Evaluating the feasibility of using smaller Large Language Models for generating impressions from findings in radiology reports

Margarita Deli-Slavova[1] and Julian Hough[1]

Department of Computer Science
School of Mathematics and Computer Science
Swansea University, Swansea SA1 8EN, UK

Abstract. Large Language Models hold the potential to revolutionise the field of radiology by automating radiology tasks and enhancing clinical decision-making. However, the substantial computational resources required for training LLMs with a large number of parameters create barriers to their use both in radiology-related research and potential clinical applications. This study evaluated the performance of two smaller large language models (Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct) in comparison to a newly introduced state-of-the-art multimodal model, LLaMA 4 Scout, for the task of impression generation in radiology. Experiments were conducted with and without in-context learning to ascertain its impact on model performance. Results show that in-context learning considerably improves the performance of all models, allowing them to achieve competitive results in the task of impression generation. These findings underscore the capabilities of small open-source LLMs and highlight the potential of advanced multimodal models for radiology-specific NLP tasks.

Keywords: NLP · Radiology · Summarisation · In-context learning.

1 Literature review

1.1 Communication and workflow in radiology - challenges and perspectives

Clear and effective communication between physicians is essential for providing good-quality patient care. Suboptimal communication among healthcare practitioners can be one of the main contributors to medical errors, which can negatively impact patient care [9]. Considering the magnitude of the medical and legal implications of poor communication, seamless communication between radiologists and referring physicians is essential. Writing reports is a considerable part of the radiologists' daily work and the main form of communication between radiologists and referring physicians [6] [11]. The 'findings' portion of the radiology report features factual observations about the study and can be extremely valuable in radiologist-to-radiologist communication [11]. The 'impressions' section

of a radiology report serves the important purpose of communicating directly to the referring physician the radiologist's interpretation of the findings [11]. This section also features differential diagnosis and recommendations.

It has been suggested that referring physicians mainly rely on the impressions section of the report and often only read the findings when the impression part is not sufficiently clear, underlining the importance of radiologists conveying relevant information as clearly as possible in the impression section [11]. The American College of Radiology (ACR) Council has elevated the role of communication to a critically important component, particularly in diagnostic radiology [22]. Radiological examinations are commonly performed at the request of referring clinicians, after which radiologists send back reports of the imaging studies [22]. It has been indicated that globally, 80–90% of radiologists do not directly communicate with patients before or after imaging studies [10]. However, with the recent rise of patient-centered care, the existing communication paradigm in radiology is likely to undergo changes. Nevertheless, factors such as high workloads in radiology departments could considerably affect the likelihood and speed of such changes taking place. However, current and future AI innovations can potentially broadly impact healthcare workloads, including radiology.

1.2 Miscommunication and errors in radiology - sources and implications

Errors and poor communication between radiologists and referring physicians can bring about medical and legal issues. According to a study by Kim&Mansfield, *under-reading*, where an examination is reported as normal despite the presence of a detectable abnormality [20], represented 42% of total errors [13]. *Interpretative errors* in radiology occur at a rate of approximately 4% [23]. Around 1 billion radiology examinations are performed annually worldwide, which equates to 40 million interpretive errors annually [23]. Up to 80% of interpretative errors are perceptual, and due to the importance of detection in radiology, failing to detect conspicuous abnormalities can lead to missed diagnoses [23]. Despite the advances in imaging modalities, humans' perceptual limitations suggest the need for the development of tools that can reduce errors and enhance patient safety.

Although not commonly addressed in studies, poorly written reports are another considerable cause of poor patient outcomes [6]. Despite being the main vessel for communication between radiologists and referring physicians, radiology reports can lack the clarity to allow the referring physician to appreciate the radiologist's interpretation, conclusions and advice [6]. The discrepancy that can occur between the message conveyed by the radiologist in the report and the interpretation of the referring physician can lead to misunderstandings with adverse consequences for the patient [6]. Miscommunications between the referring clinician and the radiologist can also arise from inadequate clinical information or expectations that do not realistically reflect the capabilities of a radiological technique [6]. Despite the importance of clinical information for radiological examinations, it is common during in-hospital care for imaging studies to take

place before a detailed medical history is acquired, which can reflect negatively on the quality of care [6].

