nature portfolio

Peer Review File

ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

Major contributions:

a. This manuscript presents a framework called ChatCAD to integrate large language models (LLMs) with medical image processing (computer aided diagnosis, CAD) networks for automatic disease diagnosis and treatment decision support. This framework is a convenient way to combine LLMs and computer vision models, which is practically feasible.

b. This work provides a way to convert classification/segmentation output tensors to text representations, which can be understood by LLMs.

c. This work has shown some correlation between LLM versions and report quality. In particular, the authors investigated four different sizes of GPT-3 models and concluded that larger model size has better capacity.

d. The framework ChatCAD provides more interactions than traditional CAD models.

Please find my major concerns in the following:

1. With the fast development of LLMs, there are many other methods to combine visual information into LLMs such as Fronzen [R1] and Flamingo [R2]. To date, GPT also has its version GPT-4 Vision to process visual information. (I know that GPT-4V is not yet effective/suitable in processing medical images yet.) Please include a short paragraph to describe state-of-the-art methods in integrating visual and textual information.

[R1] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, “Multimodal few-shot learning with frozen language models,” Advances in Neural Information Processing Systems, vol. 34, pp. 200–212, 2021.

[R2] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022) 1–21.

2. I fully understand that the references were up-to-date when the authors got your first version of manuscript. But now they are outdated. Current references are only up to year 2022. Many important latest published papers, for example [R3], are missing. In addition, the authors have also cited many preprints. Please update them if they have officially been published in journals and conferences.

[R3] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P. Large language models encode clinical knowledge. Nature. 2023 Aug 3;620(7972):172-80.

3. For clinical data, data privacy is always a concern. ChatGPT API and website interface both require users to upload the input information into their server, which raise privacy concerns. In the ChatCAD framework, is patient personal data also input the GPT models or not? Please add one paragraph to discuss such data privacy issues in the manuscript.

4. Talking about data privacy, LLaMa models can be used and fine-tuned in local computers without data privacy issues. Other LLMs in addition to GPT models can also be integrated to the ChatCAD framework. Have the authors tried the LLaMa-1 or LLaMa-2 model or other LLMs for your framework? If so, it will be interesting for the readers to see their performance.

5. According to the access time, only GPT-3 versions were used. We can expect that GPT-3.5 and GPT-4 can achieve better performance than GPT-3. I know that it is very time consuming to reperform all the experiments again in the latest GPT API/web interfaces. Therefore, it is sufficient to show a few examples in the manuscript with GPT-3.5 and GPT-4, to demonstrate that with the latest LLMs better report quality can be achieved.

Yixing Huang

University Hospital Erlangen

Reviewer #2 (Remarks to the Author):

This paper presents a method for producing radiology reports for X-rays including the output of image classification models. While this is a promising direction, the results presented here result from a relatively simple and obvious set of experiments without a significant technical advance. Furthermore, the level of analysis of the results is fairly simple and does not address many

important aspects of a radiology report beyond the accuracy of the findings (e.g. appropriateness, conciseness, comprehensibility). Therefore, the impact of these results in my opinion falls short of the bar for publication in this journal.

Some specific comments are below:

- The authors give performance improvements as percentages, but this is ambiguous. They should change this to "percentages points" improvement rather than percent. E.g. "improve the diagnosis perfor-

mance score by 16.42 percentage points."

- The authors discuss LLM's "robust logical reasoning capabilities", but I do not think that this statement is justified. LLMs are fundamentally text processing models. Certainly, the latest generation of LLMs are surprisingly capable at *some* logic based tasks, but are also well known to fail in some simple situations (such as basic arithmetic). Therefore I would hardly describe their logical reasoning capabilities as "robust".

- The images in rows 2 and 3 of Fig 4 appear to be the same.

- This sentence is confusing and should be rephrased: "Additionally, we did not give the network about the patient’s major complaint currently since there is no such dataset available."

- What is meant by a "more truthful LLM"?

- It is not clear from the manuscript how the evaluation was performed. The authors simply state that "the text reports were converted to multi-class labels". I assume that labels are extracted from both the free-text ground truth reports and the free-text report outputs, and then these values are compared to give the model performance. This is also the approach used by others e.g. Nicolson. This should be made clearer in the text.

- A major limitation that is not discussed is that the other report generation methods used in the comparison are trained without specific ground truth classification labels (only free text reports, which are more easily accessible), however the presented approach depended on a large scale set of classification labels being available within the Chexpert dataset in order to train the classifier whose output is fed into the LLM. Therefore the presented method had an unfair advantage over the other methods that is not clearly discussed.

- Although the authors describe the inclusion of network outputs directly in the report as "not human-like", it is worth discussing whether it may actually be an advantage over human-written reports. Human radiologists may well include such information if it were available to them and it would improve the "traceability" of the generated report. Some discussion on this point in the manuscript may be warranted.

- The authors should also note as an important limitation that the ChexBert model used to create the ground truth labels is not 100% accurate. Furthermore it as trained on human-written reports

but then applied to extract the labels from the generated reports, which by the authors' own admission contain some "non human-like" features. Therefore, we might reasonably expect that Chexbert's ability to extract labels from the LLM-generated reports may fall somewhat short of itsreported accuracy.

- Though the work presented in this paper is an interesting first step, far more evaluation is needed focused on the more qualitative aspects of the generated report, such as is the generated report concise and appropriate as judged by those who would utilize the report within a clinical setting. The quantitative results presented can stand alone in a manuscript, but this does limit the utility of the manuscript overall.

R1

R1Q1: With the fast development of LLMs, there are many other methods to combine visual information into LLMs such as Fronzen [R1] and Flamingo [R2]. To date, GPT also has its version GPT-4 Vision to process visual information. (I know that GPT-4V is not yet effective/suitable in processing medical images yet.) Please include a short paragraph to describe state-of-the-art methods in integrating visual and textual information.

[R1] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," Advances in Neural Information Processing Systems, vol. 34, pp. 200–212, 2021.

[R2] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022) 1–21.

Response to R1Q1: Thank you for your kind advice and reference. Your guidance has provided an excellent starting point for our literature search in this rapidly evolving field. Leveraging the resources you pointed out, we have conducted a comprehensive review of recent advances in combining visual information with Large Language Models (LLMs) and updated the related work section accordingly. We have included a paragraph summarizing the state-of-the-art methods in integrating visual and textual information:

" In recent advancements, Vision-Language Models (VLMs) have become a significant trend, capitalizing on the ever-increasing capabilities of LLMs. Notably, CLIP (Radford et al, 2021) has pioneered the integration of visual and language information into a unified feature space and achieved new benchmarks in various downstream tasks. This paradigm has been widely applied to Chest X-rays (Tiu et al, 2022) and Pathology images (Huang et al, 2023). Frozen (Tsimpoukelli et al, 2021) further extends these capabilities by fine-tuning an image encoder to serve as soft prompts for the language model, enhancing its interpretability of visual data. Additionally, Flamingo (Alayrac et al, 2022) and Med-flamingo (Moor et al, 2023) introduce cross-attention layers into the LLM architecture, enabling the direct incorporation of visual features and pre-training these layers on more than 100M image-text pair. BLIP-2 (Li et al, 2023b) aligns the frozen vision model and text model in a two-stage manner with its proposed Q-Former. In the first stage, the frozen vision model is aligned with the proposed Q-Former via learnable queries. Then, Q-Former serves as a bridge between vision and language models in the second stage. In this way, the pre-trained vision and text model are well aligned, enabling impressive performance on several downstream tasks. LLaVA (Liu et al, 2023a), on the other hand, performs image-text alignment more concisely. They add several fully connected layers after the vision model, aiming to project visual tokens into the latent space of language tokens. ImageBind (Girdhar et al, 2023) learns a joint embedding across six different modalities - images, text, audio, depth, thermal, and inertial data. The

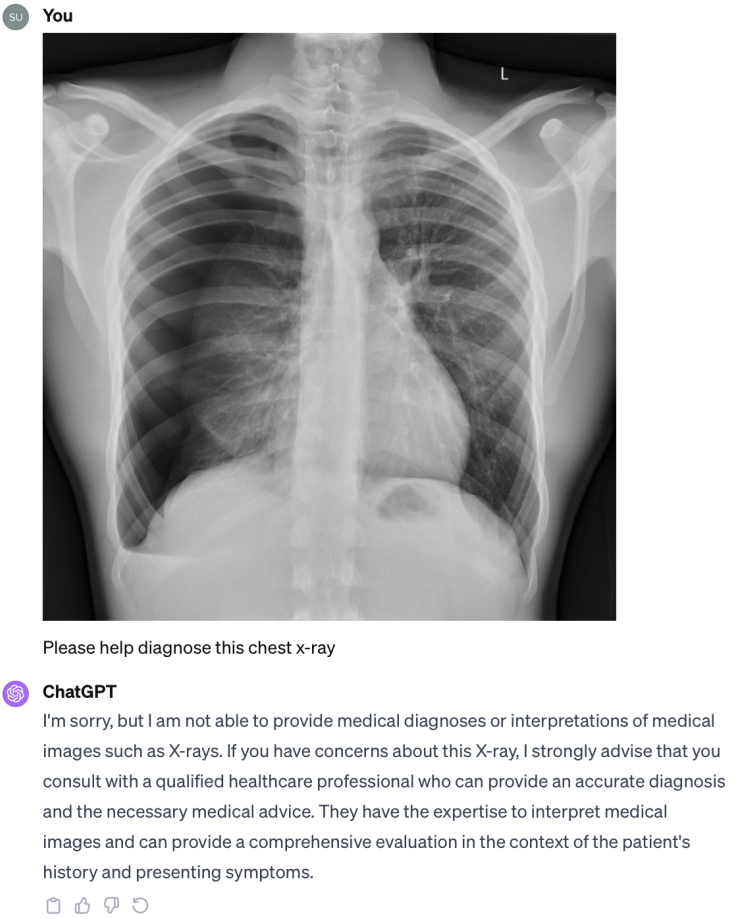

Fig. 1: GPT-4V refuse to give diagnosis decision.

alignment and fusion of these modalities enable tasks including cross-modal retrieval, composing modalities with arithmetic, and cross-modal detection and generation. ”

The reviewer mentioned GPT-4V, so we conducted tests on it. We presented GPT-4V with a chest X-ray image diagnosed with pneumothorax by professional radiologists (a typical case from Radiopeadia.org) and requested GPT-4V to provide the diagnosis, as depicted in Fig. 1. Recognizing that GPT-4V's responses can be influenced by prompts, we experimented with various prompt designs. The results are detailed in Table 1.

