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 Meta-Llama-3-8B-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., 2022) 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., 2022). 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., 2022). 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., 2022). 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., 2022) 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).

In summary, this paper makes the following 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 benchmark against open-source LLMs like Meta-Llama-3-8B-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.

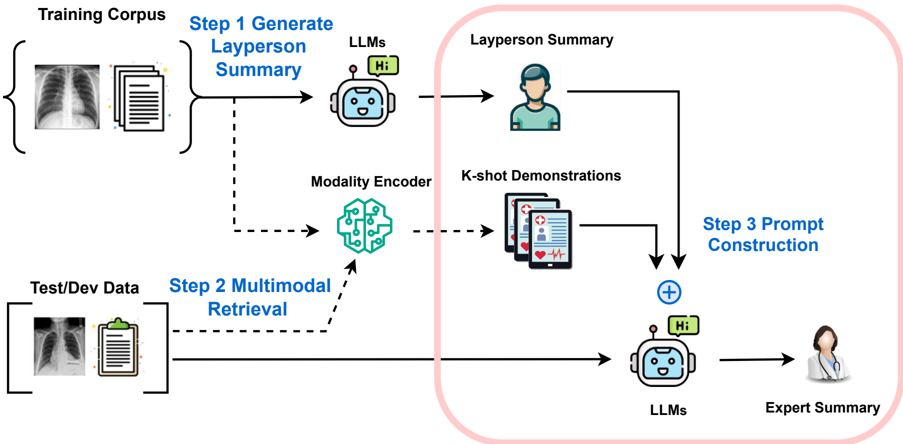

Figure 1: Overview of the LaypersonPrompt Framework. First, we generate layperson summaries from the training corpus using LLMs prompting. Then, for a test input, we use multimodal retrieval to find relevant examples. Finally, we incorporate these layperson summaries into the prompt, applying patient-doctor 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., 2023; Yao et al., 2023b; Nori et al., 2023; Ma et al., 2023). For example, Shi et al. (2023) 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,

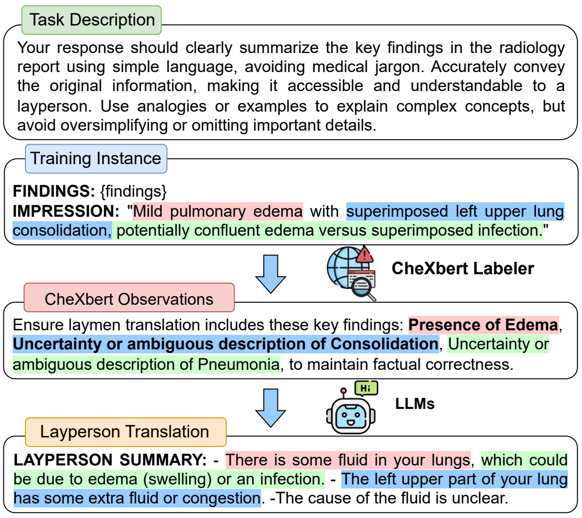

Figure 2: Step 1: Layperson Summarization of the Training Dataset. An illustration of the layperson summary prompt used to generate layperson summaries for training examples. Disease observations are highlighted in different colors. The illustration shows a single example, with Instruction and Response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

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 use a zero-shot prompting strategy enhanced with metadata from an external tool. Specifically, we employ the CheXbert labeler (Smit et al., 2020b) to extract important medical observations from radiology impressions (e.g., "No Finding", "Pneumonia", "Cardiomegaly", etc.). Using CheXbert's output, we create an additional instruction for the language model that includes these key concepts. The exact form of the prompt is shown in Figure 2. This prompt integrates the Task Instruction, Findings, Impression, and the extracted CheXbert concepts. 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.

