Simplified Rewriting Improves Expert Summarization

Xingmeng Zhao[1], **Tongnian Wang**[2], and **Anthony Rios**[2]
[1]University of Colorado Anschutz Medical Campus
[2]University of Texas at San Antonio
xingmeng.zhao@cuanschutz.edu, anthony.rios@utsa.edu

Abstract

Radiology report summarization (RRS) is critical for clinical workflows, requiring concise "Impressions" distilled from detailed "Findings." This paper proposes a novel prompting strategy that enhances RRS by introducing a layperson summary as an intermediate step. This summary helps normalize key observations and simplify complex terminology using communication techniques inspired by doctor–patient interactions. Combined with few-shot in-context learning, this approach improves the model's ability to map generalized descriptions to specific clinical findings. We evaluate our method on three benchmark datasets, MIMIC-CXR, CheXpert, and MIMIC-III, and compare it against state-of-the-art open-source language models in the 7B/8B parameter range, such as Llama-3.1-8B-Instruct. Results show consistent improvements in summarization quality, with gains of up to 5% on some metrics for prompting, and more than 20% for some models when instruction tuning.

1 Introduction

Radiology report summarization (RRS) is a compelling task for exploring natural language processing (NLP) methods in the biomedical domain from a computational perspective (Van Veen et al., 2023a). RRS involves generating concise "Impressions" from the detailed "Findings" in radiology reports. These reports are critical for diagnosis, treatment planning, and longitudinal health records, and are authored by radiologists based on imaging modalities such as X-rays, CT scans, MRI scans, and ultrasounds. The "Findings" section captures objective observations from the images, while the "Impression" section offers the radiologist's clinical interpretation and diagnostic conclusions.

In biomedical applications, the effectiveness of large language models (LLMs) often depends on domain- and task-specific adaptation through fine-tuning (Singhal et al., 2023). While LLMs demonstrate strong capabilities in natural language understanding and generation, fine-tuning models with billions of parameters, such as GPT-3, is computationally expensive and resource-intensive. To address these challenges, recent work has focused on more efficient alternatives, including parameter-efficient fine-tuning (PEFT) and prompting (Van Veen et al., 2023a,b; Liu et al., 2022), which aim to use existing model knowledge while significantly reducing computational overhead.

In contrast, prompting through in-context learning (ICL) (Brown et al., 2020; Dong et al., 2023) offers a practical and lightweight alternative to full fine-tuning for adapting LLMs to new tasks. ICL enables models to perform few-shot learning by embedding relevant information and examples directly within the prompt (Lampinen et al., 2022). With well-designed prompts, LLMs can be guided to produce accurate outputs through contextual cues and task-specific demonstrations. This process can be further improved using techniques such as Retrieval-Augmented Generation (RAG) (Wang et al., 2023), which dynamically selects relevant examples from external corpora. In biomedical applications, prompting has shown effectiveness in tasks like radiology report summarization by simplifying complex findings into concise summaries (Chen et al., 2023a). Adding explanations to ICL prompts can further improve performance on specialized tasks such as medical question answering (Nori et al., 2023; Zhang et al., 2023). However, generating explanations for summarization is more difficult due to the open-ended nature of the output and the challenge of aligning intermediate reasoning with summary objectives.

Moreover, LLMs trained on general text corpora often lack the specific knowledge required for specialized fields, limiting their performance (Yao et al., 2023a; Holmes et al., 2023). Addressing this deficiency typically involves extensive fine-tuning, which is resource-intensive and costly. While ICL

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3075–3097
December 20-24, 2025 ©2025 Association for Computational Linguistics

can help by embedding relevant information within prompts, this alone is not always sufficient (Brown et al., 2020; Dong et al., 2023). Intuitively, non-fine-tuned models are "non-experts" in the medical domain, especially smaller open-source models.

In real-world settings such as doctor–patient conversations, prior research shows that technical or scientific knowledge can be effectively communicated to non-experts through strategies like reformulation and simplification (Gülich, 2003). These techniques break down complex information into clearer, more accessible language to enhance understanding. Motivated by these communication principles, we propose a novel prompting strategy that integrates simplification with ICL to improve the performance of non-expert LLMs in specialized domains. Our method avoids costly fine-tuning (Nori et al., 2023; Zhang et al., 2023) by generating a layperson summary before the expert summary. These lay summaries are included in the in-context examples to guide the model's reasoning on new inputs. From another perspective, we treat the layperson summary as a normalization step that abstracts key clinical observations into consistent, general expressions. This is particularly useful in radiology, where variations in reporting style and terminology (Yan et al., 2023), along with a wide range of illnesses, result in high lexical variability. By mapping terms like "pneumonia" and "bronchitis" to a shared concept such as "infection of the lungs," the model can more easily recognize semantically similar cases across in-context examples, even when the terminology differs. This normalization helps the model align general descriptions in the lay summary with specific details in the Findings, improving consistency and robustness in expert summary generation (Peter et al., 2024).

Overall, this paper has threefold contributions:

1. We introduce a novel prompting approach inspired by doctor–patient communication, where a simplified (layperson) summary is generated before the expert summary. This strategy, combined with few-shot ICL, enhances RRS using general-purpose, non-expert LLMs.

2. We evaluate LLM performance on three RRS datasets: MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019), and MIMIC-III (Johnson et al., 2016), and Medical Question Summarization (MMQS) dataset (Ghosh et al., 2024). We benchmark against open-source LLMs, like Llama-3.1-8B-Instruct (AI@Meta, 2024) for comprehensive comparison.

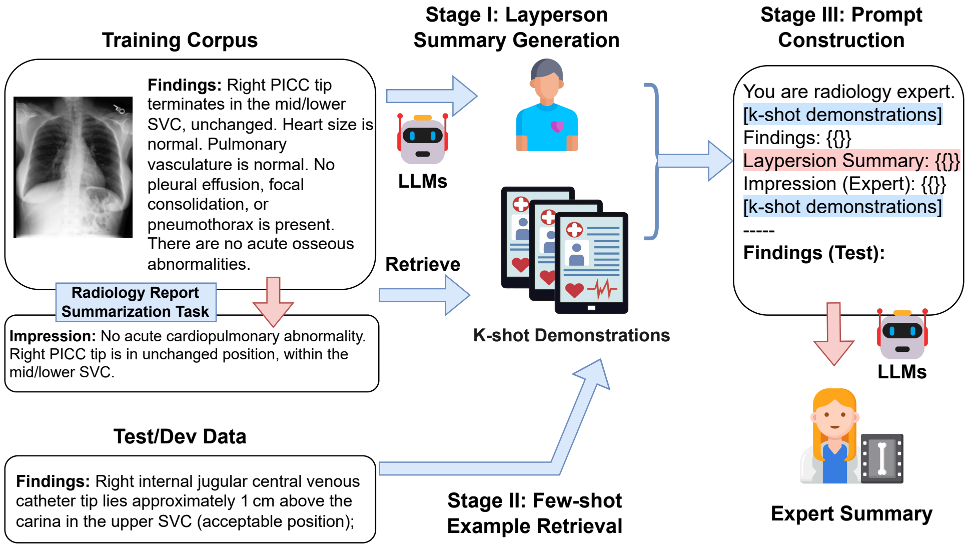
3. We also investigate whether using the layperson prompt as an instruction is effective for instruction tuning, evaluating its impact on expert impression generation.[1]

2 Related Work

LLMs for Medicine. Recent advances in LLMs have demonstrated that LLMs can be adapted with minimal effort across various domains and tasks. These expressive and interactive models hold great promise due to their ability to learn broadly useful representations from the extensive knowledge encoded in medical corpora at scale (Singhal et al., 2023). Fine-tuned general-purpose models have proven effective in clinical question-answering, protected health information de-identification (Sarkar et al., 2024), and relation extraction (Hernandez et al., 2023). Some LLMs, such as BioGPT (Luo et al., 2022) and ClinicalT5 (Lu et al., 2022), have been trained from scratch using clinical domain-specific notes, achieving promising performance on several tasks. Additionally, in-context learning with general LLMs like InstructGPT-3 (Ouyang et al., 2022), where no weights are modified, has shown good performance (Agrawal et al., 2022). They have also demonstrated the ability to solve domain-specific tasks through zero-shot or few-shot prompting and have been applied to various medical tasks, such as medical report summarization (Otmakhova et al., 2022) and medical named entity recognition (Hu et al., 2023). But, this generally only works with closed-source models such as GPT4.

Retrieval-Augmented LLMs. Retrieval augmentation connects LLMs to external knowledge to mitigate factual inaccuracies. By incorporating a retrieval module, relevant passages are provided as context, enhancing the language model's predictions with factual information like common sense or real-time news (Ma et al., 2023). Recent studies indicate that retrieval-augmented methods can enhance the reasoning ability of LLMs and make their responses more credible and traceable (Shi et al., 2024; Yao et al., 2023b; Nori et al., 2023; Ma et al., 2023). For example, Shi et al. (2024) trains a dense retrieval model to complement a frozen

[1] See the appendix for the full analysis.


Training Corpus
Findings: Right PICC tip terminates in the mid/lower SVC, unchanged. Heart size is normal. Pulmonary vasculature is normal. No pleural effusion, focal consolidation, or pneumothorax is present. There are no acute osseous abnormalities.
Radiology Report Summarization Task
Impression: No acute cardiopulmonary abnormality. Right PICC tip is in unchanged position, within the mid/lower SVC.
Stage I: Layperson Summary Generation
LLMs
Retrieve
K-shot Demonstrations
Stage III: Prompt Construction
You are radiology expert.
[k-shot demonstrations]
Findings: {{}}
Laypersion Summary: {{}}
Impression (Expert): {{}}
[k-shot demonstrations]

Findings (Test):
LLMs
Expert Summary
Test/Dev Data
Findings: Right internal jugular central venous catheter tip lies approximately 1 cm above the carina in the upper SVC (acceptable position);
Stage II: Few-shot Example Retrieval

Figure 1: Overview of the LaypersonPrompt Framework. First, we prompt LLMs to generate layperson summaries for each training example. Then, for a given test input, we use few-shot example retrieval to select relevant examples from the training set. Finally, we construct an instruction prompt by combining the retrieved examples with their layperson summaries to guide the model's reasoning for expert summary generation.