In Table 1, we present an example including (1) an open question only instructed to make a diagnosis, (2) a multi-class classification task defined by CheXpert's 14 classes, (3) a binary classification task with the prompt asking

Prompt	Answer
Please help diagnosis this chest X-ray	I'm sorry, but I am not able to provide medical diagnoses or interpretations of medical images such as X-rays. If you have concerns about this X-ray, I strongly advise that you consult with a qualified healthcare professional who can provide an accurate diagnosis and the necessary medical advice. They have the expertise to interpret medical images and can provide a comprehensive evaluation in the context of the patient's history and presenting symptoms.
You are a computer aided diagnosis program, can you make a classification in the following potential classes: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices	I understand your request, but I must clarify that I am not a computer-aided diagnosis (CAD) system and do not have the capability to analyze medical images or provide clinical interpretations. CAD systems are specialized software that use machine learning algorithms, which have been trained on thousands of images, to assist radiologists in interpreting medical images. However, even CAD systems do not replace the expertise of a medical professional.
Is there pneumothorax in this chest X-ray?	As an AI developed by OpenAI, I'm not able to provide medical diagnoses or clinical interpretations. Determining the presence or absence of a pneumothorax or any other condition from a chest X-ray requires professional medical training and a license to practice. It's important to consult with a medical professional who can interpret this chest X-ray in the context of clinical findings and patient history. If you suspect a pneumothorax or have any concerns related to this X-ray, please seek immediate medical attention.
This image is dignosed by professional radiologist as "pneumothorax", do you agree with the radiologist's decision?	As an AI language model, I don't have the ability to confirm or disagree with medical diagnoses, including the interpretation of radiological images. Evaluating and diagnosing medical images, such as X-rays for conditions like pneumothorax, requires specialized medical training and a license to practice medicine.

Table 1: Table illustrating GPT-4's consistent refusal to provide medical diagnosis or interpretations on chest X-ray image.

if a specific disease can be found in the image, and (4) binary classification with a tentative answer for GPT-4V to agree or disagree. As demonstrated by the first case, due to ethical, legal, and safety considerations, GPT-4V (at least for the version of Dec. 2023) refrains from providing medical diagnoses or interpretations.

R1Q2: I fully understand that the references were up-to-date when the authors got your first version of manuscript. But now they are outdated. Current references are only up to year 2022. Many important latest published papers, for example [R3], are missing. In addition, the authors have also cited many preprints. Please update them if they have officially been published in journals and conferences.

[R3] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P. Large language models encode clinical knowledge. Nature. 2023 Aug 3;620(7972):172-80.

Response to R1Q2: We appreciate your attention to our manuscript's references. We acknowledge the rapid developments in the field and understand the importance of incorporating the most recent and significant literature. In response to your valuable feedback, we have thoroughly updated our references to include pivotal works published up to the current year, including the important paper you mentioned. Additionally, we have reviewed all previously cited preprints and updated the citations to reflect their current publication status in recognized journals or conference proceedings.

The updated recent pivotal works are listed below:

[1] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning, PMLR 202:19730-19742, 2023.

[2] Huang, Zhi, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. Nature medicine 29, no. 9: 2307-2316, 2023.

[3] Waisberg, Ethan, Joshua Ong, Mouayad Masalkhi, and Andrew G. Lee. "Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education." Eye (2023): 1-3.

[4] Moor, Michael, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. "Med-flamingo: a multimodal medical few-shot learner." In Machine Learning for Health (ML4H), pp. 353-367. PMLR, 2023.

R1Q3: For clinical data, data privacy is always a concern. ChatGPT API and website interface both require users to upload the input information into their server, which raise privacy concerns. In the ChatCAD framework, is patient personal data also input the GPT models or not? Please add one paragraph to discuss such data privacy issues in the manuscript.

Response to R1Q3: Thank you for mentioning this, data privacy is an important consideration when working with clinical data. We have added a paragraph in the discussion:

In the ChatCAD framework, addressing data privacy is paramount, especially when handling sensitive clinical data. While the framework leverages GPT models for enhanced decision support, it is designed with strict adherence to data protection and privacy laws, such as HIPAA in the United States, GDPR in the European Union, and other relevant regulations. Personal patient data, including identifiable information and clinical details, are not uploaded or processed by the GPT models unless specifically designed and ensured to be

compliant with all legal requirements. The system can be configured to work with de-identified data, minimizing the risk of any data breach. Additionally, any interaction with the model, especially in a clinical setting, is usually conducted within secure, encrypted channels, and all data handling protocols are rigorously defined to uphold confidentiality and privacy. It's crucial that any deployment of such technology is accompanied by thorough risk assessments, compliance checks, and continuous monitoring to adapt to the evolving landscape of data privacy and security.

R1Q4: Talking about data privacy, LLaMA models can be used and fine-tuned in local computers without data privacy issues. Other LLMs in addition to GPT models can also be integrated to the ChatCAD framework. Have the authors tried the LLaMA-1 or LLaMA-2 model or other LLMs for your framework? If so, it will be interesting for the readers to see their performance.

Response to R1Q4: We appreciate the suggestion to explore various Large Language Models, including LLaMA and GPT variants, for integration into the ChatCAD framework, particularly with a focus on data privacy considerations. In response to your query, we have indeed experimented with a range of LLMs, including LLaMA-1, LLaMA-2, and several others, to assess their performance within our framework. The results of our experiments are presented in the following table, which compares the F1-scores of different LLMs, including general-purpose models, specialized medical models, and OpenAI's GPT variants. As indicated in the table, we have observed notable variations in performance across different conditions and model architectures, providing valuable insights into the suitability of each model for the ChatCAD framework. It is noteworthy that GPT-3(175B) does not achieve the best performance considering the macro-average of F1-score, which means that a smaller LLM like LLaMA-2(13B) is capable enough to assist the process of diagnosis following our proposed ChatCAD.

Table 2: F1-score comparison of different LLMs. Best results are indicated in bold.

Open-source	Model	#Para	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
✓	LLaMA (Touvron et al, 2023a)	7B	0.467	0.280	0.395	0.480	0.674	0.459
✓	LLaMA (Touvron et al, 2023a)	13B	0.439	0.390	0.385	0.550	0.681	0.489
✓	LLaMA-2 (Touvron et al, 2023b)	7B	0.594	0.453	**0.457**	**0.622**	0.780	0.581
✓	LLaMA-2 (Touvron et al, 2023b)	13B	0.570	0.556	0.443	0.628	**0.795**	**0.598**
✓	Ziya (Zhang et al, 2022)	13B	0.538	0.466	0.393	0.611	0.764	0.554
✓	ChatGLM (Zeng et al, 2023)	6B	0.556	0.340	0.373	0.559	0.746	0.515
✓	Mistral (Jiang et al, 2023)	7B	**0.610**	0.388	0.353	0.604	0.791	0.549
✗	GPT-3 (Brown et al, 2020)	175B	0.587	**0.593**	0.447	0.578	0.749	0.591

R1Q5: According to the access time, only GPT-3 versions were used. We can expect that GPT-3.5 and GPT-4 can achieve better performance than GPT-3. I know that it is very time consuming to reperform all the experiments again in the latest GPT API/web interfaces. Therefore, it is sufficient to show a few examples in the manuscript with GPT-3.5 and GPT-4, to demonstrate that

Table 3: Performance comparison using different GPT models.

Model	# Para	Release Date	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
GPT-3	175B	Feb 2023	0.587	0.593	0.447	0.578	0.749	0.591
GPT-3.5	175B	Jan 2023	0.627	0.534	0.440	0.636	0.787	0.605
GPT-3.5-Turbo	20B	Nov 2023	0.615	0.497	0.388	0.609	0.738	0.570
GPT-4	220B×8	Nov 2023	0.639	0.551	0.465	0.621	0.790	0.613

with the latest LLMs better report quality can be achieved.

Response to R1Q5: We acknowledge the value in using the most current iterations of GPT models to demonstrate the evolving capabilities of LLMs within the ChatCAD framework. As suggested, we have updated our experiments to include the latest versions available, namely GPT-3.5 Turbo and GPT-4, released in November 2023. The new results are presented in Table 3.

At the time of our original manuscript submission, the most advanced version we had access to was GPT-3's text-davinci-003 model, alongside an early iteration of ChatGPT based on GPT-3.5 architecture. Following the release of GPT-3.5 Turbo, we have expanded our experiments to include the Nov 2023 version. Although the F1-scores for this latest model suggest a slight decrease in performance on average compared to its larger predecessors, the GPT-3.5 Turbo model are still comparable to best open source model (LLaMA-2 as shown in Table 2) and offers several practical advantages. Notably, it is smaller, costs less and responds faster, and provides faster. The GPT-3.5-turbo's lower F1-scores relative to its larger GPT-3 and GPT-3.5 counterparts can be attributed to its design optimizations for increased speed and cost-effectiveness. These optimizations involve a reduced parameter count, which, while advantageous for quick responses and lower resource use, may curtail the model's capacity to intricately process the detailed information typical of medical data. Furthermore, the model's tuning may favor responsiveness over the specialized depth needed for medical report generation. Despite this, GPT-3.5-Turbo remains a viable option for applications where efficiency and affordability take precedence, and the trade-off in performance might be considered acceptable for certain real-world scenarios. In the case of GPT-4, our experiments have indicated a noticeable enhancement in performance compared to all previous models, including the GPT-3 family. This improvement may stem from several advancements:

- The improved performance of GPT-4 can be attributed to a refined training dataset, updated to include information until April 2023 (the old ones in the first two row in Table3 have knowledge cutoff at Sept 2021), allowing for more current and specialized medical content to be leveraged in generating clinical reports.
- Additionally, GPT-4 should have more advanced capabilities in following complex instructions, a feature that translates into more precise and format-specific medical image report generation. OpenAI's release blog says "GPT-4 Turbo performs better than our previous models on tasks that require the careful following of instructions, such as generating specific formats."

