CUTN_Bio at BioLaySumm: Multi-Task Prompt Tuning with External Knowledge and Readability adaptation for Layman Summarization

Bhuvaneswari Sivagnanam[1], Rivo Krishnu C H[1], Princi Chauhan[1], Saranya Rajiakodi[1]

[1]Department of Computer Science, Central University of Tamil Nadu, India

Correspondence: saranya@acad.cutn.ac.in

Abstract

In this study, we presented a prompt based layman summarization framework for the biomedical articles and radiology reports developed as part of the BioLaySumm 2025 shared task at the BioNLP Workshop, ACL 2025. For Subtask 1.1 (Plain Lay Summarization), we utilized the abstract as input and employed Meta-LLaMA-3-8B-Instruct with a Tree-of-Thought prompting strategy and obtained 13th rank. In Subtask 1.2 (Lay Summarization with External Knowledge), we adopted an extractive plus prompt approach by combining LEAD-K sentence extraction with Meta-LLaMA-3-8B-Instruct. Medical concepts were identified using MedCAT, and their definitions were taken from Wikipedia to enrich the generated summaries. Our system secured the 2nd position in this subtask. For Subtask 2.1 (Radiology Report Translation), we implemented a Retrieval-Augmented Generation (RAG) approach using the Zephyr model to convert professional radiology reports into layman terms, achieved 3rd place in the shared task.

1 Introduction

In recent years, it has become much easier for people to access scientific and medical information online. Research papers and clinical reports like radiology reports are now widely available. In particular, biomedical articles and radiology reports often use difficult terms and specialized language that is difficult for reading and understanding for most people(Tariq et al., 2024). This makes it harder for the students or general public to understand medical information and reduce the impact of the scientific research. Lay summarization, which means rewriting scientific or medical content in simple language for the general public, is a helpful solution to this problem. Previous studies have shown the value of creating patient-friendly versions of radiology reports(Tariq et al.), and have also highlighted the need to make scientific communication more suitable for different types of readers(Fonseca and Cohen, 2024). However, writing good lay summaries is still a difficult task. Large Language Models (LLMs) like GPT-3.5 and LLaMA are good at general summarization, but they often struggle to produce easy to read to summary due to the technical jargon present in the medical text and the models are not customized for the medical text(Fonseca and Cohen, 2024). In this paper, we describe our system for the BioLaySumm 2025 Shared Task(Xiao et al., 2025). We combine prompt-based language models, extractive summarization techniques, and background knowledge from external sources. Our contributions are:

1. We introduced the use of the Tree-of-Thought (ToT) prompting strategy in biomedical lay summarization to generate more readable, logically organized, and controllable summaries.
2. We leveraged Chain-of-Thought (CoT) prompting and role based prompting for lay summarization of biomedical articles improving the clarity and factual consistency without requiring large scale supervised data.
3. We developed a retrieval-augmented summarization pipeline for radiology reports by storing medical concepts and definitions from the dataset and Wikipedia in ChromaDB, enabling definition retrieval to improve clarity for lay readers.

2 Related Work

Recent studies in biomedical text summarization have focused on making complex medical artciles to be easier to understand for general audience, especially because of the fast growing number of scientific articles. Researchers have explored both extractive methods and abstractive methods to create summaries that are easier for non-experts

BioNLP 2025 Shared Tasks, pages 269–274

to read. Traditional extractive techniques such as Lead-K, TextRank, and TF-IDF help pick the most important sentences from a text. Tools like SciSpacy (Neumann et al., 2019) and MedCAT (Kraljevic et al., 2021) are useful for identifying medical terms that might need simpler explanations in the summaries. New large language models (LLMs) such as Meta's LLaMA-3-Instruct (Touvron et al., 2023) and OpenAI's GPT-3.5/4 (Achiam et al., 2023) have shown strong abilities to write summaries using prompts, even without much extra training. Models that are fine-tuned to follow instructions produce better results and clearer outputs that matches with the result user wants. Tree-of-thought prompting (Yao et al., 2023) and chain-of-thought reasoning (Wei et al., 2022) have demonstrated improvements in factuality and coherence for complex text generation tasks, including medical content. In the context of radiology report translation, recent shared tasks (e.g., BioLaySumm, MEDIQA) and benchmark datasets such as MIMIC-CXR (Johnson et al., 2019) and PadChest(Bustos et al., 2020) have helped the development of models that generate layman friendly summary of professional reports. Prior studies (You et al., 2024) have used Retrieval-Augmented Generation (RAG) to supplement missing background knowledge and improve factual accuracy. To evaluate how good the summaries are, common tools include ROUGE, BLEU, METEOR, and BERTScore for relevance, as well as readability scores like FKGL and DCRS. To check if the summary facts match the source text, tools like AlignScore and SummaC are often used.

