5cNLP at BioLaySumm2025: Prompts, Retrieval, and Multimodal Fusion

Juan Antonio Lossio-Ventura[1], Callum Chan[2], Arshitha Basavaraj[3], Hugo Alatrista-Salas[4], Francisco Pereira[1], Diana Inkpen[2]
[1]Machine Learning Core, National Institute of Mental Health, National Institutes of Health, USA
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Canada
[3]International Institute of Information Technology, Bangalore, India
[4]De Vinci Research Center, Paris, France,
juan.lossio@nih.gov, cchan073@uottawa.ca, arshitha.basavaraj@iiitb.ac.in
hugo.alatrista_salas@devinci.fr, francisco.pereira@nih.gov, dinkpen@uottawa.ca

Abstract

In this work, we present our approach to addressing all subtasks of the BioLaySumm 2025 shared task by leveraging prompting and retrieval strategies, as well as multimodal input fusion. Our method integrates: (1) zero-shot and few-shot prompting with large language models (LLMs); (2) semantic similarity-based dynamic few-shot prompting; (3) retrieval-augmented generation (RAG) incorporating biomedical knowledge from the Unified Medical Language System (UMLS); and (4) a multimodal fusion pipeline that combines images and captions using image-text-to-text generation for enriched lay summarization. Our framework enables lightweight adaptation of pretrained LLMs for generating lay summaries from scientific articles and radiology reports. Using modern LLMs, including Llama-3.3-70B-Instruct and GPT-4.1, our 5cNLP team achieved third place in Subtask 1.2 and second place in Subtask 2.1, among all submissions.

1 Introduction

BioLaySumm's third edition (Xiao et al., 2025b) introduces a new task focused on translating radiology reports into layperson-friendly language, while continuing its existing biomedical article summarization task from previous editions (Goldsack et al., 2024, 2023). Summaries are expected to include more background information and reduce technical jargon to improve accessibility.

Thus, BioLaySumm 2025 comprises two main tasks, each with two subtasks, aimed at improving biomedical communication for lay audiences. Task 1 focuses on generating accessible summaries of biomedical research articles from PLOS and eLife, either directly (Subtask 1.1) or with the integration of external knowledge sources (Subtask 1.2). Task 2 targets the translation of radiology reports into layperson-friendly language, using text alone (Subtask 2.1) or in combination with chest x-ray images (Subtask 2.2). This task was offered in both open and closed tracks, with the closed track additionally incorporating the MIMIC-CXR dataset. We opted for the closed track in our submission.

To address these tasks, we developed a unified and flexible framework that combines prompting, retrieval, and multimodal fusion techniques. It supports zero- and few-shot prompting with LLMs, dynamic few-shot selection via embedding-based nearest neighbors, retrieval-augmented generation using UMLS (Bodenreider, 2004), and multimodal processing through image-text-to-text generation for enriched lay summarization. Based on our previous experience, we adopted structured (compositional) prompting including task goals, instructions, formatting guidelines, and output specifications (Chan et al., 2025). Also, previous work shows that LLMs perform better with well-chosen in-context examples (Brown et al., 2020; Liu et al., 2021). Following (Liu et al., 2021), we chose most similar samples based on cosine similarity for few-shot prompting. We also explored varying the number and selection strategy of these examples. Moreover, we explored several LLMs of varying sizes, including Llama-3.1-8B, Llama-3.1-8B-Instruct (standard and 8-bit quantized), Llama-3.3-70B-Instruct, and GPT-4.1. A single approach was applied across all task datasets, without building data-specific models, to improve generalizability.

2 Shared Task Overview

Task 1: Lay Summarization: Participants were required to generate layperson-accessible summaries of biomedical articles from two datasets, PLOS and eLife, using two different approaches.

- **Subtask 1.1: Plain Lay Summarization**: Given an article's abstract and main text, systems had to produce a non-technical summary suitable for a general audience.

BioNLP 2025 Shared Tasks, pages 215–231

- **Subtask 1.2: Lay Summarization with External Knowledge**: This task extended Subtask 1.1 by permitting the use of additional knowledge sources (e.g., databases, medical ontologies) to enrich contextual understanding for lay readers.

Task 2: Radiology Report Translation to Layperson's Terms: This task was offered in open and closed tracks. The open track used PadChest (Bustos et al., 2020), Open-i, and BIMCV-COVID19 (de la Iglesia Vayá et al., 2020), while the closed track additionally included MIMIC-CXR (Johnson et al., 2019, 2024).

