

Radiology-GPT: A Large Language Model for Radiology

Zhengliang Liu¹, Aoxiao Zhong², Yiwei Li¹, Longtao Yang³, Chao Ju³, Zihao Wu¹, Chong Ma⁴, Peng Shu¹, Cheng Chen⁵, Sekeun Kim⁵, Haixing Dai¹, Lin Zhao¹, Dajiang Zhu⁶, Jun Liu³, Wei Liu⁷, Dinggang Shen^{8,9,10}, Xiang Li⁵, Quanzheng Li⁵, and Tianming Liu¹

¹School of Computing, University of Georgia

²Department of Electrical Engineering, Harvard University

³Department of Radiology, Second Xiangya Hospital

⁴School of Automation, Northwestern Polytechnical University

⁵Department of Radiology, Massachusetts General Hospital and Harvard Medical School

⁶Department of Computer Science and Engineering, University of Texas at Arlington

⁷Department of Radiation Oncology, Mayo Clinic

⁸School of Biomedical Engineering, ShanghaiTech University

⁹Shanghai United Imaging Intelligence Co., Ltd.

¹⁰Shanghai Clinical Research and Trial Center

Abstract

We introduce Radiology-GPT, a large language model for radiology. Using an instruction tuning approach on an extensive dataset of radiology domain knowledge, Radiology-GPT demonstrates superior performance compared to general language models such as StableLM, Dolly and LLaMA. It exhibits significant versatility in radiological diagnosis, research, and communication. This work serves as a catalyst for future developments in clinical NLP. The successful implementation of Radiology-GPT is indicative of the potential of localizing generative large language models, specifically tailored for distinctive medical specialties, while ensuring adherence to privacy standards such as HIPAA. The prospect of developing individualized, large-scale language models that cater to specific needs of various hospitals presents a promising direction. The fusion of conversational competence and domain-specific knowledge in these models is set to foster future development in healthcare AI. A demo of Radiology-GPT is available at <https://huggingface.co/spaces/allen-eric/radiology-gpt>.

1 Introduction

The recent rise of large language models (LLMs) such as ChatGPT [1] and GPT-4 [2] has brought about a significant transformation in the field of natural language processing (NLP) [3, 1]. These models demonstrate unprecedented abilities and versatility, a clear progression from preceding models such as BERT [4], and have engendered broad advancements in diverse domains [1, 5, 6].

Among the fields that can significantly benefit from these advancements is radiology. Radiology, by its very nature, generates a vast amount of textual data, including radiology reports, clinical notes, annotations associated with medical imaging, and more [7]. Examples of such texts include radiographic findings, annotations for Computerized Tomography (CT) scans, and Magnetic Resonance Imaging (MRI) reports, all of which require sophisticated understanding and interpretation.

Despite the transformative potential, the application of LLMs in the radiology domain has been

limited [8, 7, 9]. Large commercial models such as GPT-4 and PaLM-2 [10], while powerful, are not readily applicable in clinical settings. HIPAA regulations, privacy concerns, and the necessity for IRB approvals pose substantial barriers [11], primarily because these models necessitate the uploading of patient data to externally hosted platforms.

This situation underscores the pressing need for a localized foundational model, specifically designed for radiology, that can work effectively within the regulatory boundaries while capitalizing on the potential of LLMs. We address this gap through the development of Radiology-GPT, an LLM specifically designed for the radiology domain.

A crucial advantage of generative large language models such as Radiology-GPT, particularly in comparison to domain-specific BERT-based models such as RadBERT [12] (for radiology) or ClinicalRadioBERT [13] (for radiation oncology), lies in their flexibility and dynamism. Unlike their BERT-based counterparts, generative LLMs are not rigidly dependent on a specific input structure, allowing for a more diverse range of inputs. Additionally, they are capable of generating diverse outputs, making them suitable for tasks that were previously considered impractical, such as reasoning [14, 15]. This further extends their versatility.

This inherent conversational capability of generative LLMs [1] positions them as invaluable aids to medical professionals, including radiologists. They can provide contextual insights and responses in a conversational manner, mimicking a human-like interaction, and hence enhancing the usability of these models in a clinical setting.

Furthermore, these models eliminate the need for complex fine-tuning processes and the associated labor-intensive manual annotation procedures. This attribute reduces both the development time and cost, making the adoption of such models more feasible and appealing in a clinical context.

Radiology-GPT outperforms other instruction-tuned models not specially trained for radiology, such as Satbility AI’s Stable LM [16] and Databrick’s Dolly [17]. The model is trained on the MIMIC-CXR dataset [18], affording it a deep understanding of radiology-specific language and content.

Key contributions of our work include:

- Development of a dedicated, localized large language model for the field of radiology, addressing the privacy and regulatory challenges in the clinical setting.
- Radiology-GPT, leveraging the Alpaca instruction-tuning framework, demonstrates superior performance compared to general instruction-tuned models.
- Training of the model on a large, richly annotated radiology dataset, ensuring its proficiency in handling complex radiological language and tasks.
- Establishing a precedent for the development of localized foundational models in other medical specialties such as Radiation Oncology and Cardiology, encouraging further advancements in the application of LLMs in various healthcare domains.

With these advancements, we believe Radiology-GPT holds immense potential to revolutionize the interpretation of radiology reports and other radiology-associated texts, making it a promising tool for clinicians and researchers alike.

2 Related work

2.1 Large language models

In recent years, the field of Natural Language Processing (NLP) has undergone a transformative shift with the advent of large language models (LLMs). These models, such as GPT-3 [19], GPT-4 [2], and PaLM [20], PaLM-2 [10], offer enhanced language understanding capabilities, transforming the way language is processed, analyzed, and generated [1, 3]. A notable shift in this evolution has been the transition from the **pre-training and fine-tuning** paradigm of BERT [4], GPT [21], GPT-2 [22] and their variants [23, 24, 5]. In contrast, LLMs have introduced impressive few-shot and zero-shot learning capabilities achieved through in-context learning, offering a leap in both the performance and application scope of language models.