Despite the importance of good communication in radiology, radiologists do not receive sufficiently extensive formal training on the radiology report structure and the pertinent medical and legal intricacies [22]. Radiology trainees obtain report-writing skills more passively by reading radiology reports, receiving feedback and emulating the style of senior colleagues. The significance of possessing good communication and report-writing skills in the field of radiology necessitates more focus on the training received by future radiologists.

The increase in radiologists' workload, unparalleled by a necessary increase in radiologist numbers, poses another risk for errors. The concern that higher demand in radiology can bring about reduced accuracy has attracted considerable attention due to important medico-legal implications [4]. With the growing ageing population and likely further rise in demand, this issue is poised to gain even more relevance in the future. Despite the propositions for introducing workload limits, current evidence is not sufficient to substantiate concrete and suitable workload or speed limits in radiology [4].

1.3 Addressing miscommunication and errors in radiology

Various papers have proposed addressing errors and miscommunication in radiology by tackling different causal factors. Literature is increasingly shedding light on AI technology's possibilities in this context. AI technology has been proposed as a solution to reducing radiologists' trivial tasks, such as measuring lymph nodes and lung nodules [4]. Additionally, AI could be utilised to enhance trainee radiologists' training process. Also, there is a proliferating interest in how AI could be used to assist radiologists in identifying abnormalities and help improve accuracy [5]. A recent study aimed to design a system that can minimise perceptual errors by utilising eye gaze data and radiology reports [5]. The system demonstrated potential by correcting 76% of the missed abnormalities and achieving a Total-Usefulness score exceeding 0.4 [5]. However, it is worth noting that the study used a simulated error dataset, which does not accurately represent the broad range of perceptual errors in radiology and thus may affect reproducibility when applied to real-world radiology datasets.

However, at present, there is no sufficient evidence consolidating the effectiveness and utility of AI in improving radiologists' decision-making and patient outcomes [4]. Additionally, AI systems require large amounts of data, which brings about ethical concerns regarding data privacy and security, considering the sensitivity of healthcare data. Additionally, there are various regulations that AI systems used in radiology must adhere to, depending on the location. However, at present, there is no specific legislation on the use of AI in healthcare in the UK. This highlights the importance of developing appropriate legal frameworks regarding the use of AI in healthcare.

1.4 Closed-source vs Open-source LLMs in radiology

Closed-source LLMs The properties of Transformer-based models, such as impressive scalability and versatility, have consolidated their superiority for most Natural Language Processing (NLP) tasks, and those models continue to attract significant academic and industry interest in different applications [21]. Nevertheless, some applications using transformer-based models that exhibit state-of-the-art performance, such as CoPilot, ChatGPT, GPT-3 and GPT-4, are closed-source, preventing the public from accessing and modifying the source code, model weights or training data [21]. This constitutes an impediment to the progress in the field of NLP and brings about a myriad of ethical quandaries. For the most common NLP task in the domain of radiology - generating impressions from findings, Ma et al., 2024 proposed ImpressionGPT – which utilises dynamic prompts and an iterative response optimisation approach [17]. The results demonstrated that ImpressionGPT outperformed Radiology-Llama2, ChatGPT, GPT-4 and the fine-tuned GPT- 3 against the Rouge metrics on the MIMIC-CXR and OpenI datasets [17]. Similarly, when evaluating the potential of LLMs in the de-identification of clinical reports, Liu et al., 2023 showed that when compared against BERT, RoBERTa, and ClinicalBERT, GPT-4 achieves the highest de-identification accuracy [14]. The study highlighted another strength of the GPT-4 model: carefully designed prompts can confer extremely high accuracy without requiring fine-tuning.

Open-source LLMs Despite the impressive capabilities of commercial models such as GPT-4, ChatGPT, and PaLM 2 for radiology-related tasks, their utility is partially neutralised by their inability to be implemented in healthcare systems. Therefore, due to privacy and security protection guidelines and mandates, it is imperative to focus on localized large language models in the realm of healthcare. Various parameter techniques and privacy accounting methods have enabled the private fine-tuning of models that are comparable in performance to commercial models. This has created the right conditions for the LLaMA model, which is under a non-commercial license to gain prominence. For instance, a recent study introduced Radiology-Llama2, which was trained on a large radiology dataset using instruction tuning to generate radiology impressions from findings [15]. Radiologists have ascertained that impressions generated by Radiology-Llama2 surpass those of general LLMs in coherence, conciseness and clinical utility [15]. Additionally, it was shown that the Radiology-Llama2 performance exceeds that of all other models across all ROUGE metrics [15].