- **Subtask 2.1: Radiology Report Translation:** The goal was to build models to translate professional radiology reports to layperson's terms.
- **Subtask 2.2: Multimodal Radiology Report Translation:** This was a multi-modal task with the goal of achieving a lay translation of radiology reports. The input was chest x-ray images and radiology reports and the output should be a report in layperson's terms.

Datasets: All datasets were made available by the organizers on HuggingFace (Xiao et al., 2025a; Zhao et al., 2024) - except the imaging data from MIMIC-CXR used in Subtask 2.1. For Task 1, two datasets from biomedical journals, PLOS and eLife, were provided (Goldsack et al., 2022; Luo et al., 2022). For Task 2, four datasets were used: Open-i, PadChest, BIMCV-COVID19, and MIMIC-CXR (Zhao et al., 2025). Participants could choose between using only the first three (open track) or all four (closed track). The training, validation, and test splits are detailed in Appendix Tables 4 and 5.

Evaluation Metrics: Submissions were evaluated using task-specific metrics. For Task 1, summaries were assessed on relevance (ROUGE-1/2/L, BLEU, METEOR, BERTScore), readability (FKGL, DCRS, CLI, LENS), and factuality (AlignScore, SummaC). Task 2 used the same relevance metrics, similar readability measures (excluding LENS), and clinical-specific factuality metrics (CheXbert-F1, RadGraph-F1). All metrics were determined by the shared task organizers.

3 Methods

We used prompting, retrieval, and multimodal fusion with Llama and GPT models, outlined below.

TASK 1

We focused on text-to-text generation tasks, mainly using zero-shot, one-shot, and few-shot prompting. Building on our experience from previous shared tasks, we used structured (compositional) prompting, which included task goals, instructions, guidelines, and output formats (Chan et al., 2025). In this work, we extended our structured prompts by incorporating role-based instructions, directing the model to adopt specific personas, such as a teacher explaining complex concepts to students of varying ages (role prompting). We tried small models as baselines and larger models to increase performance. For instance, Llama-3.1-8B and its quantized variant support a combined input/output token limit of 8,192 tokens. Accordingly, we constrained model responses to 500 tokens and truncated input articles when necessary. Most experiments involving small models were conducted using zero-shot prompting.

Subtask 1.1

- **Zero-Shot Prompting on Initial and Final Article Segments**: To maximize the use of available tokens, we used only the beginning and end of each article. This approach was applied with small models only.
- **Zero-Shot Prompting on Summaries**: We divided long texts into chunks, summarized each chunk individually, and then combined them into a final summary.
- **Zero/One-shot Prompting on Section-Based Inputs**: Articles often contained diverse section structures. We extracted combinations such as: abstract only, abstract + introduction, abstract + discussion + conclusion (when available), or all four sections.
- **One-Shot Prompting with Random Sample**: Due to token constraints, we used a random example per prompt.
- **One-Shot Prompting with Most Similar Example**: We used cosine similarity (via Llama-3.1-8B embeddings) to find the most similar article-summary pair from the validation dataset. For long articles, we split them into chunks, computed embeddings, and averaged them. The most similar validation example was then included in the prompt.
- **Few-Shot Prompting with Five Examples (Lay Summaries Only)**: We selected five lay sum-

maries based on the most similar examples from the validation set in this few-shot prompting.

Subtask 1.2

This subtask aimed to improve upon Task 1.1 by incorporating external knowledge. It was based on the Retrieval-Augmented Generation (RAG) framework and focused on handling technical terms. Our process for this task included the following steps.

- **Extraction of Clinical Terms**: We used structured zero-shot prompting to extract technical terms from test articles.
- **Definition of Clinical Terms**: Each extracted term was queried using the Unified Medical Language System (UMLS) API. When available, the most suitable definition was selected.
- **Prompt Augmentation**: The resulting term-definition pairs were formatted and incorporated into the zero-shot, one-shot, and few-shot prompts used in Subtask 1.1. These refined prompts were then applied with larger models.

TASK 2

As in Task 1, we used prompt engineering to convert radiology reports (intended for healthcare professionals) into layperson-accessible summaries.

Subtask 2.1: Closed Track

We used the MIMIC-CXR dataset along with three public datasets: PadChest, Open-i, and BIMCV-COVID19. We used structured zero-shot and few-shot prompting approaches, incorporating examples selected either at random or based on cosine similarity of embeddings. Prompts explicitly defined the terms "radiology report" and "layman report" and included clear guidelines, as described as follows.