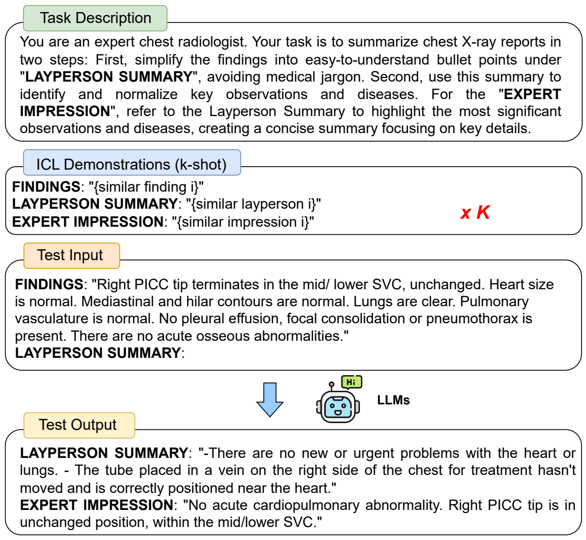

Figure 3: Step 3: Final Expert Summary Prompt Construction. Example of LaypersonPrompt. This is the final prompt after finding in-context examples to generate the final expert summary (i.e., the Impression section).

Step 2: Multimodal Demonstration Retrieval A major feature 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., 2020) 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 In this study, we evaluate our prompting method on two radiology reports summarization datasets. The MIMIC-III summarization dataset, as introduced by (Johnson et al., 2016; Chen et al., 2023), 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. On the other hand, 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 hospital (Irvin et al., 2019) to assess the out-of-domain generalization of models trained on MIMIC-CXR. We use OpenChat-3.5-7B (Wang et al., 2023a), Starling-LM-7B (Zhu et al., 2023), and Meta-Llama-3-8B-Instruct (AI@Meta, 2024) 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*

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph
	OpenChat-3.5-7B	3.98	21.74	42.74	64.98	18.34
Zero-Shot	Starling-LM-7B	3.64	21.28	42.29	64.63	17.93
	Meta-Llama-3-8B-Instruct	5.19	23.56	40.99	66.65	23.56
	OpenChat-3.5-7B	8.24	27.44	45.86	67.00	26.90
Few-Shot	Starling-LM-7B	6.79	25.85	44.71	66.76	25.13
	Meta-Llama-3-8B-Instruct	6.33	25.81	40.19	65.72	30.13
	OpenChat-3.5-7B	8.11	27.68	44.62	65.71	26.80
Few-Shot + Chexbert	Starling-LM-7B	6.29	25.57	42.96	63.56	24.09
	Meta-Llama-3-8B-Instruct	9.20	28.25	44.63	67.23	**30.48**
	OpenChat-3.5-7B	8.96	28.46	45.35	67.00	27.90
Few-Shot + Layperson	Starling-LM-7B	8.35	26.97	44.93	66.29	26.94
	Meta-Llama-3-8B-Instruct	**9.36**	**29.03**	**46.91**	**68.64**	29.96

Table 1: Overall performance across the four prompts on the MIMIC CXR in-domain test.

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph
	OpenChat-3.5-7B	2.22	25.14	47.10	68.95	10.68
Zero-Shot	Starling-LM-7B	2.18	24.61	46.49	70.36	10.50
	Meta-Llama-3-8B-Instruct	2.01	23.69	42.53	68.76	10.99
	OpenChat-3.5-7B	5.23	27.43	48.00	70.32	12.35
Few-Shot	Starling-LM-7B	4.66	26.77	47.20	70.68	11.64
	Meta-Llama-3-8B-Instruct	3.37	22.09	39.35	66.49	11.22
	OpenChat-3.5-7B	5.43	26.50	44.95	69.80	12.31
Few-Shot + Chexbert	Starling-LM-7B	3.40	23.93	44.12	64.90	10.74
	Meta-Llama-3-8B-Instruct	3.79	24.75	42.52	70.05	11.79
	OpenChat-3.5-7B	**7.74**	28.71	**48.04**	71.28	13.15
Few-Shot + Layperson	Starling-LM-7B	7.01	28.90	48.02	71.02	12.93
	Meta-Llama-3-8B-Instruct	7.47	**29.03**	47.29	**71.91**	**13.63**

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.