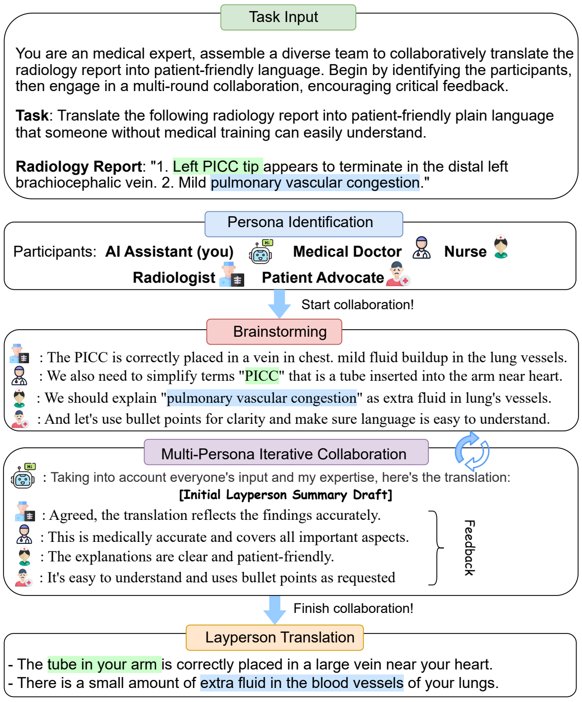
language model. By using feedback from the LLM as a training objective, the retrieval model is optimized to provide better contextual inputs for the LLM. Yao et al. (2023b) focuses on designing interactions between the retriever and the reader, aiming to trigger emergent abilities through carefully crafted prompts or a sophisticated prompt pipeline. Our approach combines retrieval-augmented methods with layperson summaries to enhance general LLMs reasoning in radiology report summarization, using patient-doctor communication techniques for better understanding and accuracy.

Communication Techniques for Laypersons. Non-experts, such as patients, have been shown to perform well on expert tasks, like medical decision-making and understanding complex topics when information is simplified using effective communication techniques (Gülich, 2003; LeBlanc et al., 2014; Allen et al., 2023; van Dulmen et al., 2007; Neiman, 2017). This simplification can also improve general LLM's performance on specialized tasks. Studies show that non-experts, with supervision, can generate high-quality data for machine learning, producing expert-quality annotations for tasks like identifying pathological patterns in CT lung scans and malware run-time similarity (O'Neil et al., 2017; VanHoudnos et al., 2017; Snow et al., 2008). Recent research has shown that LLMs can simplify complex medical documents, such as radiology reports, making them more accessible to laypersons. For instance, ChatGPT has been used to make radiology reports easier to understand, bridging the communication gap between medical professionals and patients (Jeblick et al., 2023; Lyu et al., 2023; Li et al., 2023). Ki and Carpuat (2025) also show that simplifying text is an effective way to improve machine translation quality. Inspired by these findings, we explore whether presenting expert-level information in a simpler language can improve the performance of general LLMs on tasks that typically require specialized knowledge, such as those involving medical data.

3 Methodology

In this section, we describe our prompting strategy. Figure 1 shows a high-level overview of our approach. Our strategy has three main components: 1) layperson summarization of the training dataset used as in-context examples; 2) "few-shot example retrieval," which is how we generate text embeddings to find relevant in-context examples; and 3) final expert summary prompt construction, which is how we integrate the layperson summaries and in-context examples to generate the final expert summary. We describe each component in the following subsections and how the three components are integrated into a unified prompt.

Stage I: Layperson Summary Generation. Based on Singhal et al. (2023), LLMs encode a large amount of medical knowledge during pretraining. Our layperson summarization step aims to encourage the model to actively use its inter-



You are an medical expert, assemble a diverse team to collaboratively translate the radiology report into patient-friendly language. Begin by identifying the participants, then engage in a multi-round collaboration, encouraging critical feedback.

Task: Translate the following radiology report into patient-friendly plain language that someone without medical training can easily understand.

Radiology Report: "1. Left PICC tip appears to terminate in the distal left brachiocephalic vein. 2. Mild pulmonary vascular congestion."

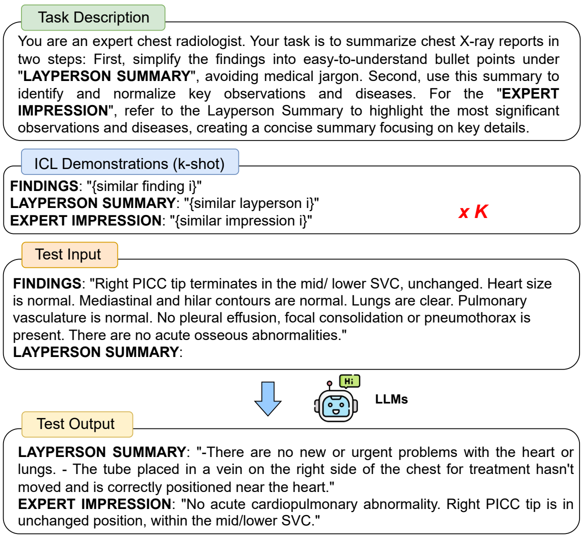
Persona Identification
Participants: **AI Assistant (you)** **Medical Doctor** **Nurse** **Radiologist** **Patient Advocate**
Start collaboration!
Brainstorming

: The PICC is correctly placed in a vein in chest. mild fluid buildup in the lung vessels.
: We also need to simplify terms "PICC" that is a tube inserted into the arm near heart.
: We should explain "pulmonary vascular congestion" as extra fluid in lung's vessels.
: And let's use bullet points for clarity and make sure language is easy to understand.

: Taking into account everyone's input and my expertise, here's the translation:
[Initial Layperson Summary Draft]
: Agreed, the translation reflects the findings accurately.
: This is medically accurate and covers all important aspects.
: The explanations are clear and patient-friendly.
: It's easy to understand and uses bullet points as requested

- The tube in your arm is correctly placed in a large vein near your heart.
- There is a small amount of extra fluid in the blood vessels of your lungs.

Figure 2: Stage I: Layperson Summarization Generation . An illustration of the layperson summary prompt used to generate layperson summaries for training examples. Disease observations are highlighted in different colors. The illustration shows a single example, with Instruction and Response sections repeated multiple times using few-shot in-context examples.



You are an expert chest radiologist. Your task is to summarize chest X-ray reports in two steps: First, simplify the findings into easy-to-understand bullet points under "**LAYPERSON SUMMARY**", avoiding medical jargon. Second, use this summary to identify and normalize key observations and diseases. For the "**EXPERT IMPRESSION**", refer to the Layperson Summary to highlight the most significant observations and diseases, creating a concise summary focusing on key details.

FINDINGS: "{similar finding i}"
LAYPERSON SUMMARY: "{similar layperson i}"
EXPERT IMPRESSION: "{similar impression i}"

FINDINGS: "Right PICC tip terminates in the mid/ lower SVC, unchanged. Heart size is normal. Mediastinal and hilar contours are normal. Lungs are clear. Pulmonary vasculature is normal. No pleural effusion, focal consolidation or pneumothorax is present. There are no acute osseous abnormalities."
LAYPERSON SUMMARY:

LAYPERSON SUMMARY: "-There are no new or urgent problems with the heart or lungs. - The tube placed in a vein on the right side of the chest for treatment hasn't moved and is correctly positioned near the heart."
EXPERT IMPRESSION: "No acute cardiopulmonary abnormality. Right PICC tip is in unchanged position, within the mid/lower SVC."

Figure 3: Stage III: Prompt Construction. Example of LaypersonPrompt for generating the final expert summary. This is the final prompt after finding in-context examples to generate the final expert summary (i.e., the Impression section).

nal knowledge to convert complex medical texts into more straightforward language, enhancing accessibility and understanding for individuals without medical expertise (Cao et al., 2020). For instance, rephrasing "pulmonary edema" as "fluid in the lungs" makes it more comprehensible. This approach not only helps to bridge the knowledge gap for laypeople but also plays an important role in helping models better understand and summarize medical content. Intuitively, by generating simplified summaries as an intermediate step, models can more effectively capture the semantic meaning of the texts (Liu et al., 2024; Sulem et al., 2018; Paetzold and Specia, 2016; Shardlow and Nawaz, 2019). In this context, we generate layperson summaries as an intermediate step for all training examples to enhance the generation of expert summaries.

To generate accurate layperson summaries, we employ a multi-round, multi-persona collaboration method inspired by the Task-Solving Agent framework (Wang et al., 2024). As shown in Figure 2, we begin with a radiology report impression and identify several expert roles, including a medical doctor, nurse, radiologist, patient advocate and AI assistant, to provide diverse insights. In the brainstorming phase, these experts clarify medical terminology and highlight key findings (e.g., "PICC," and "pulmonary vascular congestion"). Through iterative collaboration, they refine the content to ensure clarity and accuracy. Finally, the refined draft is transformed into a concise, accessible summary that effectively communicates essential medical details to patients. We then use this prompt to generate layperson summaries and store these summaries along with their corresponding Findings and Impressions as training triples, which are used as in-context examples. See Appendix **??** for a complete example of what the output looks like.

Stage II: Few-shot Example Retrieval Another key part of our system is retrieving similar examples from the training corpus to use as in-context examples. We focus on selecting a few high-quality examples to help the LLM generate more accurate and consistent summaries. To find the most relevant examples, we follow the text-only retrieval process of Wang et al. (2023). Specifically, we encode the *Findings* and *Impression* sections of each report using an encoder-decoder model, such as Clinical-T5 (Lehman and Johnson, 2023). Given an input report X, we compute its similarity to each training sample X_i using cosine similarity: $\text{Sim}(X, X_i) = \frac{E_X \cdot E_{X_i}}{\|E_X\| \|E_{X_i}\|}$, where E_X and E_{X_i} are the text embeddings. We then rank all training samples by similarity and retrieve the top-k most similar reports. Finally, we include the *Findings* and *Impression* sections from these reports as few-shot examples in the input prompt.

