Shared Task at Biolaysumm2025 : Extract then summarize approach Augmented with UMLS based Definition Retrieval for Lay Summary generation.

Aaradhya Gupta and Dr Parameswari Krishnamurthy
LTRC, International Institute of Information Technology, Hyderabad
aaradhya.gupta@research.iiit.ac.in and param.krishna.iiit.ac.in

Abstract

We present LayForge, a two-track lay summary generation system developed for the BioLaySumm2025 shared task(Xiao et al., 2025) . Task 1.1 addresses the lay summarization using only the internal content of the article, while Task 1.2 augments this process with domain knowledge such as biomedical definitions and concept explanations. BioLaySumm employs a modular architecture that leverages large language models (LLMs), a BioBERT-based named entity recognizer (NER), and the UMLS(Bodenreider, 2004) knowledge base to create readable, informative, and faithful lay summaries. Our system shows strong performance on both tasks when evaluated on the PLOS and elife subset(Goldsack et al., 2022), particularly in readability and factuality metrics. The architecture illustrates how modularity and domain adaptation can be effectively combined for accessible biomedical communication.

1 Introduction

Lay summaries are a critical bridge between dense biomedical literature and non-specialist audiences, including patients, caregivers, and policy makers. These summaries must balance clarity, completeness, and technical accuracy. The BioLaySumm2025 shared task(Xiao et al., 2025) presents two summarization challenges:

- **Task 1 (Internal-only):** Generate a lay summary using only the content of the original article.
- **Task 2 (Augmented):** Improve the lay summary by incorporating external biomedical knowledge such as terminology definitions.

We introduce **LayForge**, a flexible and extensible system designed to address both tracks. Our design is rooted in modular NLP techniques - chunk extraction, LLM-based draft generation, and iterative rewriting—with additional augmentation for Task 2 using BioBERT-based NER (Lee et al., 2019) and UMLS-based concept simplification.

Our contributions include:

- A two-tiered summarization pipeline that integrates pretrained LLMs with biomedical NER and knowledge retrieval.
- A task-specific rewriting mechanism for increasing the readability and accessibility of summaries.
- A detailed performance comparison across readability, fidelity, and factuality metrics.

2 Related Work

The BioLaySumm shared task series began in 2023 (Goldsack et al., 2023), with a follow-up edition in 2024 (Goldsack et al., 2024), laying the groundwork for consistent evaluation and dataset development in biomedical lay summarization. Our work builds on the methodologies and evaluation frameworks introduced in these earlier editions, Biomedical summarization has traditionally leveraged sequence-to-sequence architectures and domain-specific pretrained models such as BioBERT and PubMedBERT(Beltagy et al., 2020). Recent trends in summarization, including the use of large language models and retrieval-augmented generation (RAG)(Lewis et al., 2020), show promise in improving factuality and reducing hallucination. Entity-level simplification is another important strand, where domain terms are replaced or explained using biomedical ontologies. However, most prior work stops at simple substitutions, while our system integrates retrieved definitions into fluent rewrites. Instruction tuning for LLMs is also a promising avenue of research. (Tran et al., 2024) introduced a corpus of 25,005 human-crafted prompts to instruction-tune LLaMA models on biomedical tasks, yielding QA gains and generation improvements There have also been

BioNLP 2025 Shared Tasks, pages 185–189

3 System Architecture

We ensure that the system architecture is modular and easy to understand. The 2 tasks share a common pipeline in the beginning. The augmentation using the UMLS backed definitions is performed for task 2 at the end.

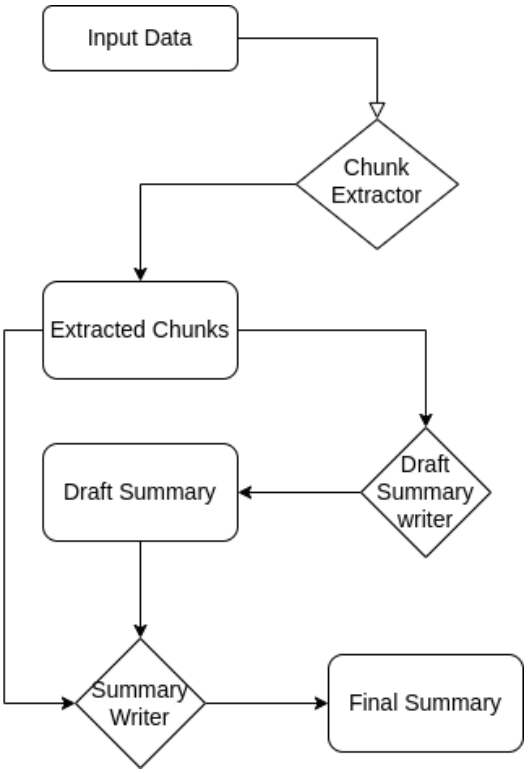

Figure 1: Task 1.1 : Framework for lay summary generation with no external information

3.1 Shared Architecture (Tasks 1 & 2)

Preprocessing and Chunking Articles are segmented into overlapping text chunks (3,000 tokens with 200-token overlap) to accommodate LLM context windows and ensure semantic continuity.

