MIRAGES at BioLaySumm2025: The Impact of Search Terms and Data Curation for Biomedical Lay Summarization

Benjamin Pong **Ju-Hui Chen** **Jonathan Jiang**
Abimael Hernandez Jimenez **Melody Vahadi**
Department of Linguistics, University of Washington, Seattle, WA, USA
{benpong, juhuic, jjiang85, abimaelh, mvahadi}@uw.edu

Abstract

Biomedical articles are often inaccessible to non-experts due to their technical complexity. To improve readability and factuality of lay summaries, we built on an extract-then-summarize framework by experimenting with novel extractive summarization strategies and employing Low Rank Adaptation (LoRA) fine-tuning of Meta-Llama-3-8B-Instruct on data selected by these strategies. We also explored counterfactual data augmentation and post-processing definition insertion to further enhance factual grounding and accessibility. Our best performing system treats the article's title and keywords (i.e. search terms) as a single semantic centroid and ranks sentences by their semantic similarity to this centroid. This constrained selection of data serves as input for fine-tuning, achieving marked improvements in readability and factuality of downstream abstractive summaries while maintaining relevance. Our approach highlights the importance of quality data curation for biomedical lay summarization, resulting in 4th best overall performance and 2nd best Readability performance for the BioLaySumm 2025 Shared Task at BioNLP 2025.

1 Introduction

Biomedical research journals contain the latest findings on public health but highly technical language prevents the general public from understanding their content, which poses a challenge to health literacy (Guo et al., 2021). One solution is creating lay summaries – short, readable versions of scientific texts that use plain language and provide contextual information to bridge knowledge gaps.

This paper presents our submission to the BioLaySumm 2025 shared task 1.1 (Xiao et al., 2025), which focuses on generating lay summaries for biomedical articles. This task builds on previous editions of the shared task introduced in 2023 (Goldsack et al., 2023) and further developed in 2024 (Goldsack et al., 2024), which emphasize the challenges of readability, factuality, and accessibility in biomedical lay summarization. We built on the success of an extract-then-summarize pipeline (You et al., 2024) by developing novel sentence selection strategies that identify the most salient content from each article, prior to summarization, using titles and key words (i.e search terms). Unlike You et al. (2024) who explored the use of keywords for definition retrieval, and (Zhou et al., 2024) who explored title infusion for prompting, we explored the impact of these search terms at the level of extractive summarization. Our system [1] aims to balance relevance, readability, and factuality.

2 Dataset

The datasets used for this task are the PLOS and eLife datasets (Goldsack et al., 2022). The PLOS dataset comprises text from articles from life sciences. The eLife dataset contains articles on life sciences and medicine. The PLOS data set contains 24,773 training instances and 1,376 validation instances, while eLife contains 4,346 training and 241 validation instances.

3 Methods

Our system includes a preliminary retrieval-based extractive summarization process, and model finetuning and inference using Meta-Llama-3-8B-Instruct [2] (AI@Meta, 2024).

3.1 Preliminary Experiment: Preprocessing and Extractive Summarization

We first investigated which extractive summarization strategy would be most useful for finetuning and downstream abstractive summarization. We removed information in parentheses and citations. To

[1] https://github.com/Abimaelh/bio-laysum.git
[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

BioNLP 2025 Shared Tasks, pages 232–239
August 1, 2025 ©2025 Association for Computational Linguistics

extract salient content, we employed seven extractive strategies using SpaCy's sentence tokenizer (Honnibal et al., 2020), BioBERT's (Lee et al., 2019) embeddings, and cosine similarity for similarity scoring.

Strategy 1 (Control): Selects the first 4096 tokens for abstractive summarization.

Strategy 2: Converts the title to an embedding, ranks sentences by cosine similarity to the title embedding and selects the top 40.

Strategy 3: Enhances Strategy 2 by concatenating keywords into the title to form an embedding before computing similarity and selecting the top 40 sentences.