Stage III: Prompt Construction The final step in our pipeline involves prompting an LLM to generate an expert summary, following the generation of layperson summaries for all training examples, and identifying relevant in-context examples for development/test instances using few-shot example retrieval. The prompt comprises three main components: 1) Task Instruction; 2) In-context learning examples (ICL Demonstrations); and 3) the test input instance. An example is shown in Figure 3.

First, the Task Instruction directs the model to generate a layperson summary, followed by an expert impression. During data preparation, we use the prompt defined in Step 1 to generate layperson summaries for the training examples. This prompt is only used at that stage and is not applied at inference time. At inference time, given a new input instance with its Findings text, we use the Clincal T5 encoder and retrieval approach described in Step 2 to retrieve up to 32 similar examples from the training set. Each retrieved example includes: (1) the Findings, (2) the corresponding layperson summary (generated during data preparation), and (3) the expert Impression. These components are concatenated to form the in-context prompt. After the in-context examples, we append the input instance's Findings section, followed by the string "Layperson Summary:". The model is then prompted to generate the layperson summary for the input instance, and immediately afterward, the expert impression, both as part of a single prompt.

Enhanced Radiology Report Summarization. We incorporate both the layperson summary, which reflects the model's internal medical knowledge, and a set of retrieved few-shot examples, which provide external knowledge, to construct the prompt used at inference time. This prompt guides the language model in generating the expert impression for a given radiology report. We hypothesize that generating a layperson summary before the expert impression helps the model standardize the content of the Findings by first translating complex clinical language into general, simplified concepts. For example, conditions such as "pneumonia" and "bronchitis" may both be expressed as "infection of the lungs" in the lay summary. This abstraction reduces variation and enables the model to identify consistent patterns that link generalized expressions to expert-level impressions. Once the lay summary is generated, the model only needs to relate these general terms to the specific details in the Findings, similar to coreference resolution. This step encourages the model to infer underlying clinical meaning that is not always stated explicitly, effectively allowing it to "read between the lines." Without this intermediate layer of abstraction, the model must directly reason over more diverse and complex language in the Findings, making the task more challenging and less consistent.

4 Experimental Results

This section covers the datasets, evaluation metrics, overall results, and error analysis.

Datasets and baseline models. In this study, we evaluate our prompting method on three radiology reports summarization datasets. The MIMIC-III summarization dataset, as introduced by (Johnson et al., 2016; Chen et al., 2023b), contains 11 anatomy-modality pairs (i.e., 11 body parts and imaging modalities such as head-MRI and abdomen-CT). The dataset consists of train, validation, and test splits of 59,320, 7,413, and 6,531 findings-impression pairs, respectively. The MIMIC-III dataset only contains radiology reports without the original images. In contrast, the MIMIC-CXR summarization dataset (Johnson et al., 2019) is a multimodal summarization dataset containing findings and impressions from chest X-ray studies and corresponding chest X-ray images. It comprises 125,417 training samples, 991 validation samples, and 1624 test samples. Furthermore, we incorporate an out-of-institution test set of 1000 samples from the Stanford hospital(CheXpert) (Irvin et al., 2019) to assess out-of-domain generalization of models trained on MIMIC-CXR. Finally, in Appendix A.2, we also evaluate on the Multimodal Medical Question Summarization dataset (a non-radiology report dataset), showing our method can generalize beyond radiology images. We evaluate model performance using Llama-3.1-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-9b-it (Team et al., 2024).

Evaluation Metrics. Performance is evaluated using the following metrics: BLEU4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), Bertscore (Zhang et al., 2020), F1CheXbert (Delbrouck et al., 2022b), and F1RadGraph (Delbrouck et al., 2022a). Intuitively, BLEU4 measures the precision, while ROUGE-L assesses the recall of the n-gram over-

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	3.37	15.72	39.74	65.39	16.45	28.13
	Mistral-7B-Instruct-v0.3	3.79	17.40	43.39	65.97	15.95	29.30
	Gemma-2-9b-it	3.47	15.66	37.23	66.80	19.28	28.49
Few-Shot	Llama-3.1-8B-Instruct	11.61	32.05	53.67	**70.23**	**30.62**	**39.64**
	Mistral-7B-Instruct-v0.3	**10.65**	29.32	**52.94**	56.14	21.84	34.18
	Gemma-2-9b-it	**11.64**	**30.82**	50.66	60.43	23.78	35.47
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**11.67**	**32.46**	**54.56**	69.03	29.16	39.38
	Mistral-7B-Instruct-v0.3	8.67	**29.81**	51.58	**66.33**	**24.91**	**36.26**
	Gemma-2-9b-it	10.16	30.71	**52.95**	**68.42**	**27.44**	**37.94**

Table 1: Overall performance on the MIMIC CXR in-domain test dataset. We **bold** all results from our framework that outperform the few-shot and zero-shot baselines for the respective model (e.g., Llama vs. Llama).

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	3.11	21.47	46.10	73.14	9.58	30.68
	Mistral-7B-Instruct-v0.3	3.30	22.34	47.65	**72.41**	9.30	31.00
	Gemma-2-9b-it	2.52	19.80	41.21	**73.48**	10.16	29.43
Few-Shot	Llama-3.1-8B-Instruct	3.75	27.73	51.41	72.72	10.65	33.25
	Mistral-7B-Instruct-v0.3	3.13	25.78	49.37	62.30	10.60	30.24
	Gemma-2-9b-it	4.00	**25.82**	48.70	63.87	**10.78**	30.63
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**6.91**	**27.81**	**51.94**	**74.01**	**11.37**	**34.41**
	Mistral-7B-Instruct-v0.3	**5.37**	**27.02**	**50.96**	**68.84**	**10.76**	**32.59**
	Gemma-2-9b-it	**4.31**	24.84	**48.85**	**69.41**	10.30	**31.54**

Table 2: Overall performance across the four prompts on the Stanford Hospital (out-of-domain) test set. The in-context examples for this dataset are from the MIMIC-CXR dataset. We **bold** all results from our framework that outperform the few-shot and zero-shot baselines for the respective model (e.g., Llama vs. Llama).

lap between the generated radiology reports and the original summaries. BERTScore calculates the semantic similarity between tokens of the reference summary and the hypothesis, where the hypothesis refers to the model-generated summary. F1CheXbert uses CheXbert (Smit et al., 2020), a Transformer-based model, to evaluate the clinical accuracy of generated summaries by comparing identified chest X-ray abnormalities in the generated reports to those in the reference reports. F1RadGraph, an F1-score style metric, leverages the RadGraph (Jain et al., 2021) annotation scheme to evaluate the consistency and completeness of the generated reports by comparing them to reference reports based on observation and anatomy entities.

Overall Results. Table 1 reports performance on the in-domain MIMIC-CXR dataset, comparing Zero-Shot, Few-Shot, and our Few-Shot + Layperson prompting strategies. The proposed Few-Shot + Layperson approach, motivated by doctor-patient communication, first prompts the model to generate a simplified summary before producing the expert-level Impression. This intermediary step consistently improves performance across metrics. For instance, Llama-3.1-8B improves in BLEU4 (11.61 → 11.67), ROUGE-L (32.05 → 32.46), and BERTScore (53.67 → 54.56). Similarly, Mistral-7B and Gemma-2-9B show improvements in average performance, with Mistral reaching 36.26 and Gemma 37.94, both surpassing their standard Few-Shot baselines. These results suggest that layperson guidance supports clearer and more consistent generation of radiology Impressions.

As shown in Table 2, the Few-Shot + Layperson approach also improves performance on the out-of-domain Stanford Hospital dataset. Llama-3.1-8B-Instruct achieves the highest F1-cheXbert (74.01) and BERTScore (51.94), while Mistral-7B and Gemma-2-9B show gains in BLEU4, ROUGE-L, and F1-RadGraph. Compared to the standard Few-Shot setup, all three models show consistent improvements in overall average score, highlighting the robustness of our approach to distribution shifts and unseen reporting styles.

The results on the MIMIC-III dataset are shown in Table 3. The Few-Shot + Layperson method consistently outperforms both Zero-Shot and standard Few-Shot prompting across most models and metrics. For instance, Llama-3.1-8B-Instruct improves BLEU4 from 7.87 to 12.68, ROUGE-L from 23.42 to 26.13, and F1-RadGraph from 20.93 to 23.98. Similarly, Gemma-2-9B-it achieves gains

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	5.34	18.77	42.62	51.83	18.43	27.40
	Mistral-7B-Instruct-v0.3	5.94	19.82	45.41	53.48	18.09	28.55
	Gemma-2-9b-it	4.58	18.34	41.07	52.03	18.06	26.82
Few-Shot	Llama-3.1-8B-Instruct	7.87	23.42	47.18	53.91	20.93	30.66
	Mistral-7B-Instruct-v0.3	9.80	**24.45**	**49.87**	**55.15**	19.33	**31.72**
	Gemma-2-9b-it	7.79	23.08	45.61	53.78	20.04	30.06
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**12.68**	**26.13**	**50.32**	**55.70**	**23.98**	**33.76**
	Mistral-7B-Instruct-v0.3	**10.47**	22.51	49.15	49.07	**19.37**	30.11
	Gemma-2-9b-it	**11.59**	**25.57**	**50.24**	**54.56**	**22.28**	**32.85**

Table 3: Overall performance across the four prompts on MIMIC III. We **bold** all results from our framework that outperform the few-shot and zero-shot baselines for the respective model (e.g., Llama vs. Llama).

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Original	Few-Shot	12.81	37.55	54.71	**67.67**	34.95	41.54
	Few-Shot + Layperson	**13.91**	**37.60**	56.76	67.46	**35.37**	**42.22**
Mask	Few-Shot	0.60	6.67	16.35	28.00	6.60	11.64
	Few-Shot + Layperson	**5.38**	**25.05**	**45.63**	**45.70**	**20.60**	**28.47**
Finetuning	Base	0.59	10.06	19.91	24.62	2.45	11.53
	Layperson	**13.90**	**40.60**	**56.49**	**42.80**	**33.65**	**37.49**

Table 4: Overall performance of the Llama-3.1-8B-Instruct model on the MIMIC-CXR valid dataset across three settings: **Original** (unaltered input), **Mask** (Findings entities replaced with gibberish), and **Finetuning** (instruction-tuned models evaluated on masked input). **Base** denotes general instruction tuning. **Layperson** indicates instruction tuning with `<think>`layperson summary`</think>` prepended before the expert summary. Bold values indicate improvements over the respective baselines.