3 Methodology

For this biomedical articles summarization task, we used the PLOS and eLife dataset provided by the organizers as described in (Goldsack et al., 2022; Luo et al., 2022; Goldsack et al., 2023, 2024). The PLOS dataset is the larger of the two, comprising 24,773 training and 1,376 validation instances, while the eLife dataset contains 4,346 training and 241 validation instances. For the radiology report summarization task, we used the close track setup which includes Open-i, PadChest, BIMCV-COVID19, along with MIMIC-CXR (Zhao et al., 2025; Xiao et al., 2025).

3.1 Plain Lay Summarization

We used the Meta-LLaMA-3-8B-Instruct model to simplify medical texts into layperson-friendly summaries along with Tree of Thought algorithm(1). To ensure high-quality simplification, we designed a prompt that instructed the model to (i) shorten and simplify long sentences, (ii) replace complex medical terms with everyday language and (iii) keep the original meaning accurate. After generating the summaries, we cleaned them using regular expressions to remove extra characters, model tags, and formatting issues. This gave us a neat version of the simplified text. For each input, we asked the model to generate two versions (n = 2). Although creating more versions (like 3–5) can give better results, we chosen two to save time. We then picked the best one using a custom scoring formula called wrb(Readbility-Bertscore based): wrb = 0.55 × br_score + 0.45 × avg_read. Here, br_score is (1 – BERTScore) × 100, and avg_read is the average of two readability scores: Flesch-Kincaid Grade Level and Dale-Chall Readability Score. We used BERTScore to check how close the simplified summary was to the original abstract. After choosing the best summary in the first round, we fed it back into the model to generate more refined outputs. This was done for up to two rounds (m = 2). If the first summary already scored well (above a quality threshold of 12), we skipped the second round to save time.

3.2 Lay Summarisation with External Knowledge

We presented a layman summary system[Figure 1] designed for biomedical articles, using two different extractive–abstractive hybrid models. The first model used Meta's LLaMA-3-8B-Instruct, combining section-wise leading sentence extraction and medical term definitions from MedCAT. The second model used GPT-3.5-turbo with TF-IDF-based sentence selection and medical terms identified using SciSpacy. Both models used specially written prompts and the GPT model used a step-by-step reasoning prompt to improve accuracy and structure. In the first setup, each article from the BioLaySumm dataset is splitted into its usual sections (like Introduction, Methods, Results, Discussion) using newlines as markers. From each section, the first 10 sentences are picked (Lead-10 method), since these usually carry the main ideas. These sentences are joined with the abstract to make a

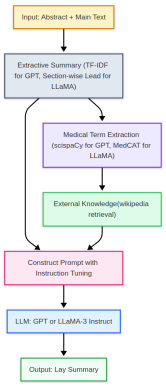

Figure 1: This flowchart illustrates the subtask 1.2 pipeline for biomedical lay summarization using external knowledge. Extracted summaries and abstract are combined with medical term definitions from sources like Wikipedia, and a LLaMA-3-Instruct model generates simplified summaries for the general public.

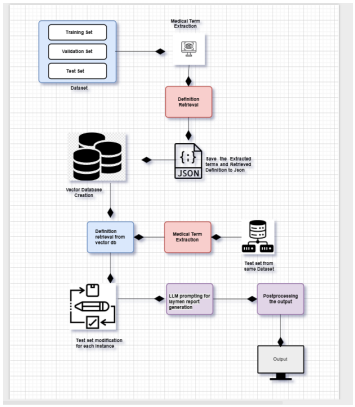

Figure 2: Pipeline for the radiology report translation

shorter input for the summarization model. To get the cleaned text, we preprocessed the text data by removing the contents in brackets like citations. To make the summary easier to understand for non-experts, we extract medical terms from the abstract and Lead-10 sebtences using MedCAT. This tool links terms to standard medical databases. We removed common English words using a dictionary and fetched simple definitions for up to 10 terms from Wikipedia. These definitions are added to the prompt to give extra background. The final prompt includes: The abstract, sentences picked from each section, up to 10 definitions of medical terms. The second setup uses GPT-3.5-turbo through OpenAI's API. Instead of section wise extractive summary, this setup used TF-IDF to pick the top 40 most important sentences from the full text. We then used SciSpacy model to find medical terms, based on known databases like UMLS and MeSH. Definitions are again retrieved from Wikipedia and added to the prompt. The GPT model uses a step-by-step reasoning prompt (Chain-of-Thought) to improve its output. This prompt asks the model to think through the abstract and sentences, explain hard terms, and then write a clear summary in simple words. This helps reduce errors and improves clarity.