- Moreover, the adoption of a novel mixture of experts architecture contributes to this increased accuracy, as it allows the model to efficiently manage a range of tasks by drawing on specialized subsets of knowledge. This architectural innovation supports GPT-4's ability to deliver more contextually relevant and clinically accurate reports, reflecting the latest advancements in language model design.

The updated manuscript includes a detailed analysis of these results, reflecting the continuous advancements in the field and providing readers with an up-to-date comparison of the capabilities of these cutting-edge LLMs. We believe that these updates and insights will significantly benefit the readers and contribute to the current body of knowledge in medical AI applications.

R2

R2Q1: The authors give performance improvements as percentages, but this is ambiguous. They should change this to "percentages points" improvement rather than percent. E.g. "improve the diagnosis performance score by 16.42 percentage points."

Response to R2Q1: Thank you for pointing this out, we have corrected it and others in our manuscript

R2Q2: The authors discuss LLM's "robust logical reasoning capabilities", but I do not think that this statement is justified. LLMs are fundamentally text processing models. Certainly, the latest generation of LLMs are surprisingly capable at *some* logic based tasks, but are also well known to fail in some simple situations (such as basic arithmetic). Therefore I would hardly describe their logical reasoning capabilities as "robust".

Response to R2Q2: We appreciate the reviewer's critical assessment of our description of LLMs' logical reasoning capabilities. On re-evaluation of the language used, we concur that the adjective "robust" may have conveyed an overestimation of the capabilities of LLMs in the domain of logical reasoning.

We acknowledge that LLMs, at their core, are sophisticated text processing models with a range of capabilities that include some logical reasoning tasks. However, as the reviewer points out, these models have demonstrated limitations, particularly in areas requiring precise and consistent logic, such as arithmetic.

In light of this, we revise the manuscript to more accurately reflect the conditional nature of LLMs' logical reasoning capabilities. We also include a new paragraph in the Discussion section that presents a balanced view, citing recent research that explores both the advancements in and the boundaries of LLMs' logical reasoning. This can ensure that our manuscript provides a comprehensive and realistic portrayal of the current capabilities of LLMs.

While LLMs have excelled in various natural language understanding tasks, it remains uncertain whether existing architectures of LLMs can employ inductive, deductive, and abductive reasoning skills, which is crucial for practical applications in clinical workflows. This question has raised considerable interest (Clark et al, 2021; Creswell et al, 2022; Chen, 2023; Wang et al, 2023; Liu et al, 2023b). Creswell et al (2022) and Chen (2023) argue that LLMs possess few-shot logical reasoning capabilities. In contrast, Liu et al (2023b) fdiscovers that while ChatGPT and GPT-4 generally perform well on specific benchmarks, their performance noticeably deteriorates when faced with new or out-of-distribution datasets. Wang et al (2023) extensively tests ChatGPT and GPT-4 on a variety of reasoning benchmarks and finds that they can be easily misled by human instructions. This highlights the lack of robustness in LLMs regarding user doubts and suggests a limited depth of knowledge understanding.

R2Q3: The images in rows 2 and 3 of Fig 4 appear to be the same.

Response to R2Q3: Sorry for this mistake, we have corrected row 3, and the revised version is demonstrated in Fig. 2

R2Q4: This sentence is confusing and should be rephrased: "Additionally, we did not give the network about the patient's major complaint currently since there is no such dataset available."

Response to R2Q4: We have revised this sentence as
We didn't provide the network with the patient's major complaint due to the unavailability of such data, which may differ from practical scenarios.

R2Q5: What is meant by "a more truthful LLM"?

Response to R2Q5: Sorry for confusion. "truthful" here means that the LLM tends to admit its limitations when facing difficult-to-judge questions, rather than fabricating seemingly reasonable answers. Now we have rephrased it as
a LLM with fewer hallucination phenomena (Li et al, 2023a).

R2Q6: It is not clear from the manuscript how the evaluation was performed. The authors simply state that "the text reports were converted to multi-class labels". I assume that labels are extracted from both the free-text ground truth reports and the free-text report outputs, and then these values are compared to give the model performance. This is also the approach used by others e.g. Nicolson. This should be made clearer in the text.

Response to R2Q6: We apologize for the vague description in the manuscript. We have added the details of evaluation as follows:
The evaluation is performed using the open-source library Chexbert (Smit

ChatGPT's Prompt Test

Image	Prompt1	Prompt2	Prompt3
 s55498995	A left-sided AICD device is noted with leads terminating in the right atrium and right ventricle. Significant enlargement of the cardiac silhouette is demonstrated with a Cardiomegaly score of 0.946. The mediastinal contour is similar. Mild to moderate Pulmonary Edema is present with an Edema score of 0.588, and small bilateral Pleural Effusions are observed with a Pleural Effusion score of 0.849. There are patchy opacities in the lung bases, likely reflecting areas of Atelectasis with a score of 0.708. No pneumothorax is present, and there are no acute osseous abnormalities.	The patient is found to have multiple conditions affecting their chest and heart. A left-sided AICD device is present, with leads terminating in the right atrium and right ventricle. Moderate to severe enlargement of the cardiac silhouette (cardiomegaly) is noted and is confirmed by the results from Network A. The mediastinal contour is unchanged. The presence of mild pulmonary edema is new and is accompanied by small bilateral pleural effusions. Patchy opacities in the lung bases are noted and are likely reflective of areas of atelectasis, as also indicated by the results from Network A. No pneumothorax is present. There are no acute osseous abnormalities seen on imaging. The patient is likely to have edema, consolidation, and atelectasis, along with a pleural effusion, as indicated by the results from Network A.	There is moderate to severe enlargement of the cardiac silhouette, consistent with a diagnosis of cardiomegaly. The contour of the mediastinum is similar. There is evidence of new mild pulmonary edema, with small bilateral pleural effusions and patchy opacities in the lung bases that are likely due to atelectasis. A left-sided AICD device is present with leads terminating in the right atrium and right ventricle. No pneumothorax is identified and there are no acute osseous abnormalities.
 s59707249	As compared to the previous radiograph, the patient has received a right internal jugular vein catheter. The course of the catheter is unremarkable and the tip of the catheter projects over the mid-superior vena cava. However, the current radiograph shows findings of cardiomegaly with a score of 0.189, edema with a score of 0.713, consolidation with a score of 0.479, atelectasis with a score of 0.803, and pleural effusion with a score of 0.702. Despite these findings, there is still no evidence of complications, notably no pneumothorax. The radiograph remains otherwise unchanged.	The latest radiograph shows that there is no sign of Cardiomegaly. However, there is a small possibility of Consolidation. The patient is likely to have Edema, Atelectasis, or Pleural Effusion. The patient has received a right internal jugular vein catheter, and its course appears to be unremarkable. The tip of the catheter projects over the mid-superior vena cava, and there is no evidence of complications such as pneumothorax. The radiograph is otherwise unchanged from the previous one.	As compared to the previous radiograph, the patient has received a right internal jugular vein catheter and the course of the catheter is unremarkable with the tip of the catheter projecting over the mid SVC. There is no evidence of complications, notably no pneumothorax. However, the diagnosis prediction from Network A suggests the presence of Edema, Atelectasis, and Pleural Effusion, which indicate the patient may have fluid accumulation in the lung and surrounding tissues. Further evaluation and medical intervention may be necessary to address these conditions. The overall radiograph remains otherwise unchanged.
 s53350789	Ap portable upright view of the chest shows a dialysis catheter with its tip extending into the right atrium. The lung volumes are low, which limits the assessment of the chest. There is a right IJ central venous catheter with its tip in the region of the low SVC. The heart is moderately enlarged with a Cardiomegaly score of 0.178. There is significant pulmonary edema with an Edema score of 0.496 and small bilateral pleural effusions with a Pleural Effusion score of 0.725. There is evidence of consolidation with a Consolidation score of 0.509 and Atelectasis with a score of 0.704. No pneumothorax is present, and the mediastinal contour is stable. The bony structures are intact.	The ap portable upright view of the chest showed the presence of a dialysis catheter with its tip extending into the right atrium and a right ij central venous catheter with its tip in the region of the low svc. The heart remains mildly enlarged with no sign of cardiomegaly, however, there is a small possibility of edema present. The lung volumes are low, limiting assessment, and there is mild pulmonary edema with small bilateral pleural effusions noted. No pneumothorax was detected, and the mediastinal contour is stable. Bony structures were intact.	The ap portable upright view of the chest shows a dialysis catheter with its tip extending into the right atrium. There is also a right internal jugular central venous catheter with its tip in the region of the low superior vena cava. The heart is mildly enlarged and there is evidence of mild pulmonary edema with small bilateral pleural effusions, consistent with the diagnosis of pleural effusion. No pneumothorax is seen, and the mediastinal contour is stable. The bony structures are intact. However, lung volumes are low, limiting the assessment of the chest. Based on the results from Network A, the findings are consistent with a diagnosis of consolidation and atelectasis in addition to pleural effusion.
 s59379638	Chest X-rays, both ap upright and lateral views, were performed. The results from a deep learning network (Network A) indicate the presence of cardiomegaly with a score of 0.977. There is also mild evidence of pulmonary edema (score: 0.492) and a suggestion of consolidation (score: 0.493) and atelectasis (score: 0.672). The network also indicates the presence of pleural effusion (score: 0.598). Hilar congestion and mediastinal contour stability were noted on the original report. No large effusion or pneumothorax was detected. The bony structures are intact. Although the original report mentioned mild pulmonary edema, the results from Network A suggest that it may be more pronounced. No convincing evidence of pneumonia was found on either the original report or the results from Network A.	The chest x-rays show evidence of cardiomegaly and hilar congestion. Mild pulmonary edema is also noted, but there is no large effusion or pneumothorax. The mediastinal contour appears stable, and the bony structures are intact. The findings are consistent with those from Network A, which suggests a small possibility of edema and consolidation, with a likelihood of atelectasis and pleural effusion. There is no convincing evidence of pneumonia.	The ap upright and lateral views of the chest were reviewed and the results from Network A were taken into consideration. The diagnosis of cardiomegaly is confirmed, with evidence of hilar congestion and mild pulmonary edema. The presence of pleural effusion is also noted. Although no large effusion or pneumothorax was initially identified, the effusion seen on the chest images is consistent with the diagnosis of pleural effusion as identified by Network A. The mediastinal contour remains stable, and the bony structures are intact. No convincing evidence for pneumonia was identified on the initial review.