	Model	Test	Hidden	MIMIC-III
Base	Llama-3.1	44.05	31.89	13.47
	Mistral-8B	42.83	27.13	31.62
	Gemma2	46.53	30.01	33.45
Layperson	Llama-3.1	**47.89**	**35.08**	**36.81**
	Mistral-8B	**47.87**	**34.91**	**32.32**
	Gemma2	**48.02**	**34.50**	**34.39**

Table 5: Average performance across three evaluation sets: "Test" (MIMIC-CXR in-domain), "Hidden" (MIMIC-CXR out-of-domain), and "MIMIC-III." Each value is the average of five metrics, such as BLEU4, ROUGEL, BERTScore, F1-cheXbert, and F1-RadGraph. Bold values mark improvements over the corresponding base model.

in BLEU4 (11.59 vs. 7.79), ROUGE-L (25.57 vs. 23.08), and F1-RadGraph (22.28 vs. 20.04). Mistral-7B also shows a competitive BLEU4 improvement (10.47 vs. 9.80). These results demonstrate that incorporating layperson summaries helps LLMs better abstract and align key medical concepts, even across datasets with different styles and terminology. Overall, the Few-Shot + Layperson strategy improves generalization, reinforcing its effectiveness for radiology report summarization in varied clinical settings. We also conduct significance testing to support these findings; detailed results are provided in the Appendix.

Model	Prompt	Acc.	Th.	Use.	Org.	Comp.	Succ.
Llama-3.1-8B	Few-Shot	4.67	**3.78**	4.35	4.96	**5.00**	4.99
	Layperson	**4.88**	3.76	**4.43**	**4.98**	4.99	**4.99**
Mistral-7B	Few-Shot	4.71	3.85	4.43	4.88	4.96	4.93
	Layperson	**4.94**	**4.39**	**4.80**	**4.99**	**5.00**	**5.00**
Gemma-2-9B	Few-Shot	4.49	3.34	4.00	4.74	4.91	4.71
	Layperson	**4.83**	**3.82**	**4.47**	**4.97**	**4.99**	**4.99**

Table 6: LLM-as-a-Judge ratings (1–5) for expert impressions on the MIMIC-CXR in-domain. Bold indicates higher score per model.

LLM-as-a-Judge Evaluation of Summary Quality. To ensure that the intermediate layperson summaries are factually sound before generating the final expert impressions, we additionally evaluate the quality of both outputs using an LLM-as-a-Judge framework. This provides a practical alternative to expert human evaluation, which is valuable but costly to scale, and has been shown to approximate clinician assessments in medical summarization (Croxford et al., 2025). For each finding, we generate two outputs with the same model: a layperson-friendly summary and an expert-style impression. We then use `medgemma-27b-text-it` (Sellergren et al., 2025) to score each output on a 1–5 Likert scale across six dimensions (1 is bad and 5 is good): accurate (factually correct), thorough (covers clinically important issues), useful (helpful for the target provider), organized (clear structure), comprehensible (easy to read), and succinct (no unnecessary text).

Model	Prompt	Acc.	Th.	Use.	Org.	Comp.	Succ.
Llama-3.1-8B	Few-Shot	**4.78**	3.54	4.26	4.94	4.99	**4.99**
	Layperson	4.58	**3.80**	**4.35**	**4.95**	**5.00**	4.97
Mistral-7B	Few-Shot	4.09	3.67	3.98	4.32	4.51	4.33
	Layperson	**4.69**	**3.67**	**4.34**	**4.87**	**4.97**	**4.95**
Gemma-2-9B	Few-Shot	4.38	3.19	3.88	4.70	4.90	4.71
	Layperson	**4.75**	**3.66**	**4.34**	**4.95**	**5.00**	**4.99**

Table 7: LLM-as-a-Judge ratings (1–5) for expert impressions on the Stanford Hospital out-of-domain test set. Bold indicates higher score per model.

Model	Prompt	Acc.	Th.	Use.	Org.	Comp.	Succ.
Llama-3.1-8B	Few-Shot	**4.85**	3.74	4.34	4.93	4.98	4.97
	Layperson	4.77	**4.20**	**4.51**	**4.95**	**5.00**	**4.96**
Mistral-7B	Few-Shot	4.47	4.06	4.34	3.97	4.76	4.76
	Layperson	**4.88**	**4.20**	**4.63**	**4.96**	**5.00**	**4.99**
Gemma-2-9B	Few-Shot	4.78	3.85	4.44	4.93	4.99	**4.98**
	Layperson	**4.91**	**3.92**	**4.52**	**4.98**	**4.99**	4.99

Table 8: LLM-as-a-Judge ratings (1–5) for expert impressions on the MIMIC-III. Bold indicates higher score per model.

The layperson summaries score very well across datasets, with perfect comprehensibility (5.00) and high accuracy (4.86–4.96), showing that the intermediate step produces summaries that are both easy to understand and clinically reliable, as shown in Table 20 in Appendix. For expert impressions, Layperson prompting improves quality over Few-Shot prompting for most models (Tables 6, 7, and 8). On the MIMIC-CXR in-domain test set, Layperson prompting increases accuracy by +0.21 for Llama-3.1-8B and +0.34 for Gemma-2-9B, and yields higher usefulness, comprehensibility, and succinctness for 8 out of 9 model–criterion comparisons. Few-Shot is slightly better only in organization for some cases (e.g., 4.96 vs. 4.98 for Llama). Overall, these results show that introducing a lay summary as an intermediate step leads to more accurate, readable, and helpful expert summaries.

Ablation Study with Chain-of-Thought (CoT). We evaluate whether incorporating an expert Chain-of-Thought (CoT) reasoning step provides additional benefit beyond standard Few-Shot prompting and our Layperson prompting framework. All experiments are conducted using the Llama-3.1-8B-Instruct model on the validation dataset. Following prior work on medical CoT for report generation (CoMT) (Jiang et al., 2025), the CoT variant first decomposes each finding into five structured fields (modality, organs, size/shape, location, and symptoms/signs) before generating the final summary. The full CoT prompt is included in the Appendix A.8. Table 9 reports results on the Mask and Out-of-Domain settings, both designed to simulate unfamiliar or unseen medical terminology. Across both conditions, Layperson prompting yields the best performance, with the largest improvements appearing under masking. This supports our hypothesis that inserting a lay-language intermediate step helps reduce jargon sensitivity and improves robustness to distribution shift.

Instruction Tuning with Layperson Prompt. We adopt the layperson prompt format for instruction tuning. Inspired by Wu et al. (2025), we treat the layperson summary as an explicit intermediate step that guides the model to "think aloud" before generating the expert summary. Specifically, the model is trained to take the Findings as input and generate a layperson summary enclosed in <think>...</think> tags, followed by the expert Impression. This encourages the model to first normalize and simplify key observations, improving both reasoning and final output quality. As shown in Table 5, instruction tuning with layperson prompts leads to consistent performance gains across all datasets and models. On average, each model improves by 3–5 points over its base counterpart, with the most notable gains on out-of-domain and low-resource settings. These results highlight the effectiveness of structured simplification as a useful training signal for improving clinical summarization. Full metric results are provided in the Appendix.

Error Analysis and Discussion. We analyze the impact of adding a layperson summary step by conducting an error analysis of the Llama-3.1-8B-Instruct model on the MIMIC-CXR validation set. Specifically, we compare two prompting strategies: the standard Few-Shot method and our proposed Few-Shot + Layperson approach. The goal is to assess whether introducing simplified, patient-friendly language before the expert impression helps the model better interpret complex or unfamiliar medical terminology. Our intuition is based on the observation that when language models encounters highly specialized or unknown terms, they may misinterpret the context or even refuse to process the request. To address this, we aim to steer the model's attention towards the underlying clinical context by embedding layperson translation in the prompt. This strategy helps the model focus on semantic content rather than being distracted by domain-specific jargon. The approach parallels how humans often rephrase complex information into simpler terms to facilitate comprehension.

Setting	Prompting Method	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Mask	Few-Shot	0.59	10.06	19.91	24.62	2.45	11.53
	Few-Shot + Expert CoT	1.82	17.69	26.62	16.49	14.35	15.39
	Few-Shot + Layperson	**5.38**	**25.05**	**45.63**	**45.70**	**20.60**	**28.47**
Out-of-domain	Few-Shot	3.75	27.73	51.41	72.72	10.65	33.25
	Few-Shot + Expert CoT	5.76	27.68	51.75	73.76	10.92	33.97
	Few-Shot + Layperson	**6.91**	**27.81**	**51.94**	**74.01**	**11.37**	**34.41**

Table 9: Performance of Few-Shot, Expert CoT, and Layperson prompting on the Mask and Out-of-Domain validation sets, which simulate unfamiliar medical terminology. All experiments use Llama-3.1-8B-Instruct. Bold values indicate the best score within each setting.

Radiology Report Summarization

Masked Radiology Finding:
"There are moderate bilateral **entity1** with overlying **entity2** and possible consolidation. **entity3** prominence of **entity4** suggests mild **entity5**."

Ground Truth Impression:
"Bilateral pleural effusions, cardiomegaly, mild edema indicating fluid overload."

Layperson Summary:
"Moderate fluid around both lungs with slight lung changes and mild inflammation. Some underlying issue with the lung tissue that's not clearly visible. "

Few-Shot: "I can't fulfill that request."

Few-Shot + Layperson:
"Bilateral lung consolidation with mild pulmonary edema and cardiomegaly."

To simulate real-world challenges where language models may encounter unfamiliar or ambiguous medical terms, we designed a robustness test by replacing key medical entities with nonsensical "gibberish" tokens. Using MedSpaCy (Eyre et al., 2022), we first identified clinical entities in the Findings section, and then substituted each with a random string (e.g., "pleural effusions" becomes "abcdefg"). This setup allows us to evaluate whether the inclusion of layperson summaries can help the model generate coherent and accurate expert impressions despite encountering unknown terminology. We hypothesize that the simplified layperson summary, combined with the surrounding context in the Findings, encourages the model to normalize unfamiliar terms into more accessible language, thereby improving its overall performance. Results of this experiment are shown in Table 4. The "Mask" section reports performance on the modified examples for both the baseline (Llama-3.1-8B-Instruct + Few-Shot) and our method (Llama-3.1-8B-Instruct + Few-Shot + Layperson). We also include performance on the original, unmodified data for comparison.

