TLPIQ at BioLaySumm: Hide and Seq, a FLAN-T5 Model for Biomedical Summarization

Melody Bechler[1], Carly Crowther[1], Emily Luedke[1], Natasha Schimka[1], Ibrahim Sharaf[1]
[1]Department of Linguistics, University of Washington, Seattle, WA, USA
{mbechler, carlyc88, eluedke, nschim2, ibshar}@uw.edu

Abstract

BioLaySumm 2025 is a shared task that aims to automatically generate lay summaries of scientific papers for a wider audience of readers without domain-specific knowledge, making scientific discoveries in the domain of biology and medicine more accessible to the general public. Our submission to the task is a FLAN-T5 base model fine-tuned on the abstract and conclusion of articles and expert-written lay summaries from the shared task's provided datasets. We find that our system performs competitively in terms of relevance, exceeds the baseline on factuality, but falls short on readability.[1]

1 Introduction

Lay summarization is the task of summarizing domain specific texts into simplified summaries non-experts can understand. In these types of summaries, complex jargon is eliminated and information is summarized in a clear and concise manner for easy readability. Biomedical literature is an example of highly technical, jargon-rich texts that are difficult to understand by those outside of the field, but are invaluable resources for interested researchers, professionals, and the general public. Unfortunately, this wealth of knowledge has limited accessibility and comprehension due to length and complexity. Lay summaries can improve science literacy, help limit the spread of misinformation, and invite interdisciplinary work (King et al., 2017).

To address these persistent issues, the Biomedical Lay Summarization task (BioLaySumm) 2025 (Xiao et al., 2025) shared task at the BioNLP Workshop at ACL 2025[2] focuses on various biomedical lay summarization tasks, from plain lay summarization to multimodal lay summarization. The Lay People in Question (TLPIQ) team focuses on plain lay summarization as a baseline model to summarize biomedical texts. This model aims to improve accessibility and understanding of these complex texts, while maintaining factuality and domain relevance.

2 Related Work

Previous work has evaluated two types of summarization: extractive and abstractive. Extractive summarization aims to select verbatim components of a document to create a summary, whereas abstractive summarization generates novel summaries. Overviews of the past two years of the task can be found in Goldsack et al. (2023) and Goldsack et al. (2024). Particularly, using an extract-then-summarize approach with TextRank (Mihalcea and Tarau, 2004) and BERT (Devlin et al., 2019), You et al. (2024) extracted text to reduce input length to separately fine-tune GPT-3.5 and a Longformer Encoder Decoder model to achieve the best performance in the task last year.

Preprocessing techniques showcased positive summarization results. Zhao et al. (2024) indicated that hard truncation and text-chunking resulted in better quality and efficiency compared to data augmentation and prompt engineering techniques. Modi and Karthikeyan (2024) utilized a preprocessing over the abstract technique to extract initial sentences from a document and remove punctuation and enclosed text to successfully increase summary readability.

Previous work has utilized smaller parameter sequence-to-sequence models with varying results. Malik et al. (2024) utilized a FLAN-T5 model with a basic prompt structure, but the lack of constraints, limited training, and context length of the model resulted in poor lay summarization output. Modi and Karthikeyan (2024) also fine-tuned a FLAN-T5-base, but focused on preprocessing over the

[1] Our code is made available in a public repository: https://github.com/nschimka/TLPIQ—BioLaySumm-2025

[2] https://aclweb.org/aclwiki/BioNLP_Workshop

BioNLP 2025 Shared Tasks, pages 196–201
August 1, 2025 ©2025 Association for Computational Linguistics

abstract and a cosine scheduler to generate lay summaries.

In this task, we train a sequence-to-sequence FLAN-T5-base [3] model with abstract extraction, instruction tuning with dataset tags, and a specialized prompt template to improve upon previous T5 lay summarization methods. Sequence-to-sequence models handle input and output sequences better than other larger models while being computationally efficient, making the T5 model a strong choice for this summarization task.

3 Data

The dataset for the shared task is from Goldsack et al. (2022) which includes articles from two different biomedical resources. The Public Library of Science[4] (PLOS) is an open-access non-profit publisher of articles from various peer-reviewed journals in a wide variety of scientific fields. eLife[5] is an open-access peer-reviewed journal of biomedical and life sciences. Of the two, PLOS is longer, with 24,773 instances for training and 1,376 for validation. eLife contains 4,346 instances for training and 241 for validation.

We performed exploratory data analysis (EDA) on the two data sets to better understand the quantitative and qualitative features of both the articles and the summaries. See Appendix A for the results of the EDA.