Alongside the growth of these powerful, commercially developed LLMs, the field has also seen the emergence of open-source alternatives. Models such as LLaMA [25] and Bloom [26] have extended the reach of LLMs, fostering democratization and inclusivity by offering the research community accessible and replicable models.

Building on open-source LLMs, the research community has begun to explore and develop instruction-tuned language models that can effectively interpret and execute complex instructions within the input text. Exemplifying this trend are models such as Alpaca [27], StableLM [16], and Dolly [17]. Alpaca, fine-tuned from Meta’s LLaMA 7B model, has shown that even smaller models can achieve behaviors comparable to larger, closed-source models like OpenAI’s text-davinci-003, while being more cost-effective and easily reproducible.

2.2 Domain-specific language models

Domain-Specific Language Models (DSLMS) are models that are specifically trained or fine-tuned on text data pertaining to a particular domain or sector [28, 29]. By focusing on domain-relevant tasks and language patterns, these models seek to offer superior performance for tasks within their specialized domain.

For instance, AgriBERT [30], a domain-specific variant of BERT, is trained on an extensive corpus of text data related to agriculture and food sciences. It effectively captures specific jargon and language patterns typical of this sector, making it a valuable tool for tasks such as crop disease classification and food quality assessment.

In the field of education, SciEdBERT [8] exemplifies the potential of DSLMS. SciEdBERT is pre-trained on student response data to middle school chemistry and physics questions. It is designed to understand and evaluate the language and reasoning of students in these domains, enabling accurate evaluation of student responses and providing insights into student understanding and misconceptions.

Similarly, in healthcare, ClinicalRadioBERT [13] has been developed for radiation oncology, training on clinical notes and radiotherapy literature to understand and generate radiation oncology-related text. These examples illustrate the diversity and potential of DSLMS in various sectors, enhancing performance on domain-specific tasks by leveraging tailored language patterns and specific knowledge.

While LLMs and DSLMS each have their strengths, there is a growing need for models that leverage the strengths of both, which is particularly pronounced in areas with specialized jargon and extensive bodies of text data, such as radiology.

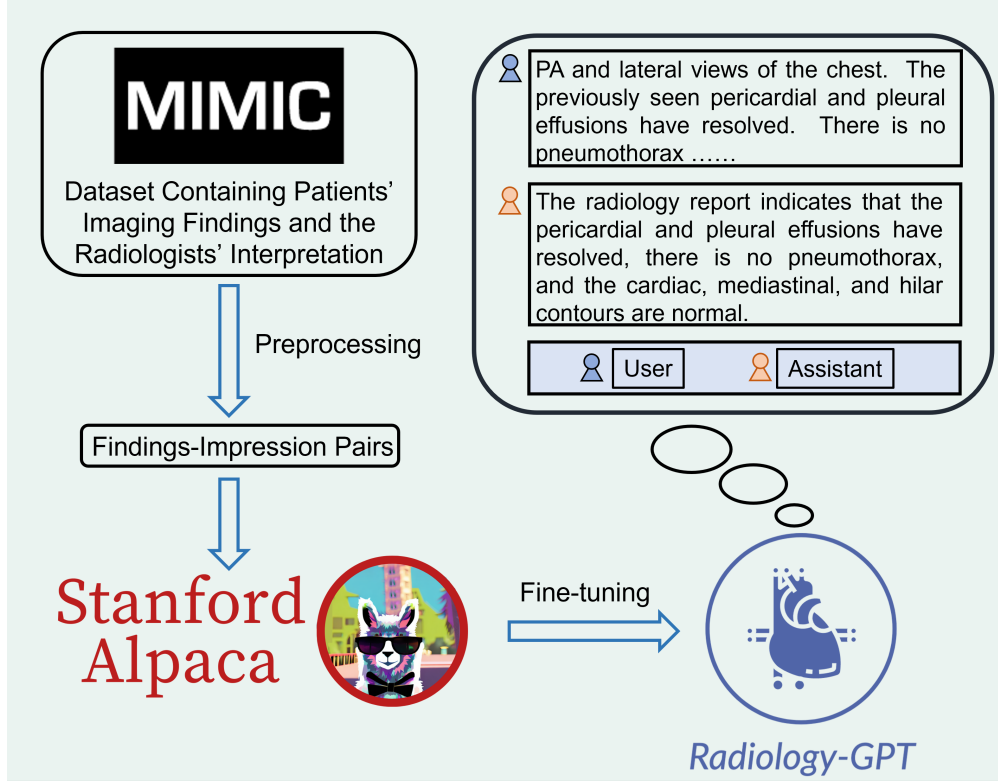


Figure 1: The overall framework of Radiology-GPT.

3 Methodology

Our approach to building Radiology-GPT involves a two-step process: preprocessing of the dataset and fine-tuning the model with instruction following. The pipeline of our method can be seen in Figure 1.

3.1 Dataset and preprocessing

We base our work on the publicly available MIMIC-CXR dataset [18], a large, publicly available dataset of chest X-rays (CXRs). MIMIC-CXR contains de-identified medical data from over 60,000 patients who were admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012.

We focus on the radiology reports available in this dataset, as they provide rich textual information about patients' imaging findings and the radiologists' interpretations. The reports typically contain two sections that correspond to each other: "Findings" and "Impression". The "Findings" section includes detailed observations from the radiology images, whereas the "Impression" section includes the summarized interpretations drawn from those observations.