- **Zero-Shot Prompting on Radiology Report**: Our baseline used a structured prompt without examples.
- **Few-Shot Prompting with Five Random Examples**: We added five example pairs of radiology reports and lay summaries, one from each dataset, plus a fifth example illustrating variations of reports containing the phrase "No significant findings". This improved factuality and relevance, but caused a slight drop in readability.
- **Few-Shot Prompting with Ten Most Similar Examples**: For each test case, we used cosine similarity on BERT-large uncased embeddings to select the ten most similar examples from the validation set (approximately 20k samples).
- **Few-Shot Prompting with Twenty Most Similar Examples**: We extended the above method to include the top 20 most similar examples. Like the ten-example approach, this relied on the validation dataset to reduce computational costs while maintaining strong performance.

Subtask 2.2

We adopted an image-text-to-text model, BLIP (Bootstrapping Language-Image Pretraining) (Li et al., 2022), that combines a Vision Transformer with a Transformer-based text decoder to generate text from images and optional textual prompts. While less advanced than newer models like BLIP-2 (Li et al., 2023) and LLaVA (Zhang et al., 2025), it offers an efficient solution for descriptive image captioning. For the experiments, we used images and corresponding radiology reports, lay summaries, and metadata from OpenI, PadChest, and BIMCV-COVID19. The Hugging Face test set included 10,537 records, though actual image counts varied (e.g., OpenI often includes two images per record), and some images were missing. After aligning the metadata with the available images, the final dataset comprised 9,865 entries. Therefore, we were unable to submit official results due to mismatches between the number of processed records and the expected count.

4 Results

We report the results of our official submissions on the test data, as evaluated by the official evaluation server. The results for Subtask 1.1, Subtask 1.2, and Subtask 2.1 are presented in Tables 1, 2, and 3, respectively.

5 Discussion

We officially submitted approaches for three Subtasks 1.1, 1.2, and 2.1. Our approaches focused on the generalization of a single method (using the same model) across different datasets. For Task 1, we used a single approach for both eLife and PLOS. Similarly, for Subtask 2.1, we adopted a unified model for MIMIC, COVID, PadChest, and OpenI. We also conducted experiments for Subtask 2.2; however, due to issues related to dataset download and size, we were unable to submit our results for evaluation. Our experiments provided several key insights regarding the performance of

		Llama-3.1 (8-bit quantized)			Llama-3.3-70B-Inst.		GPT-4.1	
Description	**Metric**	Baseline	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Relevance	ROUGE ↑	0.2701	0.2283	0.2429	**0.3349**	0.3334	0.3080	0.3056
	BLEU ↑	4.1857	2.6787	3.0217	6.0490	**6.1354**	4.2153	4.1381
	METEOR ↑	**0.2791**	0.2459	0.2575	0.2703	0.2676	0.2632	0.2584
	BERTScore ↑	0.8358	0.8239	0.8282	0.8581	**0.8586**	0.8533	0.8534
Readability	FKGL ↓	12.2884	**8.3792**	9.3130	16.6736	16.0718	15.5356	15.5398
	DCRS ↓	7.2444	**6.1730**	6.5255	10.5558	10.3976	10.3787	10.4061
	CLI ↓	11.8690	**8.4256**	9.2667	15.8282	15.3358	14.1439	14.1545
	LENS ↑	65.5266	70.7002	71.8203	74.2810	76.0519	77.2428	**77.5635**
Factuality	AlignScore ↑	0.6061	0.4526	0.4893	**0.6366**	0.6307	0.4483	0.4506
	SummaC ↑	0.5348	**0.6141**	0.6114	0.4456	0.4550	0.4202	0.4186

Table 1: Performance of the 5cNLP team for Subtask 1.1. The baseline was scored from the results of zero-shot prompting on Llama-3.1-8B-Instruct (8-bit quantized). Submissions S1 and S2 correspond to the scores of additional role-prompting experiments performed on the same Llama-3.1 model. Submissions S3 and S4 correspond to the scores of one and few shot prompting respectively on Llama-3.3-70B-Instruct. Similarly, submissions S5 and S6 correspond to the scores of one and few shot prompting on GPT-4.1.