4 Methods

4.1 Preprocessing

Because the FLAN-T5-base model has a maximum input length of 1,024 tokens, the original articles needed to be shortened significantly from the average token lengths of 6,981 tokens for PLOS and 10,428 tokens for eLife (see Figure 1 in Appendix A.1).

We segmented each article into sections using newline characters, appended the dataset-provided keywords to each input to enrich contextual information, and removed in-text citations with a regular expression.

We implemented a TF-IDF scoring function with scikit-learn's `TfidfVectorizer` class (Pedregosa et al., 2012) to find the most important sentences in each section, rank the sentences by importance, then take the first *n* allotted tokens starting at the top of the list of sentences. We found the most success allotting 50% of the input tokens to the abstract and 50% to the conclusion/discussion section (depending on each article's naming conventions).

4.2 Model

We fine-tuned FLAN-T5-base (248M parameters), an instruction-tuned variant of the T5 architecture (Raffel et al., 2020; Chung et al., 2022) to balance the compute cost and performance in our combined 30K sample biomedical corpus. Its Transformer backbone with multi-headed attention (Vaswani et al., 2017) captures long-range dependencies in scientific text, enabling accurate and accessible lay summaries.

At inference time, we steer our fine-tuned FLAN-T5-base model with a diverse controlled beam search setup to balance faithfulness, readability, and coverage. We generate up to 400 new tokens (minimum 120) beyond the input prompt to ensure complete summaries without truncation, using 8 beams divided into 4 diversity groups (diversity penalty = 0.8) to explore varied phrasings. To avoid repetition of three-gram patterns, we enforce `no_repeat_ngram_size=3` to avoid repeating n-grams and apply a mild repetition penalty of 1.2. A length penalty of 0.9 encourages more comprehensive output.

Details of our model approach can be found in Appendix B.

5 Evaluation

Evaluation for this task cover three areas: relevance of the summary to the original article, readability, and factuality.

Relevance is measured with ROUGE (1, 2, and L), BLEU, METEOR, and BERTScore; readability is measured with Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), and LENS; factuality is measured with AlignScore and SummaC.

BioLaySumm 2025 utilizes the Codabench (Xu et al., 2022) platform for participants to submit their predicted summary results. Our final model output consisted of the predicted summaries, quality scores, token counts, and input text identifiers. We created a script to retain only the predicted summaries in the appropriate submission format to then evaluate model performance.

[3] https://huggingface.co/docs/transformers/model_doc/flan-t5

[4] https://huggingface.co/datasets/BioLaySumm/BioLaySumm2025-PLOS

[5] https://huggingface.co/datasets/BioLaySumm/BioLaySumm2025-eLife

Source	ROUGE	BLEU	METEOR	BERTScore
FLAN-T5 base	0.34	7.16	0.27	0.86
llama3-8B-sft	**0.37**	**9.86**	**0.31**	0.86
qwen2.5-7B-sft	0.35	8.74	0.3	**0.87**

Table 1: Comparison of our relevance scores across evaluation metrics compared to the baselines. Best score for each metric is in bold.

Source	FKGL	DCRS	CLI	LENS
FLAN-T5 base	13.44	10.59	13.43	43.68
llama3-8B-sft	**12.21**	**9.23**	**12.98**	72.86
qwen2.5-7B-sft	12.71	9.65	13.7	**60.22**

Table 2: Readability scores across metrics

6 Results

Tables 1, 2, and 3 present our FLAN-T5 Base model's performance alongside the shared-task baselines—Llama3 (8B params) (Grattafiori et al., 2024) and Qwen2.5 (7B params) (Qwen Team, 2024)—as reported on Codabench[6].

Our combined-dataset FLAN-T5 system (248M params) achieves a ROUGE of 0.34 and a BERTScore of 0.86, compared to Llama3's ROUGE of 0.37 (BERTScore = 0.86) and Qwen2.5's ROUGE of 0.35 (BERTScore = 0.87) (Table 1). In factuality metrics (Table 3), we match or exceed these baselines, with an AlignScore of 0.76 (vs. 0.72/0.75) and SummaC of 0.64.

However, our readability scores (Table 2) reveal a larger gap: our FKGL of 13.44 and LENS of 43.68 lag behind Llama3 (12.21/72.86) and Qwen2.5 (12.71/60.22).

These results demonstrate that a lightweight 248M-parameter FLAN-T5 model can achieve relevance and factuality on par with much larger 7–8 B–parameter systems, but still requires further refinement to match their readability.

