Improving Expert Radiology Report Summarization by Prompting Large Language Models with a Layperson Summary

Anonymous ACL submission

Abstract

Radiology report summarization (RRS) is crucial for patient care, requiring concise "Impressions" from detailed "Findings." This paper introduces a novel prompting strategy to enhance RRS by first generating a layperson summary. This approach normalizes key observations and simplifies complex information using non-expert communication techniques inspired by doctor-patient interactions. Combined with few-shot in-context learning, this method improves the model's ability to link general terms to specific findings. We evaluate this approach on the MIMIC-CXR, CheXpert, and MIMIC-III datasets, benchmarking it against 7B/8B parameter state-of-the-art open-source large language models (LLMs) like Llama-3.1-8B-Instruct. Our results demonstrate improvements in summarization accuracy and accessibility, particularly in out-of-domain tests, with improvements as high as 5% for some metrics.

1 Introduction

Radiology reports summarization (RRS) is an interesting task to explore 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" and images in radiology reports. These reports, critical for patient diagnosis, treatment planning, and maintaining comprehensive records, are written by radiologists based on medical imaging techniques like X-rays, CT scans, MRI scans, and ultrasounds. The "Findings" section details objective observations from the imaging, while the "Impressions" section provides the radiologist's professional interpretation and diagnostic conclusions.

In biomedical applications, the effectiveness of large language models (LLMs) models largely depends on their adaptation through domain- and task-specific fine-tuning (Singhal et al., 2023). LLMs have shown remarkable proficiency in natural language understanding and generation, making them adaptable to various tasks. However, fine-tuning large models like GPT-3, with billions of parameters, requires substantial computational resources and high costs. To address these issues, researchers have shifted towards more efficient techniques like parameter-efficient fine-tuning (PEFT) and prompting (Van Veen et al., 2023a,b), leveraging existing model capabilities while reducing computational demands (Liu et al., 2022).

In contrast, prompting through in-context learning (ICL) (Brown et al., 2020; Dong et al., 2023) provides a practical alternative to extensive fine-tuning of LLMs. In ICL, relevant information is embedded directly within prompts, allowing LLMs to adapt to tasks with few-shot demonstrations (Lampinen et al., 2022) quickly. By carefully crafting these prompts, researchers can guide LLMs to generate accurate responses by providing clear context and examples. Techniques such as Retrieval-Augmented Generation (Wang et al., 2023b) can further improve this process. Prompting has also proven effective in converting complex radiological data into clear and concise summaries (Chen et al., 2023a). Moreover, Nori et al. (2023) found that combining ICL with explanations enhances the adaptation of general LLMs to specialized tasks, such as medical question answering, by integrating intermediate reasoning steps and thus improving problem-solving abilities (Zhang et al., 2023). However, generating explanations for summarization tasks is inherently more challenging compared to question-answering and traditional text classification.

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

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.

However, in real-world settings (e.g., in actual doctor-patient conversations), research indicates that scientific or technical knowledge can be effectively transferred to non-experts through communication techniques like reformulation and simplification, which simplifies complex information and uses straightforward language to enhance understanding (Gülich, 2003). Hence, inspired by effective doctor-patient communication methods, this paper proposes a novel prompting strategy that combines simplification techniques with ICL to enhance the performance of non-expert LLMs in specialized areas. This approach aims to improve model performance without needing costly fine-tuning (Nori et al., 2023; Zhang et al., 2023) by simplifying complex information and incorporating it through prompts before an expert summary is generated. The in-context examples have layperson/simplified language as part of them to help guide the model for a new example. From another perspective, we introduce a novel approach that first generates a layperson (non-expert) summary to *normalize* key observations. Radiologists often have distinct reporting styles, leading to variations in terminology and impacting the consistency of medical documentation (Yan et al., 2023). Additionally, the vast number of illnesses increases the variety of vocabulary encountered in reports. Normalizing terms in the layperson summary can better identify patterns between simplified summaries and detailed expert impressions, making it easier to link general terms to specific findings (Peter et al., 2024). For example, normalizing "pneumonia" and "bronchitis" to "infection of the lungs" helps the model recognize important concepts in the in-context examples, even if pneumonia is used in the test instance while bronchitis is used in the in-context examples. The LLM can then connect them back to the findings (summary).