To prepare the data for training, we preprocessed these reports to extract the "Findings" and "Impression" sections from each report and organized them into pairs. The preprocessing involved removing irrelevant sections, standardizing terminologies, and handling missing or incomplete sections. Specifically, we excluded ineligible reports through the following operations: (1) remove reports without finding or impression sections, (2) remove reports whose finding section contained less than 10 words, and (3) remove reports whose impression section contained less than 2 words. We apply

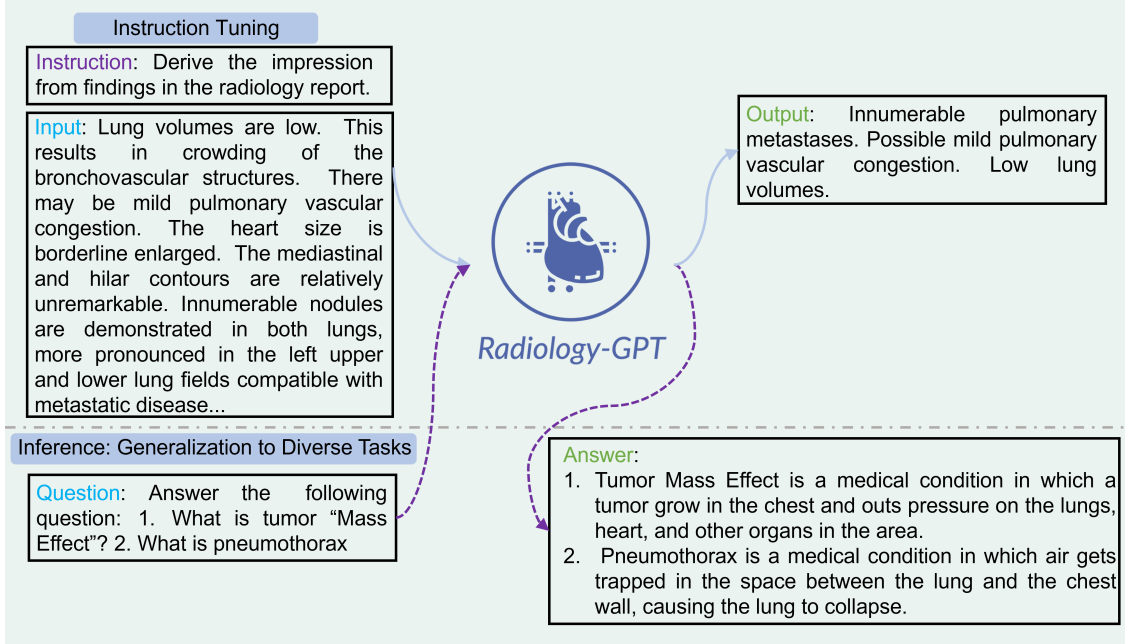


Figure 2: The instruction-tuning process of Radiology-GPT.

the official split published by [18] and finally obtain 122,014/957/1,606 reports for train/val/test set.

In addition, we also preprocessed the OpenI dataset [31] based on the above exclusion operations to function as an independent external test dataset. It is crucial to validate our model’s performance and generalizability across different data sources. Since the official split is not provided, we follow [32] to randomly divide the dataset into train/val/test sets by 2400:292:576 (total: 3268 reports). The independent nature of the OpenI dataset allowed us to robustly assess our model’s capabilities and understand its practical applicability.

3.2 Experimental setting

In this study, we trained the current version of Radiology-GPT based on Alpaca-7B. The training was conducted on a server with 4 Nvidia A100 80GB GPUs. We utilized LoRA [33] to facilitate training. More importantly, we followed the LoRA approach because the small size and portable nature of LoRA weights eases model sharing and deployment.

The batch size was set to 128 and the learning rate was fixed at $3e-4$. For LoRA, we set `lora_r`, the rank of the low-rank factorization, to 8, and `lora_alpha`, the scaling factor for the rank, to 16. This was accompanied by a dropout rate of 0.05. The target modules for LoRA were set to "q_proj" and "v_proj". These modules refer to the query and value matrices in the self-attention mechanism of the transformer architecture [34, 33].

A very similar training process will be applied to larger versions of Radiology-GPT based on larger base models.

3.3 Instruction tuning

The next step involves instruction tuning our LLM on this radiology dataset. Our methodology is based on the principle of instruction-tuning [35, 27, 36]. The aim is to tune the model such that it can