		Llama-3.3-70B-Inst.		GPT-4.1	
Description	**Metric**	**S1**	**S2**	**S3**	**S4**
Relevance	ROUGE ↑	**0.3364**	0.3350	0.3117	0.3089
	BLEU ↑	**5.9982**	5.9029	4.2778	4.1743
	METEOR ↑	**0.2764**	0.2747	0.2733	0.2659
	BERTScore ↑	**0.8576**	**0.8576**	0.8531	0.8533
Readability	FKGL ↓	16.8155	16.2979	15.7437	**15.5391**
	DCRS ↓	10.5226	**10.2896**	10.3104	10.3314
	CLI ↓	15.7708	15.2399	**14.1524**	14.2205
	LENS ↑	73.8590	75.5722	76.9570	**77.4515**
Factuality	AlignScore ↑	**0.6258**	0.6099	0.4431	0.4461
	SummaC ↑	**0.4468**	0.4455	0.4185	0.4154

Table 2: Performance of the 5cNLP team for Subtask 1.2. Submissions S1 and S2 correspond to the scores of RAG one and few shot prompting respectively on Llama-3.3-70B-Instruct. Similarly, submissions S3 and S4 correspond to the scores of RAG one and few shot prompting on GPT-4.1.

		Llama-3.3-70B-Inst.		GPT-4.1		
Description	**Metric**	**S1**	**S2**	**S3**	**S4**	**S5**
Relevance	ROUGE ↑	0.4424	0.5078	0.4679	0.5170	**0.5547**
	BLEU ↑	16.3978	23.4148	19.7649	25.0122	**28.2705**
	METEOR ↑	0.5051	0.5630	0.5169	0.5654	**0.6095**
	BERTScore ↑	0.9196	0.9317	0.9257	0.9332	**0.9371**
Readability	FKGL ↓	12.3058	9.8568	8.8586	8.5402	**8.0463**
	DCRS ↓	10.0489	9.6991	9.2135	**9.1778**	9.2373
	CLI ↓	10.1783	9.1757	8.2113	**8.1571**	8.2250
Factuality	Similarity ↑	0.8309	0.8561	0.8401	0.8591	**0.8717**
	RadGraph ↑	0.2452	0.2759	0.2566	0.2872	**0.3170**
	F1CheXbert ↑	0.7172	0.7348	0.6971	0.7162	**0.7495**

Table 3: Performance of the 5cNLP team for Subtask 2.1 across 5 submissions. **S1:** Structured zero-shot prompt with Llama-3.3-70B-Instruction model. **S2:** Structured few-shot prompt with random examples with Llama-3.3-70B Instruct model. **S3:** Structured zero-shot prompt with GPT-4.1 model. **S4:** Structured few-shot prompt with random examples with GPT-4.1 model. **S5:** Structured few-shot prompt with similarity-based examples with GPT-4.1 model.

LLMs for lay summarization of research articles and radiology reports.

Task 1 was particularly challenging due to the length of the research articles. Models often can only attend to portions of the input, potentially missing critical information—especially in longer documents. Moreover, using RAG with external sources introduces additional complexities. RAG requires a supplementary step: clinical term identification. In our approach, we extracted clinical terms through prompting, which were then used to query UMLS. We believe that explicitly incorporating a dedicated clinical entity recognition or term extraction step could significantly enhance the quality of the generated summaries.

Prompt Structure and Role Specification

Task 1: When compared to the baseline, role specification in Task 1 prompts produced responses with higher readability but lower relevance. Prompts that specified roles such as *"You will act as a teacher"* significantly improved the simplicity of the responses' language; however, the style of writing did not align with the gold standard and resulted in lower relevance scores.

Subtask 2.1: Naive, unstructured prompts, such as *"The following is a radiology report containing medical terms: <radiology-report>. I would like a brief summary of the radiology report that anyone without medical knowledge can understand, i.e., a layman report"*, performed significantly worse than structured prompts incorporating explicit role specification and output guidelines. For instance, prompts beginning with: *"You are an expert medical communicator. Your task is to..."*, consistently produced higher-quality layperson summaries, emphasizing the importance of structured role-focused instructions.

Model Scale and Performance

Task 1 and Subtask 2.1: Across both structured and unstructured prompts, larger parameter models within the same architecture demonstrated superior performance. For example, the Llama-3.3-70B-Instruct model outperformed its smaller counterpart, Llama-3.1-8B. For Subtask 1.1, this also demonstrates the larger models' ability to consider a greater amount of information and instruction. They are not constrained by the token limit as was the case with Llama-3.1-8B. For Subtask 2.1, a similar trend was observed with GPT-4.1o compared to GPT-4.1, underscoring the impact of model scale, as well as context length on translation accuracy.