Strategy 4: Inspired by the utility of singular value decomposition (SVD), for topic modeling and text summarization Steinberger and Jezek (2004), we apply SVD to group sentences by topic and select the top 40 sentences from the topics ranked closest to the gold summary.

Strategy 5: Compute the article's mean embedding and extract the top 40 sentences that are most semantically similar.

Strategy 6: Prepends title and keywords to the article and segment the article into four core sections(abstract, introduction, results, and discussion)[3]. From this condensed content, we rank sentences according to their similarity to the mean embedding of the uncondensed article, and select the top 40 sentences.

Strategy 7: The reverse of 6, where we segment the article to the same four core sections, extract the top 40 sentences and prepend the title and keywords.

The outputs of the following seven extractive strategies were summarized by Meta-Llama-3-8B-Instruct (prompt in Section 3.2) and are evaluated on the eLife validation set using Strategy 1 as a control and comparing their relative performance. The articles were trimmed to 4096 tokens for inference, due to computational constraints. Appendix A shows the evaluation results and analysis. Strategy 2 and 3 showed reasonable potential to influence downstream abstractive summarization.

[3]We simply segmented the article into chunks according to the number of section headings, used these chunks as proxies for sections and removed the chunks corresponding to Materials and Methods since they are less relevant for summarization.

3.2 Baseline: Zero-shot prompt

As our baseline, we prompted Meta-Llama-3-8B-Instruct to generate abstractive lay summaries for articles on Strategy 1 using the following zero-shot prompt template:

```
System: You are a chatbot with expertise in summarizing documents
User: Provide a lay summary of the following text: {article}
```

3.3 Meta-Llama-3-8B-Instruct Finetuning

To evaluate how the best performing extractive strategies influence downstream summarization quality, we finetuned Meta-Llama-3-8B-Instruct on the unprocessed data (Strategy 1), and top-performing Strategies 2 and 3 using Low Rank Adaptation (LoRA) (Hu et al., 2022), and compared these finetuned instances against the baseline.

The data for finetuning was prepared by randomly selecting 650 training instances from both eLife and PLOS, totaling 1300 shuffled samples for finetuning. For evaluation, we used 150 randomly selected validation samples from both datasets, totaling 300 shuffled samples.

We present the set of hyperparameters considered in Appendix B, Table 4, and refer to them as sets 1 to 3 for the rest of this paper. Our experiments are incremental, starting from finetuning on 200 samples across Sets 1 and 2. Based on our results, finetuning on Strategy 1 using Set 1 did not improve over the baseline, but finetuning on Strategy 2 and 3 boosted Readability and Factuality scores. Finetuning on Set 2 did not show improvements.

Following this near-positive results, we performed a sample-size ablation study on 1000 samples and 1300 samples using set 1, to test if sample size further improves model performance. Since our results show that a sample size beyond 1000 does not induce improvements, we conducted further experiments on hyperparameter sets 3 on 1000 training samples.

3.4 Counterfactual Data Augmentation

Prior work (Rajagopal et al., 2022) claim that training on counterfactually augmented data can improve factual consistency of general-domain abstractive summaries by inducing entity-errors, and attempt to extend this hypothesis for lay summarization. To develop the counterfactual data, we used the same 1000 training samples that were

Table 1: Finetuning evaluation scores (Systems 1-16) on 300 randomly sampled data instances from both eLife and PLOS' validation set . Entries under model configuration for systems 1-15 are interpreted as: `llama_{hyperparameter set}_{strategy}_{sample size}`