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., 2020a), 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, Few-Shot + Chexbert prompting, and our Few-Shot + Layperson prompting strategies for the radiology reports summarization task on the MIMIC-CXR dataset. The Few-Shot + Chexbert method adds disease keywords to help the model focus. In contrast, the Few-Shot + Layperson method mimics doctor-patient communication by creating a simplified summary for laypeople before generating the expert summary. We find that the Few-Shot + Layperson method yielded the best results overall. Meta-Llama-3-8B-Instruct achieved the highest scores in BLEU4 (9.36), ROUGEL (29.03), BERTScore (46.91), and F1-cheXbert (68.64), and strong performance in F1-RadGraph (29.96). OpenChat-3.5-7B and Starling-LM-7B also showed significant improvements with Few-Shot + Layperson, notably in BLEU4 and F1-RadGraph. Specifically, on OpenChat-3.5-7B, ROUGE-L, and F1-RadGraph, there were respective increases of 0.78 and 1.10 compared to not using the layperson summary. For Starling-LM-7B, these metrics rise by 1.12 and

		BLEU4	ROUGEL	BERTScore	F1-cheXbert	F1-RadGraph
Zero-Shot	OpenChat-3.5-7B	4.61	19.85	43.02	52.06	21.41
	Starling-LM-7B	4.51	19.52	42.57	51.77	21.19
	Meta-Llama-3-8B-Instruct	5.61	20.34	41.70	51.87	24.43
Few-Shot	OpenChat-3.5-7B	8.02	22.33	45.56	52.71	23.37
	Starling-LM-7B	7.95	21.98	45.05	52.49	23.01
	Meta-Llama-3-8B-Instruct	6.25	20.03	38.75	47.54	24.76
Few-Shot + Chexbert	OpenChat-3.5-7B	8.05	21.94	**45.61**	51.03	24.70
	Starling-LM-7B	9.28	22.43	44.93	49.94	22.05
	Meta-Llama-3-8B-Instruct	7.39	21.36	40.76	48.06	24.40
Few-Shot + Layperson	OpenChat-3.5-7B	8.62	**22.95**	45.31	**52.81**	24.37
	Starling-LM-7B	10.02	22.70	45.14	51.83	24.32
	Meta-Llama-3-8B-Instruct	**10.03**	21.49	45.29	50.78	**24.99**

Table 3: Overall performance across the four prompts on the MIMIC III.

2.85, respectively. These results suggest incorporating a layperson summary can enhance model performance in summarizing radiology reports.

On the Stanford Hospital test set in Table 2, the Few-Shot + Layperson prompting yields a respective increase in performance across multiple metrics. OpenChat-3.5-7B achieved the highest BLEU4 (7.74) and BERTScore (48.04), while Meta-Llama-3-8B-Instruct led in ROUGEL (29.03), F1-cheXbert (71.91), and F1-RadGraph (13.63). Starling-LM-7B also showed substantial improvements in ROUGEL (28.90 vs. 23.93) and BERTScore (48.02 vs. 44.12) compared to Few-Shot + Chexbert. These results highlight the effectiveness of using a layperson summary to enhance model performance in summarizing radiology reports on the 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, Meta-Llama-3-8B-Instruct saw increases in BLEU4 (10.03 vs. 7.39) and F1-RadGraph (24.99 vs. 24.40) compared to Few-Shot + Chexbert. 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. 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. We conducted an error analysis of the OpenChat-3.5-7B model on the MIMIC-CXR test dataset, comparing the Few-Shot + Layperson prompting strategy to Few-Shot prompting using multimodal embeddings. We analyzed performance trends across different impression lengths using ROUGE-L for text similarity and F1-RadGraph for entity accuracy and completeness of the generated radiology reports. The results are shown in Figure 4. We found that ROUGE-L scores decrease with longer impressions while F1-RadGraph scores increase. This suggests that while the model's text similarity drops with longer impressions, its accuracy in identifying specific medical entities improves. The F1-RadGraph metric benefits from the richer context and greater detail in longer impressions, allowing the model to recognize better and generate relevant diseases or observations. Our layperson summary approach normalizes terminology and links general terms to specific findings, enhancing the model's accuracy. This approach is similar to how doctors explain complex medical information to patients: by simplifying and clarifying terms, they ensure understanding and accuracy. This technique helps the model identify and generate important entities better, leading to higher F1-RadGraph scores for longer impressions.