In-Context Learning

Task 1 and Subtask 2.1: The inclusion of contextual examples within prompts further improved model performance. Few-shot prompting, particularly with dynamically selected examples based on cosine similarity from the training sample embedding space, yielded the best results. This approach ensured that the model received relevant, semantically aligned demonstrations for the given input.

Retrieval Augmented Generation

Subtask 1.2: When analyzing the impact of incorporating external knowledge, we should compare Subtask 1.1 prompts against their Subtask 1.2 counterparts (i.e., Subtask 1.1. S3 against Subtask 1.2 S1, S4 against S2, etc.). Overall, using our methods, we observed no significant performance impact when including external knowledge. This outcome can be attributed to several factors. First, the definitions included in the prompt may have been insufficient or irrelevant. Second, the provided definitions may not have added any new information beyond what the LLMs already contained.

6 Conclusion and Future Work

We proposed a framework for translating medical texts into layperson's language focusing on summarizing biomedical articles and translating radiology reports. Using state-of-the-art LLMs (e.g., Llama-3.3-70B-Instruct and GPT-4.1), our 5cNLP team ranked third in Subtask 1.2 and second in Subtask 2.1. Rankings were based on normalized averages across all evaluation metrics. Our experiments highlighted the importance of structured, role-specific prompting, model scale, and contextual example selection in optimizing LLM performance. Moreover, while LoRA fine-tuning was applied to smaller models, prompt engineering yielded better results.

Future work may include full model training, improved prompt design, and the integration of additional external knowledge sources. For Subtask 2.2, alternative strategies for multimodal fusion could be explored. The proposed framework is also adaptable to other biomedical applications, such as patient question answering, clinical decision support, and summarizing electronic health records for non-expert audiences.

Limitations

Our experiments are limited to English-language radiology reports. Experiments for other languages could reveal more challenges in generating lay summaries. We also had limited time and computational resources; therefore, our conclusions are valid only for a small number of LLMs.

Ethics

The datasets provided by the shared task organizers were carefully prepared to ensure proper use of the data, without information about the patients. We used the datasets solely for research purposes, as expected.

Acknowledgments

Research reported in this publication was supported in part by the Intramural Research Program of the National Institute of Mental Health: ZIC-MH002968 (Francisco Pereira and Juan Antonio Lossio-Ventura) and by the Natural Science and Engineering Research Council of Canada (Diana Inkpen).

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:1134–1142.

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self states from social media posts. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *Preprint*, arXiv:2006.01174.

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.

A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. 2024. MIMIC-CXR Database (version 2.1.0).

A.E.W. Johnson, T.J. Pollard, and S.J. et al. Berkowitz. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data 6, 317*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

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.

Chenghao Xiao, Kun Zhao, Xiao Wang, and Siwei Wu. 2025a. BioLaySumm Shared Task at ACL.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025b. 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.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Kun Zhao, Chenghao Xiao, Chen Tang, Bohao Yang, Kai Ye, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2024. X-ray made simple: Radiology report generation and evaluation with layman's terms. *arXiv preprint arXiv:2406.17911*.

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.1 Task 1 and 2 datasets' splits

Dataset	Training	Validation	Test
PLOS	24,773	1,376	142
eLife	4,346	241	142

Table 4: Training, validation, and test splits for the PLOS and eLife datasets for Task 1.

Dataset	Training	Validation	Test
PadChest	116,847	7,824	7,130
MIMIC-CXR	45,000	5,000	500
BIMCV-COVID19	31,364	2,042	3221
Open-i	2,243	134	186

Table 5: Training, validation, and test splits for the evaluated datasets for Task 2.

A.2 Task 1 Prompt Templates

```
### Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience and should not contain highly technical terms.
—
### Guidelines for Output:
- The summary should be in layman terms and will not include any technical terms.
- The summary should avoid using acronyms.
- Limit the output summary to 300 words.
- The output should only contain the summary and will not reference the article itself.
- Do not provide sources.
- Do not include any disclaimers.
- Do not include any information that is not relevant to the summarization.
- Do not repeat the guidelines given by the prompt.
—
### Input:
{article}

### Output:
```

Table 6: Zero-Shot Structured Prompt Template for Task 1.1, Baseline.