System	Model Configuration	R-1	R-2	R-L	BERTScore	FKGL	DCRS	CLI	SummaC	AlignScore
Baseline	No-finetuning	0.4316	0.1130	0.4015	0.8500	12.7278	10.5846	13.0331	0.5654	0.6424
1	llama_1_1_200	0.3985	0.1148	0.3701	0.8482	13.3061	11.1538	13.2315	0.5078	0.6125
2	llama_1_2_200	0.4174	0.1069	0.3906	0.8477	12.6564	10.3625	12.4354	0.5692	0.6296
3	llama_1_3_200	0.4221	0.1095	0.3936	0.8488	12.9245	10.5067	12.4350	0.5555	0.6248
4	llama_2_1_200	0.4172	0.1172	0.3846	0.8482	14.0118	11.1758	13.7989	0.5078	0.6346
5	llama_2_2_200	0.4238	0.1116	0.3949	0.8492	13.1023	10.5741	13.0970	0.5509	0.6337
6	llama_2_3_200	0.4252	0.1117	0.3946	0.8496	12.9096	10.6275	13.0385	0.5572	0.6434
7	llama_1_1_1000	0.4130	0.1125	0.3868	0.8399	12.4096	10.6496	12.0954	0.6300	0.7122
8	llama_1_2_1000	0.4057	0.1025	0.3814	0.8448	11.3494	9.9221	11.6261	0.6300	0.6557
9	llama_1_3_1000	0.4045	0.1014	0.3799	0.8441	11.0981	9.7785	11.3806	0.6300	0.6846
10	llama_1_1_1300	0.4153	0.1115	0.3886	0.8464	12.7704	11.0471	12.8860	0.7100	0.7032
11	llama_1_2_1300	0.4156	0.1138	0.3893	0.8453	12.1517	10.5510	12.4498	0.6738	0.6255
12	llama_1_3_1300	0.4112	0.1072	0.3861	0.8424	11.5830	10.1520	11.8491	0.644	0.6094
13	llama_3_1_1000	0.4157	0.1125	0.3892	0.8399	12.3269	10.7378	12.4745	0.6385	0.7514
14	llama_3_2_1000	0.4158	0.1162	0.3880	0.8427	12.8280	10.9584	12.7057	0.6720	0.6223
15	llama_3_3_1000	0.4069	0.1025	0.3814	0.8445	11.2793	9.8721	11.5129	0.6133	0.6066
16	counterfactual	0.4001	0.0989	0.3770	0.8427	11.3365	10.0515	11.6893	0.6469	0.625

used to finetune System 9[4] but selected 250 samples to be modified by employing BERN2 (Sung et al., 2022), a multitask Named Entity Recognition (NER) model to extract biomedical entity mentions from their gold summaries. These entity mentions were masked out with their corresponding categories. We used Meta-Llama-3-8B-Instruct to substitute each category with a random entity belonging to that category, followed by a finetuning experiment on a data mixture of counterfactual data (See Appendix C for training templates and prompt template).

3.5 Postprocessing: Lay definition Insertion using LLM and UMLS

To enhance the readability and relevance of the generated summaries by our best performing model in Table 1 (i.e., System 9) we added a postprocessing strategy by using LLMs and UMLS as external knowledge bases of lay definitions. The goal is to simplify some biomedical terms from the summaries, and provide contextual knowledge through definitions [5].

We used SciSpacy's Biomedical NER model (Neumann et al., 2019) to extract biomedical entity mentions, and their definitions through its connection with the Unified Medical Language System (UMLS) database. For entity mentions that are absent, we employed Meta-Llama-3-8B-Instruct to provide definitions. With this hybrid approach to definition retrieval, we constructed a term-definition dictionary for each generated summary. For each generated summary, we randomly extracted 10 pairs of terms and definitions to be incorporated into a prompt for postprocessing. The prompt templates can be found in Appendix D.

[4] This turned out to be our best performing model. See Results.

[5] Note that this experiment was conducted on test set summaries produced by our highest-achieving model, and was evaluated on the system provided by the organizers.

4 Evaluation

All experiments except postprocessing were done using a subset of the metrics given by the organizers, on 300 randomly chosen validation samples as mentioned. For relevance, ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), and BERTscore (Zhang et al., 2020) were used. For readability, Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995), and Coleman-Liau index (CLI) (Coleman and Liau, 1975) were used. Lower FKGL, DCRS and CLI scores represent superior readability. Finally, for factuality, SummaC (Laban et al., 2022) and AlignScore (Zha et al., 2023) were used. Postprocessing experiment, as well as our best performing model, were (re)evaluated on the test set using the organizers' evaluation pipeline.