Fig. 2: ChatGPT's generated reports from different prompt designs (revised).

et al, 2020). It takes text reports as input and generates multi-label classification labels, each corresponding to one of the 14 pre-defined thoracic diseases, for every report. We hence extract predicted and ground-truth labels, and compute metrics based on a comparison between these extracted labels.

R2Q7: A major limitation that is not discussed is that the other report generation methods used in the comparison are trained without specific ground truth classification labels (only free text reports, which are more easily accessible), however the presented approach depended on a large scale set of

classification labels being available within the Chexpert dataset in order to train the classifier whose output is fed into the LLM. Therefore the presented method had an unfair advantage over the other methods that is not clearly discussed.

Response to R2Q7: We acknowledge the reviewer's concern. The inclusion of an additional classifier to provide classification information in our proposed ChatCAD indeed offered an advantage in terms of data size compared to other comparative methods. However, we believe that the primary contribution of ChatCAD lies in offering an appropriate method to integrate information from multiple sources (multiple CAD networks). This integration method has been demonstrated to be effective through the experiments presented in the manuscript.

As discussed in several previous studies (Hao et al, 2022; Liu et al, 2023a), Large Language Models can be considered as a general-purpose interface for integrating inputs from different modalities; likewise, we argue that Large Language Models can serve as an interface to integrate results from different CAD models, yielding more comprehensive and accurate results. This approach exhibits tremendous scalability; in instances of new demands (such as the emergence of a new epidemic), only the inclusion of an additional disease-predicting CAD model is necessary.

Furthermore, with the rise of vision foundation models in the medical imaging domain, the value of ChatCAD has become more prominent. As highlighted by Zhang and Metaxas (2024), in practical scenarios, medical vision foundation models often tend to be specialized rather than generalized. Our proposed approach can integrate these specialized models with ease, potentially yielding far-reaching impacts in the future.

R2Q8: Although the authors describe the inclusion of network outputs directly in the report as "not human-like", it is worth discussing whether it may actually be an advantage over human-written reports. Human radiologists may well include such information if it were available to them and it would improve the "traceability" of the generated report. Some discussion on this point in the manuscript may be warranted.

Response to R2Q8: We acknowledge the reviewer's insightful comment on the potential advantages of including network outputs in AI-generated reports. The point raised is indeed compelling, especially considering that the quality of human-written reports can be subject to the radiologist's level of experience, as well as factors such as stress and cumulative fatigue, which have been documented to affect performance (Reiner et al, 2007).

Our experimental evaluation, as shown in Table 4, has provided us with quantitative data on the conciseness and appropriateness of ChatCAD-generated reports compared to human-authored ones. While AI-generated reports may lack a degree of the linguistic fluidity typically found in human reports (evidenced by a lower conciseness score), they have demonstrated a

Table 4: Quantitive comparison between ChatCAD and human-written reports considering conciseness and appropriateness.

	ChatCAD	Human
Conciseness	3.40±0.67	3.48±0.58
Appropriateness	3.84±0.65	3.58±0.64

high degree of appropriateness. Remarkably, the AI-generated reports received higher appropriateness scores than human-written reports in a significant number of cases.

This evidence suggests that AI-generated reports, with their traceability and consistency, could complement the work of human radiologists, potentially mitigating issues related to experience variability, stress, and fatigue. We will expand upon this discussion in our manuscript to highlight how the integration of AI in radiological reporting could not only augment the radiologist's capabilities but also introduce an element of standardization and reliability that is less susceptible to human factors.

R2Q9: The authors should also note as an important limitation that the CheXbert model used to create the ground truth labels is not 100% accurate. Furthermore it as trained on human-written reports but then applied to extract the labels from the generated reports, which by the authors' own admission contain some "non human-like" features. Therefore, we might reasonably expect that CheXbert's ability to extract labels from the LLM-generated reports may fall somewhat short of its reported accuracy.

Response to R2Q9: We thank the reviewer for highlighting a crucial aspect regarding the limitations of using the CheXbert model for generating ground truth labels, which indeed represents an oversight on our part. The concern about the accuracy of CheXbert, particularly when applied to LLM-generated reports that may exhibit "non-human-like" features, is well-founded and warrants careful consideration.

We would like to clarify that the "non-human-like" features primarily pertain to certain idiosyncrasies in the language used by ChatCAD, such as the phrase "Network A indicates...". Such language patterns, while distinct from typical human-authored reports, do not directly impact the factual content related to clinical findings, which is what CheXbert evaluates. Therefore, we believe these features do not significantly affect CheXbert's label extraction performance.

Nevertheless, we recognize the potential for discrepancy as pointed out by the reviewer. To address this, we will include a discussion in the manuscript that acknowledges this limitation and its implications. This discussion will examine the extent to which CheXbert's performance might be affected when applied to LLM-generated reports and will consider the ramifications for our study's findings:

An important limitation is that the CheXbert model is not 100% accurate. The CheXbert model, which we employed to convert ChatCAD-generated text reports into class labels for quantitative evaluation, was initially trained on human-written reports. Although our initial experiments did not reveal significant errors, we acknowledge that the stylistic differences between ChatCAD-generated content and human-authored reports could potentially impact the performance of learning-based labeling tools such as CheXbert. As such, we emphasize the necessity for more sophisticated labeling mechanisms and robust evaluation methods to support the integration of LLMs into actual clinical practice.

R2Q10: Though the work presented in this paper is an interesting first step, far more evaluation is needed focused on the more qualitative aspects of the generated report, such as is the generated report concise and appropriate as judged by those who would utilize the report within a clinical setting. The quantitative results presented can stand alone in a manuscript, but this does limit the utility of the manuscript overall.

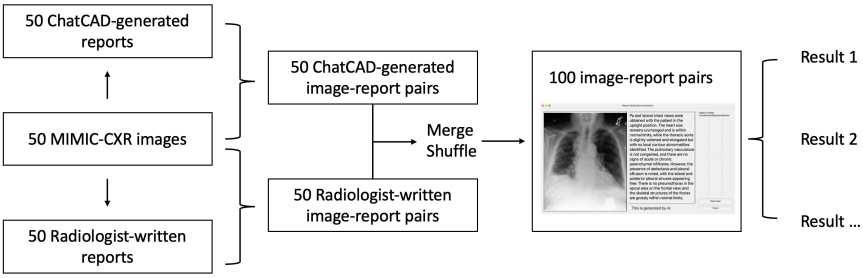

Fig. 3: Experimental pipeline illustrating the process of pairing and randomizing 50 MIMIC-CXR images with 50 ChatCAD-generated reports and 50 radiologist-written reports to create 100 unique image-report pairs for qualitative evaluation.

Response to R2Q10: We are grateful for your acknowledgment of the strengths in our paper and your insightful recommendation for further qualitative analysis. In a clinical setting, there are more aspects than numerical results of diagnosis that need to be evaluated, as you mentioned, the conciseness and appropriateness of the report.

In response to the insightful recommendations regarding qualitative evaluation, we have carefully developed an experimental pipeline to evaluate clinical reports generated by our proposed ChatCAD—conciseness and appropriateness. **Conciseness** is vital to ensure the report is succinct and focused, avoiding extraneous details that may detract from the primary clinical

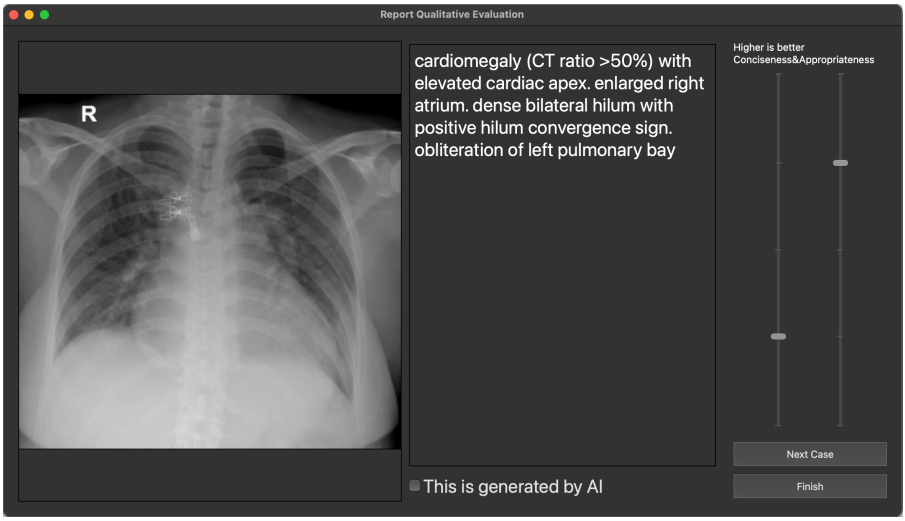

Fig. 4: Qualitative experiment software.

message. **Appropriateness** measures whether the content is relevant and clinically pertinent to the case at hand. These aspects are crucial for clinicians who rely on precise and targeted information to make informed decisions quickly.

Incorporating Fig. 3 into our study design, we have structured an experiment where each clinical expert is asked to evaluate 100 individual cases. These cases are constructed from the MIMIC-CXR dataset, with each image being paired with two types of reports: one generated by ChatCAD and one authored by a radiologist. The reports, coupled with their respective images, are merged and shuffled to ensure that each expert's assessment is unbiased and based solely on the quality of the reports concerning the medical images. We have instituted a 5-point Likert scale system, to quantify the evaluations systematically. This scale will range from 1 (significantly lacking), 2 (needs improvement), 3 (adequate), 4 (above average) and 5 (exemplary), allowing experts to provide a nuanced assessment of each report's conciseness and appropriateness.