1.5 Utilization of smaller LLMs in radiology for Impression generation

Recent research has attempted to explore the potential application of LLMs for a wide range of radiology-related tasks, from inferring disease diagnosis to generating radiology reports [12] [26]. However, automated summarisation of radiology findings (i.e. impression/conclusion generation) remains one of the

most prominent radiology-related NLP tasks in literature thanks to its potential for reducing radiologists' workload by being utilised for clinical decision-support purposes in radiology. Additionally, given its prominence in research, impression generation serves as a valuable benchmark for evaluating and comparing LLM performance in the domain of radiology.

While many studies on the use of LLM models for radiology-related tasks have tested the feasibility of larger models with more than 10 billion parameters such as GTP-4, PaLM 2, Llama-3-70B [12], there are many advantages to using smaller LLMs with up to 10 billion parameters or less. Firstly, when fine-tuning using domain-specific datasets, smaller LLMs can surpass larger general LLMs in performance on domain-specific tasks [7]. For instance, using a model with 2.7 billion parameters and an innovative training pipeline, researchers have demonstrated that their language model outperformed their larger counterparts in chest X-ray report automation [25]. Moreover, smaller LLMs require fewer computational resources, highlighting their efficiency [7]. Additionally, the complexity of architecture and the number of parameters of models are inversely correlated with interpretability and transparency, both of which are important in the medical domain [7].

This study aims to evaluate the feasibility of using smaller open-source LLMs in the most common NLP task in the radiology domain of generating impressions based on findings and compare their performance to recently released state-of-the-art large open-source LLMs.

The exploration of the feasibility of smaller LLMs for the field of radiology in this study is underpinned by the growing interest in smaller language models, which, despite having considerably smaller model sizes, can demonstrate performance similar to that of their larger counterparts. This shift in research towards smaller LLMs has been catalysed by computational constraints in fields such as academia and healthcare. Smaller foundation models, such as Llama 3 8B, which has demonstrated superior performance compared to competing models while offering considerable computational efficiency, hold immense promise for NLP applications in radiology [7]. The selection of the three models in this study is grounded in the aim of investigating the trade-offs between model size and performance in NLP tasks in radiology, as well as exploring the impact of in-context learning on performance.

To the best of our knowledge, there are no published papers in which the chosen models have been tested on this task using the OpenI dataset.

2 Methodology

2.1 Data

The OpenI dataset will be used for this study [8]. The first one thousand reports from the dataset were selected for the experiments on generating impressions from findings. Reports were excluded if they do not have either findings or impression sections.

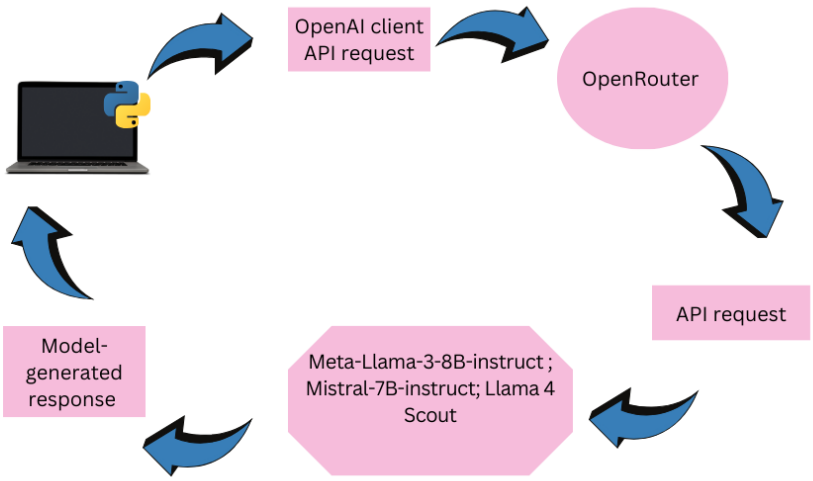

Fig. 1. Model-interaction process

2.2 Models

For this study, two open-source models with less than 10 billion parameters will be used, namely meta-llama/Meta-Llama-3-8B-Instruct and mistralai/Mistral-7B-Instruct. They will be compared to the recently released state-of-the-art Llama 4 Scout model for the task of impression generation. The Llama 4 Scout model features a Mixture of Experts architecture, echoing a trend towards utilising attention-compatible mechanisms to confer higher performance and training and inference efficiency on LLMs [18]. The Mixture of Experts architecture enables the activation of only a subset of all parameters during inference depending on scores assigned by a gating network [18]. Powerful commercial LLMs, such as GPT-4, have been speculated to employ a Mixture of Experts architecture [18]. The models were accessed using OpenRouter, which provides an interface for accessing and experimenting with different models. Firstly, an API request using the OpenAI client API is created, and then OpenRouter routes the request to the specified models, which generate a response and send it back, as per Figure 1. This will be implemented using Python.