Overall, this paper has threefold contributions:

1. We introduce a novel prompting approach inspired by doctor-patient communication techniques that generate a simplified (layperson) summary before the expert summary. This strategy, combines with a few-shot ICL with the layperson summary, enhances RRS using 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 one multimodal medical question summarization (MMQS) dataset (Ghosh et al., 2024). We also benchmark against open-source LLMs like Llama-3.1-8B-Instruct (AI@Meta, 2024) for comprehensive comparison.
3. We conduct a comprehensive analysis to determine the optimal modality for ICL. We also examine the required number of examples and the impact of layperson summaries on impressions and evaluate model performance on inputs of different lengths.[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

[1] See the appendix for complete analysis.

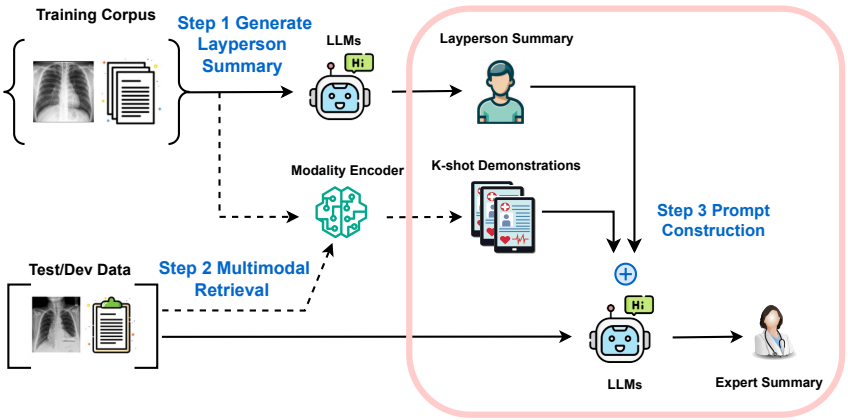

Figure 1: Overview of the LaypersonPrompt Framework. First, we generate layperson summaries from the training corpus using LLMs prompting. Then, for a test input, we use multimodal retrieval to find relevant examples. Finally, we incorporate these layperson summaries into the prompt, applying patient-doctor communication techniques to improve the model's reasoning.

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 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 demonstrate 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). Inspired by these findings, we explore whether presenting expert-level information in 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) "multimodal demonstration retrieval," which is how we generate 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.

Step 1: Layperson Summarization of the Training Dataset. Layperson summarization involves converting 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

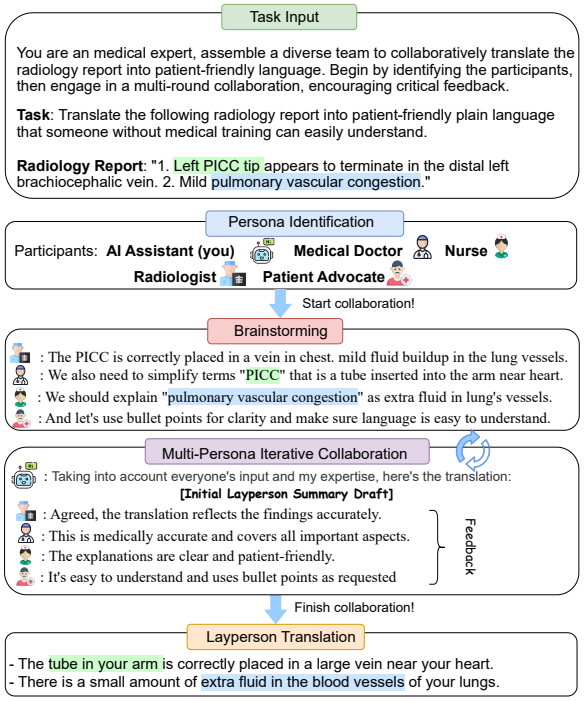

Figure 2: Step 1: Layperson summarization of the Training Dataset. An illustration of the layperson summary prompt used to generate layperson summaries for training examples. Disease observations are highlighted in different colors. The illustration shows a single example, with Instruction and Response sections repeated multiple times using few-shot in-context examples.

