RainCityNLP at BioLaySumm2025: Extract then Summarize at Home

Jen Wilson **Avery Bellamy** **Rachel Edwards** **Michael Pollack** **Helen Salgi**
Department of Linguistics, University of Washington
(jenwils, elshire, rke4, uwmpp, hvs278)@uw.edu

Abstract

As part of the BioLaySumm shared task at ACL 2025, we developed a summarization tool designed to translate complex biomedical texts into layperson-friendly summaries. Our goal was to enhance accessibility and comprehension for patients and others without specialized medical knowledge. The system employed an extractive-then-abstractive summarization pipeline. For the abstractive component, we experimented with two models: Pegasus-XSum and a Falcons.ai model pre-trained on medical data. Final outputs were evaluated using the official BioLaySumm 2025 metrics. To promote practical accessibility, we completed all experimentation on consumer-grade hardware, demonstrating the feasibility of our approach in low-resource settings.

1 Introduction

The BioLaySumm shared task *Lay Summarization of Biomedical Research Articles and Radiology Reports @ BioNLP Workshop, ACL 2025* (Xiao et al., 2025) is conducting its third iteration this year. The goal of the shared task is to improve techniques for summarizing biomedical texts in non-scientific lay-terms, in order to increase the accessibility and understanding of medical texts for patients and others who are not in the medical field. We used the data from the shared task as well as their evaluation methods to create and evaluate our models and referenced previous participants' work for inspiration. We used an extractive-then-abstractive summarization technique. Beginning with extractive summarization and followed by training both the Pegasus-XSum model and the Falconsai/medical_summarization model to produce abstractive summaries. As a step towards future iterations of summarization, we have also created a dictionary of medical terms translated to lay-terms for injection.[1]

[1]https://github.com/michael-pollack/573Project.git

2 Related Work

Our pipeline of extractive-to-abstractive summarization was inspired by previous iterations of this workshop (Goldsack et al., 2023), (Goldsack et al., 2024) and the winning paper from 2024 (You et al., 2024). Our work is also influenced by the datasets used in this task (eLife and PLOS), which were developed by (Goldsack et al., 2022) and (Luo et al., 2022).

3 Description of Data

The dataset 'BioLaySumm2025-PLOS' consists of 26,291 rows and the dataset 'BioLaySumm2025-eLife' consists of 4,729 rows. Each row consists of the following information: the original text of a biomedical article, a gold-standard lay-terms summary, a list of section headings, a list of keywords, the year of publication, and the article title. Both datasets are already split into training, validation, and test.

We created a lay-term dictionary to add lay-term injection to our pipeline in the future. The dictionary consists of medical terms and their corresponding lay-term alternative based on a Stanford Glossary of medical terms (Stanford Research Compliance Office, n.d.). We were careful to start definitions with a consonant if the original word began with a consonant, and extended this to vowels. This premeditated measure was taken to facilitate smoother substitutions in the future with lay-term injections in the abstractive summaries.

3.1 Pre-Processing Data

Analysis showed that there are a large number of citations in academic text, which tend not to contribute significantly to the actual meaning of the document and are laden with complicated punctuation that affected our sentence tokenizer. We removed all information enclosed in parentheses

BioNLP 2025 Shared Tasks, pages 190–195

using regex and acknowledge that it removes more than just citations.

Since TF-IDF relies on vocabulary counts to calculate the importance of words, it is beneficial to remove stopwords and lemmatize the data first to reduce vocabulary size and establish obvious connections between different morphological variations of the same word. We used the built-in NLTK list of English stopwords, as well as our own short list of custom stopwords to target and remove stopwords from the data. NLTK's WordNet Lemmatizer was used to lemmatize remaining words in the document. The data we used to create the extractive summaries consists of both the clean lemmatized data resulting from these preprocessing techniques as well as the un-lemmatized version.

4 Model

4.1 Total Pipeline Overview

We use data cleaning and TF-IDF for preprocessing and the creation of extractive summaries. The extractive summaries are fed into an abstractive summary model.

4.2 TF-IDF

Term Frequency – Inverse Document Frequency (TF-IDF) (Sparck Jones, 1988) gives each word in a document a score based on importance to the document's meaning relative to the collection of documents. We chose TF-IDF because it allows us to numerically calculate the importance of words and sentences in a systematic way, thereby allowing us to rank which sentences should appear in the final extractive summary.

We used Scikit-Learn's prebuilt TF-IDF vectorizer model with the cleaned and lemmatized Elife training data as input to calculate the numerical importance of every word in each document in the data set. This produces a set of (word, vector) pairs for each document, where the larger the vector number, the higher the importance of the word. We calculated the importance of each sentence within a document by summing the TF-IDF scores of each word in the current sentence and dividing by the sentence's total word count. A higher score means that the sentence has a greater relevance to the meaning of the document.

We then return the 40% top-scoring sentences as an extractive summary.