Our findings show that the Few-Shot + Layperson approach consistently outperforms the baseline. In particular, when key terms are masked, baseline performance drops sharply (ROUGE-L: 37.55 → 6.67), indicating its difficulty handling unfamiliar input. In contrast, our method is more robust under these conditions (ROUGE-L: 37.60 → 25.05), suggesting that the added layperson summary helps the model generalize better by guiding it toward the core clinical meaning. When encountering such unknown or nonsensical terms, the standard Few-Shot model often fails to generate a meaningful summary and instead requests clarification. For example, it usually simply state, "I can't fulfill that request." We provide an example below:

Additionally, instruction tuning with layperson prompts (**Layperson**) achieved strong performance on masked inputs, outperforming all baselines with a notable jump in all metrics (e.g., BLEU4: 13.90 vs. 0.59 for Base and ROUGEL: 10.06 vs. 40.60). These results show that the layperson prompting strategy, which guides the model to first generate a simplified summary, helps improve its reasoning ability and generalization when dealing with unfamiliar or ambiguous clinical input. This leads to more accurate expert summaries even when key clinical entities are masked or ambiguous.

5 Conclusion

This paper introduces a simple and effective prompting strategy inspired by doctor–patient communication. The method guides the model to first generate a layperson summary before the expert impression, helping it reason through and organize complex clinical content. It consistently improves performance across MIMIC-CXR, CheXpert, and MIMIC-III, especially on out-of-domain data. We also extend this idea to instruction tuning by using <think>...</think> tags to simulate intermediate reasoning. This strategy improves generalization and robustness for both prompting and instruction tuning. Future work will explore improved prompt design, better token efficiency, and larger models with extended context capacity.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

Limitation

Our approach relies on layperson-style intermediate summaries, but we do not include an ablation directly comparing them to expert-style or neutral explanations. Prior work has studied expert reasoning prompts, but without a controlled comparison it remains unclear whether improvements come from simplification specifically or from the presence of any structured intermediate step.

The method assumes that simplifying findings into general concepts helps models normalize domain-specific terminology. While our quantitative and LLM-as-a-judge results support this intuition, we do not conduct a detailed qualitative error analysis to identify cases where simplification may blur clinically important distinctions or introduce subtle semantic drift.

The framework also depends on the quality of lay summaries generated during preprocessing. Errors in these summaries, such as omissions or misinterpretations, propagate into the expert impression stage. Because our evaluation focuses on final summaries, we cannot fully separate how intermediate inaccuracies affect downstream performance.

The method is evaluated primarily on radiology datasets with structured findings and predictable terminology. Its effectiveness on less standardized clinical narratives or specialties with higher linguistic variability remains uncertain. Finally, although layperson prompting improves robustness to unseen or masked terminology, we do not explore failure modes where simplification interacts poorly with rare conditions, ambiguous findings, or highly technical descriptions.

Ethics Statement

In this work, we have introduced our Layperson Summary Prompting strategy, inspired by doctor-patient communication techniques. This approach aims to simplify complex medical findings into layperson summary first, then uses this simplified information to generate accurate expert summaries. However, it is important to address the ethical implications of using LLMs in this context. LLMs used for radiology report summarization can produce errors or biased outputs if the training data is of low quality or representative. These models also can be wrong, and such biases can lead to unfair outcomes and exacerbate health disparities. Therefore, radiologists should use AI-generated summaries as supportive tools, retaining control over clinical decisions. AI should be seen as an information resource to reduce time and cognitive effort, aiding in information retrieval and summarization, rather than as an interpretative agent providing clinical decisions or treatment recommendations.

Additionally, integrating AI into clinical practice raises significant ethical considerations regarding patient privacy, data security, and informed consent. Using large volumes of sensitive patient data for training AI models necessitates stringent measures to protect patient rights and ensure data confidentiality. Ethical principles such as fairness, accountability, and transparency should guide the deployment of AI technologies in healthcare. These principles help ensure that AI systems are used responsibly and that the benefits of AI are distributed equitably among all stakeholders. Furthermore, potential risks associated with AI implementation include perpetuating existing biases, privacy breaches, and the misuse of AI-generated data, necessitating careful consideration and proactive management (Yildirim et al., 2024).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Katherine A Allen, Victoria Charpentier, Marissa A Hendrickson, Molly Kessler, Rachael Gotlieb, Jordan Marmet, Emily Hause, Corinne Praska, Scott Lunos, and Michael B Pitt. 2023. Jargon be gone–patient preference in doctor communication. *Journal of Patient Experience*, 10:23743735231158942.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.

2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023a. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023b. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy. Association for Computational Linguistics.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2022. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: CLIP and LLM synergy for multimodal question summarization in healthcare. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22031–22039. AAAI Press.

Elisabeth Gülich. 2003. Conversational techniques used in transferring knowledge between medical experts and non-experts. *Discourse studies*, 5(2):235–263.

Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR.

Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T Sio, Lisa A McGee, Jonathan B Ashman, Xiang Li, Tianming Liu, Jiajian Shen, et al. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, 13.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *ArXiv preprint*, abs/2303.16416.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI*

Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 590–597. AAAI Press.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, pages 1–9.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Dayeon Ki and Marine Carpuat. 2025. Automatic input rewriting improves translation with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10829–10856, Albuquerque, New Mexico. Association for Computational Linguistics.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas W LeBlanc, Ashley Hesson, Andrew Williams, Chris Feudtner, Margaret Holmes-Rovner, Lillie D Williamson, and Peter A Ubel. 2014. Patient understanding of medical jargon: a survey study of us medical students. *Patient education and counseling*, 95(2):238–242.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text (version 1.0.0). *PhysioNet*.

Hanzhou Li, John T Moon, Deepak Iyer, Patricia Balthazar, Elizabeth A Krupinski, Zachary L Bercu, Janice M Newsome, Imon Banerjee, Judy W Gichoya, and Hari M Trivedi. 2023. Decoding radiology reports: Potential application of openai chatgpt to enhance patient understanding of diagnostic reports. *Clinical Imaging*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yan Liu, Yazheng Yang, and Xiaokang Chen. 2024. Improving long text understanding with knowledge distilled from summarization model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11776–11780. IEEE.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Andrea B Neiman. 2017. Cdc grand rounds: improving medication adherence for chronic disease management—innovations and opportunities. *MMWR. Morbidity and mortality weekly report*, 66.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *ArXiv preprint*, abs/2311.16452.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, Antonio Jimeno Yepes, and Jey Han Lau. 2022. M3: Multi-level dataset for multi-document summarisation of medical studies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3887–3901, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alison Q O'Neil, John T Murchison, Edwin JR van Beek, and Keith A Goatman. 2017. Crowdsourcing labels for pathological patterns in ct lung scans: can non-experts contribute expert-quality ground truth? In *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 6th Joint International Workshops, CVII-STENT 2017 and Second International Workshop, LABELS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10–14, 2017, Proceedings 2*, pages 96–105. Springer.

Gustavo H. Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3761–3767. AAAI Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maryke Peter, Stacy Maddocks, Clarice Tang, and Pat G Camp. 2024. Simplicity: Using the power of plain language to encourage patient-centered communication. *Physical therapy*, 104(1):pzad103.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not always enough. *ArXiv preprint*, abs/2402.00179.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Sandra van Dulmen, Emmy Sluijs, Liset Van Dijk, Denise de Ridder, Rob Heerdink, and Jozien Bensing. 2007. Patient adherence to medical treatment: a review of reviews. *BMC health services research*, 7:1–13.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023a. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, et al. 2023b. Clinical text summarization: adapting large language models can outperform human experts. *ArXiv preprint*, abs/2309.07430.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Nathan VanHoudnos, William Casey, David French, Brian Lindauer, Eliezer Kanal, Evan Wright, Bronwyn Woods, Seungwhan Moon, Peter Jansen, and Jamie Carbonell. 2017. This malware looks familiar: Laymen identify malware run-time similarity with chernoff faces and stick figures. In *10th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, pages 152–159.

Tongnian Wang, Xingmeng Zhao, and Anthony Rios. 2023. UTSA-NLP at RadSum23: Multi-modal retrieval-based chest X-ray report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 557–566, Toronto, Canada. Association for Computational Linguistics.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Tong Wu, Chong Xiang, Jiachen T Wang, G Edward Suh, and Prateek Mittal. 2025. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. 2023. Style-aware radiology report generation with RadGraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688, Singapore. Association for Computational Linguistics.

Jing Yao, Wei Xu, Jianxun Lian, Xiting Wang, Xiaoyuan Yi, and Xing Xie. 2023a. Knowledge plugins: Enhancing large language models for domain-specific recommendations. *ArXiv preprint*, abs/2311.10779.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho de Castro, Shruthi Bannur, Stephanie L. Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew P. Lungren, Javier Alvarez-Valle, Aditya V. Nori, and Anja Thieme. 2024. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 444:1–444:22. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness harmlessness with rlaif.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2022. Exact paired-permutation testing for structured test statistics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4894–4902.

A Appendix

A.1 Baseline and Implementation Details

For our baseline approach, we adopt a prefixed zero-shot prompting strategy (Duan et al., 2019; Zhao and Schütze, 2021), which prepended a brief instruction to the beginning of a standard null prompt. We use the instruction, “You are an expert chest radiologist. Your task is to summarize the radiology report findings into an impression with minimal text”. This instruction provides the model with a fundamental context for the RRS task. Immediately following the instruction, we append the specific findings from the report and then prompt the model with “IMPRESSION:” to initiate the generation process. Additionally, we investigate the effectiveness of few-shot ICL prompts with up to 32 similar examples, using the same template as our Few-Shot prompting method, which is not incorporating the intermediate reasoning step (i.e., without the layperson summary).