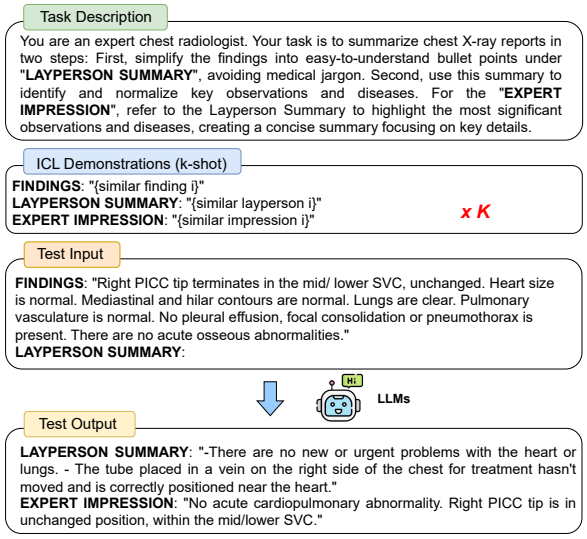

Figure 3: Step 3: Final Expert Summary Prompt Construction. Example of LaypersonPrompt. This is the final prompt after finding in-context examples to generate the final expert summary (i.e., the Impression section).

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., 2024a). 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 A.3 for a complete example of what the output looks like.

Step 2: Multimodal Demonstration Retrieval Another essential part of our system is finding similar examples in the training dataset for each test example to use as in-context examples. In our approach, we focus on substantially improving the performance of LLMs with a few well-chosen examples to generate more accurate and standardized summaries. Selecting the right examples is a critical task in few-shot learning, as it greatly affects the effectiveness of the LLMs. To ensure the selection of the most relevant examples, we follow the multimodal retrieval procedure outlined by Wang et al. (2023b), which is fine-tuned with radiology reports and chest X-ray images. According to their approach, we retrieve the top-k similar radiology report based on different modalities, i.e., chest X-ray images, text findings, and multi-modal data (combining findings and images) from a medical corpus using a pre-trained multi-modal encoder. Then, we include the findings and impressions of the top k of the most similar report as input in our final prompt.

Formally, given an input instance x_i consisting

of a text input w and image m, our goal is to retrieve the most similar examples $\{x_1, \ldots, x_{\mathcal{N}(x_i)}\}$, where $\mathcal{N}(x_i)$ represents the top k similar examples to x_i. To achieve this, we employ a multimodal image-text retrieval model that uses separate encoders for text and image modalities alongside a multimodal encoder for integrating their embeddings. Specifically, the image is processed through a pre-trained Vision Transformer (ViT) model (Dosovitskiy et al., 2021) to generate image embeddings. Since some findings correspond to multiple images, we average all image embeddings corresponding to the same findings. Next, we adapt a pre-trained Transformer encoder-decoder model, such as Clinical-T5 (Lehman and Johnson, 2023), to handle multimodal inputs. Specifically, we pass the findings as input to the T5 encoder and initialize its hidden state with the averaged image embeddings. The final EOS token from the T5 encoder is used as the multimodal embeddings. Note that this model cannot be used as-is with the initial pre-trained models. Instead, we train this model where the T5 encoder outputs are passed to the T5 decoder to generate the impressions. After training the joint model, we remove the decoder, and only the embeddings are used later.