Conversely, the decrease in ROUGE-L scores for longer impressions may be due to the increased complexity and diversity of information. Longer impressions often include more nuanced and varied details, which may not exactly match the ground truth in phrasing or specific word choice. As a result, the text similarity measured by ROUGE-L tends to be lower for longer impressions. This metric is sensitive to the exact wording and word sequence, so the greater length and complexity can result in more divergence from the reference text.

Interestingly, when evaluated with the ROUGE-L metric, the Few-Shot + Layperson prompting approach performs best on shorter texts (29.11 vs. 27.41). For example, consider the short impression:

Ground Truth Short Impression: "No acute cardiopulmonary process."

Layperson Summary: "The lungs look normal with no issues like infection or fluid build-up. The heart and nearby structures are normal, and there are no broken bones."

Few-Shot: "No acute thoracic pathology."

Few-Shot + Layperson (Our): "No acute cardiopulmonary process."

The example above provides a qualitative analysis of the models' performance, with given findings. The Few-Shot method summarized the findings as "No acute thoracic pathology," which is too broad compared to the precise ground truth "No acute cardiopulmonary process." However, our method accurately matched the ground truth by first creating a layperson summary and then summarizing it, ensuring clarity and precision. This approach's effectiveness in capturing short, concise statements makes it better at closely matching the ground truth.

Despite the increasing complexity, the Few-Shot + Layperson prompting continues to outperform Few-Shot prompting in longer impressions. An example of a long impression is:

Ground Truth Long Impression: "Right basilar opacity is probably atelectasis, but could represent early or developing pneumonia in the appropriate clinical setting."

Layperson Summary: "The right lower part of the lung has a subtle cloudy area, which might be a small lung collapse but could also be early or developing pneumonia, depending on the patient's condition."

Few-Shot: "Normal cardiomegaly with subtle lung base opacity, possibly atelectasis or early pneumonia."

Few-Shot + Layperson (Our): "1. Subtle opacity at the right lung base, possibly representing atelectasis, but could be early or developing pneumonia in the appropriate clinical setting."

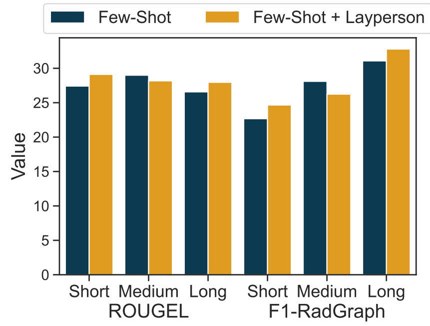

Figure 4: Error Analysis on MIMIC-CXR Test Dataset: Performance Comparison of OpenChat-3.5-7B Model across Different Impression Lengths.

For long impression, the Few-Shot method is less precise than the ground truth, adding unnecessary details like "Normal cardiomegaly" and missing the position terms "right". In contrast, our method simplifies complex findings into layperson terms and then translates them back into accurate expert summaries. For example, "Right basilar opacity is probably atelectasis, ... early or developing pneumonia" becomes "The right lower lung looks cloudy, likely a small collapse or early pneumonia." This layperson summary is then accurately converted to "Subtle opacity at the right lung base, possibly atelectasis or early pneumonia," ensuring clarity and precision. The improvement with longer texts is likely due to the extra context they provide, similar to detailed doctor-patient explanations.

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 computational demands and context token limitations of the 7B model, particularly with longer, more complex medical reports. Future work will focus on optimizing token usage within these constraints and exploring larger models with expanded context capacities. By leveraging the principles of effective doctor-patient communication, our method aims to enhance non-expert LLMs performance in specialized fields without requiring 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., 2024). 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 arXiv: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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2022. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. *arXiv preprint arXiv:2211.08584*.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023. 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. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

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 arXiv:2303.16416*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

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.

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.

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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

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.

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 arXiv: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 Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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 Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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.

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 arXiv: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.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

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. 2020a. 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.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020b. 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.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y 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.

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.

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 arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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, et al. 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.