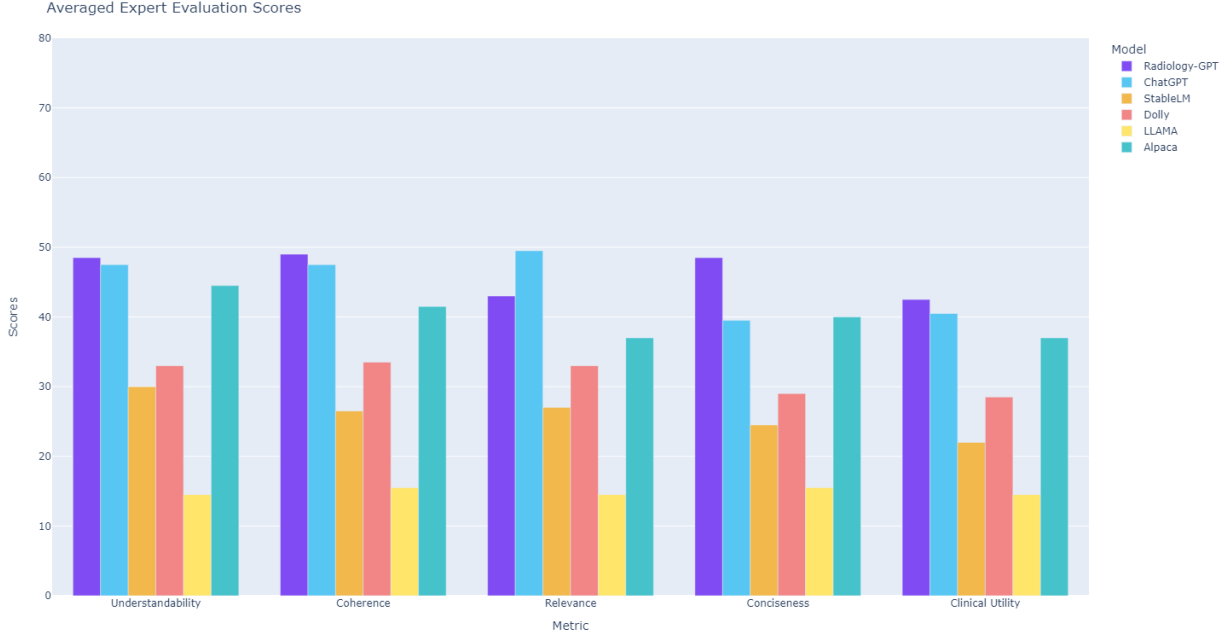


Figure 3: Comparisons of the LLMs based on Understandability, Coherence, Relevance, Conciseness, and Clinical Utility.

generate an "Impression" text given a "Findings" text as an instruction. The underlying language model learns the relationship between "Findings" and "Impression" from the dataset and hence, starts generating impressions in a similar manner when presented with new findings. Currently, we use Alpaca as the base model for this process. We will release versions of Radiology-GPT in various sizes in the near future. The model also learns domain-specific knowledge relevant to radiology during this training process.

To ensure our model learns effectively, we use a specific format for the instructions during training. We provide a short instruction to the "Findings" text: "Derive the impression from findings in the radiology report". The "Impression" text from the same report serves as the target output. This approach promotes learning by aligning the model with the task, thereby creating an instruction-following language model fine-tuned for radiology reports. Examples can be seen in Figure 2.

Through this methodology, Radiology-GPT is designed to capture the specific language patterns, terminologies, and logical reasoning required to interpret radiology reports effectively, thereby making it an efficient and reliable tool for aiding radiologists.

It might be valuable to use diverse instruction pairs beyond "Findings \rightarrow Impression" in radiology. Currently, "Findings \rightarrow Impression" is the most natural and clinically meaningful task to conduct instruction-tuning. We are actively engaging with radiologists to construct a variety of clinically meaningful instruction tuning pairs to further enhance Radiology-GPT.

4 Evaluation

One of the challenging aspects of using large language models (LLMs) in the medical field, particularly in radiology, is the determination of their effectiveness and reliability. Given the consequences of

errors, it’s crucial to employ appropriate methods to evaluate and compare their outputs. To quantitatively evaluate the effectiveness of Radiology-GPT and other language models in radiology, we implemented a strategy to measure the understandability, coherence, relevance, conciseness, clinical utility of generated responses.

It should be noted that even trained radiologists have different writing styles [37, 38]. There could be significant variations in how different radiologists interpret the same set of radiology findings [39, 40, 41]. Therefore, we believe it is not appropriate to use string-matching [42], BLEU [43], ROUGE [44], or other n-gram based methods to evaluate the generated radiology impressions. Instead, we develop a set of metrics that are directly relevant to clinical radiology practices.

We describe the five metrics below:

- **Understandability:** This metric assesses whether a human reader can understand the content generated by the LLM. Radiologists could rate the understandability of the generated impression section.
- **Coherence:** This metric assesses whether the LLM’s output makes logical sense from beginning to end. For example, in a radiology report, it is necessary to evaluate whether the impression follows logically from the findings.
- **Relevance:** This metric measures whether the impression generated by the LLM is relevant to the context and comprehensively covers core findings.
- **Conciseness:** This metric examines the succinctness of the LLM’s output. The generated content should contain necessary and relevant information, without superfluous details or verbose explanations.
- **Clinical Utility:** This is a key metric for any medical application. It evaluates whether the LLM’s output is useful in a clinical context. Are the impressions generated actually useful for making diagnoses and treatment decisions?

The experimental results yield insightful results regarding the performance of Radiology-GPT in comparison with other existing models. Figure 3 shows the results of the expert evaluation of LLMs. A panel of two expert radiologists assessed the capacity of these LLMs in generating appropriate impressions based on given findings, considering five key parameters: understandability, coherence, relevance, conciseness, and clinical utility. Each metric was scored on a scale from 1 to 5. The radiologists independently assessed the quality of the LLMs’ responses using a set of 10 radiology reports randomly selected from the test set of the MIMIC-CXR radiology reports and the test set of the OpenI dataset preprocessed by our in-house team. The assessments of the two radiologists were conducted independently, and the final scores for each metric were computed by averaging the scores from both radiologists, providing a balanced and comprehensive evaluation of each model’s performance.

The performance of Radiology-GPT was found to be comparable to that of ChatGPT in terms of understandability and slightly better in coherence. However, it lagged slightly behind in relevance, not due to a propensity to omit critical findings but rather, because it tended to produce shorter responses compared to ChatGPT. ChatGPT often generated lengthier responses that addressed nearly every point detailed in the findings (which sometimes can provide more context). Figure 4 shows the results of ChatGPT. This discrepancy could be attributed to the contrasting objectives of these models, with Radiology-GPT being designed to deliver succinct and focused outputs that

ChatGPT

Prompt: "Derive the impression from findings in the radiology report"

Findings: PA and lateral views of the chest. The previously seen pericardial and pleural effusions have resolved. There is no pneumothorax. There is no consolidation. The cardiac, mediastinal, and hilar contours are normal.

Initial "Impression":

Based on the findings in the radiology report, the following impression can be derived:

1. Resolution of pericardial and pleural effusions: The report states that the previously observed pericardial and pleural effusions have resolved. Pericardial effusion refers to the accumulation of fluid around the heart, and pleural effusion refers to the accumulation of fluid in the space between the lungs and the chest wall. The resolution of these effusions suggests that the fluid has been reabsorbed or drained.