We conduct experiments with three open-source LLMs: Llama-3.1-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-9b-it (Team et al., 2024). All experiments were conducted using two Nvidia A6000 GPUs. For the few-shot model, the average running time is around 2 hours. In contrast, the Few-Shot + Layperson models have an average running time of around 8 hours. Processing the MIMIC data with 24 examples takes approximately 36 hours. In our work, all these models have been implemented using the Hugging Face framework (Wolf et al., 2019). Specifically, the Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Gemma-2-9b-it are reported to perform strongly in common sense reasoning and problem-solving ability (Zhu et al., 2023). To select the best parameters in our study, we employed ROUGE-L and F1RadGraph metrics on the validation set. These metrics help determine the most effective parameter settings for the model. The ROUGE-L metric focuses on the longest common subsequence and is particularly suitable for evaluating the quality of text summaries. On the other hand, the F1RadGraph is specifically designed to assess the accuracy of extracting and summarizing key information from radiology reports by analyzing entity similarities.

For optimizing our model's hyper-parameters, we employed a random search strategy on valid dataset. This involved experimenting with various settings: the number of prepended similar examples was varied across a set 2, 8, 12, 16, 24, 32, and these examples were matched using different modality embeddings (text, image, or multimodal), all while employing the same template. We find that for the Llama-3.1-8B-Instruct, the best performance is achieved with 32 examples for both Few-Shot and Few-Shot + Layperson prompting methods. Additionally, we experimented with temperature settings ranging from 0.1 to 0.9, top p values set between 0.1 and 0.6, and top k values of 10, 20, and 30. Through this exploratory process, we identified the most effective settings as a temperature of 0.2, a top p value of 0.5, and a top k setting of 20. We adopt the same hyperparameters for all experiments. These settings yielded the best results in our evaluations. It's significant to note the impact of the “temperature” parameter on the diversity of the model's outputs. Higher temperature values add more variation, introducing a greater level of randomness into the content generated. This aspect is especially valuable for adjusting the output to meet specific requirements for creativity or diversity.

To ensure compatibility with the model's capabilities, we restricted the length of the prompt (which includes the instruction, input, and output instance) to 7800 tokens. This limit was set to prevent exceeding the model's maximum sequence length of 8,192 tokens for Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Gemma-2-9b-it. In cases where prompts exceeded this length, they were truncated from the beginning, ensuring that

Method	Prompt tokens (avg)	Completion tokens (avg)	Total tokens (avg)
Few-Shot prompt	1853.1	31.1	1884.2
Few-Shot + Expert CoT	1951.5	112.8	2064.3
Layperson prompt	2107.8	70.0	2177.8

Table 10: Token usage per example on MIMIC-CXR-Valid with Llama-3.1-8B-Instruct.

essential information and current findings were preserved. Moreover, we constrained the generated output to a maximum of 256 tokens to strike a balance between providing detailed content and adhering to the model's constraints. This approach was key in optimizing the effectiveness of summarization within the operational limits of the 7B models. Table 11 shows the prompt lengths for different numbers of examples used in our study. For the MIMIC-III dataset, using 32 examples exceeds the 7800 token limit, so we opted to use only 16 examples.

To quantify the computational overhead of different prompting strategies, we report per-example token usage on MIMIC-CXR-Valid using Llama-3.1-8B-Instruct (Table 10). Layperson prompting incurs a moderately higher prompt length than Few-Shot but requires substantially fewer completion tokens than Expert CoT, yielding a favorable trade-off in end-to-end cost. Similar trends were observed for other model families (e.g., Mistral and Gemma), indicating that the relative efficiency advantages of Layperson prompting are model-agnostic.

A.2 Medical Question Summarization Results

Furthermore, we assess an additional dataset, the Multimodal Medical Question Summarization (MMQS) Dataset, introduced by Ghosh et al. (2024). This dataset contains 3,015 multimodal medical queries, each accompanied by visual cues and expert-annotated gold summaries that reference various body parts (e.g., skin, eyes, ears). As shown in Table 12, we observe consistent patterns across the three prompting regimes.

When we compare the standard Few-Shot setting to our Few-Shot+Layperson approach, every model sees a clear uplift in performance. Llama-3.1-8B-Instruct's average score jumps from 30.43 to 42.09, a gain of 11.66 points (a 38 percent relative improvement). This improvement is driven by more than doubling its BLEU, 4 score—from 10.09 to 20.84, as well as substantial increases in ROUGE-L (31.52 to 43.54) and BERTScore (49.68 to 61.89). In practice, the layperson prompt helps Llama-3.1 generate summaries that are not only lexically richer but also more semantically aligned and fluent.

Mistral-7B-Instruct-v0.3 also benefits, albeit more modestly: its average climbs from 35.55 to 38.34 (an increase of 2.79 points, or roughly 7.8 percent). Its BLEU-4 rises from 13.57 to 16.83, ROUGE-L from 37.16 to 39.80, and BERTScore from 55.93 to 58.40. These gains demonstrate that even already strong few-shot models can produce clearer, more coherent medical summaries when guided to use non-expert language.

The largest relative boost is seen with Gemma-2-9b-it, whose average soars from 23.45 to 37.55, an absolute gain of 14.10 points, equivalent to a 60 percent improvement. Its BLEU-4 score leaps from 4.01 to 16.93, ROUGE-L climbs from 26.14 to 38.25, and BERTScore jumps from 40.21 to 57.47. This dramatic uplift underscores how layperson-focused prompting can unlock substantial performance gains for models that struggle under a standard few-shot regime.

Overall, while Few-Shot priming delivers a strong baseline boost for all three models, our Few-Shot + Layperson method consistently amplifies that effect—especially in BLEU-4 and BERTScore—and narrows the performance disparities. By tailoring summaries to align with non-expert understanding, we achieve more accurate, coherent, and accessible medical question summaries across the board.

We further assess whether the benefits of the lay-intermediate step generalize beyond radiology report summarization. As shown in Table 13, Layperson prompting consistently improves summary quality across most dimensions. For example, with Llama-3.1-8B, Layperson prompting yields higher accuracy (4.91 vs. 4.57), thoroughness (4.58 vs. 3.97), and usefulness (4.54 vs. 4.21). Similar gains are observed for Gemma-2-9B, including improvements in usefulness (4.42 vs. 4.41) and organization (5.00 vs. 4.93). Mistral-7B shows the smallest gap, with Few-Shot slightly higher in accuracy (4.84 vs. 4.92) but Layperson scoring the strongest across the other five criteria (e.g., usefulness 4.47 vs. 4.33). Overall, these results suggest that the lay summary step enhances downstream expert summarization quality even in multimodal medical question settings.

		2	**8**	**12**	**16**	**24**	**32**
MIMIC-CXR	Few-Shot	643	1285	1713	2141	2994	3850
	Few-Shot + Layperson	889	1826	2452	3084	4333	5587
MIMIC-III	Few-Shot	1035	2500	3474	4451	6405	8359
	Few-Shot + Layperson	1340	3277	4565	5856	8442	11025

Table 11: Average Token of Prompts.

		BLEU4	**ROUGEL**	**BERTScore**	**Average**
Zero-Shot	Llama-3.1-8B-Instruct	1.25	8.19	16.41	8.62
	Mistral-7B-Instruct-v0.3	3.17	15.93	35.31	18.14
	Gemma-2-9b-it	1.22	9.72	20.01	10.32
Few-Shot	Llama-3.1-8B-Instruct	10.09	31.52	49.68	30.43
	Mistral-7B-Instruct-v0.3	13.57	37.16	55.93	35.55
	Gemma-2-9b-it	4.01	26.14	40.21	23.45
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**20.84**	**43.54**	**61.89**	**42.09**
	Mistral-7B-Instruct-v0.3	**16.83**	**39.80**	**58.40**	**38.34**
	Gemma-2-9b-it	**16.93**	**38.25**	**57.47**	**37.55**

Table 12: Performance of models on Multimodal Medical Question Summarization (MMQS) Dataset.

Model	**Prompt**	**Acc.**	**Th.**	**Use.**	**Org.**	**Comp.**	**Succ.**
Llama-3.1-8B	Few-Shot	4.57	3.97	4.21	4.91	4.98	4.97
	Layperson	**4.91**	**4.58**	**4.54**	**4.98**	**4.99**	**4.99**
Mistral-7B	Few-Shot	4.84	4.08	4.33	4.97	4.98	4.98
	Layperson	**4.92**	**4.43**	**4.47**	**4.99**	**5.00**	**5.00**
Gemma-2-9B	Few-Shot	4.81	**4.34**	4.41	4.93	4.99	4.96
	Layperson	**4.90**	4.15	**4.42**	**5.00**	**5.00**	**5.00**

Table 13: LLM-as-a-Judge evaluation (1–5 Likert) of expert-style summaries on MMQS. Bold indicates the higher score per model between Layperson vs. Few-Shot prompting.

A.3 Significance Testing Results.

To assess the robustness of our method, we applied paired permutation tests following established NLP practices (Dror et al., 2018; Zmigrod et al., 2022). We compared the Layperson prompting strategy against the standard Few-Shot baseline across all datasets and metrics. Our updated tests confirm that the Layperson method yields consistent and statistically significant improvements:

On the MIMIC-CXR dataset (Table 1), the Layperson method outperforms the baseline in 9 out of 15 metric comparisons. On the Stanford hidden test set (Table 2), it shows consistent gains across 13/15 comparisons. On the MIMIC-III dataset (Table 3), the method achieves improvements in 12 out of 15 cases. In the masked robustness experiment simulating corrupted clinical input (Table 4), the Layperson strategy performs better in all 5 metrics. Finally, on the multimodal medical question summarization dataset (Table 12), it outperforms the baseline in 9 out of 9 comparisons. Across a total of 59 metric comparisons, our approach outperforms the baseline in 48 cases, 45 of which are statistically significant ($p < 0.05$). These results confirm the effectiveness of the Layperson prompting strategy in enhancing model reasoning and generalization across diverse clinical scenarios.

A.4 Ablation Study on Specialized Medical LLMs vs. General LLMs

We examine whether the gains from Layperson prompting hold for domain-specialized medical LLMs, rather than being limited to general-purpose models. We evaluate a medical model, MMed-Llama-3-8B (Qiu et al., 2024), alongside the general instruction-tuned Llama-3.1-8B-Instruct on the CheXpert out-of-domain radiology summarization benchmark. For both models, we test three prompting regimes: Few-Shot, Few-Shot with Expert CoT, and Few-Shot with Layperson prompting.

