communications engineering

Article

https://doi.org/10.1038/s44172-024-00271-8

Interactive computer-aided diagnosis on medical image using large language models

Sheng Wang[1,2,3,6], Zihao Zhao[1,6], Xi Ouyang[3], Tianming Liu[4], Qian Wang[1,5] ✉ & Dinggang Shen[1,3,5] ✉

Computer-aided diagnosis (CAD) has advanced medical image analysis, while large language models (LLMs) have shown potential in clinical applications. However, LLMs struggle to interpret medical images, which are critical for decision-making. Here we show a strategy integrating LLMs with CAD networks. The framework uses LLMs' medical knowledge and reasoning to enhance CAD network outputs, such as diagnosis, lesion segmentation, and report generation, by summarizing information in natural language. The generated reports are of higher quality and can improve the performance of vision-based CAD models. In chest X-rays, an LLM using ChatGPT improved diagnosis performance by 16.42 percentage points compared to state-of-the-art models, while GPT-3 provided a 15.00 percentage point F1-score improvement. Our strategy allows accurate report generation and creates a patient-friendly interactive system, unlike conventional CAD systems only understood by professionals. This approach has the potential to revolutionize clinical decision-making and patient communication.

Large Language Models (LLMs), such as OpenAI's GPT series[1], are advanced artificial intelligence systems that have demonstrated remarkable results in natural language processing[2]. Trained on vast amounts of text data, LLMs have the potential to revolutionize various industries, including marketing, education, and customer service. Notably, in the medical domain, LLMs like ChatGPT[3] have showcased their potential as valuable tools for providing medical knowledge and advice. For example, ChatGPT has successfully passed part of the US medical licensing exams, illustrating its capacity to augment medical professionals in delivering care[4]. Some recent studies[5,6] have primarily investigated the potential application of LLMs in medical education. However, despite their impressive progress in natural language processing, LLMs' ability to understand visual information in computer vision tasks remains a challenge. Addressing this limitation is crucial, especially in the medical field, where medical images play a significant role in supporting clinical decisions.

Focusing on the visual aspect, medical image computer-aided diagnosis (CAD) networks have achieved significant success in supporting clinical decision-making processes in the medical field[7–12]. These networks leverage advanced deep learning algorithms to analyze medical images, such as X-rays, CT scans, and MRIs, and then provide valuable insights to support clinical decision-making. Unlike LLMs, CAD networks have been designed specifically to handle the complexities of visual information in medical images, making them well-suited for tasks such as disease diagnosis[13], lesion segmentation[14], and report generation. These networks have been trained on large amounts of medical image data, allowing them to learn to recognize complex patterns and relationships in visual information that are specific to the medical field.

In recent advancements, Vision-Language Models (VLMs) have become a significant trend, capitalizing on the ever-increasing capabilities of LLMs. Notably, CLIP[15] has pioneered the integration of visual and language information into a unified feature space and achieved promising performance in various downstream tasks. This paradigm has been widely applied to Chest X-rays[16] and Pathology images[17]. Frozen[18] 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[19] and Med-flamingo[20] introduce cross-attention layers into the LLM architecture, enabling the direct incorporation of visual features and pre-training these layers on more than 100 M image-text pair. BLIP-2[21] 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

[1]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China. [2]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. [3]Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China. [4]School of Computing, University of Georgia, Athens, GA, USA. [5]Shanghai Clinical Research and Trial Center, Shanghai, China. [6]These authors contributed equally: Sheng Wang, Zihao Zhao ✉e-mail: qianwang@shanghaitech.edu.cn; dinggang.shen@gmail.com

way, the pre-trained vision and text model are well aligned, enabling impressive performance on several downstream tasks. LLaVA[22], 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[23] learns a joint embedding across six different modalities - images, text, audio, depth, thermal, and inertial data. The alignment and fusion of these modalities enable tasks including cross-modal retrieval, composing modalities with arithmetic, and cross-modal detection and generation.

The aim of this paper is to provide a scheme that bridges current LLMs and CAD models. In this scheme, namely ChatCAD, the image is first fed into multiple networks, i.e., an image classification network, a lesion segmentation network, and a report generation network as depicted in Fig. 1a. The results produced by classification or segmentation are a vector or a mask, which cannot be understood by LLMs. Therefore, we transform these results into the text representation form as shown in the middle panel of Fig. 1. These text-form results will then be concatenated together as a prompt "*Refine the report based on results from Network A and Network B*" for the LLM. The LLM then summarizes the results from all the CAD networks. As the example in this figure, the refined report combines the findings from all three networks to provide a clear and concise summary of the patient's condition, highlighting the presence of pneumonia and the extent of the infection in the left lower lobe. In this way, the LLM could correct errors in the generated report based on the results from CAD networks. As shown in Fig. 2, experiment shows that our scheme could improve the diagnosis performance score of the state-of-the-art report generation methods by 16.42% points. A major benefit of our approach is the utilization of LLM's robust logical reasoning capabilities to combine various decisions from multiple models provided by multiple vendors. This allows us to update CAD model individually. For instance, in response to an emergency outbreak such as COVID-19, we can add a pneumonia classification model (differentiating between community-acquired pneumonia and COVID-19[24]) using very few cases without affecting the other models.