A.3 Task 2 Prompt Templates

A.4 Task 2.1 Experiments

We conducted additional experiments comparing structured and unstructured prompts using both zero-shot and few-shot approaches with randomly selected examples. Table 16 summarizes these results, which were generated using a subset of the validation data. Due to a technical error, we couldn't compute F1CheXbert scores for experiments E1 and E2.

Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience and should not contain highly technical terms.
—
Role:
You will act as a middle school teacher who is explaining the article to a group of grade 7 students who are 12 years old and who require simple language to understand your summarization.
—
Guidelines for Output:
- The summary should be in layman terms and will not include any technical terms.
- The summary should avoid using acronyms.
- Limit the output summary to 300 words.
- The output should only contain the summary and will not reference the article itself.
- Do not provide sources.
- Do not include any disclaimers.
- Do not include any information that is not relevant to the summarization.
- Do not repeat the guidelines given by the prompt.
—
Input:
{article}

Output:

Table 7: Zero-Shot Structured Role Prompt Template for Task 1.1, S1.

Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience and should not contain highly technical terms.
—
Role:
You will act as a secondary school teacher who is explaining the article to a group of grade 9 students who are 15 years old and who require simple language to understand your summarization.
—
Guidelines for Output:
- The summary should be in layman terms and will not include any technical terms.
- The summary should avoid using acronyms.
- Limit the output summary to 300 words.
- The output should only contain the summary and will not reference the article itself.
- Do not provide sources.
- Do not include any disclaimers.
- Do not include any information that is not relevant to the summarization.
- Do not repeat the guidelines given by the prompt.
—
Input:
{article}

Output:

Table 8: Zero-Shot Structured Role Prompt Template for Task 1.1, S2.

Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience, avoiding technical jargon unless briefly defined.
—
Instructions:
- Use the example below as a guide, matching its structure and writing style in your summary.
- The summary should be in layman terms.
- Briefly define any technical terms that must be included.
- Do not reference the original article or include disclaimers.
- Exclude any information not relevant to the summary.
- Do not provide sources
- Do not repeat the guidelines given by the prompt
- Avoid repeating information unnecessarily
—
Guidelines for Output:
- Format: Clear, flowing prose in one paragraph.
- Content: Capture the essential meaning and logic of the article as if the summary itself were a brief version of the full text.
- Audience: General readers without specialized knowledge of the topic.
—
Example:
Article:
{example_article}

Summary:
{example_summary}
—
Now, summarize the following article based on the given criteria and using the same style of the example:

Article:
{article}

Summary:

Table 9: One-Shot Structured Prompt Template for Task 1.1, S3 and S5.

Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience, avoiding technical jargon unless briefly defined.
—
Instructions:
- Use the example below as a guide, matching its structure and writing style in your summary.
- The summary should be in layman terms.
- Briefly define any technical terms that must be included.
- Do not reference the original article or include disclaimers.
- Exclude any information not relevant to the summary.
- Do not provide sources
- Do not repeat the guidelines given by the prompt
- Avoid repeating information unnecessarily
—
Guidelines for Output:
- Format: Clear, flowing prose in one paragraph.
- Content: Capture the essential meaning and logic of the article as if the summary itself were a brief version of the full text.
- Audience: General readers without specialized knowledge of the topic.
—
Examples:
-> Example 1
Summary:
{example_summary_1}

-> Example 2
Summary:
{example_summary_2}

-> Example 3
Summary:
{example_summary_3}

-> Example 4
Summary:
{example_summary_4}

-> Example 5
Summary:
{example_summary_5}
—
Now, summarize the following article based on the given criteria and using the same style of the example:

Article:
{article}

Summary:

Table 10: Few-Shot Structured Prompt Template for Task 1.1, S4 and S6.

Task:
Your task is to perform layman summarization of the following biomedical article by succintly summarizing the article in a way that is easy to understand for a general audience, avoiding technical jargon unless briefly defined.
—
Definitions:
Use the following definitions to better understand and summarize the article.
{definitions}
—
Instructions:
- Use the example below as a guide, matching its structure and writing style in your summary.
- The summary should be in layman terms.
- Briefly define any technical terms that must be included.
- Do not reference the original article or include disclaimers.
- Exclude any information not relevant to the summary.
- Do not provide sources
- Do not repeat the guidelines given by the prompt
- Avoid repeating information unnecessarily
—
Guidelines for Output:
- Format: Clear, flowing prose in one paragraph.
- Content: Capture the essential meaning and logic of the article as if the summary itself were a brief version of the full text.
- Audience: General readers without specialized knowledge of the topic.
—
Example:
Article:
{example_article}

Summary:
{example_summary}
—
Now, summarize the following article based on the given criteria and using the same style of the example:

Article:
{article}

Summary:

Table 11: One-Shot Structured Prompt Template for Task 1.2, S1 and S3.