Step 3: Expert Summary 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 multimodal demonstration 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 specifies that the model should create a layperson summary followed by an expert impression. Detailed guidelines are provided for generating both the layperson summary and the expert impression. It is important to note that the layperson summary is generated as part of this prompt for the input instance before generating the expert impression. The prompt defined in Step 1 is only used for the training examples. Next, given the input instance's Findings text and radiology image, we use the same multi-modal encoder and retrieval approach described in Step 2 to find relevant in-context examples from the training dataset. We generate a sequence of up to 32 in-context demonstrations. After identifying the relevant training examples, we append each training instance's Findings, layperson summary, and Impression to generate the sequence of in-context examples. Finally, we append the Findings section of the text instance and the string "Layperson Summary:". The model will first generate the layperson summary followed by the final expert Impression.

Why does generating a layperson summary before the expert impression work? Models can produce general information (e.g., "Infection of the lungs" for "pneumonia") in the layperson summary, which helps to standardize the content in the Findings before creating the Impression. This means different illnesses can be simplified to the same concept (e.g., "bronchitis" can also be simplified to "Infection of the lungs"). The idea is that the model can find common patterns in these general (layperson) expressions that correlate with the expert Impression, as long as the Findings have similar content. After generating the layperson summary, the model only needs to connect the general terms in the summary to the specific details in the Findings, similar to coreference resolution. Without the layperson summary, the model must directly find patterns in the more varied Findings section, making the task more complex.

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. Additionally, we incorporate an out-of-institution multimodal test set of 1000 samples from the Stanford

	Model	BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	6.24	22.63	46.53	67.21	19.88	32.50
	OpenChat-3.5-0106-gemma	6.30	22.14	43.48	66.53	17.06	31.10
	Ministral-8B-Instruct-2410	5.22	21.79	42.55	67.94	18.22	31.14
Few-Shot	Llama-3.1-8B-Instruct	8.39	31.32	50.94	68.66	28.93	37.65
	OpenChat-3.5-0106-gemma	10.67	29.51	49.02	63.27	27.15	35.92
	Ministral-8B-Instruct-2410	10.81	28.60	51.50	68.16	24.86	36.79
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**11.85**	30.24	**52.77**	68.16	27.42	**38.09**
	OpenChat-3.5-0106-gemma	**11.02**	**30.29**	**51.99**	**65.22**	26.49	**37.00**
	Ministral-8B-Instruct-2410	**11.34**	**30.19**	**52.88**	**68.95**	**27.33**	**38.14**

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).

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	2.39	23.38	48.10	71.94	9.08	30.98
	OpenChat-3.5-0106-gemma	3.70	25.07	47.54	63.35	8.47	29.63
	Ministral-8B-Instruct-2410	4.62	27.04	47.21	70.51	10.05	31.89
Few-Shot	Llama-3.1-8B-Instruct	4.27	26.94	48.52	73.13	9.63	32.50
	OpenChat-3.5-0106-gemma	3.81	22.61	45.61	62.43	8.94	28.68
	Ministral-8B-Instruct-2410	6.01	28.83	51.03	71.79	11.69	33.87
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**7.44**	**29.62**	**54.40**	**74.41**	**11.14**	**35.40**
	OpenChat-3.5-0106-gemma	**5.32**	**26.31**	**49.39**	**65.41**	**10.02**	**31.29**
	Ministral-8B-Instruct-2410	**7.84**	**30.11**	**52.57**	**73.95**	**11.71**	**35.24**

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).

hospital(CheXpert) (Irvin et al., 2019) to assess the 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 use Llama-3.1-8B-Instruct (AI@Meta, 2024), Ministral-8B-Instruct-2410 (Jiang et al., 2023), and OpenChat-3.5-0106-gemma (Wang et al., 2023a) in our experiments to compare model performance.