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 arXiv: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 arXiv: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.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. 2024. Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv: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 arXiv:1910.03771*.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. 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.

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 arXiv: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*.

Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, et al. 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*, pages 1–22.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

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.

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: OpenChat-3.5-7B (Wang et al., 2023a), Starling-LM-7B (Zhu et al., 2023), Meta-Llama-3-8B-Instruct (AI@Meta, 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 OpenChat-3.5-7B, Starling-LM-7B, and Meta-Llama-3-8B-Instruct are reported to perform strongly in common sense reasoning and problem-solving ability (Zhu et al., 2023). OpenChat-3.5-7B is built on the Mistral 7B with conditioned reinforcement learning fine-tuning, and Starling-LM-7B is built on OpenChat-3.5-7B with reinforcement learning from AI feedback. 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. 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 OpenChat-3.5-7B model and Meta-Llama-3-8B-Instruct, the best performance is achieved with 32 examples for both Few-Shot and Few-Shot + Layperson prompting methods. In contrast, the Starling-LM-7B model exhibits optimal performance with 32 examples when using the Few-Shot prompt and 24 examples for the Few-Shot + Layperson prompt. 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 OpenChat-3.5-7B, Starling-LM-7B and Meta-Llama-3-8B-Instruct. 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.

A.2 Discussion and Model Analysis

A natural question that arises is, "Does integrating a larger number of examples in Few-Shot + Layperson prompting lead to better overall performance?". To answer this question, we explore the relationship between performance and the number of examples integrated. To better quantify the contributions of different components in our model, we conducted ablation studies focusing on various prompt types and modality embeddings for the radiology reports summarization task. Using the MIMIC-CXR validation dataset, we evaluated the performance of three models, OpenChat-3.5-7B, Starling-LM-7B, and Meta-Llama-3-8B-Instruct across a range of configurations. Our analysis focuses on understanding the effectiveness of embedding matches for different modalities (including image, text, and multimodal), as well as determining the optimal number of examples needed for effective summarization. The results of these ablations on the MIMIC-CXR validation set are presented in Figure 5, Figure 6, and Figure 7. Specifically, we note that Few-Shot + Layperson prompting with multimodal embedding matched examples slightly outperforms the image and text embedding matched ones. For all OpenChat-3.5-7B, Starling-LM-7B, and Meta-Llama-3-8B-Instruct employing the LaypersonPrompt demonstrates performance enhancements compared to the original prompt.

Furthermore, as we increase the number of examples, the performance continues to rise, demonstrating that prompting the model with more in-context examples improves performance. However, we can also observe a slight performance decrease in some cases after reaching 24 examples. These findings suggest that while multimodal embeddings provide a robust framework for summarization, there is a complex relationship between the number of examples and performance gains. Our studies highlight the importance of multimodal context and suggest a diminishing return for additional examples in text and image modalities beyond a certain point. This insight is critical for optimizing the efficiency and accuracy of our summarization model when processing radiology data.

Table 4 shows the prompt lengths corresponding to various numbers of examples used in our study. We aim to explore how the length of prompts affects model performance. Initially, models with shorter context lengths were explored, like LLaMA-2-7B (Touvron et al., 2023), but their performance in summarizing radiology reports was limited due to context length constraints of 4,096 tokens. Because these limitations significantly impacted their ability to perform in-context learning effectively, these models were not chosen for our study. Instead, models capable of processing more extended contexts, like OpenChat-3.5-7B, up to 8,192 tokens, were selected to handle better the extensive information needed for accurate radiology report summarization.