Task:
Your task is to perform layman summarization of the following biomedical article by succinctly summarizing the article in a way that is easy to understand for a general audience, avoiding technical jargon unless briefly defined.

—
Definitions:
Use the following definitions to better understand and summarize the article.
{definitions}
—
Instructions:
- Use the example below as a guide, matching its structure and writing style in your summary.
- The summary should be in layman terms.
- Briefly define any technical terms that must be included.
- Do not reference the original article or include disclaimers.
- Exclude any information not relevant to the summary.
- Do not provide sources
- Do not repeat the guidelines given by the prompt
- Avoid repeating information unnecessarily
—
Guidelines for Output:
- Format: Clear, flowing prose in one paragraph.
- Content: Capture the essential meaning and logic of the article as if the summary itself were a brief version of the full text.
- Audience: General readers without specialized knowledge of the topic.
—
Examples:
-> Example 1
Summary:
{example_summary_1}

-> Example 2
Summary:
{example_summary_2}

-> Example 3
Summary:
{example_summary_3}

-> Example 4
Summary:
{example_summary_4}

-> Example 5
Summary:
{example_summary_5}
—
Now, summarize the following article based on the given criteria and using the same style of the example:

Article:
{article}

Summary:

Table 12: Few-Shot Structured Prompt Template for Task 1.2, S2 and S4.

Task:
You are an expert medical communicator. Your role is to translate radiology reports, originally written for healthcare professionals, into language that an average person without a medical background can understand. The rewritten report should preserve all essential medical findings and implications suitable for the general public. Note that you must avoid redundancy.

—

Definitions:
- Radiology Report: A medical document that describes findings from imaging studies such as X-rays, CT scans, or MRIs.
- Layman Report: A simplified, non-technical explanation suitable for someone with no formal medical education.

—

Guidelines:
- Generated number of tokens: Try to match the number of tokens of the original Radiology Report, adjusting as needed based on report complexity.
- Avoid Speculation: Do not add interpretations beyond what is stated in the original report.
- Maintain a Reassuring and Neutral Tone: Use clear, calm, and factual language.
- Structure: Present the information in a single, coherent paragraph.
- The single paragraph can be composed of one or several sentences.
- Adhere to Reported Diagnoses: Only summarize what is already reported; do not include diagnoses not explicitly stated. - Avoid redundancy.

—

Guidelines for Output:
- Format: Clear and concise prose.
- Redundancy: Avoid repeating information unnecessarily.
- Length: Should closely match the number of tokens in the original Radiology Report, adjusting as needed based on report complexity.
- Audience: A general reader with no medical background or clinical training.

—

Analyze the Following Radiology Report Based on the Given Criteria:

Radiology Report:
{radiology_report}

Response (Layman Report):

Table 13: Zero-Shot Structured Prompt Template for Task 2.1, S1 and S3

Task:
You are an expert medical communicator. Your role is to translate radiology reports, originally written for healthcare professionals, into plain language that an average person without a medical background can understand. The rewritten report should preserve all essential medical findings and implications suitable for the general public. Note that you must avoid redundancy.

—
Definitions:
- Radiology Report: A medical document that describes findings from imaging studies such as X-rays, CT scans, or MRIs.
- Layman Report: A simplified, non-technical explanation suitable for someone with no formal medical education.

—
Guidelines:
- Generated number of tokens: Try to match the number of tokens of the original Radiology Report, adjusting as needed based on report complexity.
- Avoid Speculation: Do not add interpretations beyond what is stated in the original report.
- Maintain a Reassuring and Neutral Tone: Use clear, calm, and factual language.
- Structure: Present the information in a single, coherent paragraph.
- The single paragraph can be composed of one or several sentences.
- Adhere to Reported Diagnoses: Only summarize what is already reported; do not include diagnoses not explicitly stated.
- Avoid redundancy.

—
Guidelines for Output:
- Format: Clear and concise prose.
- Redundancy: Avoid repeating information unnecessarily.
- Length: Should closely match the number of tokens in the original Radiology Report, adjusting as needed based on report complexity.
- Audience: A general reader with no medical background or clinical training.