2.3 Generating impressions from findings task and assessing the impact of in-context learning

To test the performance of Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct models against Llama 4 Scout, findings and impression sections from the reports are extracted, and a prompt is sent to the model, providing the findings and asking it to generate impressions based on the findings. The responses from the model are then returned. ROUGE scores are computed for all separate files and overall. Following that, in-context learning is applied to the next set of experiments. In-context learning, which is also referred to as few-shot prompting,

consists of providing examples of target input-output pairs in the prompt [14]. Five radiology reports from the remaining OpenI dataset were provided as examples, and the outputs were compared to those without in-context learning. System prompts are commonly used to improve the performance of AI systems, specifying the role of the system and steering the LLM in the desired direction. The system prompt that is passed to all LLMs for all experiments for this task is: "You are a radiologist expert skilled at writing radiology reports". For these experiments, the temperature is set to 0.4, as the medical context implies higher determinism than creativity.

2.4 Evaluation

For the task of generating impressions based on findings, the outputs of the models will be evaluated using ROUGE metrics. The ROUGE metrics are obtained by comparing the original radiology report impression to the corresponding impression generated the LLMs for each report. Finally, the average ROUGE scores for all files are computed. Three different ROUGE measures are utilised in this project: ROUGE-1, ROUGE-2, and ROUGE-L. The ROUGE metrics are frequently used for quantitatively assessing model generated text summarisations by comparing their similarity to human-generated ones based on overlapping n-grams(number of consecutive words) [19].

The results of all three models with and without in-context learning will be compared against each other and other models for the same task using the same dataset. The values that the ROUGE scores can take are between 0 and 1. A higher score indicates better performance on a summarisation task [3].

3 Results

The experiments were carried out, and the overall ROUGE scores for the two models with and without in-context learning were calculated by summing individual scores and dividing them by the number of files. The table with the scores can be seen in Fig. 2.

The results show that when in-context learning was introduced, the performance of Meta Llama-3-8B-instruct exceeded that of Mistral-7B-instruct. With in-context learning, Llama 4 Scout outperformed both Mistral-7B-instruct and Llama-3-8B-instruct. These results were compared to those of Radiology-Llama2 [15]. Radiology-Llama2, a domain-specific model based on Llama 2 architecture and trained on a large dataset of radiology reports, achieved 0.4185 in ROUGE-1, 0.2569 in ROUGE-2, and 0.4087 in ROUGE-L when evaluated using the OpenI dataset [15], which Llama 4 Scout with in-context learning marginally exceeded. Although the Meta-Llama-3-8B-instruct and Mistral-7B-instruct models with and without in context learning underperformed against Radiology-Llama2, they outperformed other general LLMs tested in that study on the OpenI dataset such as Llama 2 (0.0848 in ROUGE-1, 0.0205 in ROUGE-2, and 0.0712 in ROUGE-L), PaLM 2(0.1386 in ROUGE-1, 0.0477 in ROUGE 24 2, and 0.1194 in ROUGE-L)

	Meta-Llama-3-8B-instruct	Meta-Llama-3-8B-instruct(in-context learning)	Mistral-7B-instruct	Mistral-7B-instruct (in-context learning)	Llama 4 Scout	Llama 4 Scout (in-context learning)
ROUGE-1	0.1605	0.4329	0.1872	0.3431	0.1558	0.4470
ROUGE-2	0.0572	0.2715	0.0647	0.1987	0.0605	0.2965
ROUGE-L	0.1396	0.4183	0.1644	0.3259	0.1368	0.4241

Fig. 2. overall ROUGE scores

and GPT-4 (0.1171 in ROUGE-1, 0.0343 in ROUGE-2, and 0.0975 in ROUGE-L).