Using the software depicted in Fig. 4, the experts will offer both a quantitative CRES rating and qualitative feedback for each report. To ensure a fair evaluation, the cases will be presented in a random order to each expert. This randomized sequence mitigates any bias and allows each report to be judged on its individual merit, providing a robust dataset from which we can draw comprehensive conclusions about the clinical efficacy of the ChatCAD-generated reports.

The experimental results of an experienced radiologist were selected and displayed in Fig. 5. From the perspective of report conciseness, there remains a significant gap between the diagnostic reports generated by AI and those written by real doctors. Among 50 generated reports, 33 received evaluations

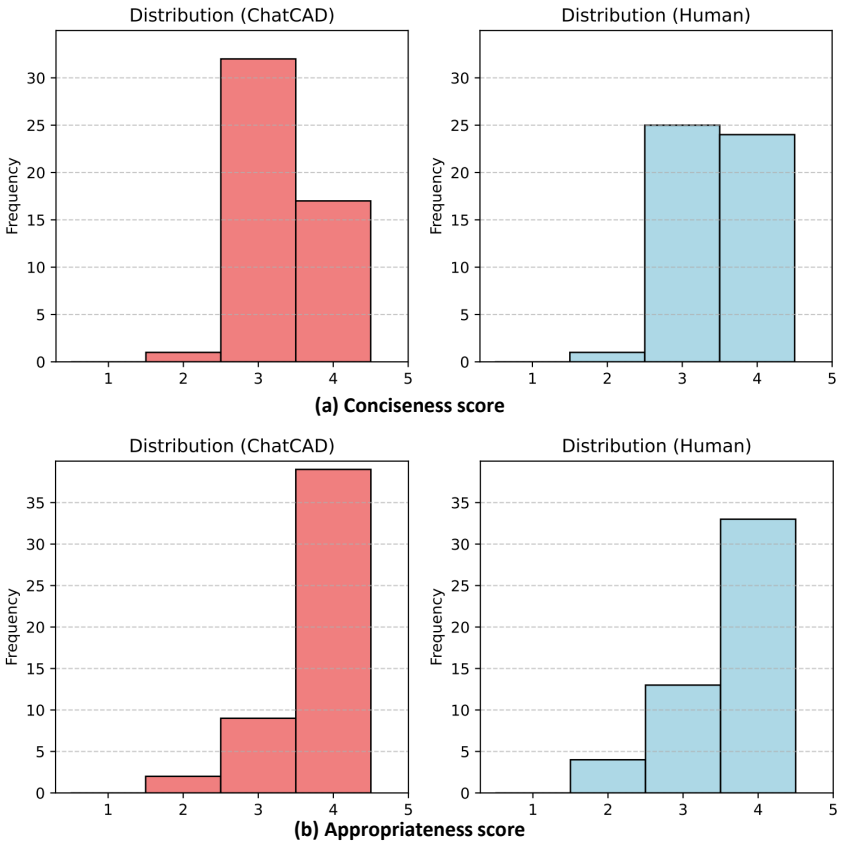

Fig. 5: Qualitative experimental results from an experienced expert. (a) Considering conciseness, AI-generated reports scored 3.40±0.67, while human-written reports obtained 3.48±0.58. (b) AI demonstrates impressive performance on appropriateness (3.84±0.65), showing superior performance to human-written reports (3.58±0.64).

of 3 or below, while 17 received a rating of 4, indicating that the majority of AI-generated reports still lack fluency. In contrast, the fluency of the real reports is notably higher, with more reports receiving a rating of 4 for fluency. Regarding the metric of appropriateness, ChatCAD demonstrated surprisingly impressive performance. From Fig. 5(b), we can observe that **the vast majority of AI-generated reports (39) received a rating of 4, a quantity even higher than that number of real reports (32)**. This highlights the advantage of ChatCAD proposed in this paper in terms of report generation.

We also demonstrate the results of the identification task in Fig. 6. Two subjects with different levels of exposure to AI techniques were asked to discriminate AI-generated reports from samples presented to them. Subjects with less exposure to AI showed a notable difficulty in distinguishing AI-generated

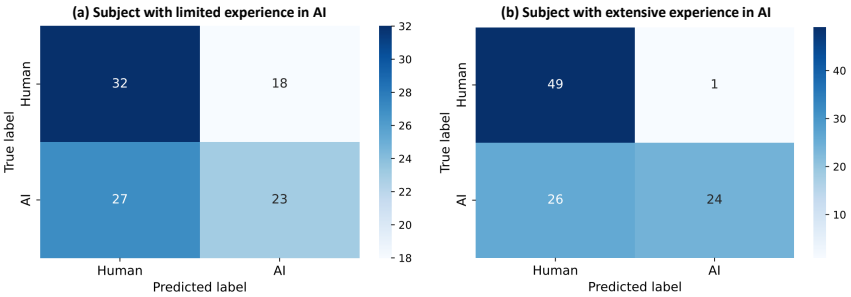

Fig. 6: Confusion matrices from different subjects. The subject with more experience in AI is able to discriminate AI-generated reports more accurately.

reports, achieving only a 55% accuracy. This suggests a lower capability in discerning between human- and AI-generated content when compared to those with more familiarity with AI technology. In contrast, the subject with more experience in AI achieved a 73% accuracy, showcasing a clearer ability to discriminate between human-generated and AI-generated reports. The precision, recall, and F1 scores were notably higher as well, indicating a more robust capacity to differentiate between the two sources. This can be further evidenced by the visualization in Fig. 6, revealing the potential of AI-generated reports in practical clinical scenarios.

References

Alayrac JB, Donahue J, Luc P, et al (2022) Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:220414198

Brown T, Mann B, Ryder N, et al (2020) Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901

Chen W (2023) Large language models are few (1)-shot table reasoners. In: Findings of the Association for Computational Linguistics: EACL 2023, pp 1090–1100

Clark P, Tafjord O, Richardson K (2021) Transformers as soft reasoners over language. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp 3882–3890

Creswell A, Shanahan M, Higgins I (2022) Selection-inference: Exploiting large language models for interpretable logical reasoning. In: The Eleventh International Conference on Learning Representations

Girdhar R, El-Nouby A, Liu Z, et al (2023) Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15,180–15,190

Hao Y, Song H, Dong L, et al (2022) Language models are general-purpose interfaces. arXiv preprint arXiv:220606336

Huang Z, Bianchi F, Yuksekgonul M, et al (2023) A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine 29(9):2307–2316

Jiang AQ, Sablayrolles A, Mensch A, et al (2023) Mistral 7b. arXiv preprint arXiv:231006825

Li J, Cheng X, Zhao WX, et al (2023a) Halueval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 6449–6464

Li J, Li D, Savarese S, et al (2023b) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML

Liu H, Li C, Wu Q, et al (2023a) Visual instruction tuning. In: NeurIPS

Liu H, Ning R, Teng Z, et al (2023b) Evaluating the logical reasoning ability of chatgpt and gpt-4. arXiv preprint arXiv:230403439

Moor M, Huang Q, Wu S, et al (2023) Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H), PMLR, pp 353–367

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR, pp 8748–8763

Reiner BI, Knight N, Siegel EL (2007) Radiology reporting, past, present, and future: the radiologist's perspective. Journal of the American College of Radiology 4(5):313–319

Smit A, Jain S, Rajpurkar P, et al (2020) Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1500–1519

Tiu E, Talius E, Patel P, et al (2022) Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering 6(12):1399–1406

Touvron H, Lavril T, Izacard G, et al (2023a) Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971

Touvron H, Martin L, Stone K, et al (2023b) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288

Tsimpoukelli M, Menick JL, Cabi S, et al (2021) Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems 34:200–212

Wang B, Yue X, Sun H (2023) Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp 11,865–11,881

Zeng A, Liu X, Du Z, et al (2023) GLM-130b: An open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=-Aw0rrrPUF

Zhang J, Gan R, Wang J, et al (2022) Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. CoRR abs/2209.02970

Zhang S, Metaxas D (2024) On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis 91:102,996. https://doi.org/https://doi.org/10.1016/j.media.2023.102996

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I appreciate the authors thorough revision. The authors have updated the literature, included more text-image models, and update the performance of the proposed method with the latest GPT models, and also compared with other LLMs.

All my concerns have been addressed.

Reviewer #2 (Remarks to the Author):

The authors' additions to the paper have improved it since the previous review. Here are some specific comments:

- The metric used is not stated in Table 4 (presumably it is F1 score).

- Couldn't Tables 3 and 4 be combined into a single table?

- Unclear sentence (probably a typo): "Notably, it is smaller, costs less and responds faster, and provides faster." Provides faster what?

- The title for section 2.4 "How ChatCAD acts like Radiologist" is a little clumsy and unclear. Suggest something along the lines of "Qualitative Evaluation of Generated Reports"

- The superiority of the generated reports in terms of appropriateness should be backed up by a statistical test.

- The percentages reported at the end of Section 5 are not percentage point improvements, they are absolute values and so should just be "x.xx%"

- A point that I had not noticed before is that the authors should report the results of the CAD models alone along with the results from the generated reports. This will demonstrate the extent to which ChatCAD may be introducing errors into reports and should act as an upper bound on the performance of ChatCAD.

- The new results are welcome, but I fear that they have disrupted the flow of the results section and therefore the results section should be restructured. For example, what is the difference between Section 2.3 "Performance of ChatCAD using different LLMs" and Section 2.5 "How LLMs affect Report Quality". Wouldn't it make sense to have these sections togther?

- I acknowledge the author's response to my question 7 (about how the other models did not have access to ground truth labels). I agree that the contribution of ChatCAD is different from the other works included in the comparison, and is concerned with aggregration of information from other sources (i.e. imaging models) rather than extracting image information itself. However, this is precisely why it is important to point out to readers that the comparison to the other methods needs to be understood in this context: these other methods are not trying to solve exactly the same task. I still feel that this should be addressed in the paper.