5 Results and Analysis

5.1 Experimental results

We report the results of our experiments in Table 1. The system that we submitted to the leaderboard for BioLaySumm2025 is system 9.

Table 2: Comparison of best system with and without post-processing. These systems were evaluated on the test set using the evaluation pipeline provided by the organizers of Biolaysumm. We submitted our best performing model, (i.e., system 9.)

System	ROUGE	BLEU	METEOR	BERTScore	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
Best (submission)	0.2877	4.6323	0.2305	0.8461	11.7109	8.4596	11.9899	71.2714	0.6811	0.6047
$\text{Best}_{+postprocessing}$	0.2498	3.1827	0.2021	0.8345	12.9000	8.3068	11.7878	61.9381	0.6026	0.5916

5.1.1 Impact of hyperparameters and sample-size ablation study

The results of systems 1 to 3 show that finetuning on filtered articles using hyperparameter set 1 generally outperform the zero-shot baseline in terms of readability, while finetuning on unfiltered ones did not. Hyperparameter set 2 did not improve model performance. Our ablation study on hyperparameter set 1 shows that increasing the sample size to 1000 for finetuning has a larger positive effect on both readability and factuality (Systems 7-9) compared to the baseline, but a sample size beyond that did not (Systems 10-12). However, finetuning does not seem to improve relevance scores across the board.

5.1.2 Impact of extractive summarization strategies

The effect of extractive summarization strategies is compounded on by the effect of sample size. System 2 outperforms system 3 in terms of readability and factuality, suggesting that keywords are inert. However, when the sample size increases to 1000, while they both outperform the baseline, system 9 outperformed system 8 in readability and factuality. This suggests that while the title is capable of extracting pivotal sentences in the article, the impact of keywords scales with data volume.

5.2 Impact of data augmentation

As expected from (Rajagopal et al., 2022), finetuning on counterfactually augmented data showed improvements in SummaC score, but a slight decrease in relevance and readability scores (Compare systems 9 and 16). This experiment verifies the reproducibility of (Rajagopal et al., 2022)'s work on using counterfactual data augmentation improves factuality for summarization with tradeoffs in relevance. In addition, our experiment sparks the promise of extending their methodology to the context of biomedical lay summarization. We leave this exploration to future work.

5.3 Impact of Post-processing using definition insertions

As presented in Table 2, our result for post-processing surprisingly showed marginal improvements in readability scores (DCRS and CLI), and a drop in other evaluation metrics. We speculate that while definition insertions helped with text simplification, the NER model is flawed in that it also extracts non-technical terms like "blood" and "human". Redundant definitions of these terms could have been incorporated into the summary, hence affecting factual consistency, and inducing verbosity.

5.4 Results of Final System Submission

Table 2 shows the results of our best performing model, which we submitted to the leaderboard. Our model was ranked 4th on the leaderboard, and achieved 2nd place in terms of Readability scores.

6 Discussion and Conclusion

Our study highlights the trade-offs in biomedical lay summarization between input selection, model fine-tuning, and postprocessing. Strategically curating input–particularly by leveraging document titles and keywords–can significantly improve the readability of generated summaries. Finetuning Meta-Llama-3-8B-Instruct on such targeted content surpasses using unfiltered inputs.

A comparison of extractive strategies reveals that title-based selection performs better with smaller training sets, while the inclusion of keywords becomes more effective as the models handle more data, suggesting that keywords provide additional semantic information that enhances generalization, particularly in data-rich settings across different topics.