Top-k Sentence Extraction For each chunk, salient sentences are extracted using an LLM (LLaMA 3-70B) prompted to select informative statements. The resulting sentence pool contains all the key findings and methods.

Draft Generation We conditioned the LLM with article metadata, keywords, and extracted sentences to generate a draft lay summary. Prompts guide the model to assume a "science teacher" persona to ensure accessibility.

Iterative Rewriting Two rewriting passes are applied:

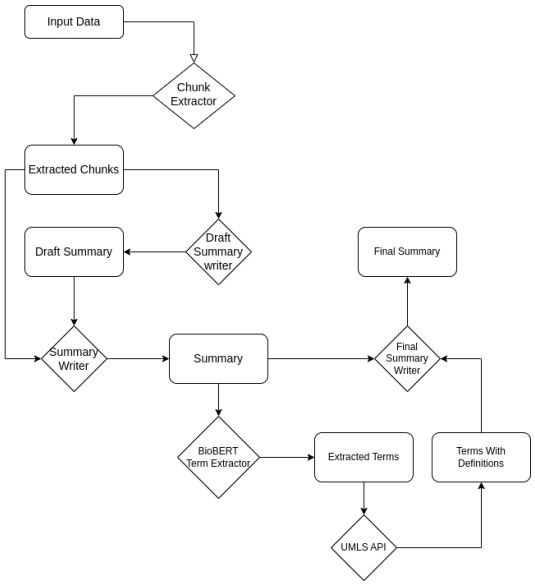

Figure 2: Task 1.2 : Framework for lay summary generation with external information

- **Reader Rewrite:** A persona-based prompt enhances flow and readability.
- **Jargon Softening:** Phrases are simplified and clarified, guided by syntactic and lexical heuristics.

3.2 Knowledge Augmentation (Task 2 Only)

Domain NER A BioBERT model fine-tuned for NER identifies and extracts biomedical terms that are not layperson friendly in the summary.

Definition Retrieval Each detected term is passed to a UMLS-backed lookup API, which retrieves lay definitions.

Definition-Guided Rewrite These definitions are incorporated into the summary through guided LLM prompts, either by appending explanations or paraphrasing terms inline.

3.3 Model Selection

We chose LLaMA 3–70B as our backbone because it offers a strong balance between model capacity and computational cost, while remaining fully open-source under a permissive license. In preliminary experiments (not shown), LLaMA 3–70B outperformed smaller variants (e.g., 13B) on zero-shot biomedical QA benchmarks. Additionally, its 8K-token context window accommodates long article chunks without resorting to expensive retrieval passes, which was critical for processing 3,000-token windows in our pipeline.

4 Implementation Details

We implemented LayForge in Python, using Langchain for orchestrating LLM calls and LangGraph for managing pipeline state. Sentence extraction and summarization use the Groq-hosted LLaMA 3-70B model. BioBERT NER is handled using the SimpleTransformers library.

UMLS queries are made via a RESTful endpoint returning short, simplified definitions. All components are containerized and run using Google Colab with GPU acceleration for efficiency.

4.1 Handling UMLS Definition Ambiguity

When a detected term has multiple definitions in UMLS, our lookup strategy resolves ambiguity by:

- **Source prioritization:** We only accept definitions whose rootSource is in (MSH, PDQ, NCI, MEDLINEPLUS), in that order.
- **Conciseness heuristic:** If multiple definitions remain, we choose the one with the fewest tokens, assuming brevity aids lay understanding.
- **Fallback:** If no preferred definition is found, we leave the term unchanged and rely on the LLM's paraphrasing step to "soften" it.

5 Experimental Setup

We evaluate both tasks using the BioLaySumm2025 PLOS and elife datasets.

Evaluation Metrics Various evaluation metrics were used to evaluate performance of the system in different fields.(Luo et al., 2022)

- **Readability:** FKGL, DCRS, CLI and LENS.
- **Content Fidelity:** ROUGE-L, BLEU-4, METEOR, BERTScore.
- **Factuality:** SummaC and AlignScore.