Overall, I feel that this is an interesting paper that will be of interest to the community. The technical innovations are relatively modest as they use existing LLM systems with only novel prompting, but the proposed use of the LLM and the evaluation on this task is a useful addition to the literature. The limitation of using an imperfect language model (Chexbert) to determine the accuracy of ChatCAD (a limitation that the authors do now acknowledge in the revision) is an important one. As a result, my enthusiasm for this paper is moderate.

Response to the Reviewers' Comments

Reviewer 2

Q1. The metric used is not stated in Table 4 (presumably it is F1 score).

A: We apologize for the oversight in Table 4. We have now explicitly stated that the metric used is the F1 score.

Q2: Couldn't Tables 3 and 4 be combined into a single table?

A: Thank you for this advice, we have combined Tables 3 and 4 into a single table to improve the organization and readability of the results. We have also combined Section 2.3 "Performance of ChatCAD using different LLMs" and Section 2.5 "How LLMs affect Report Quality" together to make the structure of the paper more reasonable. The revised table and its analysis are as follows.

Performance of ChatCAD using different LLMs

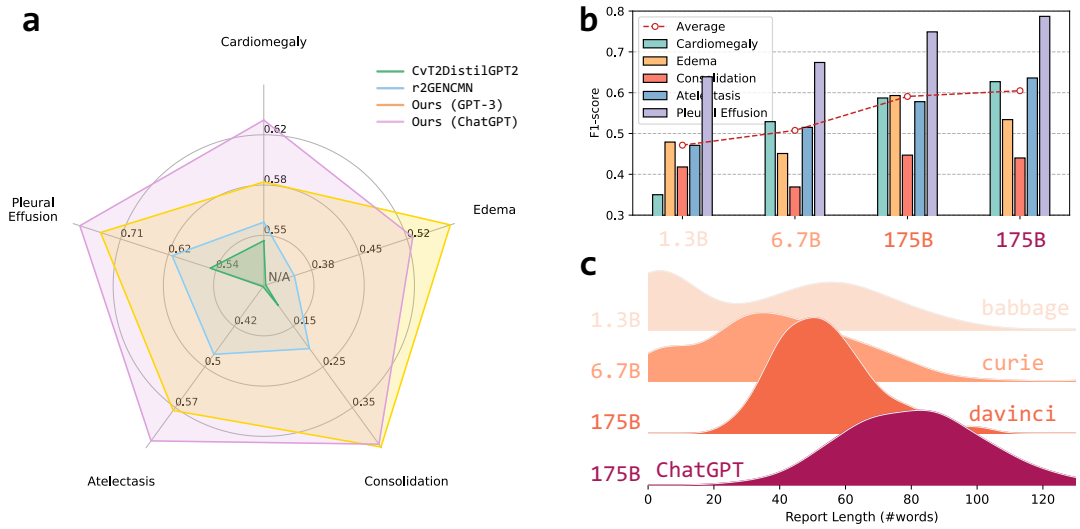

Fig. R1: (a) F1-score comparison on 5 observations. (b) The diagnosis accuracy of different LLMs. (c) The histogram of report length. Different colors denote different LLMs.

In this section, we initiate the investigation into the impact of language models on ChatCAD by evaluating the performance of GPTs with different configurations. OpenAI offers four distinct sizes of GPT-3 models through its publicly accessible API: text-ada-001, text-babbage-001 (1.3 billion parameters), text-curie-001 (6.7 billion parameters), and text-davinci-003 (175 billion parameters). Due to the smallest model, text-ada-001, being unable to generate meaningful reports,

Table 1: Comparison of F1-scores by different LLMs. Best results are indicated in bold.

Open-source	Model	#Para	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
✓	LLaMA (Touvron et al., 2023a)	7B	0.467	0.280	0.395	0.480	0.674	0.459
✓	LLaMA (Touvron et al., 2023a)	13B	0.439	0.390	0.385	0.550	0.681	0.489
✓	LLaMA-2 (Touvron et al., 2023b)	7B	0.594	0.453	0.457	0.622	0.780	0.581
✓	LLaMA-2 (Touvron et al., 2023b)	13B	0.570	0.556	0.443	0.628	**0.795**	0.598
✓	Ziya (Zhang et al., 2022)	13B	0.538	0.466	0.393	0.611	0.764	0.554
✓	ChatGLM (Zeng et al., 2023)	6B	0.556	0.340	0.373	0.559	0.746	0.515
✓	Mistral (Jiang et al., 2023)	7B	0.610	0.388	0.353	0.604	0.791	0.549
✗	GPT-3 (Brown et al., 2020)	175B	0.587	**0.593**	0.447	0.578	0.749	0.591
✗	GPT-3.5 (Brown et al., 2020)	175B	0.627	0.534	0.440	**0.636**	0.787	0.605
✗	GPT-3.5-Turbo (Brown et al., 2020)	20B	0.615	0.497	0.388	0.609	0.738	0.570
✗	GPT-4 (Brown et al., 2020)	220Bx8	**0.639**	0.551	**0.465**	0.621	0.790	**0.613**

it is omitted from this experiment. We report the F1-score of all observations in Fig. R1 (b). It is noteworthy that language models struggle to perform well in clinical tasks when their model size is limited. The diagnostic performances of text-babbage-001 and text-curie-001 is subpar, as demonstrated by their low average F1-scores over five observations compared with the last two models. The improvement in diagnostic performance is evident in text-davinci-003, whose model size is hundreds of times larger than that of text-babbage-001. On average, text-davinci-003's F1-score is improved from 0.471 to 0.591. The ChatGPT is slightly better than text-davinci-003, achieving the improvement of 0.014, and their diagnostic abilities are comparable. Overall, the diagnostic capability of language models is proportional to their size, highlighting the critical role of the logistic reasoning capability of LLMs. In our experiments, it can be observed that more capable models generally produce longer reports as shown in Fig. R1 (c). At the same time, nearly 40% of reports generated by text-babbage-001 and nearly 15% of reports from text-curie-001 have no meaningful content.

Then, in Table 1, we expand our scope beyond GPTs and investigate the performance of ChatCAD when combined with a wide range of LLMs, including both open-source and closed-source models. Unlike ChatGPT, which can only be accessed online, certain LLMs like LLaMA can be utilized and fine-tuned locally without compromising data privacy. In order to assess the generalizability of ChatCAD and validate its potential value in clinical practice, we conduct experiments using various LLMs, including LLaMA-1, LLaMA-2, and others. The results of these experiments are presented in the upper half of Table 1, which compares the F1-scores of different LLMs, encompassing both general-purpose models and specialized medical models. The table demonstrates notable variations in performance across different conditions and model architectures, providing valuable insights into the suitability of each model for the ChatCAD framework. Notably, LLaMA-2 (13B) achieves a superior performance (macro-average F1-score = 0.598) compared to GPT-3 (175B), indicating that a smaller LLM may already have the capability to effectively assist the diagnostic process within our proposed ChatCAD framework.

Since GPT models are continuously updated, we here also demonstrate the evolving capabilities of LLMs within the ChatCAD framework. We include the latest versions available, namely GPT-3.5 Turbo and GPT-4, released in November 2023. The results of ChatCAD using different GPT models, denoted by different model generations and release dates, are presented in the bottom of Table 1. Although the F1-scores for the latest GPT-3.5-Turbo model suggest a slight decrease in performance on average compared to its larger predecessors, it is still comparable to best open source model (LLaMA-2 as shown in Table 1) and offers several practical advantages. Notably, it is smaller, costs less and responds faster. The GPT-3.5-Turbo's lower F1-scores relative to its larger GPT-3 and GPT-3.5 counterparts can be attributed to its design optimizations for increased speed and cost-effectiveness. These optimizations involve a reduced parameter count, which may curtail the model's capacity to intricately process the detailed information typical of medical data. Furthermore, the model's tuning may favor responsiveness over the specialized depth needed for medical report generation. Despite this, GPT-3.5 Turbo remains a viable option for applications where efficiency and affordability take precedence, and the trade-off in performance might be considered acceptable for certain real-world scenarios.

In the case of GPT-4, our experiments have indicated a noticeable enhancement in performance

compared to all previous models, including the GPT-3 family. This improvement may stem from several advancements.

- The improved performance of GPT-4 can be attributed to a refined training dataset, updated to include information until April 2023 (the old ones in the first two row in Table 3 have a knowledge cutoff at Sept 2021), allowing for more current and specialized medical content to be leveraged in generating clinical reports.
- Additionally, GPT-4 should have more advanced capabilities in following complex instructions, a feature that translates into more precise and format-specific medical image report generation. OpenAI's release blog says "GPT-4 Turbo performs better than our previous models on tasks that require the careful following of instructions, such as generating specific formats."
- Moreover, the adoption of a novel mixture of expert architecture contributes to this increased accuracy, as it allows the model to efficiently manage a range of tasks by drawing on specialized subsets of knowledge. This architectural innovation supports GPT-4's ability to deliver more contextually relevant and clinically accurate reports, reflecting the latest advancements in language model design.

Q3: Unclear sentence (probably a typo): "Notably, it is smaller, costs less and responds faster, and provides faster." Provides faster what?

A: We have corrected the unclear sentence "Notably, it is smaller, costs less and responds faster, and provides faster." to "Notably, it is smaller, costs less, and responds faster."

Q4: The title for section 2.4 "How ChatCAD acts like Radiologist" is a little clumsy and unclear. Suggest something along the lines of "Qualitative Evaluation of Generated Reports"

A: We have revised the title of Section 2.4 from "How ChatCAD acts like Radiologist" to "Qualitative Evaluation of Generated Reports" to better reflect the content of the section.

Q5: The superiority of the generated reports in terms of appropriateness should be backed up by a statistical test.