Our ablation study shows that increasing finetuning sample size from 200 to 1000 improves performance across readability scores (FKGL, Dale-Chall), factuality (SummaC and AlignScore), but increasing sample sizes up to 1300 samples plateaus (System 10-12) or slightly reverses gains, possibly due to noise from lower quality training samples. These findings emphasize that high quality extractive pre-processing can have a more positive impact than increasing fine-tuning sample volume alone in domain-specific summarization tasks.

Regarding hyperparameters, Set 1 was consistently effective, especially when used with 1000 samples (e.g., Systems 7-9). Raising LoRA rank to 13 and increasing the effective batch size (Set 3) yielded only marginal improvements (e.g., System 13 vs. System 7), suggesting limited benefit from increasing model capacity under our current setup.

However, we do not see improvements in Relevance scores across the board, possibly due model capacity and hyperparameter issues. Another reason for this is, improved readability may have oversimplified the summaries, resulting in information loss.

Overall, our results demonstrate that thoughtful input design and targeted fine-tuning are critical for effective biomedical lay summarization. Our future work may explore adaptive extractive techniques and multiphase generation pipelines to further enhance summary clarity and trustworthiness.

7 Limitations

Our study has several limitations that inform opportunities for future work. First, we only evaluated decoder-only LLM-based architectures–specifically Meta-Llama-3-8B-Instruct–and did not explore neural encoder-decoder models, such as T5 or BART, which are commonly used for summarization tasks. This architectural constraint explains the limited improvement in BERTscore and Relevance Scores, which often favor outputs more closely aligned with gold summaries at the token or phrase level. Secondly, while our resource-constrained hyperparameter search identified workable configurations, future work should prioritize expanded hyperparameter optimization to fully exploit the model's capacity. Thirdly, our counterfactual data augmentation experiment, requires more complexity and development to investigate the tradeoffs between relevance and factuality. Aforementioned, in our postprocessing step, using NER to extract technical biomedical terms fails to sufficiently exclude non-technical medical terminologies, which may have contributed to redundant additions and edits to the summaries. Furthermore, randomly selecting 10 term-definitions does not circumvent this issue. Future work in this direction should consider more discriminate ways to filter out non-technical terms from biomedical texts, so that actual technical terms can be easily identified for simplification. Finally, while our system did reasonably well for readability, we did not explicitly investigate the effect of readability control (Luo et al., 2022) since the degree of simplicity is subjected to each individual's demands and technical expertise.

References

AI@Meta. 2024. Llama 3 model card.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

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

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 BioLay-Summ 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, 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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

J. Peter Kincaid, Robert P. Fishburne, 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. Technical report, Naval Air Station Memphis: Chief of Naval Technical Training. Research Branch.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Dheeraj Rajagopal, Siamak Shakeri, Cicero Nogueira dos Santos, Eduard Hovy, and Chung-Ching Chang. 2022. Counterfactual data augmentation improves factuality of abstractive summarization. *Preprint*, arXiv:2205.12416.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 7th International Conference ISIM*.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical namedentity recognition and normalization tool.

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 *The 24nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLay-Summ: An extract-then-summarize approach augmented with wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. 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. In *International Conference on Learning Representations*.

Jieli Zhou, Cheng Ye, Pengcheng Xu, and Hongyi Xin. 2024. Team YXZ at BioLaySumm: Adapting large language models for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 818–825, Bangkok, Thailand. Association for Computational Linguistics.

A Evaluation Results for preliminary experiments

Based on our evaluation of the preliminary experiment using the metrics described in Section 4, we observed that Strategy 2 did the best for readability scores, clearly surpassing the baseline, while Strategy 4 (SVD topic modeling) did the best for BERTscore. As for ROUGE-L, ROUGE-1, and SummaC, Strategy 2 and 3 did comparable to the control. The low scores for factuality are to be expected (Zhou et al., 2024) from just LLM prompting techniques without further finetuning. But the scores for Strategies 2 and 3 follow the control. Hence, we chose Strategies 1, 2 and 3 for finetuning. Full evaluation results can be found below in Table 3.

B Hyperparameters

Table 4 shows the hyperparameters that we used for our experiments.