6 Results and Discussion

Our results show that augmentation with external definitions significantly improves readability metrics, with FKGL decreasing by over 3 points and DCRS/CLI also showing similar gains. The LENS metric confirms slightly longer outputs, likely due to inserted definitions and the model being more verbose to avoid using technical terms

Metric	Task 1	Task 2
ROUGE	0.32	0.29
BLEU	5.45	4.32
METEOR	0.29	0.26
BERTScore	0.85	0.85
FKGL	14.56	11.15
DCRS	10.01	8.36
CLI	15.36	11.93
LENS	71.51	81.50
AlignScore	0.69	0.61
SummaC	0.50	0.53

Table 1: Performance across BioLaySumm track as per the Leaderboard

Interestingly, although Task-2 reduces ROUGE and BLEU slightly, this can be attributed to paraphrasing and definition insertion changes that promote lay understanding at the cost of n-gram overlap. Semantic paraphrase or added explanatory phrase tend to reduce these metrics despite improving readability and fidelity(Xu et al., 2016; Ganesan, 2018; Wang et al., 2022; Elliott et al., 2019).

SummaC(Tian et al., 2022) improves in Task2, suggesting factual information is more clearly and accurately preserved with domain support. AlignScore, however, drops slightly, indicating a potential loss in extractive alignment.

7 Conclusion and Future Work

We presented LayForge, a modular, domain-aware pipeline for lay summary generation in biomedical research. Our two-track approach enables detailed comparison between internal-only and knowledge-augmented strategies.

In future work, we aim to incorporate dynamic chunk prioritization, retrieval-based augmentation, and human-in-the-loop feedback. We also plan to evaluate generalizability across different journals and medical domains. We also see great potential in Reinforcement Learning from Human Feedback (RLHF)(Natarajan et al., 2024) to directly optimize summary-quality objectives (e.g. readability, factuality) by defining appropriate reward functions.

Acknowledgments

We thank the BioLaySumm organizers and annotators for their support. This work is supported by the LTRC Lab at IIIT Hyderabad

Step	Text
Original Paragraph	"DPAGT1-CDG is a Congenital Disorder of Glycosylation (CDG) that ..."
Top-k Sentences	1. DPAGT1-CDG is a Congenital Disorder of Glycosylation (CDG) that lacks effective therapies. 2. DPAGT1 synthesizes dolichol-PP-GlcNAc, which is the first step in N-glycosylation. 3. We used a Drosophila rough eye model of DPAGT1-CDG with an improperly developed, small eye phenotype.
Draft Simple Passage	"DPAGT1-CDG is a rare genetic disorder that affects the synthesis of dolichol-PP-GlcNAc, a crucial step in N-glycosylation. This disorder leads to developmental delay, muscle weakness, and seizures, among other symptoms. To better understand the disease and identify potential therapeutic options..."
Retrieved Definitions	developmental delay: 'Failure to meet, or late achievement of developmental milestones.' dolichol-PP-GlcNAc: 'A compound involved in the first step of N-glycosylation.' muscle weakness: 'A reduction in the strength of one or more muscles.'
Final Rewrite	"DPAGT1-CDG is a rare genetic disorder that affects the production of a crucial molecule, essential for a process called N-glycosylation. This disorder causes symptoms including delayed developmental milestones, muscle weakness, and seizures. To study the disease and identify treatments, researchers used a fruit fly model with small, improperly developed eyes and tested 1,520 existing drugs from a library of FDA- or EMA-approved compounds..."

Table 2: Example of the Extract–Draft–Rewrite Flow

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2020. Pubmedbert: Domain-specific language model pretraining for biomedical natural language processing. In *Findings of EMNLP*, pages 3615–3620.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Dan Elliott and 1 others. 2019. Bleu is not suitable for the evaluation of text simplification. Technical blog post, ResearchGate.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

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. *Preprint*, arXiv:2408.08566.

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.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

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.

Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. 2024. Human-in-the-loop or ai-in-the-loop? automate or collaborate? *Preprint*, arXiv:2412.14232.

Tian Tian, Emma Reiter, and 1 others. 2022. Summac: Revisiting nli-based models for inconsistency detection in summaries. *Transactions of the Association for Computational Linguistics (TACL)*, 10:350–365.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *Preprint*, arXiv:2310.19975.

Zhuohan Wang and 1 others. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of EMNLP*, pages 208–221.

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.

Wei Xu, Courtney Napoles, Claire O'Connor, Chris Callison-Burch, and Ellie Loper. 2016. Optimizing statistical machine translation for text simplification. In *Proceedings of NAACL-HLT*, pages 85–95.