Table 15 shows that Layperson prompting improves performance for both model types. For Llama-3.1-8B-Instruct, the Average score increases from 33.25 to 34.41 (+1.16 over Few-Shot and

Model	Dataset	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph
llama3	MIMIC-CXR	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)
llama3	MIMIC-CXR-Hidden	*** (0.0000)	*** (0.0000)	(0.1232)	** (0.0026)	(0.0874)
llama3	MIMIC-III	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)
llama3	MMQS	*** (0.0000)	*** (0.0000)	*** (0.0000)	N/A	N/A
ministral	MIMIC-CXR	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)
ministral	MIMIC-CXR-Hidden	*** (0.0000)	*** (0.0000)	*** (0.0000)	** (0.0017)	** (0.0013)
ministral	MIMIC-III	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)
ministral	MMQS	*** (0.0000)	*** (0.0000)	*** (0.0000)	N/A	N/A
gemma	MIMIC-CXR	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)	*** (0.0000)
gemma	MIMIC-CXR-Hidden	*** (0.0000)	*** (0.0000)	(0.6401)	* (0.0188)	(0.1950)
gemma	MIMIC-III	*** (0.0000)	*** (0.0000)	* (0.0111)	(0.0806)	*** (0.0000)
gemma	MMQS	*** (0.0000)	*** (0.0000)	*** (0.0000)	N/A	N/A

Table 14: Significance markers: *** $p < .001$, ** $p < .01$, * $p < .05$; note that the MMQS dataset is not radiology related, so F1-chexbert and F1-radgraph are not applicable.

+0.44 over Expert CoT). The improvement is more pronounced for the medical model, rising from 23.30 to 31.00 (+7.70 over Few-Shot and +4.40 over Expert CoT), with consistent gains across all metrics, including clinically grounded ones (CheXbert-F1 and RadGraph-F1). While the medical model benefits more from Layperson prompting, its best score (31.00) remains below that of the general model (34.41), echoing prior findings that strong base LLMs already encode substantial medical knowledge and that domain-adaptation alone may provide limited or inconsistent gains (Jiang et al., 2025). These results suggest that lay-style reformulation provides a complementary robustness benefit for both general and specialized LLMs, likely because even medically fine-tuned models still align more strongly with general-language patterns than with highly specialized clinical phrasing.

A.5 Comparison of Encoder–Decoder vs. Sentence-Transformer Encoders

We compare two embedding models for selecting few-shot examples: an encoder–decoder model (Clinical-T5) and a sentence-transformer model (PubMedBERT) (Van Veen et al., 2024). We first examine the overlap between the retrieved few-shot examples and observe substantial agreement: 18% of the selected examples are identical (high overlap) and an additional 40% share medium overlap, indicating that the two encoders frequently retrieve similar support examples. We then evaluate both embedding choices under the same Few-Shot + Layperson prompting setup on the validation set. As shown in Table 16, performance is highly similar across metrics. Clinical-T5 achieves an average score of 42.22, while PubMedBERT achieves 42.73. These findings suggest that both embedding models perform comparably for this downstream task. We adopt Clinical-T5 in our main experiments because it has demonstrated strong performance in prior radiology summarization work, including top-ranked results in the RadSum23 shared task (Wang et al., 2023).

A.6 More Error Analysis

Notably, when encountering unknown terms, the standard Few-Shot model tends to produce longer summaries that often rephrase or repeat content from the Findings and occasionally introduce hallucinated information. In contrast, our Few-Shot + Layperson approach explicitly guides the model to focus on simplified, high-level clinical meaning. As shown in the example below, the Few-Shot model failed to capture the key observation and even fabricated a term such as “xenosign.” In comparison, our approach correctly generated the summary “No acute cardiopulmonary process. Normal heart size,” which closely matches the ground truth. An illustrative example is provided below:

Model	Prompting Regime	BLEU-4	ROUGE-L	BERTScore	CheXbert-F1	RadGraph-F1	Average
Llama-3.1-8B-Instruct	Few-Shot	3.75	27.73	51.41	72.72	10.65	33.25
	Few-Shot + Expert CoT	5.76	27.68	51.75	73.76	10.92	33.97
	Few-Shot + Layperson	**6.91**	**27.81**	**51.94**	**74.01**	**11.37**	**34.41**
MMed-Llama-3-8B	Few-Shot	3.41	16.68	37.84	51.08	7.49	23.30
	Few-ShShot + Expert CoT	4.26	22.06	41.48	57.26	7.95	26.60
	Few-Shot + Layperson	**5.91**	**25.69**	**48.89**	**64.73**	**9.79**	**31.00**

Table 15: Comparison of general vs. medical LLMs under three prompting regimes on the CheXpert Out-of-Domain evaluation. Layperson prompting improves both model types, with larger gains for the medical LLM.

Model	BLEU4	ROUGEL	BERTScore	F1-CheXbert	F1-RadGraph	Average
Few-Shot-ClinicalT5 + Layperson	13.91	37.60	56.76	67.46	35.37	42.22
Few-Shot-PubMedBERT + Layperson	13.16	43.58	57.54	59.42	39.96	42.73

Table 16: Comparison of Clinical-T5 vs. PubMedBERT embeddings for few-shot example retrieval under Layperson prompting.

Radiology Report Summarization

Masked Radiology Finding:
"Frontal and lateral views of the **entity1**. The **entity2** are clear of focal consolidation or **entity3**. Opacities at the cardiophrenic angles bilaterally are thought to represent prominent **entity4**."

Ground Truth Impression:
"No acute cardiopulmonary process."

Layperson Summary:
"The chest X-rays show clear lungs without any solid masses. Small opacities at the lung edges likely represent normal fat or muscle. Overall, the heart and chest structures appear normal."

Few-Shot:
"Clear lungs without consolidation. A subtle finding, termed "xenosign", is noted at the left lung base. Bilateral opacities are likely due to fat."

Few-Shot + Layperson:
"No acute cardiopulmonary process. Normal heart size."

To better understand this behavior, we conducted an error analysis on masked inputs. Importantly, the layperson summaries in this setting are also generated from the masked findings, ensuring that the model cannot rely on unmasked clinical terms when producing the simplified summary. This isolates the effect of the lay-intermediate step rather than advantaging it with access to the original terminology. We find that even when key medical tokens are obscured, lay-level reformulation helps the model infer the intended meaning. This aligns with a simple intuition: if a sentence such as "John Doe is [MASK] down the street with his new shoes" retains enough contextual cues, readers—and LLMs—can often recover the missing concept. Radiology findings behave similarly; once phrased in everyday language, the core clinical intent becomes easier for models to interpret and summarize faithfully.

A.7 Instruction Tuning Experiments

Here are our instruction-tuning results (Tables 17, 18, and 19). The layperson-tuned models outperform their base versions on every metric across in-domain, out-of-domain tasks. This demonstrates that layperson-focused fine-tuning yields more accurate, easy-to-understand medical summaries.

A.8 Expert Chain-of-Thought Prompting

To assess whether explicit expert reasoning improves clinical summarization, we introduce an Expert Chain-of-Thought (CoT) prompting variant. As illustrated in Figure 4, the model is instructed to first break down the radiology findings into structured clinical attributes before generating the final impression. This five-step scaffold is adapted from prior work on medical CoT for report generation (Jiang et al., 2025) and is intended to emulate how radiologists synthesize observations.

A.9 Layperson Summary Evaluation Prompting

To assess the quality of layperson-friendly summaries, we use a structured evaluation prompt that instructs the model to rate summaries along six dimensions: accuracy, thoroughness, usefulness, organization, comprehensibility, and succinctness. The evaluator is required to return scores on a 1–5 Likert scale in a fixed JSON format to ensure consistency and reproducibility across evaluations. The full evaluation prompt is provided in Figure 5.

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Base	Llama-3.1	18.15	40.73	58.40	67.33	35.65	44.05
	Mistral-7B	13.90	39.53	58.36	67.13	35.23	42.83
	Gemma2	20.15	43.27	61.02	70.05	38.16	46.53
Instruction-Tuned (Layperson)	Llama-3.1	**19.33**	**44.04**	**62.53**	**74.49**	**39.07**	**47.89**
	Mistral-7B	**20.67**	**43.15**	**62.01**	**75.13**	**38.37**	**47.87**
	Gemma2	19.66	**44.02**	**62.10**	**75.44**	**38.88**	**48.02**

Table 17: Instruction-tuning performance on the MIMIC-CXR in-domain test set. "Base" refers to the pretrained models without further tuning; "Instruction-Tuned (Layperson)" are the same models fine-tuned on layperson-style prompts. Bold indicates where instruction tuning yields gains over the base model.

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Base	Llama-3.1	5.33	28.08	50.56	63.93	11.55	31.89
	Ministral-8B	2.16	23.86	45.88	54.45	9.29	27.13
	Gemma2	4.79	27.98	48.59	57.77	10.94	30.01
Instruction-Tuned (Layperson)	Llama-3.1	**5.34**	**32.27**	**54.52**	**70.57**	**12.71**	**35.08**
	Ministral-8B	**5.64**	**32.42**	**54.84**	**69.51**	**12.15**	**34.91**
	Gemma2	**5.37**	**31.75**	**53.58**	**69.41**	**12.37**	**34.50**

Table 18: Instruction-tuning performance on the MIMIC-CXR hidden (out-of-domain) test set. "Base" refers to the pretrained models without further tuning; "Instruction-Tuned (Layperson)" are the same models fine-tuned on layperson-style prompts. Bold indicates where instruction tuning yields gains over the base model.

A.10 Expert Summary Evaluation Prompt

To evaluate the quality of expert-oriented radiology summaries, we adopt a structured LLM-as-a-judge prompt that scores outputs across the same six dimensions as the layperson evaluation—accuracy, thoroughness, usefulness, organization, comprehensibility, and succinctness—using a 1–5 Likert scale. This prompt is designed to assess technical precision and clinical appropriateness of the expert summary. The full expert evaluation prompt is shown in Figure 6.