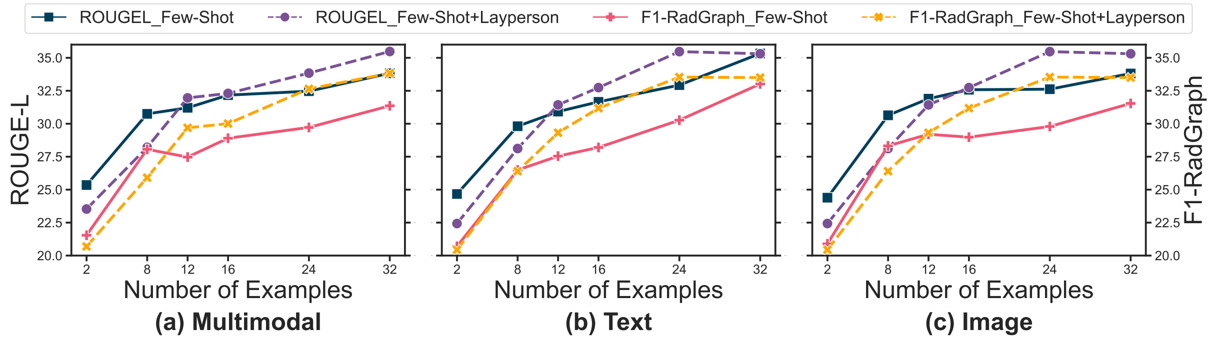

Figure 5: Validation results vs. the number of in-context examples across various prompt types and modality embeddings on OpenChat-3.5-7B.

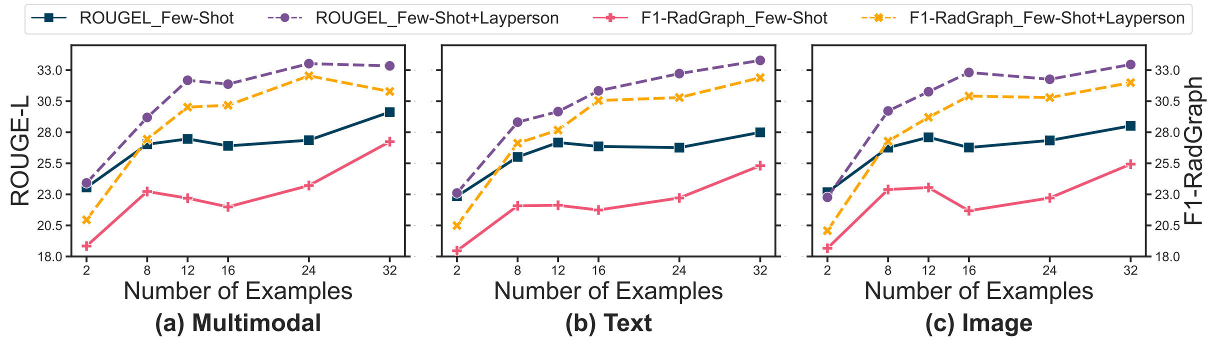

Figure 6: Validation results vs. the number of in-context examples across various prompt types and modality embeddings on Starling-LM-7B.

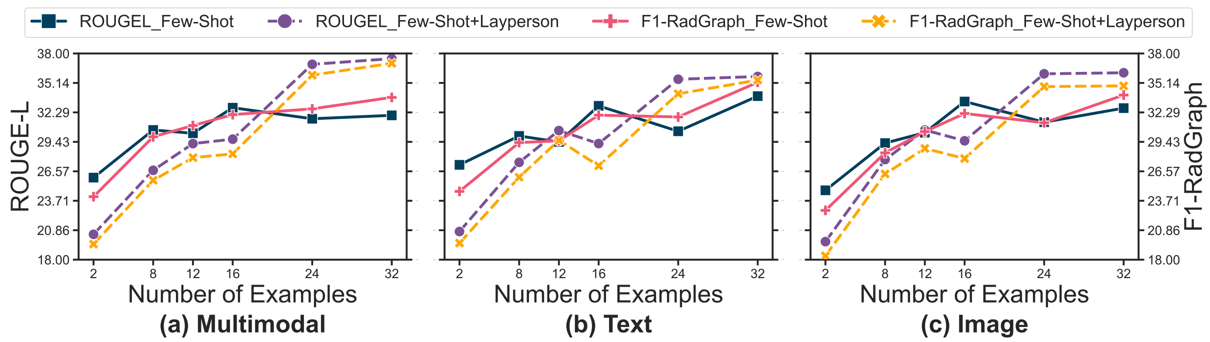

Figure 7: Validation results vs. the number of in-context examples across various prompt types and modality embeddings on Meta-Llama-3-8B-Instruct.

		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 4: Average Token of Prompts.