These results demonstrate the potential of Meta-Llama-3-8B-instruct, Mistral-7B-instruct, and Llama 4 Scout for radiology-specific tasks. Despite the LLaMA 4 Scout model (with 17B active and 109B total parameters) [1] demonstrating a slight performance advantage over the smaller LLMs such as Meta-Llama-3-8B-instruct and Mistral-7B-instructed, their high computational efficiency is an important consideration for future research into adapting open-source LLMs for NLP tasks in radiology through fine-tuning. It is worth mentioning that Radiology-Llama2 was trained on a large corpus of radiology data, while none of the models in this study were. Fine-tuning these models could significantly improve performance on the impressions-from-findings task and other radiology-related tasks. However, this study showed the impact of in-context learning and system prompting on performance. However, it is essential to note that, in this study, only a thousand radiology reports were assessed, while in the Radiology-Llama2 study, hundreds more.

4 Discussion

4.1 ROUGE metrics – strengths and weaknesses

Despite the utility and ubiquity of ROUGE metrics for evaluation of model-generated summaries, they entail some general and radiology-task-specific weaknesses. For instance, terminology variations and paraphrasing can cause perturbations in the performance of ROUGE metrics [3]. This weakness arises from the fact that ROUGE metrics do not account for semantic similarity [3]. It has been shown that substituting 20% of words with synonyms using 5 sentence summary results in 5-7% decrease in ROUGE scores [3]. This is particularly relevant in the field of radiology due to the variations. An example for this variation provided by Zhu et al., 2024 is that a radiology report can state: “the cardiac silhouette is

enlarged", while a model-generated summary can sate "moderate to severe cardiomegaly is re-demonstrated" [27]. While both suggest the same abnormality – cardiomegaly, ROUGE metrics may give a lower score due to not recognising the semantic similarity [27]. This has precluded some researchers evaluating the performance of NLP models for radiology tasks from using ROUGE metrics in favour of more qualitative measures [16].

Despite the weaknesses of ROUGE metrics, however, they remain useful and could be used alongside qualitative measures based on expert evaluation, as it has been shown that there is a correlation between semi-quantitative radiologist evaluations and ROUGE metrics [16]. The use of ROUGE metrics on common benchmark datasets allows for replicating, validating, and comparing study results. For future studies on NLP tasks in radiology, however, it may be beneficial to explore the combined use of ROUGE, semantic-aware metrics [27], and semi-quantitative evaluations by radiologists [19].

Recently, LLMs have also been proposed as an evaluation approach for radiology report generation. GPT-4, for instance, has been shown to reach evaluation performance comparable to that of radiologists while circumventing the drawbacks of both ROUGE metrics and manual expert evaluation [24]. Additionally, distilled models have demonstrated promising results as an automatic evaluation approach for radiology report generation while offering greater computational efficiency [24]. Future research efforts on evaluating model-generated radiology reports and summaries could focus on validating the evaluation efficiency of distilled models. Establishing effective automatic evaluation methods could accelerate advancements in NLP applications in radiology. Future research could also focus on developing an effective evaluation framework for model-generated radiology summaries.

4.2 Between in-context learning and fine-tuning

There are many benefits to selecting a smaller LLM as a base model to be further fine-tuned for the task of generating impressions from findings in radiology. One of the limitations that led to these techniques not being used for this study was the lack of access to sufficient data. The OpenI dataset is considerably small for training, testing and validation. Another reason why fine-tuning could not be used in this project was related to computational resource limitations. Nevertheless, it is recommended to experiment with in-context learning before considering the utilization of fine-tuning as it can help deduce if fine-tuning would improve model performance [2].

The considerable improvement that Meta-Llama-3-8B-instruct demonstrated following in-context learning and system prompting on the most common radiology summarisation task suggests that fine-tuning would improve model performance. Future research should explore fine-tuning of smaller LLMs for radiology-specific tasks.

4.3 Potential of using multimodal models

The horizon of Transformer-based model applications continues to expand, which is facilitated by innovations such as the introduction of multimodal models, which can process and learn to relate text and other modes of data, such as images or audio, simultaneously [21]. Such models can be employed for tasks such as image-to-text and audio-to-text generation [21]. The accumulating advancements in multimodal models signify their potential to be adapted for various tasks in different fields. For instance, transformer-based models that integrate text and image modalities, such as clinical reports and accompanying images, could be valuable for healthcare fields such as radiology.