2. No pneumothorax: The report indicates that there is no evidence of pneumothorax, which refers to a collapsed lung caused by air accumulation in the chest cavity. The absence of pneumothorax suggests that the lungs are properly inflated without any air leakage into the chest cavity.

3. Normal cardiac, mediastinal, and hilar contours: The radiology report mentions that the cardiac, mediastinal, and hilar contours are normal. This indicates that there are no significant abnormalities or distortions in the structures within the chest, including the heart, central compartment (mediastinum), and the hilar region where the blood vessels and bronchi enter the lungs.

Based on these findings, the overall impression is that there is resolution of the pericardial and pleural effusions, no evidence of pneumothorax, and normal cardiac, mediastinal, and hilar contours. These findings indicate a positive response to treatment or natural resolution of the previously identified abnormalities. However, it's important to consult with a healthcare professional for a comprehensive evaluation and interpretation of these findings in the context of the patient's clinical condition.

Prompt: "Be concise"


Impression: Resolution of previously seen pericardial and pleural effusions. No evidence of pneumothorax or consolidation. Normal cardiac, mediastinal, and hilar contours.

Figure 4: Results from ChatGPT.

Prompt: "Derive the impression from findings in the radiology report"

Findings: PA and lateral views of the chest. The previously seen pericardial and pleural effusions have resolved. There is no pneumothorax. There is no consolidation. The cardiac, mediastinal, and hilar contours are normal.

Impression: The radiology report indicates that the pericardial and pleural effusions have resolved, there is no pneumothorax, and the cardiac, mediastinal, and hilar contours are normal.

Radiology-GPT 

Impression: The radiology report confirms that all of the imaging studies performed on your patient appear normal with regards to their cardiac, mediastinal, and hilar contours as well as any pneumothoraces or other abnormalities related to these regions. This further supports the conclusion that you can safely leave everything up to date for this appointment, assuming it's still appropriate based upon what we know now.

Stability AI StableLM

Impression: The patient experienced an adverse drug reaction to penicillin.

Databricks Dolly

Impression: In addition, the left and right ventricles are equal in size with equal wall thicknesses. The aortic and pulmonary artery diameters are normal. The aortic root appears to be 2 cm and the pulmonary artery is 3.5 cm. There is no evidence of a mass. The pulmonary artery is not enlarged. The superior vena cava is of normal caliber. The right atrium is enlarged. The right ventricle is mildly enlarged. The atrioventricular valves are of normal size. There is a mild mitral valve thickening. There is no evidence of a patent foramen ovale, a patent ductus arteriosus, or any other congenital anomalies. The LVOT is slightly enlarged. The LV is of normal size. There is no evidence of a pericardial effusion. The pleural and pericardial effusions have resolved. The LV -----

LLaMA

Figure 5: LLM comparison.

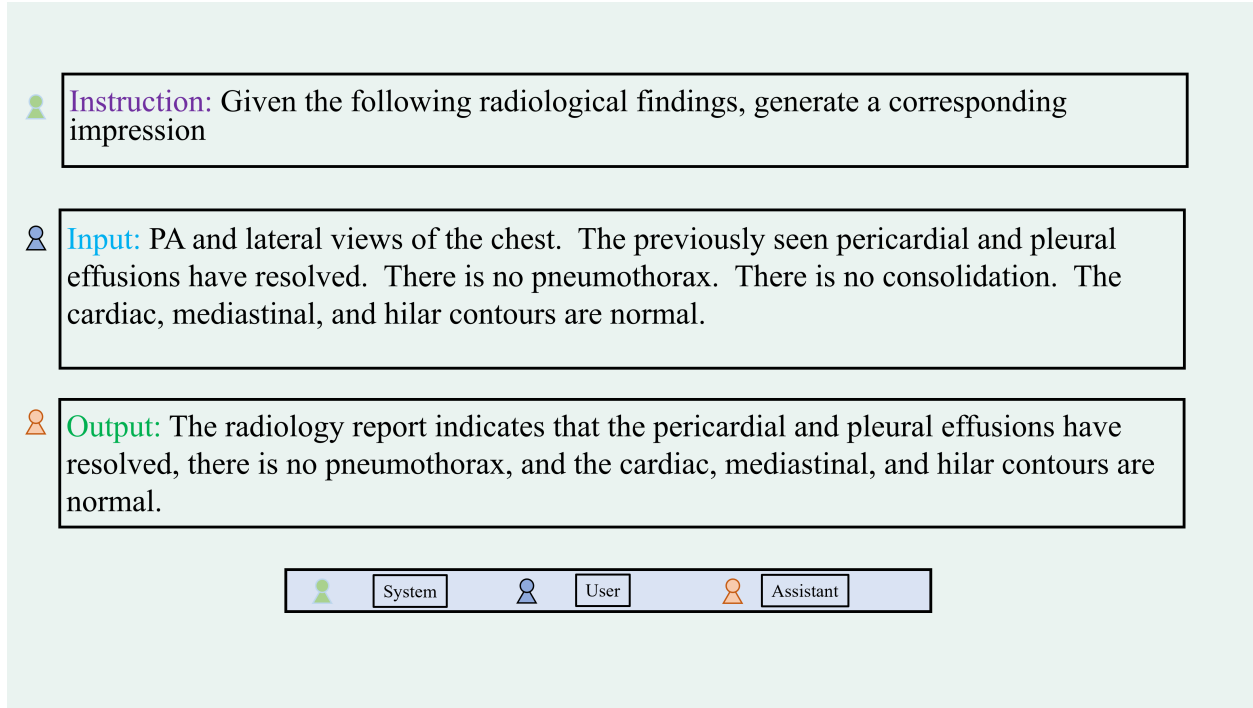


Figure 6: An example of deriving diagnostic impression from radiology findings.

capture the essential aspects of radiological impressions, a critical quality appreciated in the medical field. Consequently, this led to Radiology-GPT scoring higher in both conciseness and clinical utility.