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 overlap 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 show the performance of Zero-Shot prompting, Few-Shot prompting, and our Few-Shot + Layperson prompting strategies for the radiology reports summarization task on the MIMIC-CXR dataset. The Few-Shot + Layperson method mimics doctor-patient communication by creating a simplified summary for laypeople before generating the expert summary. We find that incorporating the layperson intermediary step yields consistent improvements over the standard Few-Shot approach: for example, Llama-3.1-8B-Instruct's BLEU4 score rises from 8.39 to 11.85, OpenChat-3.5-0106-gemma's ROUGE-L and F1-cheXbert scores increase from 29.51 and 63.27 to 30.29 and 65.22 respectively, and Ministral-8B-Instruct-2410 exhibits enhancements with its BERTScore improving from 51.50 to 52.88 and F1-cheXbert from 68.16 to 68.95. Overall, the averaged performance across all models and metrics indicates that the

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph	Average
Zero-Shot	Llama-3.1-8B-Instruct	6.79	21.84	44.00	52.82	19.30	28.95
	OpenChat-3.5-0106-gemma	6.65	20.83	44.50	51.46	16.64	28.02
	Ministral-8B-Instruct-2410	8.39	23.47	46.94	53.79	18.83	30.28
Few-Shot	Llama-3.1-8B-Instruct	7.23	23.45	46.84	49.82	22.20	29.91
	OpenChat-3.5-0106-gemma	10.24	22.59	45.27	51.13	19.82	29.81
	Ministral-8B-Instruct-2410	8.72	24.14	47.03	53.81	19.28	30.60
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**13.58**	**25.23**	**49.63**	**55.53**	**22.64**	**33.32**
	OpenChat-3.5-0106-gemma	**11.93**	**23.62**	**47.50**	**52.73**	**21.33**	**31.42**
	Ministral-8B-Instruct-2410	8.27	23.49	46.63	**69.01**	**19.57**	**33.39**

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	41.538
	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**

Table 4: Overall performance across masked and original findings with the Llama-3.1-8B-Instruct model. Results are for MIMIC-CXR. Bolded results highlight our framework's improvements over the traditional few-shot approach.

Few-Shot + Layperson strategy outperforms the conventional Few-Shot approach, highlighting the benefit of integrating layperson communication in enhancing the clarity and effectiveness of radiology report summarization.

On the Stanford Hospital test set in Table 2, the Few-Shot + Layperson prompting yields a respective increase in performance across multiple metrics. Ministral-8B-Instruct-2410 achieved the highest BLEU4 (7.84) and ROUGEL (30.11), while Llama-3.1-8B-Instruct led in BERTScore (54.40) and F1-cheXbert (74.41). OpenChat-3.5-0106-gemma also showed substantial improvements in ROUGEL (26.31 vs. 25.07) and BERTScore (49.39 vs. 45.61) compared to its Few-Shot performance. Moreover, the average scores computed across all metrics consistently increased under the Few-Shot + Layperson setting for every model. These results highlight the effectiveness of using layperson summaries to enhance model performance in summarizing radiology reports on out-of-domain dataset.

The results of the comparison on the MIMIC-III dataset are detailed in Table 3. Our model demonstrates robust performance, indicating its capability to generalize across varied medical datasets. Specifically, Llama-3.1-8B-Instruct saw increases in BLEU4 (13.58 vs. 7.23) and F1-RadGraph (25.23 vs. 23.45) when comparing the Few-Shot + Layperson method to the standard Few-Shot approach. In summary, across all three datasets, it is evident that the Few-Shot + Layperson method shows noticeable improvements, especially on the out-of-domain test set, with the overall average consistently outperforming other methods. Incorporating an intermediate layperson summary, which mimics patient–doctor communication, introduces a step for "easy-to-hard" reasoning. This approach enhances the model's accuracy and its ability to generalize across different datasets in medical imaging and report summarization.

Error Analysis and Discussion. We conducted an error analysis of the Llama-3.1-8B-Instruct model on the MIMIC-CXR valid dataset to compare two prompting strategies: the Few-Shot method and our Few-Shot + Layperson approach. The core idea behind this experiment is to determine whether guiding the model with simpler, more accessible language helps it handle complex or unfamiliar medical terminology more effectively. Our intuition is based on the observation that when a model encounters highly specialized or unknown terms, it may misinterpret the context or even refuse to process the request. We aim to steer the model's attention towards the underlying clinical context by embedding layperson explanations in the prompt rather than getting caught up in obscure jargon. This mirrors how humans often simplify complex information to enhance understanding.