4.3 Pegasus-XSum

Pegasus is an abstractive text summarization model developed by Google Research (Zhang et al., 2020). It is based on the Transformer architecture and was specifically pre-trained for summarization tasks using a "gap-sentence" technique, where whole sentences are masked and the model learns to predict them from the remaining text. This model was chosen because it could be fine-trained on our hardware.

4.4 T5 for Medical Text Summarization

Parallel to Pegasus-XSum, we also used the Falconsai/medical_summarization model (Wolf et al., 2020). This T5 Large for Medical Text Summarization model is fine-tuned specifically for medical domain summarization tasks. This model was selected for its strong performance on domain-specific texts and its ability to run efficiently on consumer-grade hardware, making it suitable for reproducible and accessible NLP research.

4.5 Computing Limitations

Limited access to high-end computing made it unrealistic to fine-tune hyper-parameters during the data validation. This is discussed in Section 7.1.

5 Evaluation

Relevance is measured using ROUGE (1, 2, and L), BLEU, METEOR, and BERTScore. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. BLEU (Bilingual Evaluation Understudy) is a measurement of an automatic translation and a human written translation of the same material. METEOR (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. For BERTScore, a neural evaluation metric uses contextual embeddings from pre-trained language models (like BERT) to calculate similarity scores between candidate and reference texts.

Readability is measured using Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), and LENS. Factuality is measured using AlignScore, and SummaC.

6 Results

For the medical summarization model, we used 'Summarize the following scientific article" as the prompt followed by the summary of the TF-IDF. For the pegasus models, we pass the TF-IDF summary through the model.

The results table of our summarizations are shown in 1.

6.1 Relevancy

As shown our evaluation data, medical_summarization scored highest for relevance. This model was developed specifically for summarization of medical text and it follows reasonably that it would score the highest. The extractive summaries coming in second in relevancy, beating out the Pegasus-XSum model (another specifically trained model for summarization) The extractive summaries were based mainly on frequency of word occurrence, using TF-IDF, which may make sense in context as a word that occurs frequently in the documents but is not an extremely general word of English, has a high likelihood of being relevant.

6.2 Readability

The Pegasus-XSum model dominated in the results for Readability in the FKGL evaluation metric, with the fine-tuned version of the model performing extremely well under the LENS evaluation metric. For DCRS the Pegasus-XSum model seems to perform slightly better than the rest, with the fine-tuned version actually performing the worst, and for CLI all three models other than the fine-tuned pegasus model perform at extremely similar levels, with the medical_sumarization model just barely performing a bit better than the rest. Readability tends to focus on word complexity and overall clarity of a summary. The Pegasus model is trained on medical texts, along with a wider variety of text to produce summarizations. This wider expanse of data could contribute to its readability scores as opposed to relevance because it is trained to create well-made abstractive summaries.

6.3 Factuality

Our original extractive summaries performed better than other models using both evaluation metrics. This is a relatively unsurprising result as the extractive summaries utilize the original sentences from the documents. Consequently, the summaries will be more factual than for the pegasus or medical_summarization models because the text is coming straight from the source.

7 Discussion

7.1 Accessible AI

In this section, we discuss how medical summarization systems can be made more accessible to a broader range of users. While recent advancements in medical NLP have demonstrated impressive capabilities, they often come with steep computational requirements, limiting their practical use outside well-resourced research institutions. Bednarczyk et al.(2025) report that the success of using an LLM for summarization relies on the computational resources available and future research should "evaluate the economic impact of deployment to ensure that LLM adoption is both technically and financially sustainable in clinical settings." We argue that accessibility - in both economic and practical terms - is essential if these technologies are to benefit clinicians, medical researchers, and public health professionals operating in low-resource environments or institutions without dedicated computing clusters.

7.2 Economic Accessibility

The growing trend toward large-scale models has created a barrier to entry for many who wish to apply or replicate state-of-the-art NLP techniques. We quickly ran into computational resource barriers because our plans of replicating previous work required computing resources that we did not have. As a result, our final choice of models and data processing were simpler to run, and can be used by people who do not have access to high-end computing resources. Our approach eliminates the need for expensive GPU clusters that are often used in academic settings.

All experiments in this study were conducted on one of two laptops. We used either a laptop with an Intel 13th Gen Core i9-13900H CPU, 32GB of RAM and a NVIDIA GeForce RTX 4060 Laptop GPU with 8GB RAM. This configuration, while not trivial, remains within reach of many individuals with limited budgets and does not rely on a distributed GPU cluster or a cloud-based API that incurs costs.

By using moderately sized, open-access models, such as Pegasus-XSum, and optimizing evaluation tools, we demonstrate that it is feasible to run sum-

	Relevance				Readability				Factuality	
model	**ROUGE**	**BLEU**	**METEOR**	**BERTScore**	**FKGL**	**DCRS**	**CLI**	**LENS**	**AlignScore**	**SummaC**
fine tuned pegasus-xsum	0.2119	1.8929	0.1453	0.8395	11.3870	8.6415	11.6500	45.9017	0.7654	0.6347
extractive summaries	0.2123	2.6039	0.2823	0.8276	19.7827	9.6537	15.4580	9.5067	0.9494	0.7874
pegasus-xsum hugging face	0.1816	2.0111	0.2122	0.8039	62.7306	13.4768	15.5227	9.1487	0.9029	0.6957
falconsai/medical_summarization	0.2845	4.8686	0.2405	0.8396	16.7401	11.6596	16.2369	9.4109	0.6118	0.6531

Figure 1: Evaluation Results

marization pipelines and evaluate results locally. This makes the system viable for clinicians, medical students, or NGOs who may wish to generate or verify lay summaries in real-world medical communication contexts.