3.3 Radiology Report Translation

To generate layman-friendly summaries from complex radiology reports, we developed a structured pipeline [Figure 2] integrating biomedical entity recognition, semantic definition retrieval, large language model prompting and post-processing. We began with the BioLaySumm/LaymanRRG-closed track dataset and used the full dataset to construct a comprehensive dictionary of medical definitions. Each report was processed using the SciSpacy model to extract medical terms and their definition was retrieved from Wikipedia using their API. We then encoded each definition using the BAAI/bge-base-en sentence transformer and indexed them in a ChromaDB vector database, facilitating fast and semantically meaningful retrieval. For inference, we focused on the test split of the dataset. Each test report underwent entity extraction, and the corresponding terms were matched with the most semantically relevant definitions from our vector database. These definitions, combined with the original report, were formatted into a context-rich prompt and passed to the zephyr-7b-beta model using Hugging Face's text-generation pipeline. The model, deployed in half-precision (FP16) on GPU, produced concise and coherent lay summaries. To ensure clarity and fluency, we applied a post-processing step to the generated summaries. This included removing redundant phrases, fixing formatting inconsistencies, and refining grammar and sentence flow. Each report's full pipeline output including the original report, extracted terms, definitions used, generated summary, and the final cleaned summary was saved in a CSV file. This allowed for incremental saving and supported recovery in case of interruptions, ensuring robustness and reproducibility of the process.

4 Result and Discussion

The performance of our submission were presented in the Table 1

Task	Model used	ROUGE	BLEU	METEOR	BERTScore	FKGL	DCRS	CLI	LENS(Task1)	AlignScore(Task1)	SummaC(Task1)
SubTask 1.2	GPT 3.5	0.2961816563	4.081113931	0.2281900631	0.8549233545	13.36619718	10.25193662	14.74401408	80.00152005	0.6890650382	0.5070385472
SubTask 1.2	LLaMA-3-8B-Instruct	0.2894793205	4.127686909	0.2682454109	0.8340255209	9.892605634	7.869929577	11.38640845	75.58860065	0.5547797396	0.737774146
SubTask 1.1	LLaMA-3-8B-Instruct with TOT	0.2681919676	3.24775969	0.2263706053	0.8484312852	10.52429577	8.835915493	11.43105634	84.14457219	0.5888629015	0.5489283793
SubTask 1.1	Preprocessed Abstract	0.3281246986	7.120012357	0.2833030102	0.8612545194	16.90774648	11.35848592	17.51320423	40.15254495	0.9937086519	0.9464885324
Task	**Model used**	**ROUGE**	**BLEU**	**METEOR**	**BERTScore**	**Similarity(Task2)**	**FKGL**	**DCRS**	**CLI**	**F1chexbert(Task2)**	**Radgraph(Task2)**
SubTask 2.1_close	zephyr-7b-beta	0.4038681644	14.89689754	0.427866722	0.9128268815	0.7975429296	7.358711512	8.527076782	7.360385788	0.7041964506	0.2162386679

Table 1: Evaluation results demonstrates the performance of our submission across Subtasks 1.1 and 1.2 of biomedical lay summarization, and Subtask 2.1 of radiology report translation.

4.1 Lay Summarisation with External Knowledge

We tested our system on 142 biomedical articles from the BioLaySumm2025-PLOS test set. Both models (LLaMA-3 with Lead-10 + MedCAT and GPT-3.5 with TF-IDF + SciSpacy + Chain-of-Thought) were run on the same dataset for comparison. We evaluated the results using relevance, readability and factuality metrics. GPT-3.5 performed better than LLaMA-3 in readability, factual correctness, and overall ROUGE-L scores, especially when it explained complex terms using step-by-step reasoning. In general, GPT-3.5 created summaries that explained key findings and terms more clearly for general readers. LLaMA-3 sometimes skipped important context. The combination of TF-IDF extraction with reasoning-based prompts worked especially well when extra background knowledge was needed for understanding.