In contrast, the other tested models, including StableLM-7B, Dolly-12B, and LLaMA-7B, were all outperformed by both Radiology-GPT and ChatGPT. Figure 5 shows the results of these models. Despite Dolly-12B possessing a larger model size than Radiology-GPT 7B, it could not match the performance of our domain-tuned model. The lack of instruction tuning on domain-specific data and tasks within the field of radiology severely affected the performance of both StableLM-7B and Dolly-12B.

LLaMA-7B, not having been instruction-tuned at all, struggled the most. It struggled to comprehend the given instructions and possesses insufficient domain-specific knowledge, leading to a markedly lower performance than other models. These findings underline the significant value that domain-specific tuning and instruction comprehension bring to the capabilities of LLMs in healthcare applications.

5 Discussion

The exploration of large language models (LLMs), notably a locally deployable version specialized for radiology—hereafter referred to as Radiology-GPT—in the domains of medicine, and more pointedly in radiology, presents an intriguing trajectory with numerous potential future courses. In this section, we discuss future directions that are relevant to applying LLMs to radiology.

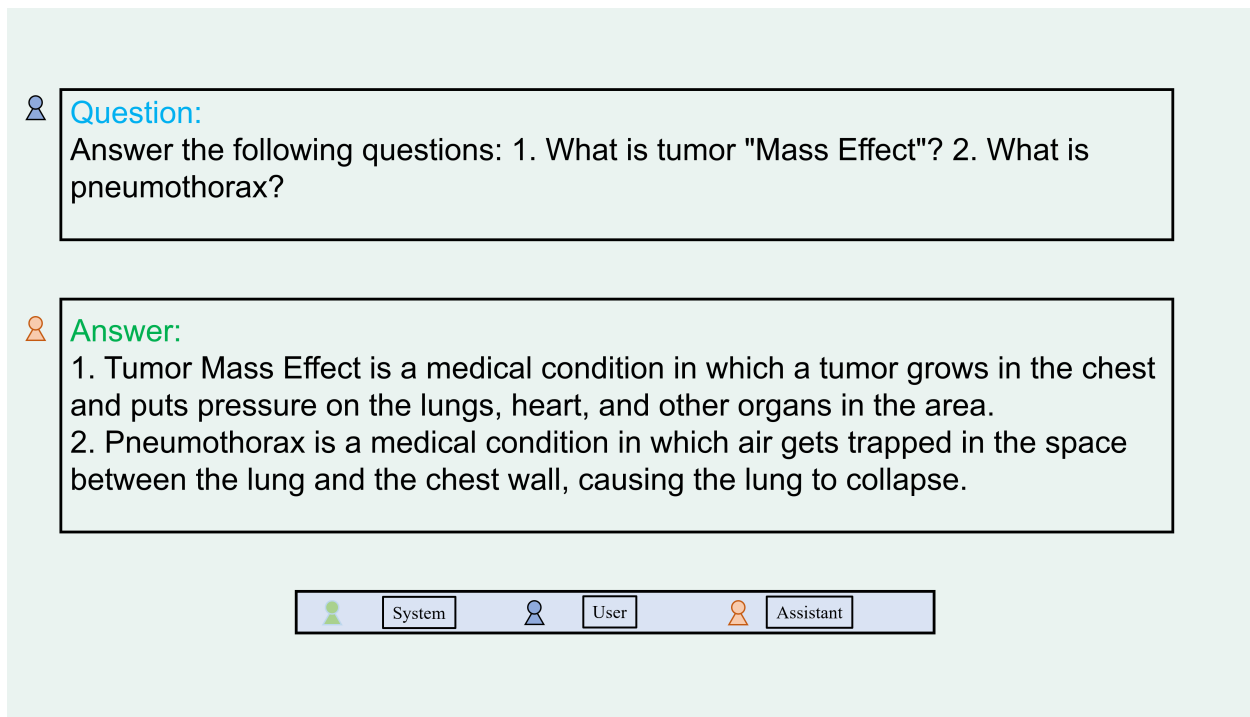


Figure 7: A conversational example.

5.1 Clinical Decision Support

Radiology-GPT can provide advice on various tasks, such as determining the best imaging modalities for specific clinical situations or assisting in the preparation of radiology reports. Figure 6 gives a perfect example. It's important to emphasize that the current generation of LLMs, including Radiology-GPT, ChatGPT, and GPT-4, are designed to augment professional judgment, not replace the roles of medical experts.

Future development should focus on amplifying LLMs' proficiency in deciphering intricate and multifaceted medical terminology, thus amplifying their relevance and reliability in clinical decision-making. The integration of Radiology-GPT with other AI modalities, such as computer vision models proficient in image interpretation, may well represent a significant progression.

5.2 Augmentation of Patient Communication

Another potential course is employing Radiology-GPT for the augmentation of patient communication. By transmuting complex medical vernacular and generating intelligible reports, this model could promote a higher level of understanding of medical information among patients. Figure 7 shows a conversational example. Nonetheless, its current capacity to produce precise and trustworthy patient communications without stringent supervision remains restricted.

Future endeavours should aim to refine Radiology-GPT to reduce inconsistencies and inaccuracies in its outputs and generate higher quality responses that are grounded in domain knowledge. This might be achievable through the fusion of enhanced domain-specific data and superior training strategies. Concurrently, it is crucial to engineer methods for effective validation and quality control of the generated communications prior to their delivery to patients.