Another advantage of LLMs to CAD models is that their extensive and robust medical knowledge can be leveraged to provide interactive explanations and medical advice as we illustrate in Fig. 1b. For example, based on an image and generated report, patients can inquire about appropriate treatment options (left panel) or define medical terms such as "airspace consolidation" (middle panel). Or with the patient's chief complaint (right panel), LLMs can explain why such a symptom happens. In this manner, patients can gain a deeper understanding of their symptoms, diagnosis, and treatment more efficiently. It can efficiently help patients to reduce consultation costs with clinical experts. As the performances of CAD models and LLMs become increasingly improved and these models can be jointly trained in the future, the proposed scheme has the potential to improve the quality of radiology reports and enhance the feasibility of online healthcare services.

Results

Diagnostic accuracy of generated reports

In this paper, we evaluate the performance of the combination of a report generation network (R2GenCMN[25]) and a classification network (PCAM[26]). The result is compared to the baseline R2GenCMN[25], CvT2DistilGPT2[27], and PCAM[26]. 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), are reported in Table 1.

The strength of our method is clearly shown in Table 1. It has obvious advantages in RC and F1, and is only weaker than R2GenCMN in terms 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

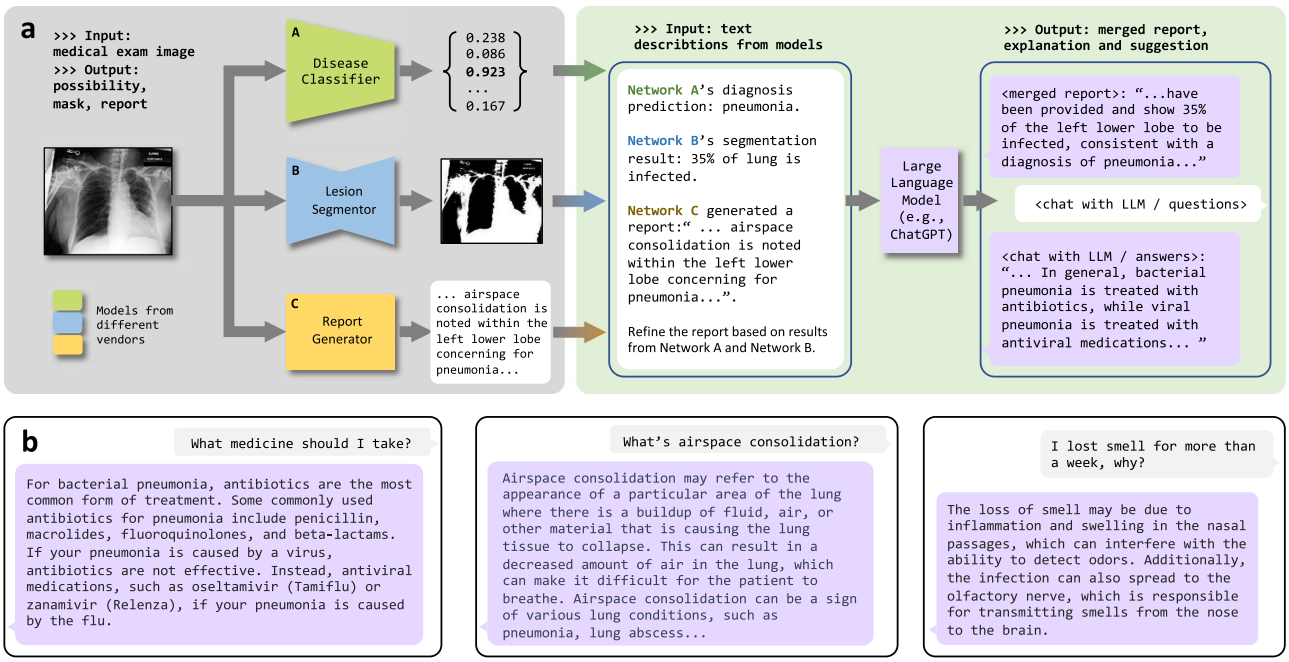

Fig. 1 | ChatCAD: an AI-assisted medical diagnosis and advice system. a Overview of our proposed strategy. The image is processed by various networks to generate diverse outputs, which are then transformed into text descriptions. The descriptions, served as a link between visual and linguistic information, are combined as inputs to a large language model (LLM). With its ability to reason and its knowledge of the medical field, the LLM can provide a condensed report. **b** Interactive explanations and medical advice from ChatCAD.