Studies utilizing multimodal models for report generation and error detection and correction in radiology have demonstrated promising results [26] [5]. However, both studies have highlighted weaknesses relating to the use of multimodal models or data used for such models. For example, for one of the studies, although the proposed multimodal model outperformed other baseline models for report generation, it did not perform well in generating sentences that it had never seen before in the training data [26]. This could negatively affect the performance of such models on real-world datasets. This issue is further exacerbated by the fact that there are only two large publicly available radiology report datasets both of which for X-ray-based reports only. This study demonstrated the multimodal Llama 4 Scout's high performance for impression generation following in-context learning, suggesting that with further fine-tuning, this model could emerge as a leading model for natural language processing (NLP) applications in radiology. However, its more complex architecture and a greater number of parameters will require higher computational resources.

4.4 Conclusions

This study attempted to ascertain the feasibility of using LLMs with fewer than 10 billion parameters, namely Meta-Llama-3-8B-instruct and Mistral-7B-instruct, for radiology-related NLP tasks and explore how in-context learning would affect model performance. The evaluation of the model outputs from the summarisation task of generating impressions based on findings demonstrated that in-context learning and system prompting led to the smaller LLMs achieving similar performance to the larger Llama 4 Sout and outperforming powerful general LLMs such as GPT-4 for the same task [15].

Acknowledgements. This work was partly funded by the UKRI EPSRC CDT for Enhancing Human Interactions and Collaborations with Data and Intelligence Driven Systems, EP/S021892/1. Hough's work is partly funded by the EPSRC FLUIDITY project, EP/X009343/1.

Disclosure of Interests. The authors have no competing interests.

References

1. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, https://ai.meta.com/blog/llama-4-multimodal-intelligence/
2. To fine-tune or not to fine-tune, https://ai.meta.com/blog/when-to-fine-tune-llms-vs-other-techniques/
3. Akter, M., Bansal, N., Karmaker, S.K.: Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 1547–1560. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.122
4. Alexander, R., Waite, S., Bruno, M.A., Krupinski, E.A., Berlin, L., Macknik, S., Martinez-Conde, S.: Mandating Limits on Workload, Duty, and Speed in Radiology. Radiology **304**(2), 274–282 (Aug 2022). https://doi.org/10.1148/radiol.212631
5. Awasthi, A., Le, N., Deng, Z., Wu, C.C., Van Nguyen, H.: Enhancing Radiological Diagnosis: A Collaborative Approach Integrating AI and Human Expertise for Visual Miss Correction (Jun 2024). https://doi.org/10.48550/arXiv.2406.19686
6. Brady, A.P.: Error and discrepancy in radiology: inevitable or avoidable? Insights into Imaging **8**(1), 171–182 (Dec 2016). https://doi.org/10.1007/s13244-016-0534-1
7. Chen, L., Varoquaux, G.: What is the Role of Small Models in the LLM Era: A Survey (Sep 2024). https://doi.org/10.48550/arXiv.2409.06857
8. D, D.F., Md, K., Mb, R., Se, S., L, R., S, A., Gr, T., Cj, M.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association : JAMIA **23**(2) (Mar 2016). https://doi.org/10.1093/jamia/ocv080
9. Fatahi, N., Krupic, F., Hellström, M.: Difficulties and possibilities in communication between referring clinicians and radiologists: perspective of clinicians. Journal of Multidisciplinary Healthcare **12**, 555–564 (2019). https://doi.org/10.2147/JMDH.S207649
10. Gutzeit, A., Heiland, R., Sudarski, S., Froehlich, J.M., Hergan, K., Meissnitzer, M., Kos, S., Bertke, P., Kolokythas, O., Koh, D.M.: Direct communication between radiologists and patients following imaging examinations. Should radiologists rethink their patient care? European Radiology **29**(1), 224–231 (Jan 2019). https://doi.org/10.1007/s00330-018-5503-2
11. Hartung, M.P., Bickle, I.C., Gaillard, F., Kanne, J.P.: How to Create a Great Radiology Report. RadioGraphics (Oct 2020). https://doi.org/10.1148/rg.2020200020
12. Kim, S.H., Schramm, S., Adams, L.C., Braren, R., Bressem, K.K., Keicher, M., Zimmer, C., Hedderich, D.M., Wiestler, B.: Performance of Open-Source LLMs in Challenging Radiological Cases – A Benchmark Study on 4,049 Eurorad Case Reports (Sep 2024). https://doi.org/10.1101/2024.09.04.24313026
13. Kim, Y.W., Mansfield, L.T.: Fool Me Twice: Delayed Diagnoses in Radiology With Emphasis on Perpetuated Errors. American Journal of Roentgenology **202**(3), 465–470 (Mar 2014). https://doi.org/10.2214/AJR.13.11493
14. Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P., Zeng, F., Sun, L., Liu, W., Shen, D., Li, Q., Liu, T., Zhu, D., Li, X.: DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4 (Dec 2023). https://doi.org/10.48550/arXiv.2303.11032
15. Liu, Z., Li, Y., Shu, P., Zhong, A., Yang, L., Ju, C., Wu, Z., Ma, C., Luo, J., Chen, C., Kim, S., Hu, J., Dai, H., Zhao, L., Zhu, D., Liu, J., Liu, W., Shen, D., Liu, T., Li, Q., Li, X.: Radiology-Llama2: Best-in-Class Large Language Model for Radiology (Aug 2023). https://doi.org/10.48550/arXiv.2309.06419

16. Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., Wu, Z., Ma, C., Shu, P., Chen, C., Kim, S., Dai, H., Zhao, L., Zhu, D., Liu, J., Liu, W., Shen, D., Li, Q., Liu, T., Li, X.: Tailoring Large Language Models to Radiology: A Preliminary Approach to LLM Adaptation for a Highly Specialized Domain. In: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (eds.) Machine Learning in Medical Imaging. pp. 464–473. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-45673-2_46
17. Ma, C., Wu, Z., Wang, J., Xu, S., Wei, Y., Zeng, F., Liu, Z., Jiang, X., Guo, L., Cai, X., Zhang, S., Zhang, T., Zhu, D., Shen, D., Liu, T., Li, X.: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. IEEE Transactions on Artificial Intelligence **5**(8), 4163–4175 (Aug 2024). https://doi.org/10.1109/TAI.2024.3364586, arXiv:2304.08448 [cs]
18. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large Language Models: A Survey, http://arxiv.org/abs/2402.06196
19. Nishio, M., Matsunaga, T., Matsuo, H., Nogami, M., Kurata, Y., Fujimoto, K., Sugiyama, O., Akashi, T., Aoki, S., Murakami, T.: Fully automatic summarization of radiology reports using natural language processing with large language models. Informatics in Medicine Unlocked **46**, 101465 (Jan 2024). https://doi.org/10.1016/j.imu.2024.101465
20. Onder, O., Yarasir, Y., Azizova, A., Durhan, G., Onur, M.R., Ariyurek, O.M.: Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. Insights into Imaging **12**(1), 51 (Apr 2021). https://doi.org/10.1186/s13244-021-00986-8
21. Patwardhan, N., Marrone, S., Sansone, C.: Transformers in the Real World: A Survey on NLP Applications. Information **14**(4), 242 (Apr 2023). https://doi.org/10.3390/info14040242
22. Pinto, F., Capodieci, G., Setola, F.R., Limone, S., Somma, F., Faggian, A., Romano, L.: Communication of Findings of Radiologic Examinations: Medicolegal Considerations. Seminars in Ultrasound, CT and MRI **33**(4), 376–378 (Aug 2012). https://doi.org/10.1053/j.sult.2012.01.014
23. Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., Reede, D.: Interpretive Error in Radiology. AJR. American journal of roentgenology **208**(4), 739–749 (Apr 2017). https://doi.org/10.2214/AJR.16.16963
24. Wang, Z., Luo, X., Jiang, X., Li, D., Qiu, L.: LLM-RadJudge: Achieving Radiologist-Level Evaluation for X-Ray Report Generation (Apr 2024), http://arxiv.org/abs/2404.00998
25. Wu, J., Kim, Y., Shi, D., Cliffton, D., Liu, F., Wu, H.: SLaVA-CXR: Small Language and Vision Assistant for Chest X-ray Report Automation (Sep 2024). https://doi.org/10.48550/arXiv.2409.13321
26. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 457–466. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_52
27. Zhu, Q., Chen, X., Jin, Q., Hou, B., Mathai, T.S., Mukherjee, P., Gao, X., Summers, R.M., Lu, Z.: Leveraging Professional Radiologists' Expertise to Enhance LLMs' Evaluation for Radiology Reports (Feb 2024). https://doi.org/10.48550/arXiv.2401.16578, arXiv:2401.16578 [cs]