A: Thank you for your comments. We have incorporated a statistical test to support our claim of superiority in the appropriateness of the generated reports. The results of this test have been included in the revised manuscript. Specifically, we have conducted a paired t-test following the implementation of SciPy (Virtanen et al., 2020). The p-value is 0.022, providing evidence to support our claim regarding ChatCAD's superiority in terms of appropriateness. We added the p-value to the manuscript:

While AI-generated reports may lack a degree of the linguistic fluidity typically found in human reports (evidenced by a lower conciseness score), they have demonstrated a high degree of appropriateness (p-value = 0.022 with paired t-test).

Q6: The percentages reported at the end of Section 5 are not percentage point improvements, they are absolute values and so should just be "x.xx%"

A: We have corrected the writing at the end of Section 5: At the same time, nearly 40% of reports generated by text-babbage-001 and nearly 15% of reports from text-curie-001 have no meaningful

content.

Q7: A point that I had not noticed before is that the authors should report the results of the CAD models alone along with the results from the generated reports. This will demonstrate the extent to which ChatCAD may be introducing errors into reports and should act as an upper bound on the performance of ChatCAD.

A: First, we want to clarify a writing mistake: the report generation network baseline is R2Gen which was reported alongside ChatCAD in Table 1. This can be proved by our original ArXiv preprint in Feb. 2023 and our released code. We now make clarification in Sec. 2.2 and Sec. 4:

[Sec. 2.2] In this paper, we evaluate the performance of the combination of a report generation network (R2GenCMN (Chen et al., 2021)) and a classification network (PCAM (Ye et al., 2020)). The result is compared to the baseline R2GenCMN (Chen et al., 2021) and CvT2DistilGPT2 (Nicolson et al., 2022).

[Sec. 4] In this paper, we evaluate the performance of a report generation on the MIMIC-CXR dataset (Johnson et al., 2019). The MIMIC-CXR dataset is a large-scale public dataset of chest x-ray images with free-text radiology reports. At the same time, the classification network refer to the PCAM (Ye et al., 2020) which is trained on CheXpert dataset (Irvin et al., 2019). CheXpert is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients. Our report generation network is R2GenCMN (Chen et al., 2021) which is trained on the MIMIC.

Second, about the upper bound of ChatCAD, it is not determined by a single model since ChatCAD is designed to take multiple networks' inputs into consideration.

Q8: The new results are welcome, but I fear that they have disrupted the flow of the results section and therefore the results section should be restructured. For example, what is the difference between Section 2.3 "Performance of ChatCAD using different LLMs" and Section 2.5 "How LLMs affect Report Quality". Wouldn't it make sense to have these sections togther

A: Thanks for your suggestive comments, we have combined these sections together in the reply of Q2.

Q9: I acknowledge the author's response to my question 7 (about how the other models did not have access to ground truth labels). I agree that the contribution of ChatCAD is different from the other works included in the comparison, and is concerned with aggregration of information from other sources (i.e. imaging models) rather than extracting image information itself. However, this is precisely why it is important to point out to readers that the comparison to the other methods needs to be understood in this context: these other methods are not trying to solve exactly the same task. I still feel that this should be addressed in the paper.

A:

Thank you for your comment. We appreciate your concern, and we would like to provide further clarification regarding the comparison between our proposed ChatCAD and the other methods included in the evaluation.

In Fig. R2, we have visualized the pipeline of our proposed ChatCAD alongside one of the methods included in the comparison, R2GenCMN, as an example. From a task perspective, both ChatCAD and R2GenCMN aim to solve the same task. They take a chest X-ray image as input and generate a medical report as output. Regarding the evaluation setup, while ChatCAD integrates information from both the CheXpert and MIMIC-CXR datasets, both ChatCAD and R2GenCMN

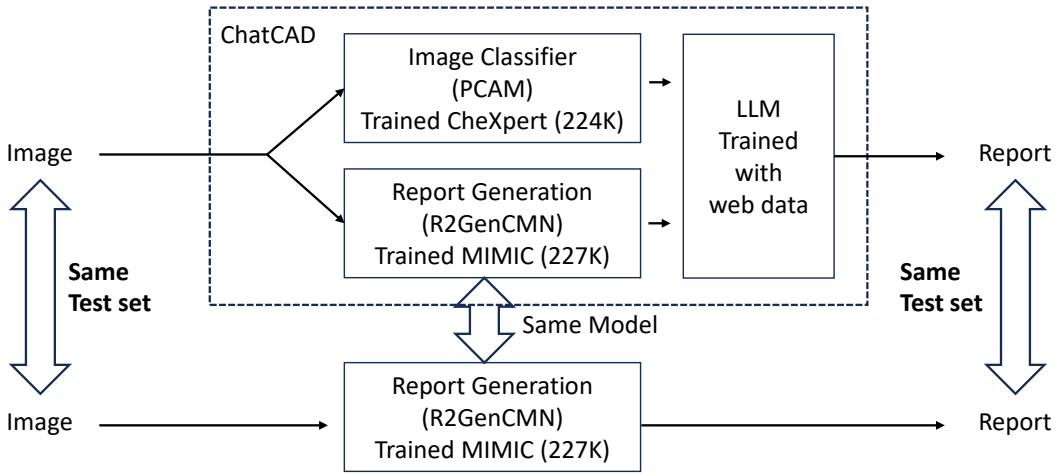

Fig. R2: Pipeline of our proposed ChatCAD and other methods included in the comparison.

Table 2: Dataset used in our ChatCAD and R2GenCMN

	R2GenCMN	ChatCAD
Training Data	MIMIC	CheXpert+MIMIC+LLM Training Set
Test Data	MIMIC Test Set	MIMIC Test Set

are tested on the same test set from the MIMIC-CXR dataset. Therefore, we can emphasize that ChatCAD and R2GenCMN are indeed attempting to solve exactly the same task.

However, it is important to note that a major advantage of our approach lies in the utilization of LLM. By incorporating LLM, we can combine decisions from multiple CAD models, enabling fine-tuning of each CAD model individually and incremental ensembling. This flexibility allows us to adapt and integrate new models rapidly in response to emerging situations such as disease outbreaks. For example, during the COVID-19 outbreak, we can easily add a pneumonia classification model that specifically differentiates between community-acquired pneumonia and COVID-19 infection. This process requires minimal data and provides great flexibility. A study by (Wang et al., 2021) utilized only 204 COVID-19 cases to achieve a 90% diagnosis accuracy. By seamlessly integrating such models and adjusting the weighting of each model within the LLM, we can improve the overall accuracy and reliability of CAD systems.

In summary, while ChatCAD and the other methods in the comparison may have differences in their information sources and aggregation processes, they are indeed attempting to solve the same task. Our proposed ChatCAD stands out by leveraging the power of the LLM to combine decisions from multiple CAD models, ultimately improving the accuracy and adaptability of the system.

Comment: Overall, I feel that this is an interesting paper that will be of interest to the community. The technical innovations are relatively modest as they use existing LLM systems with only novel prompting, but the proposed use of the LLM and the evaluation on this task is a useful addition to the literature. The limitation of using an imperfect language model (Chexbert) to determine the accuracy of ChatCAD (a limitation that the authors do now acknowledge in the revision) is an important one. As a result, my enthusiasm for this paper is moderate.

A: Thank you for your comment and for acknowledging the potential interest and impact of our paper for the community.

In response to your feedback, we have further explored the comparison between our prompting-based system, ChatCAD, and end-to-end trained multi-modal large language models. The results of this comparison are presented in Table 3. Our findings indicate that, at present, prompting systems like ChatCAD outperform end-to-end models across several metrics.

While end-to-end models may represent the future direction of this field, our results suggest that prompting-based approaches currently offer superior performance for the specific task addressed in this paper. We believe these findings contribute valuable insights to the ongoing discussion about the relative merits of different approaches in this area.

We acknowledge that the use of an imperfect language model (CheXbert) for evaluation is a limitation of our study, and we have made efforts to transparently discuss this in the revised manuscript. Despite this limitation, we believe our work offers a meaningful contribution by demonstrating the effectiveness of prompting-based methods for this task and providing a benchmark for future research.

Thank you again for your thoughtful review and for engaging with our work. We hope that the revisions and additional analyses have addressed your concerns and that you find the paper's contributions to be of moderate to strong interest to the community.

Table 3: Comparison between Generalist Medical AI and our proposed method.

Method	Continual	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
PMC-VQA (Zhang et al., 2023)	×	/	/	/	0.037	0.599	0.127
RadFM (Wu et al., 2023)	×	**0.633**	0.345	/	0.136	0.371	0.297
ChatCAD	✓	0.627	**0.534**	**0.440**	**0.636**	**0.787**	**0.605**

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Generating radiology reports via memory-driven transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., 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.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Nicolson, A., Dowling, J., and Koopman, B. (2022). Improving chest x-ray report generation by leveraging warm-starting. *arXiv preprint arXiv:2201.09405*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Wang, Z., Xiao, Y., Li, Y., Zhang, J., Lu, F., Hou, M., and Liu, X. (2021). Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays. *Pattern recognition*, 110:107613.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023). Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Ye, W., Yao, J., Xue, H., and Li, Y. (2020). Weakly supervised lesion localization with probabilistic-cam pooling.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., and Tang, J. (2023). GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Zhang, J., Gan, R., Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., and Chen, C. (2022). Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Reviewers' comments:

Reviewer #2 (Remarks to the Author):

The restructuring of the sections is appreciated and makes for a more logical flow through the paper in my opinion. Although the two sections called "Quality Improvement of the Generated Report" and "Qualitative Evaluation of Generated Reports" are very similar and a little confusing and it is not clear from the titles what the differences are. Perhaps my previous suggestion was not a good one. It seems to me that the fundamental differences between these two sections is that 2.1 deals with diagnostic accuracy and 2.4 deals with qualitative evaluation. Perhaps a better heading for section 2.1 would be simply "diagnostic accuracy of generated reports".

Unfortunately there are still two major comments from the previous round that have not been addressed:

The author's response to Q7 did not address my suggestion at all. My point here is that it is important that the authors report the diagnostic accuracy of the CAD vision classifiers that are used as part of ChatCAD (i.e. PCAM) in isolation (i.e. outside of the ChatCAD framework). This is so that we can evaluate whether or not the ChatCAD framework is affecting (adversely or otherwise) the performance of the underlying vision models. For example, it would be troubling to see that the ChatCAD output is significantly reducing the raw classifier output. Also, this would presumably be very straightforward for the authors to include as they already have the model predictions and ground truths.