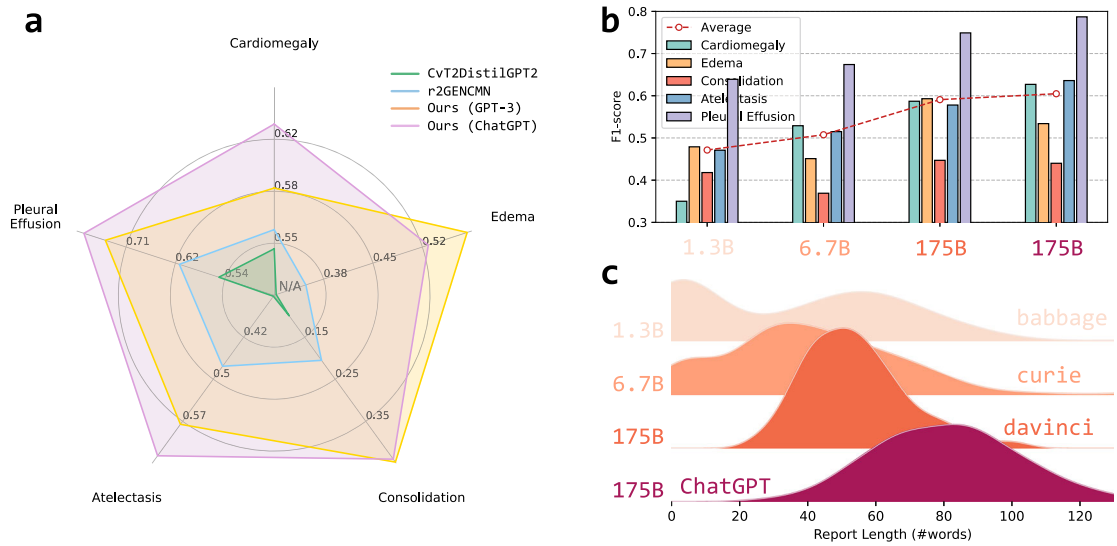

Fig. 2 | Performance evaluation of large language models in medical diagnosis. a F1-score comparison on 5 observations. **b** The diagnosis accuracy of different LLMs. **c** The histogram of report length. Different color denotes different LLMs.

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. The other strength of our method lies in its scaling performance.

It is 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. It would also be beneficial to explain the results 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 S1 in Supplementary). 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.

Qualitative analysis on prompt designs

The process of ChatCAD, as shown in Fig. 1a, is a straightforward procedure consisting of the following steps: Firstly, examination images, such as X-rays, are inputted into pre-trained CAD models to obtain results. These results are then transformed, often in tensor format, into natural language. Next, language models are employed to summarize the findings and establish a conclusion. Additionally, the results obtained from the CAD models are utilized to facilitate a conversation regarding symptoms, diagnosis, and treatment. In order to investigate the impact of prompt design on report generation, we have developed prompts, which are depicted in Fig. 3.

Reports generated from Prompt #2 and Prompt #3 are generally acceptable and reasonable in most cases as one can observe in Fig. S1 and Fig. S2 in Supplementary. "Network A" is frequently referenced in the generated reports. Some prompt tricks, e.g., "*Revise the report based on results from Network A but without mentioning Network A*", can be applied to remove its mention. We do not utilize these tricks in current experiments.

Performance of ChatCAD using different LLMs

Different from ChatGPT, which can only be accessed via online request, language models such as LLaMA can be used and fine-tuned in local computers without data privacy issues. To evaluate generalizability of ChatCAD and also to validate its potential value in clinical practice, we have experimented with a range of LLMs, including LLaMA-1, LLaMA-2, and several others. The results of our experiments are presented in Table 2, which compares F1-scores of different LLMs, including general-purpose models, specialized medical models, and OpenAI's GPT variants. As indicated in the table, there are 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 according to the macro-average of F1-score, which means that a smaller LLM such as LLaMA-2 (13B) is capable enough to assist the process of diagnosis following our proposed ChatCAD.

Since GPT models are continuously updated, we here also demonstrate the evolving capabilities of LLMs within the ChatCAD framework. We include the latest available versions, 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 2.

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 the best the open source model (LLaMA-2 as shown in Table 2) 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 optimization 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 such as 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, including information until April 2023 (the old ones in the 8th and 9th rows of Table 2 have some 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 capability in following complex instructions, a feature that translates into more precise and format-specific medical image report generation. OpenAI's

Table 1 | Comparison of diagnostic accuracy with state-of-the-art 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**

PR stands for precision. RC stands for recall and F1 stands for F1-score. Best performance are indicated in bold.