Source	AlignScore	SummaC
FLAN-T5 base	**0.76**	**0.64**
llama3-8B-sft	0.72	0.64
qwen2.5-7B-sft	0.75	0.64

Table 3: Factuality scores across metrics

7 Discussion

In our error analysis, we identified two main shortcomings of the combined-dataset model. First, eLife summaries were sometimes truncated mid-sentence; key findings would abruptly end because the model had internalized a compression ratio dominated by the shorter PLOS summaries (see Figure 3 in Appendix A.1). Second, despite our diverse beam search and generation strategies, occasional technical terms still slipped through, subtly raising both Flesch–Kincaid and LENS scores and detracting from true lay readability.

Looking ahead, we see three promising directions. First, training separate, dataset-specific models would let each learn its own optimal compression ratio and vocabulary constraints, eliminating length-bias effects. Second, a two-stage pipeline, initially generating a faithful summary and then passing it through a lightweight simplification model or rule-based rewriter, could ensure factual accuracy while improving clarity. Finally, integrating a post hoc lexical simplification step, via curated synonym lists or a small neural simplifier, would remove residual jargon and bring reading levels down to our grade 8-9 target. Together, these refinements promise to restore full-sentence integrity and markedly boost readability without sacrificing domain fidelity.

Limitations

While our system achieves strong relevance and factuality scores, it exhibits several limitations that affect its overall performance—particularly in terms of readability. First, the use of a single model trained on both PLOS and eLife datasets introduced a compression mismatch: summaries generated from longer eLife articles were occasionally truncated mid-sentence, likely due to the model internalizing an average summary length skewed by the shorter PLOS samples. This resulted in incomplete outputs and diminished coherence for eLife inputs.

Second, despite instruction-tuning and con-

[6] Accessed May 22, 2025

trolled decoding strategies, technical vocabulary and complex syntax persisted in some outputs. This limited the model's ability to consistently produce content aligned with the target 8th–9th grade reading level, as evidenced by elevated FKGL and LENS scores.

Moreover, due to time constraints, we did not explore more advanced strategies such as multi-stage summarization, dataset-specific modeling, or post-hoc simplification pipelines. These approaches may have mitigated the readability issues while preserving factual accuracy.

Finally, all evaluations rely on automatic metrics. While useful for benchmarking, they may not fully capture nuance in accessibility, clarity, or human comprehension—factors that are especially critical in the biomedical lay summarization context.

Acknowledgments

This work was produced as part of a class project for the Professional MS in Computational Linguistics during Spring 2025 at the University of Washington.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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, 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 *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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Stuart RF King, Emma Pewsey, and Sarah Shailes. 2017. Plain-language Summaries of Research: An inside guide to eLife digests. *eLife*, 6(e25410).

Hemang Malik, Gaurav Pradeep, and Pratinav Seth. 2024. HGP-NLP at BioLaySumm: Leveraging LoRA for Lay Summarization of Biomedical Research Articles using Seq2Seq Transformers. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 831–836, Bangkok, Thailand. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Satyam Modi and T Karthikeyan. 2024. Eulerian at BioLaySumm: Preprocessing Over Abstract is All You Need. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 826–830, Bangkok, Thailand. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine Learning in Python. *CoRR*, abs/1201.0490.

Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hai Tran, Zhengyun Yang, Zeya Yao, and Hong Yu. 2024. BioInstruct: Instruction Tuning of Large Language Models for Biomedical Natural Language Processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Daniel Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-Aware Decoding Strategies for Abstractive Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, 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.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLaySumm: An Extract-then-summarize Approach Augmented with Wikipedia Knowledge for Biomedical Lay Summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

Ruijing Zhao, Siyu Bao, Siqin Zhang, Jinghui Zhang, Weiyin Wang, and Yunian Ru. 2024. Ctyun AI at BioLaySumm: Enhancing Lay Summaries of Biomedical Articles Through Large Language Models and Data Augmentation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 837–844, Bangkok, Thailand. Association for Computational Linguistics.

A Data

A.1 Text Length

Figure 1 shows the distribution of tokens per article in both datasets. We found the PLOS articles to be shorter on average with a mean of 6981 tokens per article. The eLife articles were longer with an mean of 10,428 tokens with a greater variability in length.

Figure 2 compares the number of tokens across the gold standard summaries for the two datasets. A similar trend appears, with the PLOS lay summaries containing fewer (mean of 195) tokens than the eLife lay summaries (mean of 386), and the eLife distribution is again wider.