A.11 Layperson Summary Prompting

This template guides the AI, doctors, and patient advocates to work together in clear steps to turn a technical radiology report into simple bullet points. At each stage, experts check the accuracy and suggest plain-language fixes so the final summary is both correct and easy to understand, as shown in Figure 7

Expert Chain-of-Thought Prompt

You are a radiology expert. Your task is to summarize the radiology report findings into an **IMPRESSION**.

Think step by step and first decompose the findings into the following components, using *one short sentence each*:

- **Modality**
- **Organs**
- **Size/Shape**
- **Location**
- **Symptoms/Signs**
- **Health condition**

Findings: {{finding}}
Chain-of-Thought: {{cot reasoning}}
IMPRESSION: {{impression}}

Figure 4: Expert CoT prompt used in our ablation. The model is guided to first generate structured clinical reasoning before producing the final impression.

		BLEU4	ROUGEL	BERTScore	Average
Base	Llama-3.1	3.30	18.03	33.69	18.34
	Mistral-7B-Instruct-v0.3	3.81	19.91	38.45	20.72
	Gemma-2-9b-it	7.66	24.39	35.64	22.56
Instruction-Tuned (Layperson)	Llama-3.1	**28.07**	**48.45**	**65.40**	**47.31**
	Mistral-7B-Instruct-v0.3	**5.54**	**29.32**	**48.10**	**27.65**
	Gemma-2-9b-it	**15.89**	**33.23**	**46.79**	**31.97**

Table 19: Instruction-tuning performance on the MMQS Dataset (ClipSyntel). "Base" refers to the pretrained models without further tuning; "Instruction-Tuned (Layperson)" are the same models fine-tuned on layperson-style prompts. Bold indicates gains over the base model.

Dataset	n	Acc.	Thorough	Useful	Org.	Comp.	Succ.
MIMIC-CXR	125,402	4.95	3.80	4.61	4.98	5.00	5.00
MIMIC-III	59,005	4.96	3.65	4.67	4.92	5.00	5.00
MMQS	2,110	4.86	4.25	4.69	4.98	5.00	4.99

Table 20: LLM-as-a-Judge evaluation of layperson summaries across three datasets using a 1–5 Likert scale.

Layperson Evaluation System Prompt

You are evaluating a layperson-friendly summary of a chest X-ray impression. Compare the summary against the original impression and rate it on the following criteria using a 1–5 Likert scale.

Provide your response in this exact JSON format:

```
{
  "accurate": <1-5>,
  "thorough": <1-5>,
  "useful": <1-5>,
  "organized": <1-5>,
  "comprehensible": <1-5>,
  "succinct": <1-5>
}
```

CRITERIA DEFINITIONS WITH GRADING

accurate: The summary is true and free of incorrect information. 1 = Multiple major errors with overt falsifications or fabrications 2 = A major error in assertion occurs with an overt falsification or fabrication 3 = At least one assertion contains a misalignment that is taken from a source note but stated in the wrong context, including incorrect specificity in diagnosis or treatment 4 = At least one assertion is misaligned to the provider source or timing but still factual in diagnosis or treatment 5 = All assertions can be traced back to the notes

thorough: The summary is complete and documents all issues important to the patient. 1 = More than one pertinent omission occurs 2 = One pertinent and multiple potentially pertinent omissions occur 3 = Only one pertinent omission occurs 4 = Some potentially pertinent omissions occur 5 = No pertinent or potentially pertinent omissions occur

useful: All information in the summary is helpful to the target provider. 1 = No assertions are pertinent to the target user 2 = Some assertions are pertinent to the target user 3 = Assertions are pertinent but the level of detail is inappropriate (too detailed or not enough) 4 = No non-pertinent assertions; some assertions are potentially pertinent 5 = No non-pertinent assertions and level of detail is appropriate

organized: The summary is structured in a way that helps the reader understand clinical course. 1 = Assertions disorganized; incoherent grouping 2 = Some ordering issues or incoherent grouping 3 = Structure unchanged from input 4 = Logical ordering or grouping (not both) 5 = Logical ordering and grouping applied

comprehensible: Clarity and ease of understanding. 1 = Overly complex language or unfamiliar terminology throughout 2 = Some overly complex or unfamiliar terminology 3 = Language mostly unchanged from input; missed opportunities to simplify 4 = Mostly plain language with some simplifications 5 = Fully plain, well-structured language familiar to the target user

succinct: Brevity without loss of meaning. 1 = Very wordy with redundancy 2 = Multiple redundant assertions 3 = At least one redundant assertion 4 = No redundancy, but could be shorter 5 = Minimal words used without redundancy

Figure 5: Full system prompt used for LLM-based evaluation of layperson summary quality.

Expert Evaluation System Prompt

You are evaluating an expert technical summary of a chest X-ray impression. Compare the summary against the original impression and rate it on the following criteria using a 1–5 Likert scale.

Provide your response in this exact JSON format:

```
{
  "accurate": <1-5>,
  "thorough": <1-5>,
  "useful": <1-5>,
  "organized": <1-5>,
  "comprehensible": <1-5>,
  "succinct": <1-5>
}
```

CRITERIA DEFINITIONS WITH GRADING

accurate: The summary is true. It is free of incorrect information. 1 = Multiple major errors with overt falsifications or fabrications 2 = A major error in assertion occurs with an overt falsification or fabrication 3 = At least one assertion contains a misalignment that is stated from a source note but the wrong context, including incorrect specificity in diagnosis or treatment 4 = At least one assertion is misaligned to the provider source or timing but still factual in diagnosis, treatment, etc. 5 = All assertions can be traced back to the notes

thorough: The summary is complete and documents all of the issues of importance to the patient. 1 = More than one pertinent omission occurs 2 = One pertinent and multiple potentially pertinent occur 3 = Only one pertinent omission occurs 4 = Some potentially pertinent omissions occur 5 = No pertinent or potentially pertinent omission occur

useful: All the information in the summary is useful to the target provider. 1 = No assertions are pertinent to the target user 2 = Some assertions are pertinent to the target user 3 = Assertions are pertinent to target provider but level of detail inappropriate (too detailed or not detailed enough) 4 = Not adding any non-pertinent assertions but some assertions are potentially pertinent to target user 5 = Not adding any non-pertinent assertions and level of detail is appropriate to targeted user

organized: The summary is well-formed and structured in a way that helps the reader understand the patient's clinical course. 1 = All assertions presented out of order and groupings incoherent (completely disorganized) 2 = Some assertions presented out of order OR grouping incoherent 3 = No change in order or grouping (temporal or systems/problem based) from original input 4 = Logical order or grouping (temporal or systems/problem based) for all assertions but not both 5 = All assertions made with logical order and grouping (temporal or systems/problem based) – completely organized

comprehensible: Clarity of language. The summary is clear, without ambiguity or difficult sections. 1 = Overly complex, inconsistent language; unfamiliar terminology throughout 2 = Some overly complex, inconsistent, or unfamiliar terminology 3 = Language mostly unchanged from input; missed opportunities for simplification 4 = Some improvements in structure or terminology for clarity 5 = Clear, well-structured language aligned with expert audience

succinct: Economy of language. The summary is brief, to the point, and without redundancy. 1 = Too wordy across all assertions with redundancy in syntax and semantics 2 = More than one assertion has contextual semantic redundancy 3 = At least one assertion has contextual semantic redundancy or multiple syntactic assertions 4 = No syntactic redundancy and at least one could be more concise 5 = Fewest words possible with no redundancy

Figure 6: Full system prompt used for LLM-based evaluation of expert summary quality.

Task Instruction

When faced with a task, begin by identifying the participants who will contribute to solving the task. Then, initiate a multi-round collaboration process until a final solution is reached. The participants will provide critical comments and detailed suggestions whenever necessary to ensure the translation is accurate and easily understandable.
Here is a example:
Example Task: Translate the following radiology report into plain language that someone without medical training can easily understand.

Radiology Report:

- Left PICC tip appears to terminate in the distal left brachiocephalic vein.
- Mild pulmonary vascular congestion.
- Interval improvement in aeration of the lung bases with residual streaky opacity likely reflective of atelectasis.
- Interval resolution of the left pleural effusion.

Participants: AI Assistant (you); Medical Doctor; Patient Advocate; Radiologist; Nurse; Respiratory Therapist

Start collaboration!
Radiologist: The PICC line is correctly placed in a large vein in the chest. There's mild fluid buildup in the lung vessels. The lower parts of the lungs are improving but still show some areas not fully expanded. The fluid around the left lung has resolved.
Medical Doctor: Yes, let's break it down point by point to ensure accuracy.
Respiratory Therapist: We should explain "atelectasis" as areas of the lung not fully inflated and "pulmonary vascular congestion" as extra fluid in the lung's blood vessels.
Nurse: We also need to simplify terms like "PICC," and "brachiocephalic vein". Clarify that the "PICC line" is a tube inserted into the arm that reaches a large vein near the heart.
Respiratory Therapist: Also, explain "pulmonary vascular congestion" and "pleural effusion" in simple terms related to lung function.
Patient Advocate: And let's use bullet points for clarity and make sure the language is easy to understand.
AI Assistant (you): Understood. Taking into account everyone's input and my expertise, here's the translation:

- The tube in your arm is correctly placed in a large vein near your heart.
- There's a small amount of extra fluid in the blood vessels of your lungs.
- The lower parts of your lungs are getting better, but some small areas aren't fully expanded yet.
- The fluid that was around your left lung has cleared up.

Medical Doctor: This is medically accurate and covers all important aspects.
Radiologist: Agreed, the translation reflects the findings accurately.
Nurse: The explanations are clear and patient-friendly.
Respiratory Therapist: Good job explaining lung-related terms in simple language.
Patient Advocate: It's easy to understand and uses bullet points as requested.
Finish collaboration!
Final answer:

- The tube in your arm is correctly placed in a large vein near your heart.
- There's a small amount of extra fluid in the blood vessels of your lungs.
- The lower parts of your lungs are getting better, but some small areas aren't fully expanded yet.
- The fluid that was around your left lung has cleared up.

Now, identify the participants and collaboratively solve the following task step by step. After **Finish collaboration!**, remember to conclude your final solution in this exact format: **"Final answer: [Your solution here]"**

Task: Translate the following radiology report into patient-friendly plain language that someone without medical training can easily understand.

Radiology Report: "{radiology_report}"

Figure 7: Layperson Summarization Generation prompt