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.

Qualitative evaluation of generated reports

In a clinical setting, there are more aspects than the above-mentioned classification metrics that need to be evaluated. As a result, we have carefully developed an experimental pipeline to evaluate clinical reports generated by our proposed ChatCAD from two perspective: conciseness and appropriateness. Conciseness is vital to ensure the report being succinct and focused, avoiding extraneous details that may detract from the primary clinical 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 the experimental pipeline demonstrated in Supplementary Information into our study design (Fig. S3), 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 another 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 (as demonstrated in Fig. S4 in Supplementary), 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. The experts will offer both quantitative rating and qualitative feedback for each report.

The experimental results of an experienced radiologist are selected and displayed in Fig. 4. 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 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. 4a, b, we can observe that the vast majority of AI-generated reports (39) received a rating of 4, a quantity even higher than the number of real reports (32). This highlights the advantage of ChatCAD proposed in this paper in terms of report generation. Considering conciseness, ChatCAD-generated reports scored 3.40 ± 0.67, while human-written reports obtained 3.48 ± 0.58. ChatCAD demonstrates impressive performance on appropriateness (3.84 ± 0.65), showing superior performance to human-written reports (3.58 ± 0.64).

We also demonstrate results of the identification task in Fig. 4c, f. 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 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 more robust capacity in differentiating between the two sources. This can be further evidenced by the visualization in Fig. 4c, revealing the potential of AI-generated reports in practical clinical scenarios.

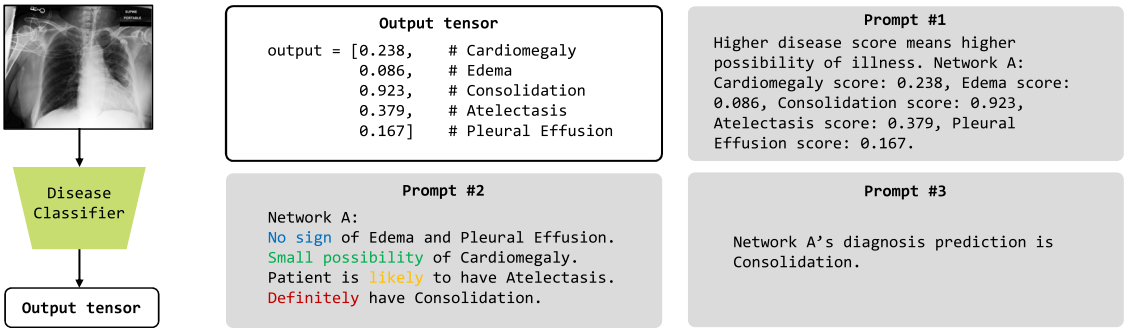

Fig. 3 | Prompt designing. Three different prompt designs were proposed to bridge between tensor and text.

Table 2 | Comparison of F1-scores by different large language models (LLMs)

Open-source	Model	#Para	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
✓	LLaMA [42]	7B	0.467	0.280	0.395	0.480	0.674	0.459
✓	LLaMA [42]	13B	0.439	0.390	0.385	0.550	0.681	0.489
✓	LLaMA-2 [43]	7B	0.594	0.453	0.457	0.622	0.780	0.581
✓	LLaMA-2 [43]	13B	0.570	0.556	0.443	0.628	**0.795**	0.598
✓	Ziya[44]	13B	0.538	0.466	0.393	0.611	0.764	0.554
✓	ChatGLM [45]	6B	0.556	0.340	0.373	0.559	0.746	0.515
✓	Mistral [46]	7B	0.610	0.388	0.353	0.604	0.791	0.549
✗	GPT-3[1]	175B	0.587	**0.593**	0.447	0.578	0.749	0.591
✗	GPT-3.5[1]	175B	0.627	0.534	0.440	**0.636**	0.787	0.605
✗	GPT-3.5-Turbo[1]	20B	0.615	0.497	0.388	0.609	0.738	0.570
✗	GPT-4[1]	220Bx8	**0.639**	0.551	**0.465**	0.621	0.790	**0.613**

Best results are indicated in bold.