7.3 Practical Reproducibility and Implementation

An important tenet of scientific research is reproducibility. Methods that can be implemented on accessible hardware can more easily be reproduced by scholars and others who want to learn. An analysis by Belz et al. (2021) demonstrates that replication and reproducibility are critical to scientific research, yet "it is surprisingly hard to achieve, 70% of scientist reporting failure to reproduce someone else's results, and more than half reporting failure to reproduce their own."

In this study, reproducibility was hindered by the complexity of previous configurations, conflicting dependencies, and reliance on costly computing environments. By deliberately choosing lightweight tools and open-source resources, we designed a summarization and evaluation pipeline that can be easily shared, executed, and adapted.

7.4 The Downside of Small Computing

While our emphasis on accessibility enables broader participation in this task, it also introduces notable limitations. Fine-tuning Pegasus-XSum on our consumer-grade hardware required approximately 40 hours, significantly slowing experimentation cycles. These experimentation cycles were slow and had to be run sequentially, instead of in parallel as could be done on a distributed GPU cluster. Due to hardware constraints, we were unable to explore larger or more recent models which would likely to produce results that score higher on the leaderboard.

Time and resource constraints prevented us from fine-tuning with separate validation data, limiting our ability to tune hyper-parameters effectively. These trade-offs demonstrate the challenges faced by low-resource researchers and scientists while highlighting the need to develop lightweight, efficient models that perform well without requiring extensive investments in hardware.

8 Conclusion

Our work demonstrates that medical summarization is achievable even with limited computational resources. By leveraging models like Pegasus-XSum, we were able to develop and evaluate effective summarization systems on a standard laptop, highlighting the potential for accessible and reproducible research in this space. Our findings support the idea that meaningful contributions to biomedical NLP can be made without relying on large-scale infrastructure, paving the way for more inclusive and resource-efficient approaches to language technology.

Limitations

While our system demonstrates results in generating readable and relevant lay summaries of biomedical texts using consumer-grade hardware, several limitations should be acknowledged.

Dataset Diversity

This work relies exclusively on two open-acess datasets: BioLaySumm2025-eLife and BioLaySumm2025-PLOS, both of which consist entirely of English language documents from a set of biomedical articles. As such, our model's generalizability to other medical domains or languages is untested.

Lay-Term Dictionary Coverage

The lay-term lexicon that we created, while a valuable resource for term injection, is limited in scope. It is derived from a single source (Stanford Research Compliance Office) and does not cover all relevant terminology. It requires an additional step and is not part of the summarization pipeline.

Pre-Processing

Our pre-processing decisions, particularly the removal of all parenthetical content using regular expressions, may have inadvertently discarded meaningful information. Although our rationale was that parenthetical content typically contains citations or supplementary material, this approach may have led to the loss of scientific details.

Experimentation Bottlenecks

Because all experiments were conducted on consumer-grade hardware without parallel GPU resources, experimentation had to proceed sequentially and in a time-consuming manner. This significantly limited our ability to iterate on model design or integrate new features (such as lay-term injection).

Validation and Fine-Tuning

Time and hardware constraints prevented us from fine-tuning using dedicated validation data. This limited our capacity to adapt the models. As a result, our models may not be optimally calibrated for the data distributions.

Ethical Considerations

The goal of this project and of the shared task as a whole is to expand the reach of biomedical text and make this information more approachable to people outside of the medical field. However, it is important to acknowledge that this task is not without its risks. For example, a flawed summarization system has the potential to give false information or omit important details from the original text, which is fundamentally opposed to the goal of the project. Additionally, it is important to include a diverse selection of texts when training a model of this kind, in order to reduce biases and create a model that can adapt and be used for a variety of new documents.

For this type of project, it is important to know that private information is not included in training documents, as that would be a violation of the privacy of individuals. The data included in our project was provided by the creators of the BioLaySumm shared task and comes from an open-access publisher (PLOS) and journal (eLife) so this is not a major concern for us.

References

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, and Christian Lovis. 2025. Scientific evidence for clinical text summarization using large language models: Scoping review. *Journal of Medical Internet Research*, 27:e68998.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. 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.

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.

Karen Sparck Jones. 1988. *A statistical interpretation of term specificity and its application in retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Stanford Research Compliance Office. n.d. Definitions & lay glossary of medical terms. https://researchcompliance.stanford.edu/panels/hs/for-all-researchers/definitions. Accessed: 2025-05-23.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

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 24th 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 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.