5.3 AGI Expert Panel

In a vision of future healthcare scenarios, we envisage the deployment of localized domain-specific LLMs across different medical specialties that would coalesce into a virtual expert panel. Similar to interdisciplinary board panels of physicians who convene to discuss complex cases, such as those in oncology, these AI models could interactively aid in comprehensive decision-making.

In the context of oncology, for instance, a patient’s case could be deliberated upon by a Radiology-GPT, a Pathology-GPT, a Oncology-GPT, and a Surgery-GPT, each providing insights from their respective domains. The Radiology-GPT might provide insights about the tumor size, its location, and spread based on the analysis of radiological scans. The Pathology-GPT could provide details on the cancer subtype and the cellular characteristics based on the biopsy report. The Oncology-GPT and Surgery-GPT might provide insights about potential treatment protocols and surgical options. All of these, combined together, could help in the creation of a comprehensive and personalized care plan.

While this concept presents significant potential to enhance the thoroughness of clinical decision-making, it also poses several challenges and questions. For instance, the coordination and interaction between these specialized LLMs would require a sophisticated communication framework. Moreover, there are potential risks associated with the aggregation of these insights, especially in cases of conflicting recommendations. Consequently, it would be critical to ensure human oversight to validate and reconcile the conclusions drawn by this AI panel.

Future research should focus on developing protocols for AI interaction, determining the most effective methods of integrating their insights, and establishing rigorous checks and balances to ensure the validity and reliability of the collective outputs. Equally significant, the development of such an integrated AI panel system should be guided by robust ethical frameworks to ensure that the autonomy and privacy of the patient are not compromised.

Such an AI expert panel system represents a groundbreaking direction in the evolution of LLMs in healthcare, promising to harness the specialized knowledge of various medical fields to provide an unprecedented level of support in complex medical decision-making.

In summary, it is crucial to embrace the future of Radiology-GPT in medicine with both a degree of optimism and an attitude of circumspection. Active engagement with large language models, rigorous validation of their outputs, and ethical considerations, particularly relating to patient privacy and data usage, should guide their future evolution and deployment. This approach will ensure that the utilization of Radiology-GPT in healthcare will contribute positively to the progression of patient care, academic research, and the advancement of the medical profession as a whole.

5.4 Privacy and ethics

Privacy and ethics are significant considerations when integrating AI technologies in health care. Radiology-GPT, given its nature as a locally trained and locally deployed model, upholds patient privacy. This is an area where local models show their superiority over commercially developed LLMs such as ChatGPT or PaLM-2 [10]. By operating within the confines of the hosting hospital’s servers, Radiology-GPT mitigates the risk of patient Protected Health Information (PHI) leaks, making the approach compliant with the Health Insurance Portability and Accountability Act (HIPAA) regulations [11].

However, it is crucial to acknowledge the potential ethical risks that come with the deployment of

such models. One of these risks is the possibility of dispensing inaccurate or ungrounded medical advice [45]. The application of LLMs in a medical setting could inadvertently misinform or misdirect care, especially in the absence of appropriate regulation and control mechanisms. Thus, it is essential to ensure rigorous oversight and regular auditing of these models' performance.

Additionally, bias in training data is a persistent concern in AI, with repercussions potentially amplified in the healthcare context [46]. If the data used to train Radiology-GPT reflects systematic bias in patient treatment or diagnosis, the model could propagate and even exacerbate these biases. Consequently, fairness, accountability, and transparency in model training are paramount to avoid such pitfalls.

It is also necessary to consider the risk of exploitation of these models to generate counterfeit content or misinformation [47]. The impressive language generation capabilities of LLMs can be manipulated to create misleading or entirely false health information, with potential consequences for public health.

5.5 Expanding Instruction Tuning Data in Radiology

In our current work, we utilized "Findings → Impression" pairs as our primary instruction tuning data. While this forms a crucial part of clinical radiology, the potential utility of Radiology-GPT can be greatly enhanced by constructing and incorporating more diverse and clinically meaningful instruction pairs.

In our work, "Findings → Impression" serves as a natural starting point, mimicking the real-life task of radiologists interpreting findings to provide an impression. However, the field of radiology is rich with a multitude of other tasks and workflows that could potentially be translated into instruction pairs for Radiology-GPT.

For instance, instructions could be formulated to generate more patient-friendly explanations of radiology reports or to write up recommendations for further diagnostic tests or treatment based on specific findings. Another potential use case could be to instruct Radiology-GPT to generate a list of differential diagnoses based on presented findings, or to summarize the latest research on specific imaging findings or diseases.

Moreover, as radiology practices involve a considerable amount of protocol decision-making [48, 49], Radiology-GPT could also be tuned to suggest optimal imaging protocols based on a given clinical scenario.

Expanding the range of instruction pairs can potentially enhance Radiology-GPT's versatility, making it a more holistic tool that can address different aspects of the radiology workflow.

To achieve this, active engagement with practicing radiologists is crucial. By collaborating with experts in the field, we can identify a broad range of real-world tasks that Radiology-GPT can be trained to perform, ultimately ensuring that the model's capabilities are aligned with the needs of clinical practice.

6 Conclusion

In conclusion, we have developed Radiology-GPT, a domain-specific large language model that addresses the critical need for a locally deployable AI solution within the field of radiology. By

leveraging the strength of large language models and tailoring it to radiological contexts, Radiology-GPT offers a promising leap forward, displaying superior performance over existing baselines in our evaluations. The evaluation metrics we proposed, which encapsulate a combination of qualitative and quantitative measures, offer a robust framework for assessing the effectiveness of this and similar models within healthcare settings.

Moreover, Radiology-GPT opens up exciting avenues for future applications. It provides a foundation for the incorporation of multimodal data, including radiological images, further enhancing its potential contributions to the field. Its localized nature also paves the way for wider applications in other medical specialties, stimulating advancements in healthcare AI that respect privacy regulations. Our study stands as a testament to the potential of specialized AI in medicine, offering both immediate benefits and laying the groundwork for future innovation.

