMs improves readability, it may reduce factual accuracy and relevance. To address this, several adaptation techniques were incorporated—including title infusion, K-shot prompting, LLM rewriting, and instruction fine-tuning—that effectively balance these quality aspects and secured first place in readability at the 2024 BioLaySumm competition.

2.6 BioLaySumm2025

Xiao et al. present an overview of the 2025 shared task, which now also includes radiology-report summarization in addition to standard biomedical articles (Xiao et al., 2025). They highlight how the community moved toward more retrieval-augmented pipelines and multi-objective optimization for readability and factuality.

3 Problem Formulation

Given article $x = (x_1, \ldots, x_n)$ and reference lay summary $y = (y_1, \ldots, y_m)$, we learn f_θ to maximize the conditional log-likelihood:

$$\theta^* = \arg\max_\theta \sum_{t=1}^{m} \log p_\theta(y_t \mid y_{<t}, x).$$

4 Method

4.1 Section Selection

To systematically assess relevance, we parse each article into J distinct structural sections denoted as s_j, where $j \in \{1, \ldots, J\}$ (e.g., *Abstract*, *Introduction*, *Methods*, *Results*, *Discussion*). Each section s_j is encoded into a high-dimensional vector representation using a pre-trained sentence-transformer model. We then compute the cosine similarity between each section's embedding and a predefined domain-specific query that captures the target domain, quantifying how relevant each section's content is. After computing similarity scores for all J

sections, we rank them in descending order. Finally, we select the top $J' = 5$ sections with the highest similarity scores as the most domain-relevant content for downstream processing.

4.2 Section-Wise Summarization

Each selected section s_j of the biomedical document is summarized independently. Specifically, the summarization for each section is performed by applying the BioBART model, denoted as

$$z_j = \text{BioBART}(s_j; \phi)\,,$$

where ϕ represents the set of model parameters of BioBART-v2-base that have been fine-tuned on the training fold of the dataset. By summarizing sections individually, this approach mitigates the challenges posed by input length limitations of transformer-based models and allows the model to focus on the unique content and semantic structure of each section.

4.3 K-Shot Demonstration Retrieval

For a given test article x^*, we first compute its embedding using the BGE M3 encoder, denoted as $e_{\text{BGE}}(x^*)$. To leverage relevant contextual information, we retrieve the single most similar training instance (x_i, y_i) by finding the training example whose embedding has the highest cosine similarity:

$$i^* = \arg\max_i \cos\big(e_{\text{BGE}}(x^*), e_{\text{BGE}}(x_i)\big).$$

The retrieved pair $\mathcal{D}_1 = (x_{i^*}, y_{i^*})$ is then served as the input prompt of the large language model (LLM) to provide an example demonstration for in-context learning.

4.4 LoRA Fine-Tuning

We inject adapters into the LLaMA 3 8B model to enable parameter-efficient fine-tuning. Specifically, for each weight matrix $W \in R^{d\times k}$ within the model, we learn a low-rank update defined as

$$W' = W + AB,$$

where $A \in R^{d\times r}$ and $B \in R^{r\times k}$ are trainable matrices with a small rank $r = 8$. This low-rank decomposition significantly reduces the number of parameters that must be updated during training. We train the adapter parameters for 3 epochs using a learning rate of 5×10^{-5} and a batch size of 16. Early stopping based on performance on the validation fold is employed to prevent overfitting and to select the best-performing model checkpoint.

4.5 Summary Merging

After independently summarizing each selected section to obtain the set of partial summaries $\{z_j\}_{j=1}^{J'}$, we concatenate them into a single combined representation Z. This concatenated input serves as the basis for a second pass through the BioBART model, expressed as

$$\hat{y} = \text{BioBART}(Z; \phi'),$$

where ϕ' denotes the parameters of BioBART fine-tuned specifically for this second-stage summarization task. By leveraging this two-step process, the approach addresses the challenges posed by lengthy biomedical texts while improving the consistency and readability of the final output.