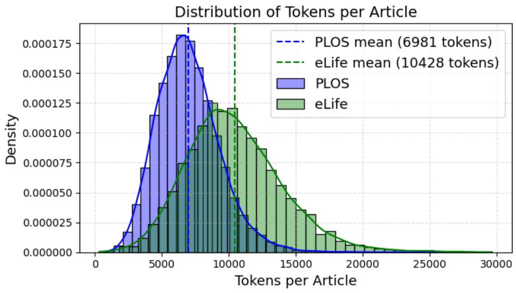

Figure 1: Distribution of tokens per article in the PLOS and eLife datasets.

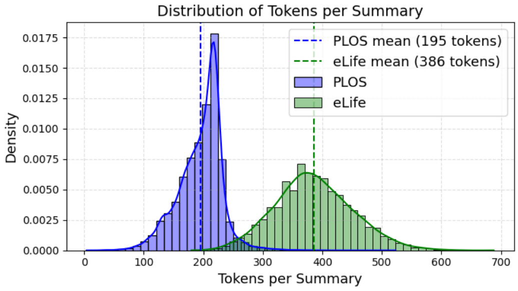

Figure 2: Distribution of tokens per summary in the PLOS and eLife datasets.

A.2 Section Relevance

You et al. (2024) compared each section of an article's relevance to the summary via cosine similarity. Across both datasets, they found the abstract, background, and conclusion to be the most relevant to the summary, in that order.

The existing dataset does not retain the section headings in place in the article text. The are extracted into a section headings list for each instance. We found that the article could be split on '\n' into a list of the different sections. We compared the listed sections across all instances and found that across PLOS instances, 100% contained an abstract, 99.85% contained an introduction, and 95.83% contained a discussion section (with another 3.53% containing a combined results/discussion section). Across eLife instances, 100% contained an abstract, 99.33% contained an introduction, and 98.62% contained a discussion. The compression ratio refers to the difference in length of an article and its lay summary. Figure 3 demonstrates that on average, the PLOS articles are less compressed than the eLife articles. While the eLife summaries are still longer on average than PLOS summaries, their articles are much longer, necessitating more compression of their information into a summary.

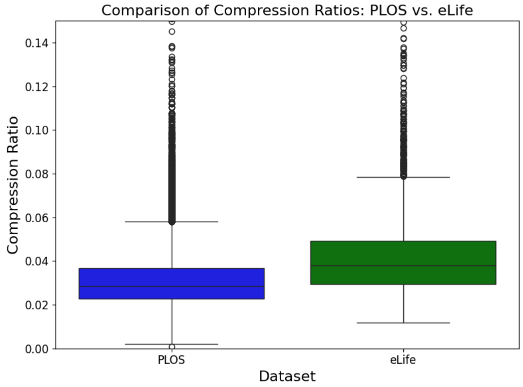

Figure 3: Distribution of compression ratios (article token length divided by summary token length).

B Model Settings

We framed the task using the prompt: *"Create a lay summary of this scientific research for a general audience who has no background in biology,"* leveraging Flan-T5's instruction-tuning capabilities. This approach aligns with recent work showing prompt-based task framing enhances performance in biomedical applications (Tran et al., 2024). We structured inputs with source-specific tags (e.g., <plos> [TITLE]. . . [ABSTRACT]. . .) as lightweight semantic cues. Input documents were truncated to 1024 tokens, with output summaries capped at 400 tokens.

Training used AdamW with a learning rate of 3e-5, weight decay of 0.01, and warmup ratio of 0.1. We employed a batch size of 12 without gradient accumulation, using PyTorch with expandable segment configuration for memory efficiency. Early stopping was applied with a patience of 2 evaluation steps. For generation, we utilized beam search with 4 beams, shown to produce more faithful summaries than sampling-based approaches (Wan et al., 2023).

The gradual decrease of both training and validation loss indicate that our model was able to learn and generalize effectively, as shown in Figure 4.

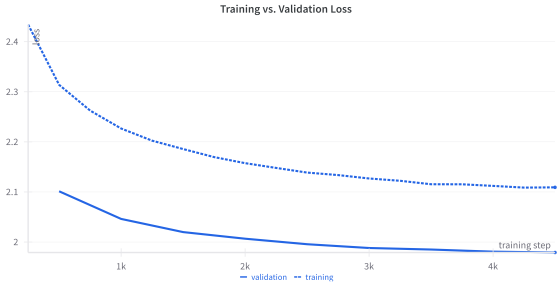

Figure 4: Comparison between the training and validation loss values across all model runs.