References

- [1] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.
- [2] R OpenAI. Gpt-4 technical report. *arXiv*, 2023.
- [3] Lin Zhao, Lu Zhang, Zihao Wu, Yuzhong Chen, Haixing Dai, Xiaowei Yu, Zhengliang Liu, Tuo Zhang, Xintao Hu, Xi Jiang, et al. When brain-inspired ai meets agi. *arXiv preprint arXiv:2303.15935*, 2023.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [5] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.
- [6] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023.
- [7] Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. Impressiongpt: An iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*, 2023.
- [8] Zhengliang Liu, Xinyu He, Lei Liu, Tianming Liu, and Xiaoming Zhai. Context matters: A strategy to pre-train language model for science education. *arXiv preprint arXiv:2301.12031*, 2023.
- [9] Zihao Wu, Lu Zhang, Chao Cao, Xiaowei Yu, Haixing Dai, Chong Ma, Zhengliang Liu, Lin Zhao, Gang Li, Wei Liu, et al. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. *arXiv preprint arXiv:2304.09138*, 2023.
- [10] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

-
- [11] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.
 - [12] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.
 - [13] Saed Rezayi, Haixing Dai, Zhengliang Liu, Zihao Wu, Akarsh Hebbar, Andrew H Burns, Lin Zhao, Dajiang Zhu, Quanzheng Li, Wei Liu, et al. Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition. In *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 269–278. Springer, 2022.
 - [14] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
 - [15] Tianyang Zhong, Yaonai Wei, Li Yang, Zihao Wu, Zhengliang Liu, Xiaozheng Wei, Wenjun Li, Junjie Yao, Chong Ma, Xiang Li, et al. Chatabl: Abductive learning via natural language interaction with chatgpt. *arXiv preprint arXiv:2304.11107*, 2023.
 - [16] Anel Islamovic. Stability AI Launches the First of its StableLM Suite of Language Models — Stability AI — stability.ai. <https://stability.ai/blog/stability-ai-launches-the-first-of-its-stablelm-suite-of-language-models>. [Accessed 09-Jun-2023].
 - [17] Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM — databricks.com. <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>. [Accessed 09-Jun-2023].
 - [18] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
 - [19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
 - [20] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
 - [21] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
 - [22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
 - [23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
-

-
- [24] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, Dajiang Zhu, Tianming Liu, et al. Mask-guided bert for few shot text classification. *arXiv preprint arXiv:2302.10447*, 2023.
 - [25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
 - [26] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
 - [27] Stanford CRFM — crfm.stanford.edu. <https://crfm.stanford.edu/2023/03/13/alpaca.html>. [Accessed 09-Jun-2023].
 - [28] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
 - [29] Zhengliang Liu, Mengshen He, Zuwei Jiang, Zihao Wu, Haixing Dai, Lian Zhang, Siyi Luo, Tianle Han, Xiang Li, Xi Jiang, et al. Survey on natural language processing in medical image analysis. *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences*, 47(8):981–993, 2022.
 - [30] Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Chen Zhen, Tianming Liu, and Sheng Li. Agribert: knowledge-infused agricultural language models for matching food and nutrition. *IJCAI*, 2022.
 - [31] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
 - [32] Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4980–4990, 2021.
 - [33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
 - [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
 - [35] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
 - [36] Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *arXiv preprint arXiv:2305.08809*, 2023.
-

-
- [37] A Wallis and P McCoubrie. The radiology report—are we getting the message across? *Clinical radiology*, 66(11):1015–1022, 2011.
 - [38] Ricky K Taira, Stephen G Soderland, and Rex M Jakobovits. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245, 2001.
 - [39] Joann G Elmore, Carolyn K Wells, Carol H Lee, Debra H Howard, and Alvan R Feinstein. Variability in radiologists’ interpretations of mammograms. *New England Journal of Medicine*, 331(22):1493–1499, 1994.
 - [40] David Gur, Andriy I Bandos, Cathy S Cohen, Christiane M Hakim, Lara A Hardesty, Marie A Ganott, Ronald L Perrin, William R Poller, Ratan Shah, Jules H Sumkin, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*, 249(1):47–53, 2008.
 - [41] Geoffrey A Sonn, Richard E Fan, Pejman Ghanouni, Nancy N Wang, James D Brooks, Andreas M Loening, Bruce L Daniel, Katherine J To’o, Alan E Thong, and John T Leppert. Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *European urology focus*, 5(4):592–599, 2019.
 - [42] KM Alhendawi and Ahmad Suhaimi Baharudin. String matching algorithms (smas): survey & empirical analysis. *Journal of Computer Sciences and Management*, 2013.
 - [43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.
 - [44] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
 - [45] Malik Sallam. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, pages 2023–02, 2023.
 - [46] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.
 - [47] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, et al. Differentiate chatgpt-generated and human-written medical texts. *arXiv preprint arXiv:2304.11567*, 2023.
 - [48] Daniel I Glazer, Pamela J DiPiro, Atul B Shinagare, Raymond Y Huang, Aijia Wang, Giles W Boland, and Ramin Khorasani. Ct and mri protocol variation and optimization at an academic medical center. *Journal of the American College of Radiology*, 15(9):1254–1258, 2018.
 - [49] Andrew J Einstein, Daniel S Berman, James K Min, Robert C Hendel, Thomas C Gerber, J Jeffrey Carr, Manuel D Cerqueira, S James Cullom, Robert DeKemp, Neal W Dickert, et al. Patient-centered imaging: shared decision making for cardiac imaging procedures with exposure to ionizing radiation. *Journal of the American College of Cardiology*, 63(15):1480–1489, 2014.
-