—
Examples:

-> Example 1
Radiology Report:
{example_radiology_report_1}
Response (Layman Report):
{example_layman_report_1}

-> Example 2
Radiology Report:
{example_radiology_report_2}
Response (Layman Report):
{example_layman_report_2}

-> Example 3
Radiology Report:
{example_radiology_report_3}
Response (Layman Report):
{example_layman_report_3}

-> Example 4
Radiology Report:
{example_radiology_report_4}
Response (Layman Report):
{example_layman_report_4}

-> Example 5
Radiology Report:
{example_radiology_report_5}
Response (Layman Report):
{example_layman_report_5}

—
Analyze the Following Radiology Report Based on the Given Criteria:

Radiology Report:
{radiology_report}

Response (Layman Report):

Table 14: Few-Shot Structured Prompt Template with random examples for Task 2.1, S2 and S4.

Task:
You are an expert medical communicator. Your task is to translate radiology reports, originally written for healthcare professionals, into plain language that an average person without a medical background can understand. The rewritten report, referred to as Layman Report, should preserve all essential medical findings and implications suitable for the general public.

—
Definitions:
- Radiology Report: A medical document that describes findings from imaging studies such as X-rays, CT scans, or MRIs.
- Layman Report: A simplified, non-technical explanation suitable for someone with no formal medical education.

—
Instructions:
- Use the examples below as a guide, matching their structure and writing style in your layman report.
- The rewritten report should be in layman terms.
- Briefly define any technical terms that must be included.
- Try to match the number of tokens of the original Radiology Report, adjusting as needed based on report complexity.
- Maintain a Reassuring and Neutral Tone: Use clear and factual language.
- Structure: Present the information in a single, coherent paragraph.
- The single paragraph can be composed of one or several sentences.
- Adhere to Reported Diagnoses: Only rewrite what is already reported; do not include diagnoses not explicitly stated.

—
Guidelines for Output:
- Format: Clear, flowing prose in one paragraph.
- Content: Capture the essential meaning and logic of the radiology report.
- Length: Should closely match the number of tokens in the original Radiology Report, adjusting as needed based on report complexity.
- Audience: A general reader with no medical background or clinical training.

—
Examples:

-> Example 1
Radiology Report:
{similar_example_radiology_report_1}
Response (Layman Report):
{similar_example_layman_report_1}

-> Example 2
Radiology Report:
{similar_example_radiology_report_2}
Response (Layman Report):
{similar_example_layman_report_2}

...
...
...

-> Example 10
Radiology Report:
{similar_example_radiology_report_10}
Response (Layman Report):
{similar_example_layman_report_10}

—
Now, rewrite the following radiology report based on the given criteria and using the same style of the examples:

Radiology Report:
{radiology_report}

Layman Report:

Table 15: Few-Shot Structured Prompt Template with Cosine Similarity-Based Examples for Task 2.1, S5.

		Llama-3.3-70B-Instruct				GPT-4.1o	
Description	**Metric**	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
Relevance	ROUGE ↑	0.3070	0.3593	0.3899	**0.4806**	0.3590	0.4434
	BLEU ↑	8.2984	11.7237	14.8005	**20.3253**	12.8078	17.5307
	METEOR ↑	0.4132	0.4583	0.4645	**0.5352**	0.3949	0.4723
	BERTScore ↑	0.8875	0.8886	0.9141	0.9170	0.9136	**0.9238**
Readability	FKGL ↓	9.3505	10.3292	9.7895	**7.9905**	10.6400	9.5130
	DCRS ↓	**8.9585**	10.3663	9.4053	9.3723	10.3322	10.1358
	CLI ↓	**8.4965**	10.1607	9.1438	9.0414	9.9672	9.8145
Factuality	Similarity ↑	0.7045	0.7320	0.7888	0.7964	0.7697	**0.8112**
	RadGraph ↑	0.1512	0.1722	0.2167	0.1938	0.2109	**0.2252**
	F1CheXbert ↑	-	-	0.7200	0.7100	**0.7700**	0.7450

Table 16: Task 2.1 experiments run on Llama-3.3-70B-Instruct and GPT 4.1o models. **E1:** Unstructured zero-shot prompt. **E2:** Unstructured few-shot prompt with 5 random examples. **E3:** Structured zero-shot prompt. **E4:** Structured few-shot prompt with 5 random examples. **E5:** Structured zero-shot prompt. **E6:** Structured few-shot prompt with 5 random examples.