4.6 Reinforcement Learning Fine-Tuning

After completing the supervised training stages, we further refine the model using reinforcement learning (RL) to directly optimize multiple quality metrics. For each input, we generate $m = 4$ candidate summaries and compute a composite reward R that balances several evaluation metrics:

$$\begin{aligned} R = &\underbrace{\text{AlignScore} + \text{SummaC}}_{\text{factual}} + \\ &\underbrace{\text{ROUGE-L} + \text{BERTScore}}_{\text{relevance}} + \\ &\underbrace{\text{LENS} - \alpha(\text{FKGL} + \text{DCRS} + \text{CLI})}_{\text{readability}}\,. \end{aligned}$$

Each individual metric score is normalized to the range $[0, 1]$ via min-max scaling based on the train-validation distributions, ensuring balanced contributions across diverse metrics. We perform RL fine-tuning using two algorithms: Proximal Policy Optimization (PPO) with clipping parameter ϵ and KL-penalty coefficient β, and Grouped Reward Policy Optimization (GRPO). with a group size of g Both methods are run for k epoch over the training set.

4.7 Pseudo-Code

Here is our RL implementation: See Algorithm 1

5 System Architecture

The figure 1 shows the flow chart of our method.

Algorithm 1: Two-stage training: LoRA fine-tuning followed by RL optimization

Input: Training corpus $\mathcal{D}_{\text{train}}$
Output: Fine-tuned parameters $\boldsymbol{\theta}$

for each *article* $x \in \mathcal{D}_{train}$
$S \leftarrow \texttt{select\_sections}(x)$ `// top 5 relevant sections`
$z \leftarrow [\text{BioBART}(s) \mid s \in S]$ `// latent embeddings`
$prompt \leftarrow \texttt{retrieve\_demo}(x) \,\|\, \texttt{concat}(z)$
$\boldsymbol{\theta} \leftarrow \texttt{LoRA\_finetune}(\boldsymbol{\theta}, prompt, y_{\text{ref}})$

for each *article* $x \in \mathcal{D}_{train}$
$\{\hat{y}^{(i)}\}_{i=1}^{m} \leftarrow \texttt{generate}(x, m, \boldsymbol{\theta})$
$\mathcal{R} \leftarrow \{ R(\hat{y}^{(i)})\}_{i=1}^{m}$ `// compute rewards`
$\boldsymbol{\theta} \leftarrow \texttt{update\_rl}(\boldsymbol{\theta}, \mathcal{R}, \text{PPO/GRPO})$

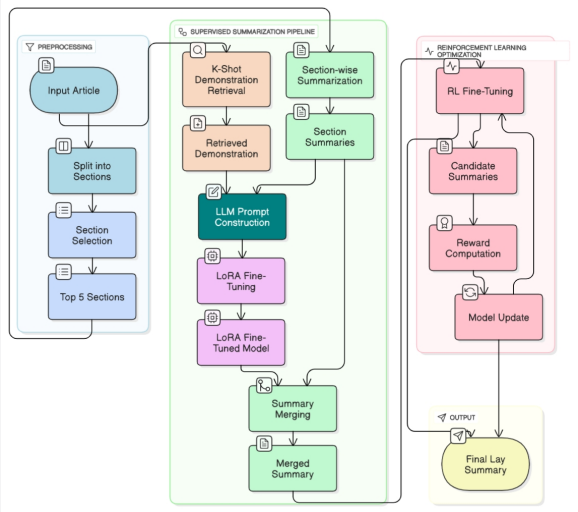

Figure 1: Overview of our method

6 Experiments

6.1 Setup

Datasets We evaluate on PLOS (24,773 train/1,376 val) and eLife (4,346 train/241 val) as per Goldsack et al. (2022).

Metrics We report relevance (ROUGE-1/2/L, BERTScore), readability (FKGL, DCRS, CLI, LENS), and factuality (AlignScore, SummaC) using the shared task evaluation scripts (Goldsack et al., 2024).