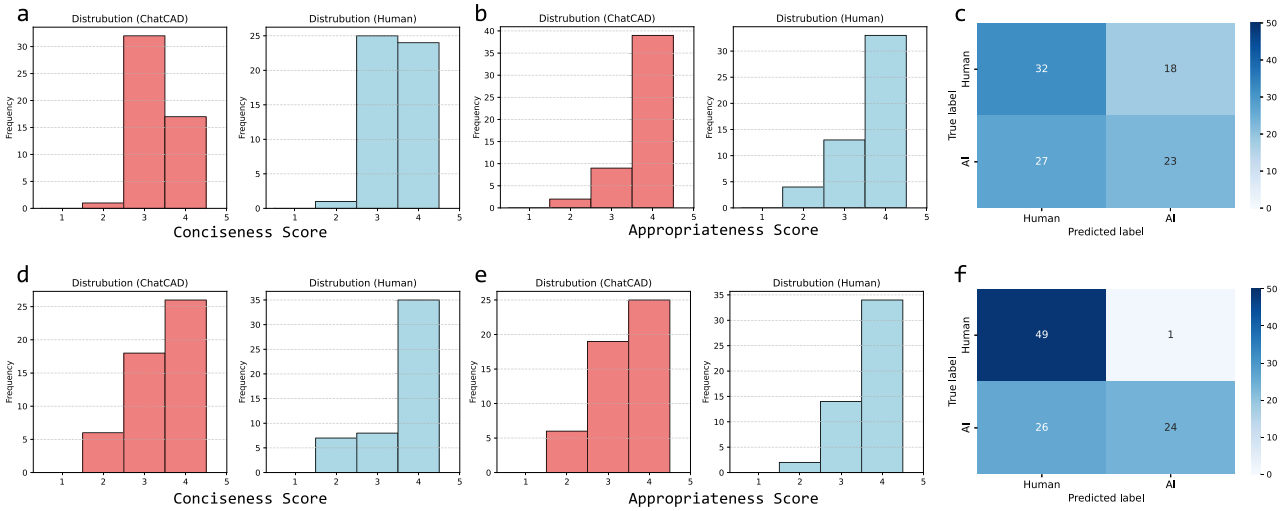

Fig. 4 | Qualitative experimental results from an experienced expert. a–c Are from an experienced clinical expert with limited experience using LLMs. While (**d**), (**e**), and (**f**) are from a trainee with extensive experience using LLMs. **a** Conciseness score comparison. **b** Appropriateness score comparison. **c** Confusion matrix from a subject with limited experience in AI, showing its performance in determining whether the report was generated by ChatCAD. **d** Conciseness score comparison. **e** Appropriateness score comparison. **f** Confusion matrix from a subject with extensive experience in AI, showing its performance in determining whether the report was generated by ChatCAD.

In summary, our experimental evaluation, as shown in Fig 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 high degree of appropriateness ($p = 0.022$ with paired t-test). 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.

How model size affect report quality

In this section, we compare the performance of different LLMs for report generation. OpenAI provides four different 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). The smallest text-ada-001 can not generate meaningful reports and is therefore not included in this experiment. We report the F1-score of all observations in Fig. 2b. 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 are 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. The details can be observed in Table 3. 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. 2c. 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.

Interactive and understandable CAD

A major advantage of our approach is the utilization of LLM to combine various decisions from multiple CAD models. This allows us to fine-tune each CAD model individually and ensemble them incrementally. For instance (c.f. Fig. 5a), in response to an emergency outbreak such as COVID-19, we can add a pneumonia classification model that differentiates between community-acquired pneumonia and COVID-19 infection. This process requires very few data and thus is very flexible. For example, [28] used 204 COVID-19 cases and reached 90% points diagnosis accuracy. The final report will then highlight the effectiveness of our approach in improving the overall accuracy and reliability of CAD systems, as well as its potential for rapid adaptation to emerging situations such as disease outbreaks. By leveraging LLM, we can seamlessly integrate new models and adjust the weighting of each model to achieve optimal performance.

The proposed ChatCAD also offers several benefits, including its ability to utilize LLM's extensive and reliable medical knowledge to provide interactive explanations and advice. As shown in Fig. 5e, f, two examples of the interactive CAD are provided, with one chat discussing pleural effusion and the other addressing edema and its relationship to swelling.

Through this approach, patients can gain a clearer understanding of their symptoms, diagnosis, and treatment options, leading to more efficient and cost-effective consultations with medical experts. As language models continue to advance and become more accurate with access to more trustworthy medical training data, ChatCAD has the potential to significantly enhance the quality of online healthcare services.

Discussion

In this paper, we explore a framework, ChatCAD, introducing LLMs in CAD. The proposed method, however, still has limitations to be solved.

First, LLM-generated reports are not human-like in a certain way. LLM is likely to output sentences like "Network A's diagnosis prediction is consistent with the findings in the radiological report" or "The findings from Network A's diagnosis prediction are supported by the X-ray". This is reflected in natural language similarity metrics when we compare them to our baseline method. ChatCAD improved the diagnosis accuracy but dropped the BLEU score[29]. We didn't provide the network with the patient's major complaint due to unavailability of such data, which may differ from practical scenarios. We believe the LLMs can process more complex information than what we currently provide. Better datasets and benchmarks are needed.

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.