Across all sets, we applied the AdamW optimizer, a LoRA dropout rate of 0.1, a LoRA alpha of 16 and a linear learning rate scheduler.

C Prompt Templates for Counterfactual Data Augumentation

We provide the following prompt templates for counterfactual data augmentation process.

The {text} refers to the gold summary and the * represents the entity mention that has been masked out and replaced with the entity category. The prompt below replaces * with a random entity mention of that category, and its output is an entity-error-induced gold summary:

```
System: You are a chatbot with knowledge in medical terms and their definitions in context.
User: The following text contains words enclosed in *́Ṫhese words are categories for biomedical entities. Replace the words with randomly chosen biomedical entities from your wealth of knowledge, and then enumerate a list of the replacements. {text}'
```

The output of the above is used for finetuning, where the model is trained to recognize factual deviance:

Table 3: Preprocessing Methods Performance Metrics Comparison

Preprocessing	R-1	R-2	R-L	BERTScore	FKGL	DCRS	CLI	SummaC	AlignScore
1 (baseline)	0.4031	0.1017	0.3779	0.8412	13.0715	10.9113	13.6171	0.5893	0.4521
2	0.4048	0.0967	0.3788	0.8417	12.8065	10.4800	13.2654	0.3885	0.3988
3	0.4070	0.0978	0.3802	0.8420	13.0214	10.5933	13.5069	0.3767	0.4147
4	0.4089	0.1020	0.3830	0.8423	12.9394	10.7197	13.6397	0.3464	0.3727
5	0.3929	0.0945	0.3675	0.8387	12.8777	10.7964	13.2739	0.3912	0.3625
6	0.3799	0.0894	0.3541	0.8310	13.8133	10.9826	13.1877	0.3523	0.3588
7	0.3851	0.0926	0.3580	0.8366	14.4809	11.0687	13.8176	0.3665	0.3444

Table 4: Hyperparameters

Hyperparameters	Set 1	Set 2	Set 3
Learning Rate	2×10^{-5}	8×10^{-6}	2×10^{-5}
Batch Size	4	8	4
Epochs	3	5	3
Grad. Accumulation	2	1	2
r (Rank)	10	10	13
LoRA dropout	0.1	0.1	0.1
lr scheduler	linear	linear	linear
Optimizer	AdamW	AdamW	AdamW

```
<|begin_of_text|><|start_header_id|>
system<|end_header_id|> You are
a chatbot with expertise in
summarizing documents. <|eot_id|>
<|start_header_id|>user
<|end_header_id|>
Provide a wrong lay
summary of this article:
{preprocessed article} <|eot_id|>
<|start_header_id|>assistant
<|end_header_id|>
Wrong lay Summary:
{entity-error-induced gold summary}
<|eot_id|>
```

Note that the template above is only for the 250 samples that were selected for counterfactual augmentation. For the rest of the 750 samples, we had the <assistant> prompt to indicate "lay summary". A mixture of original data and counterfactual data is used as training data for this finetuning experiment.

D Prompt Templates for Postprocessing Step

The prompt template used to extract definitions of entity mentions from Meta-Llama-3-8B-Instruct is as follows: `System: "You are an expert who can provide informative and lay definitions to biomedical terms."` `User: Provide only the definition of the biomedical term: term'`

As mentioned in the main text, term-definition dictionaries were constructed and incorporated into a prompt to generate a postprocessed summary. The prompt template used is:

```
System: "You are an expert biomedical editor skilled at simplifying complex medical terms for a lay audience. Use the provided dictionary to replace technical terms with their lay definitions while preserving the original meaning."
User: **Biomedical Lay Definitions Dictionary:** {term_dictionary}
*Task:** - Read the following summary: {summary}
- Replace all technical terms in the summary with their lay definitions from the dictionary.
- Do not add or remove key information.
- If a term isn't in the dictionary, retain the original term.
*Return only the paraphrased summary in one line, without any commentary**
```