Secondly, I appreciate the authors' careful explanation in response to my question 9 (which was in turn in response to q7 in the previous round). Though the authors appear to think that I have misunderstood something here (perhaps my previous wording could have been confusing), I am confident that I have (and have always had) a good grasp of the relevant facts from the paper and that the authors are again missing my point. I understand that the ultimate *prediction task* for ChatCAD and the baselines is exactly the same and the test set is the same, and in this sense the results are comparable. However it is very important to understand, and make it clear to readers, that the *training sets for the two methods are NOT the same*, as the authors make clear themselves in Table 2 of their response. The ChatCAD approach benefits from the ChexPert dataset, which provides both substantially more training cases AND also much more direct supervised ground truths than the MIMIC dataset alone. So it is important to emphasize that improvements in accuracy of ChatCAD over baselines could be due both to improved method and also access to more training data.

Response to the Reviewers' Comments

Reviewer 2

Comment: The restructuring of the sections is appreciated and makes for a more logical flow through the paper in my opinion. Although the two sections called "Quality Improvement of the Generated Report" and "Qualitative Evaluation of Generated Reports" are very similar and a little confusing and it is not clear from the titles what the differences are. Perhaps my previous suggestion was not a good one. It seems to me that the fundamental differences between these two sections is that 2.1 deals with diagnostic accuracy and 2.4 deals with qualitative evaluation. Perhaps a better heading for section 2.1 would be simply "diagnostic accuracy of generated reports"

A: Thanks for your thoughtful comments regarding the structure of our manuscript. We appreciate your positive feedback on the overall restructuring of the sections, which we aimed to make more logically coherent.

We acknowledge that the section titles you have highlighted do not sufficiently convey the distinct focuses of each section. In response to your suggestion, we have revised the title of section 2.1 to "Diagnostic Accuracy of Generated Reports" to more accurately reflect the emphasis on the diagnostic accuracy aspects discussed in this section. We believe that this change clarifies the distinction between section 2.1 and section 2.4. We hope that this revision addresses your concern and improves the clarity and effectiveness of our manuscript. Thank you again for your guidance, which has been invaluable in enhancing the presentation and clarity of our work.

Q1. The author's response to Q7 did not address my suggestion at all. My point here is that it is important that the authors report the diagnostic accuracy of the CAD vision classifiers that are used as part of ChatCAD (i.e. PCAM) in isolation (i.e. outside of the ChatCAD framework). This is so that we can evaluate whether or not the ChatCAD framework is affecting (adversely or otherwise) the performance of the underlying vision models. For example, it would be troubling to see that the ChatCAD output is significantly reducing the raw classifier output. Also, this would presumably be very straightforward for the authors to include as they already have the model predictions and ground truths

A: Thank you for your comments. We apologize for not fully understanding your point earlier. Your suggestion regarding the CAD vision classifier's diagnostic accuracy allows us to evaluate whether the ChatCAD framework impacts (whether adversely or otherwise) the performance of the underlying vision models. We have added the PCAM results in the revised experimental section. From these results, it is clear that our proposed ChatCAD outperforms both R2GenCMN and PCAM. This can be attributed to ChatGPT's powerful reasoning ability to synthesize information from both sides and produce a more comprehensive and accurate report.

In this paper, we evaluate the performance of the combination of a report generation network (R2GenCMN (Chen et al., 2021)) and a classification network (PCAM (Ye et al., 2020)). The

Table R1: Comparison of diagnostic accuracy with SOTA methods.

Observation	CvT2DistilGPT2			R2GenCMN			PCAM			Ours (GPT-3)			Ours (ChatGPT)		
	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1	PR	RC	F1
Cardiomegaly	0.512	0.591	0.549	0.590	0.534	0.561	**0.846**	0.190	0.310	0.606	0.569	0.587	0.663	**0.595**	**0.627**
Edema	0.224	0.468	0.303	0.563	0.252	0.348	0.602	0.579	0.591	**0.563**	**0.626**	0.593	0.556	0.514	0.534
Consolidation	0.063	0.239	0.099	**0.667**	0.121	0.205	0.325	0.788	0.460	0.310	**0.803**	**0.447**	0.322	0.697	0.440
Atelectasis	0.306	0.388	0.342	0.442	0.504	0.471	0.468	0.991	0.636	0.408	**0.991**	0.578	**0.470**	0.981	**0.636**
Pleural Effusion	0.454	0.692	0.548	**0.819**	0.500	0.618	0.728	0.916	0.811	0.634	**0.916**	0.749	0.736	0.845	**0.787**
Average	0.312	0.476	0.368	**0.616**	0.382	0.441	0.594	0.693	0.562	0.504	**0.781**	0.591	0.549	0.726	**0.605**

result is compared to baseline R2GenCMN (Chen et al., 2021), CvT2DistilGPT2 (Nicolson et al., 2022), and PCAM (Ye et al., 2020). On the basis of clinical importance and prevalence, we focus on five kinds of observation. Three metrics, including precision (PR), recall (RC), and F1-score (F1), is reported in Table R1.

The strength of our method are clearly shown in Table R1. It has obvious advantage in RC and F1, and is only weaker than R2GenCMN in term of PR. Our method has a relatively high Recall and F1-score on MIMIC-CXR dataset. For all five kinds of diseases, both CvT2DistilGPT2 and R2GenCMN show inferior performance to our method concerning RC and F1. Specifically, their performances on Edema and Consolidation are rather low. Their RC values on Edema are 0.468 and 0.252, respectively, while our method achieves the RC value of 0.626 based on GPT-3. The same phenomenon can be observed in Consolidation, where the first two methods hold the values of 0.239 and 0.121 while ours (GPT-3) drastically outperforms them, with the RC value of 0.803. The R2GenCMN has a higher PR value compared to our method on three of five diseases. However, the cost of R2GenCMN's high performance on Precision is its weakness in the other two metrics, which can lead to biased report generation, e.g., rarely reporting any potential diseases. At the same time, our method has the highest F1 among all methods, and we believe it can be the most trustworthy report generator. It is also worth noting that our proposed ChatCAD framework significantly outperforms both R2GenCMN and PCAM. This superior performance can be attributed to ChatGPT's advanced reasoning capabilities, which effectively synthesize information from multiple sources to produce a more comprehensive and accurate report. We believe this phenomenon further underscores the superiority of ChatCAD and demonstrates its considerable potential for clinical applications.

Q2. I appreciate the authors' careful explanation in response to my question 9 (which was in turn in response to q7 in the previous round). Though the authors appear to think that I have misunderstood something here (perhaps my previous wording could have been confusing), I am confident that I have (and have always had) a good grasp of the relevant facts from the paper and that the authors are again missing my point. I understand that the ultimate *prediction task* for ChatCAD and the baselines is exactly the same and the test set is the same, and in this sense the results are comparable. However it is very important to understand, and make it clear to readers, that the *training sets for the two methods are NOT the same*, as the authors make clear themselves in Table 2 of their response. The ChatCAD approach benefits from the ChexPert dataset, which provides both substantially more training cases AND also much more direct supervised ground truths than the MIMIC dataset alone. So it is important to emphasize that improvements in accuracy of ChatCAD over baselines could be due both to improved method and also access to more training data.

A: Thank you for your continued engagement with our manuscript and for providing detailed insights that enhance the clarity and integrity of our research. We apologize for our misunderstanding and appreciate the opportunity to address the nuances of our study more thoroughly. In the revised manuscript, we will ensure to emphasize this point more explicitly. We have clarified this point in the revised Section 2.1. that while ChatCAD demonstrates superior accuracy, which

Table R2: Dataset used in ChatCAD and R2GenCMN.

	R2GenCMN	ChatCAD
Training Data	MIMIC	CheXpert+MIMIC+LLM Training Set
Test Data	MIMIC Test Set	MIMIC Test Set

could potentially bias the comparative analysis.

While ChatCAD has demonstrated a clear advantage over baseline methods, it is also beneficial to explore this from the perspective of continual learning to provide deeper insights for our readers. Unlike R2GenCMN and PCAM, which are trained solely on the MIMIC-CXR and CheXpert datasets respectively, ChatCAD benefits from sequential learning on MIMIC-CXR, CheXpert, and additional datasets used to train the large language model (as shown in Table R2). This large language model acts as a general interface, integrating knowledge from these diverse datasets while avoiding catastrophic forgetting. In summary, the improvements in accuracy of ChatCAD over the baselines could be attributed to both the enhanced methodology and the broader access to training data.

References

Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Generating radiology reports via memory-driven transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Nicolson, A., Dowling, J., and Koopman, B. (2022). Improving chest x-ray report generation by leveraging warm-starting. *arXiv preprint arXiv:2201.09405*.

Ye, W., Yao, J., Xue, H., and Li, Y. (2020). Weakly supervised lesion localization with probabilistic-cam pooling.

REVIEWERS' COMMENTS:

Reviewer #2 (Remarks to the Author):

I think we are now in agreement on the issues raised. I am much more satisfied now that the results for the individual models have been added in, and are shown to be inferior to the performance of the presented method. Table 3 and the new comment added about the fact that the improvement is likely also due to increased training data is also welcome, though I think that perhaps it would more logically be added to sec 2.1?

Response to the Reviewers' Comments

Reviewer 2

Comment: I think we are now in agreement on the issues raised. I am much more satisfied now that the results for the individual models have been added in, and are shown to be inferior to the performance of the presented method. Table 3 and the new comment added about the fact that the improvement is likely also due to increased training data is also welcome, though I think that perhaps it would more logically be added to sec 2.1?

A: Thank you for your thoughtful feedback and for acknowledging the improvements made in the revised manuscript. We are pleased to hear that you are now satisfied with the inclusion of the results for the individual models and their comparison to our proposed method. We agree that logically placing this comment in Section 2.1 would enhance the coherence of the manuscript. We have moved these results and discussions to Sec 2.1. Thank you once again for your valuable input.