Our experiments demonstrate the significant impact of language model size on diagnostic accuracy. Larger, more advanced LLM with fewer hallucination phenomena[30] may improve the accuracy and report quality further. However, the role of vision classifiers has not yet been explored, and additional research is necessary to determine if models such as ViT[31] or SwinTransformer[32], which boast larger parameters, can deliver improved results. On the other hand, LLMs can also be used to help the training of vision models, such as correcting outputs of vision models using related medical knowledge learned in LLMs.

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.

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[33–37]. References[34,35] argue that LLMs possess few-

Table 3 | F1-score comparison of different-size LLMs

Model	Size	Cardiomegaly	Edema	Consolidation	Atelectasis	Pleural Effusion	Average
text-babbage-001	~1.3B	0.350	0.479	0.418	0.471	0.639	0.471
text-curie-001	~6.7B	0.529	0.451	0.369	0.515	0.674	0.508
text-davinci-003	~175B	0.587	**0.593**	**0.447**	0.578	0.749	0.591
ChatGPT	~175B	**0.627**	0.534	0.440	**0.636**	**0.787**	**0.605**

Best performance are indicated in bold.

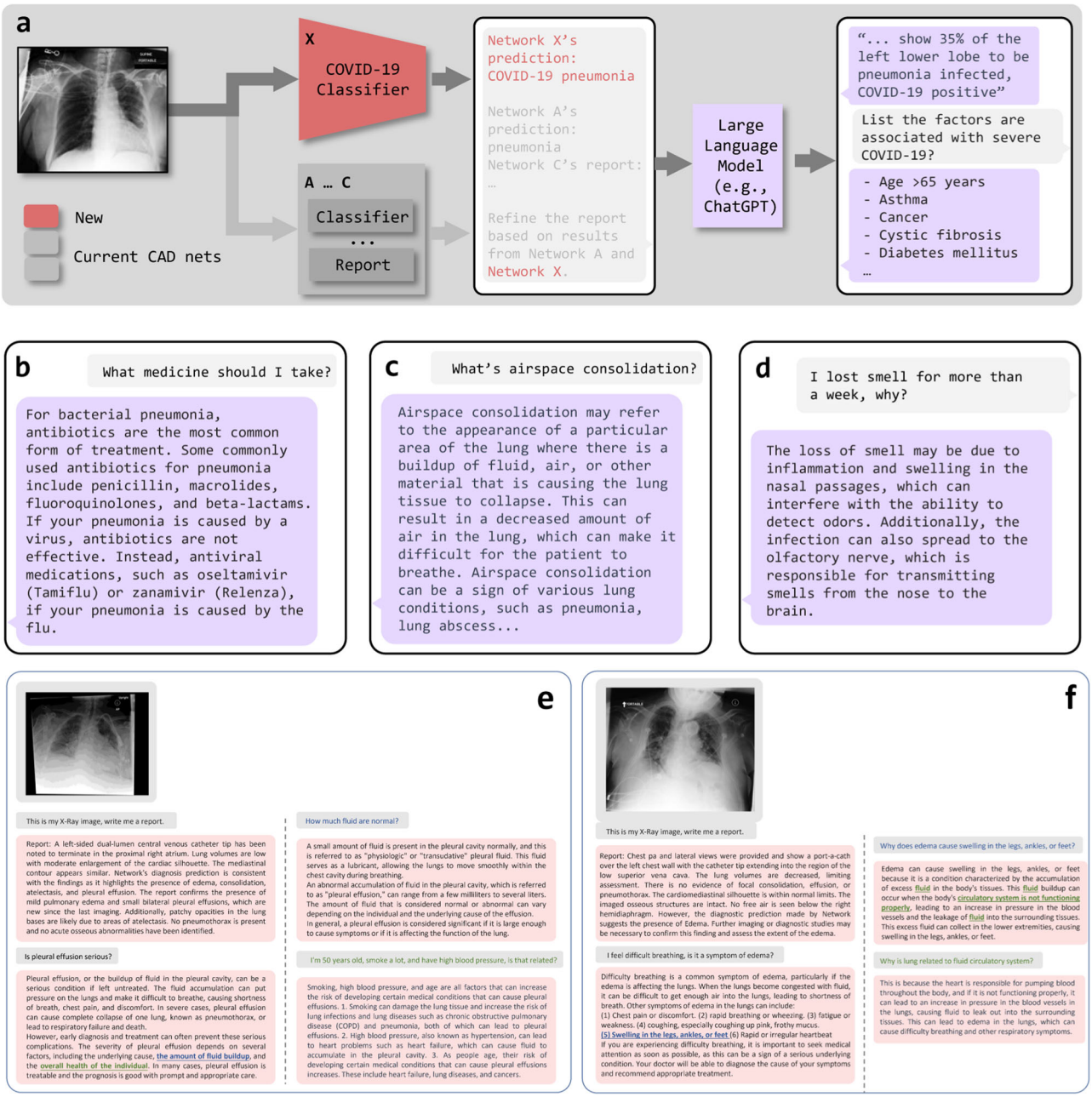

Fig. 5 | Extensibility, knowledge integration, and interactivity of ChatCAD system. a ChatCAD can seamlessly integrate new CAD models. **b–d** Leveraging LLM's comprehensive medical knowledge base to offer dynamic explanations and tailored advice. **e**, **f** Two examples showcasing the *interactive* CAD capabilities of our framework in conjunction with ChatGPT. In (**e**), the blue text represents a follow-up task concerning the fluid observed in the x-ray image, while the green text pertains to the relationship between an individual's health condition and pleural effusion. In (**f**), the blue text focuses on the topic of swelling and its underlying causes, while the green text presents a follow-up question regarding how lungs impact the circulatory system of fluids.

shot logical reasoning capabilities. In contrast[37], discovers 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. Reference[36] extensively test ChatGPT and GPT-4 on a variety of reasoning benchmarks, and find 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.

Moreover, the specifics of this paper have not been discussed with any clinical professionals, and therefore it still lacks rigor in many places. We will need to collaborate with clinical experts and conduct further research to ensure accuracy and reliability.

Methods

Dataset and implementation

In this paper, we evaluate the performance of a report generation on the MIMIC-CXR dataset[38], which is a large-scale public dataset including chest x-ray images and free-text radiology reports. At the same time, the classification network here refers to the PCAM[26], which is trained on CheXpert dataset[39]. Note that 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[25] which is trained on the MIMIC-CXR.

The reports from the LLMs are tested on the official test set of the MIMIC-CXR dataset. In particular, 300 cases are randomly selected,

including 50 cases of Cardiomegaly, 50 cases of Edema, 50 cases of Consolidation, 50 cases of Atelectasis, 50 cases of Pleural effusion, and 50 cases with no findings. The evaluation is performed using the open-source library CheXbert[40]. 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 comparison between these extracted labels.