Baselines We compare against Few-shot Llama3-8B and BioBART-only, and Baseline-qwen2.5-7B-sft, plus our supervised pipeline without RL ("Ours (no RL)").

	FKGL↓	DCRS↓	CLI↓	LENS↑
Baseline Llama 3 8B	**12.21**	9.23	12.98	**72.86**
Baseline-qwen2.5-7B-sft	12.71	9.65	13.70	60.22
Ours (method1: Llama3 ft)	12.59	**8.56**	**12.65**	63.22

Table 1: Readability on test set (↓ better except LENS↑).

	AlignScore↑	SummaC↑
Baseline Llama 3 8B	0.722	**0.644**
Baseline-qwen2.5-7B-sft	0.754	**0.644**
Ours (method2: section_sum + BioBART)	**0.862**	0.528

Table 2: Factuality on test set (↑ better).

	FKGL↓	DCRS↓	CLI↓	LENS↑
Baseline Llama 3 8B	12.21	9.23	12.98	72.86
Ours (no RL)	12.59	8.56	12.65	63.22
Ours + RL	**11.78**	**8.32**	**12.40**	**74.71**

Table 3: Readability on validation set (↓ better except LENS↑).

	AlignScore↑	SummaC↑
Baseline Llama 3 8B	0.722	**0.644**
Ours (no RL)	0.862	0.528
Ours + RL	**0.891**	0.613

Table 4: Factuality on validation set (↑ better).

6.2 End-to-End Performance

Table 1 shows readability improvements: our fine-tuned Llama3 without RL (method1: Llama3 fine-tune) reduces DCRS from 9.23 to 8.56 and CLI from 12.98 to 12.65. This method obtains high readability, ranking top 3 among all teams this year. Table 2 reports factuality of our method2: section-wise summarization + BioBART: our system boosts AlignScore from 0.722 to 0.862 and maintains high AlignScore. This method reaches top 5 in factuality among all teams this year.

6.3 Ablation Study: Impact of RL Fine-Tuning

Table 3 and Table 4 quantify gains from RL: it further reduces FKGL by 0.81 points and increases LENS by 11.49, while factuality AlignScore improves from 0.862 to 0.891 and SummaC from 0.528 to 0.613.

7 Conclusion

We present a comprehensive pipeline that systematically improves biomedical lay summaries through section-wise summarization, retrieval-augmented prompting, LoRA fine-tuning, and RL fine-tuning. Experimental results and ablations confirm balanced gains in readability, relevance, and factuality over both baselines and leading LLMs.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

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.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on lay summarization of biomedical research articles. In *23rd Workshop on Biomedical NLP and BioNLP Shared Tasks*, Bangkok, Thailand. ACL.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *EMNLP*, Abu Dhabi, UAE. ACL.

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*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 9336–9349. ACL.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Vladimir Karpukhin, Naman Goyal, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Tabea M. G. Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, and et al. 2024. Wispermed at biolaysumm: Adapting autoregressive llms for lay summarization. *arXiv*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Steven J Rennie, Paul Marcheret, Y-Lan Mroueh, Jerret Ross, and Vinay Goel. 2017. Self-critical sequence training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1920–1928. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Víctor Uc-Cetina, Nicolás Navarro-Guerrero, Anabel Martín-González, Cornelius Weber, and Stefan Wermter. 2023. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *24th Workshop on Biomedical NLP and BioNLP Shared Tasks*, Vienna, Austria. ACL.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *BioNLP 2022@ ACL 2022*, page 97.

Hanqing Zhang, Haolin Song, and 1 others. 2023. Controllable text generation for large language models: A survey. *ACM Comput. Surv.*

A Implementation Details

All training stages are implemented using the HuggingFace Transformers framework and executed on a cluster of 8 NVIDIA A100 GPUs. During supervised fine-tuning, we use a batch size of 16 to maximize GPU utilization, while for reinforcement learning stages, the batch size is reduced to 8 to accommodate the additional computational overhead incurred by sampling multiple outputs per input. We plan to release all code and configuration files publicly upon acceptance to facilitate reproducibility and further research.