4.2 Radiology Report Translation

In the radiology report translation task, we developed an approach using a retrieval-augmented generation (RAG) strategy with the Zephyr model to produce summaries that are not only accurate but also easier to understand for non-experts. The summaries generated through this method gave high relevance. The scores of readability metrics confirmed that the simplified texts were written at an accessible level, making them more understandable to the general public. From a clinical perspective, our approach maintained a good balance between simplifying the language and retaining important medical content.

5 Conclusion

In this work, we explored prompt-based summarization techniques to convert complex biomedical articles and radiology reports into simple, easy-to-understand summaries for general readers. We used a combination of extractive methods (such as Lead-10 and TF-IDF) and large language models like Meta-LLaMA-3-Instruct and GPT-3.5. For Subtask 1.1, we focused on section-wise summarization, while in Subtask 1.2, we added medical definitions retrieved using MedCAT and Wikipedia to enrich the knowledge gaps. In Subtask 2.1, we applied a Retrieval-Augmented Generation (RAG) approach with the Zephyr model to generate layman summaries from professional radiology reports. Our approach produced strong results across the shared task subtasks, showing the effectiveness of combining external knowledge, extractive summarization, and instruction-tuned language models. Even though our system produced good results, it still has has some limitations. First, the quality of extracted summaries using Lead or TF-IDF depends on the structure of the original article. If the article is not organized properly, important information might be missed. Second, retrieving accurate definitions from Wikipedia or other public sources may introduce inconsistencies, not having clear explanations or no explanations. Each and every retrieval of wikipedia definitions takes too much time and that limited the medical definition to only 10 terms. Finally, while chain-of-thought prompting improved factuality in GPT-based generation, it occasionally produced longer or slightly off-topic outputs when trying with llama that requires further refinement. In future work, we plan to improve summarization by fine-tuning the models instead of prompt tuning on more diverse medical datasets. We will also be focusing on controllable summarization and exploring the ways to get definitions from other medical resources that can increase the terms count and layman definitions. We want to incorporate user feedback mechanisms to assess how helpful the generated summaries are for real patients and the general public.

6 Declaration of AI usage

We used generative AI tools like chatGPT for paraphrasing, grammar checking while writing this article. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Marcio Fonseca and Shay B Cohen. 2024. Can large language model summarizers adapt to diverse scientific communication goals? *arXiv preprint arXiv:2401.10415*.

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 the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*.

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.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

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. *arXiv preprint arXiv:1902.07669*.

Amara Tariq, Sam Fathizadeh, Gokul Ramaswamy, Shubham Trivedi, Aisha Urooj, Nelly Tan, Matthew T Stib, Bhavik N Patel, and Imon Banerjee. 2024. Patient centric summarization of radiology findings using large language models. *medRxiv*, pages 2024–02.

Amara Tariq, Shubham Trivedi, Aisha Urooj, Gokul Ramasamy, Sam Fathizadeh, Matthew Stib, Nelly Tan, Bhavik Patel, and Imon Banerjee. Patient-centric summarization of radiology findings using two-step training of large language models. *ACM Transactions on Computing for Healthcare*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

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.

Kun Zhao, Chenghao Xiao, Sixing Yan, William K. Cheung, Kai Ye, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2025. X-ray made simple: Lay radiology report generation and robust evaluation. *Preprint*, arXiv:2406.17911.

A Appendix A: Algorithm

Algorithm 1 Tree of Though Based Iterative Text Simplification with Readability and Content Preservation

Input: Abstract
max_rounds $\leftarrow 2$
branches_per_round $\leftarrow 2$
current_text $\leftarrow$ complex_text
for round_num = 0 to max_rounds - 1 **do**
    candidates $\leftarrow$ [ ]
    for i = 0 to branches_per_round - 1 **do**
        generated $\leftarrow$ SIMPLIFY_TEXT_WITH_LLAMA(current_text)
        simplified $\leftarrow$ CLEAN_SIMPLIFIED_OUTPUT(generated)
        avg_read $\leftarrow$ READABILITY_SCORES(simplified)
        input_fc $\leftarrow$ BERTSCORE(simplified, complex_text) ▷ or AlignScore / SummaC
        br_score $\leftarrow (1 - \text{input\_fc}) \times 100$
        wrb $\leftarrow 0.55 \times$ br_score $+ 0.45 \times$ avg_read
        candidates.append([simplified, wrb, readability, input_fc])
    end for
    best_candidate $\leftarrow$ candidate with lowest wrb
    if best_candidate.wrb $\leq$ target_grade **then**
        return best_candidate.simplified
    else
        current_text $\leftarrow$ best_candidate.simplified
    end if
end for
return current_text