The LLMs are updating constantly to include more new knowledge and events, leading to the improvement of their reasoning capability. The GPT-3 model used in this paper is *text-davinci-003* which was released by OpenAI on Feb. 2023 based on InstructGPT[41]. The maximum length of the output sentences is set to 1024 and the temperature is set to 0.5. The ChatGPT[3] model used is the *Jan-30-2023* version.

Bridge the gap between image and text

As shown in Fig. 1a, ChatCAD's process is simple and consists of the following steps: (1) Input examination images (e.g., X-Ray) into pre-trained CAD models to obtain results; (2) Convert these results (often in tensor form) into natural language; (3) Use language models to summarize the findings and draw a conclusion; (4) Utilize the results from the CAD models to engage in a conversation regarding symptoms, diagnosis, and treatment. This section focuses on the second step, i.e., how to effectively design the prompt that translates the output results (usually in tensor form) into natural language.

A natural way of prompting is to show all five kinds of pathology and their corresponding scores. We first tell the LLM "Higher disease score means higher possibility of illness" as the basic rule in order to avoid some misconceptions. Then, we represent this network (assumed as the first network) prediction of each disease as "Network A: ${**disease**} score: ${**score**}". Finally, we end the prompt with "Refine the report based on the results from Network A" if a report generation network is available as shown in Fig. 1a. If there is no report generation network, this part of the prompt will be "Write a Chest X-Ray radiology report based on the results from Network A".

We then notice that the LLMs are heavily influenced by this type of prompt, usually repeating all the numbers in the refined report. Reports generated from this prompt are very different from radiologists' reports since concrete diagnostic scores are not frequently used in clinical settings. To align with the language commonly used in clinical reports, we propose to transform the concrete scores into descriptions of disease severity, which will divide the scores into four categories: "No sign" (0.0–0.2), "Small possibility" (0.2–0.5), "Likely" (0.5–0.9), and "Definitely" (0.9 and above). These categories will be used to describe the likelihood of each of the five observations. We finally tested a more concise one that reports diseases with diagnosis scores higher than 0.5 in the prompt. If no prediction is made among all five diseases, the prompt will be "No Finding". We found both the "severity descriptions" and concise one have similar performance, so we used the concise one for the short prompt thus faster inference and lower cost. An example is illustrated in Fig. 3.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this work can be obtained from the code repository: https://github.com/zhaozh10/ChatCAD and MIMIC-CXR dataset: https://physionet.org/content/mimic-cxr/2.0.0/.

Code availability

All code used in this work can be obtained from the following publicly accessible GitHub page: https://github.com/zhaozh10/ChatCAD.