In our experiment, we simulated real-world challenges by replacing key medical terms with random, nonsensical "gibberish" entities. Specifically, we used MedSpacy (Eyre et al., 2022) to identify medical entities. These were then replaced with random strings, e.g., "pleural effusions" can be re-

placed with "abcdefg." This method tests whether layperson instructions can guide the model to generate clear and concise summaries, even when confronted with entirely unfamiliar terminology. We hypothesized that, combined with the other context in the findings, the additional layperson summary would encourage the model to normalize difficult terms into simpler, plain language, improving its overall performance. We report the results of this study in Table 4. The section titled "Mask" shows the results of applying the baseline (Llama-3.1-8B-Instruct + Few-shot prompting) and our method (Llama-3.1-8B-Instruct + Few-shot + Layperson) on the modified examples, where entities in the initial findings have been replaced with gibberish. We also report the original results on the unmodified data for reference. Our results show that the Few-Shot + Layperson approach consistently outperforms the standard Few-Shot method. When modifying words, the few-shot method performs poorly compared to the original data (37.55 vs. 6.67 ROUGE-L). However, our method can more robustly adapt to the modified vocabulary (37.60 vs. 25.05 ROUGE-L). This is particularly evident in the masked setting, where key medical terms are replaced with gibberish. In these cases, the Few-Shot model's performance drops substantially, highlighting its struggle to handle unfamiliar terminology without additional guidance.

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:

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."

Notably, when encountering an unknown term, the Few-Shot model often rephrases details, resulting in longer summaries that sometimes repeat findings or even include hallucinated information. In contrast, our Layperson approach actively guides the model to simplify complex information. As shown in the following example, the Few-Shot model failed to capture the key observation and even fabricated a term like "xenosign," our Few-Shot + Layperson model successfully generated the summary "No acute cardiopulmonary process. Normal heart size," which closely aligns with the ground truth. An example of this is found below:

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."

5 Conclusion

This paper introduces a novel prompting approach inspired by doctor-patient communication techniques. By first generating a simplified (layperson) summary before creating the expert summary and combining this with few-shot in-context learning, we aim to improve the summarization of radiology reports using general LLMs. Evaluations across three datasets (MIMIC-CXR, CheXpert, and MIMIC-III) show that this method improves performance, especially in out-of-domain tests.

However, this approach faces challenges due to the 7B model's computational demands and context token limitations, especially with complex medical reports. Future work will optimize token usage and explore larger models with expanded context. By applying effective doctor-patient communication principles, our method aims to improve non-expert LLM performance in specialized fields without extensive fine-tuning.

6 Limitation

While our approach shows improvements in radiology report summarization (RRS), several limitations must be considered. First, the evaluation metrics used, such as ROUGE-L, do not always correlate well with human evaluations, necessitating cautious interpretation of the results (Wang et al., 2024b). Our study primarily relies on these automated metrics, which can overlook important nuances that human experts might catch. The absence of comprehensive human evaluations further limits the assessment of practical effectiveness. Incorporating detailed evaluations by human experts is crucial for accurately measuring model performance in real-world clinical settings in future research, as human assessments provide insights into the clinical relevance and accuracy of summaries that automated metrics may miss.

Additionally, the use of 7B parameter open-source models may not be optimal. More powerful closed models, like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023), often perform better in summarization tasks. Including results from these advanced models could provide a more comprehensive comparison and potentially challenge the necessity of the intermediate layperson summary step. Furthermore, the computational demands and context token limitations of the 7B model present significant challenges, particularly with longer and more complex medical reports. This restricts the model's ability to process extensive and detailed information effectively. Differences in the quality and consistency of radiology reports from different datasets can also affect performance due to inconsistencies in terminology and reporting styles. Moreover, the current interaction between humans and non-expert LLMs can be improved. Incorporating communication techniques similar to doctor-patient interactions will enhance the human-AI experience by making complex information more accessible and understandable. This improvement aims to make LLMs more practical and effective for expert-level tasks in various areas, bridging the gap between specialized knowledge and everyday understanding.

7 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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

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.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

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.

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.

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

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

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.

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.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *ArXiv preprint*, abs/2309.11235.

Tongnian Wang, Xingmeng Zhao, and Anthony Rios. 2023b. 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. 2024a. 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.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. 2024b. Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. *ArXiv preprint*, abs/2404.00998.

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.

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.

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), Ministral-8B-Instruct-2410 (Jiang et al., 2023), and OpenChat-3.5-0106-gemma (Wang et al., 2023a). 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, OpenChat-3.5-0106-gemma, and Ministral-8B-Instruct-2410 are reported to perform strongly in common sense reasoning and problem-solving ability (Zhu et al., 2023). OpenChat-3.5-0106-gemma is built on the highest-performing Gemma model with conditioned reinforcement learning fine-tuning. 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, Ministral-8B-Instruct-2410, and OpenChat-3.5-0106-gemma. In cases where prompts exceeded this length, they were truncated from the beginning, ensuring that 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 5 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.

A.2 Multimodal 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 refer-

		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 5: Average Token of Prompts.

		BLEU4	ROUGEL	BERTScore	Average
Zero-Shot	Llama-3.1-8B-Instruct	3.26	16.87	32.86	17.66
	OpenChat-3.5-0106-gemma	5.28	17.46	31.15	17.96
	Ministral-8B-Instruct-2410	5.82	23.36	40.55	23.24
Few-Shot	Llama-3.1-8B-Instruct	7.29	38.38	55.69	33.79
	OpenChat-3.5-0106-gemma	19.86	43.03	59.47	40.79
	Ministral-8B-Instruct-2410	13.88	35.68	53.86	34.47
Few-Shot + Layperson	Llama-3.1-8B-Instruct	**16.71**	**39.77**	**58.75**	**38.41**
	OpenChat-3.5-0106-gemma	17.08	42.90	59.33	39.77
	Ministral-8B-Instruct-2410	**14.96**	**38.23**	**55.54**	**36.24**

Table 6: Performance of models on Multimodal Medical Question Summarization (MMQS) Dataset.

ence various body parts (e.g., skin, eyes, ears). As shown in Table 6, we observe similar trends across the models. Notably, the Few-Shot + Layperson approach also works effectively for simple healthcare summarization in this context.

Across all settings, Ministral-8B-Instruct-2410 achieves the highest performance in the zero-shot setting with an average score of 23.24, outperforming both Llama-3.1-8B-Instruct (17.66) and OpenChat-3.5-0106-gemma (17.96). This suggests that Ministral is better suited for out-of-the-box summarization without additional context. However, absolute performance remains low across all zero-shot models, indicating the difficulty of the task without demonstrations.

In the few-shot setting, OpenChat-3.5-0106-gemma achieves the best overall performance with an average score of 40.79, outperforming Ministral (34.47) and Llama-3.1 (33.79). This improvement highlights the effectiveness of in-context learning, particularly for models with strong generalization capabilities. Notably, OpenChat-3.5 exhibits a substantial boost in BLEU4 (19.86) and ROUGEL (43.03), suggesting its ability to generate more lexically and structurally faithful summaries.

Introducing layperson-focused prompting further enhances performance. Llama-3.1-8B-Instruct shows the most significant improvement, increasing its average score from 33.79 to 38.41, with BLEU4 improving from 7.29 to 16.71 and BERTScore reaching 58.75, its highest across all settings. Chat-3.5-0106-1268-gemma still maintains the highest overall performance (39.77 average score), but the gap between models narrows with the layperson prompt. The results suggest that simplifying language to align with non-expert understanding benefits model output fluency and coherence.

Overall, these findings indicate that few-shot prompting substantially improves model performance, with Chat-3.5-0106-1268-gemma being the most effective model in this setting. Moreover, the few-shot + layperson approach provides further gains, demonstrating its utility for medical question summarization tasks where clarity and accessibility are crucial.

A.3 Layperson Summary Prompting

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}"