Received: 24 October 2023; Accepted: 20 August 2024;
Published online: 17 September 2024

References

1. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. (2019).
3. OpenAI. Chatgpt: Optimizing language models for dialogue https://openai.com/blog/chatgpt/ (2023).
4. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
5. Waisberg, E., Ong, J., Masalkhi, M. & Lee, A. G. Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education. *Eye* **38**, 639–641 (2024).
6. Abd-Alrazaq, A. et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med. Educ.* **9**, e48291 (2023).
7. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
8. Cheng, J.-Z. et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci. Rep.* **6**, 24454 (2016).
9. Wang, M. et al. Identifying autism spectrum disorder with multi-site FMRI via low-rank domain adaptation. *IEEE Trans. Med. Imaging* **39**, 644–655 (2019).
10. Fan, Y. et al. Multivariate examination of brain abnormality using both structural and functional MRI. *NeuroImage* **36**, 1189–1199 (2007).
11. Jie, B., Liu, M. & Shen, D. Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease. *Med. Image Anal.* **47**, 81–94 (2018).
12. Liu, M., Zhang, D., Shen, D. & Initiative, A. D. N. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* **35**, 1305–1319 (2014).
13. Wang, S., Ouyang, X., Liu, T., Wang, Q. & Shen, D. Follow my eye: using gaze to supervise computer-aided diagnosis. *IEEE Trans. Med. Imaging* **41**, 1688–1698 (2022).
14. Zhao, X. et al. RCPS: rectified contrastive pseudo supervision for semi-supervised medical image segmentation. *IEEE J. Biomed. Health* **28**, 251–261 (2024).
15. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763 (PMLR, 2021).
16. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
17. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
18. Tsimpoukelli, M. et al. Multimodal few-shot learning with frozen language models. *Adv. Neural Inf. Process. Syst.* **34**, 200–212 (2021).
19. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
20. Moor, M. et al. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367 (PMLR, 2023).
21. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML* (2023).
22. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36** (2024).

23. Girdhar, R. et al. Imagebind: one embedding space to bind them all. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15180–15190 (IEEE, 2023).
24. Ouyang, X. et al. Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia. *IEEE Trans. Med. Imaging* **39**, 2595–2605 (2020).
25. Chen, Z., Shen, Y., Song, Y. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proc. Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (ACL Anthology, 2021).
26. Ye, W., Yao, J., Xue, H. & Li, Y. Weakly supervised lesion localization with probabilistic-cam pooling 2005.14480 (2020).
27. Nicolson, A., Dowling, J., & Koopman, B. Improving chest X-ray report generation by leveraging warm starting. *Artif. Intell. Med.* **144**, 102633 (2023).
28. Wang, Z. et al. Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest x-rays. *Pattern Recognit.* **110**, 107613 (2021).
29. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (ACL Anthology, 2002).
30. Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proc. Conference on Empirical Methods in Natural Language Processing* 6449–6464 (EMNLP, 2023).
31. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations* (2021).
32. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
33. Clark, P., Tafjord, O. & Richardson, K. Transformers as soft reasoners over language. In *Proc. 29th International Conference on International Joint Conferences on Artificial Intelligence* 3882–3890 (2021).
34. Creswell, A., Shanahan, M. & Higgins, I. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *Proc. Eleventh International Conference on Learning Representations* (ICLR, 2022).
35. Chen, W. Large language models are few (1)-shot table reasoners. In *Proc. Findings of the Association for Computational Linguistics: EACL 2023* 1090–1100 (ACL Anthology, 2023).
36. Wang, B., Yue, X. & Sun, H. Can Chatgpt defend its belief in truth? evaluating LLM reasoning via debate. In *Proc. Findings of the Association for Computational Linguistics: EMNLP 2023* 11865–11881 (EMNLP, 2023).
37. Liu, H. et al. Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439* (2023).
38. Johnson, A. E. et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
39. Irvin, J. et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
40. Smit, A. et al. Combining automatic labelers and expert annotations for accurate radiology report labeling using Bert. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1500–1519 (EMNLP, 2020).
41. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
42. Touvron, H. et al. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
43. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
44. Zhang, J. et al. Fengshenbang 1.0: being the foundation of Chinese cognitive intelligence. *CoRR***abs/2209.02970** (2022).
45. Zeng, A. et al. GLM-130b: an open bilingual pre-trained model. In *Proc. 11th International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=-Aw0rrrPUF (2023).
46. Jiang, A. Q. et al. Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (grant numbers 62131015, U23A20295, 82394432), the STI 2030-Major Projects (No. 2022ZD0209000), Shanghai Municipal Central Guided Local Science and Technology Development Fund (grant number YDZX20233100001001), Science and Technology Commission of Shanghai Municipality (grant number 21010502600), and The Key R&D Program of Guangdong Province, China (grant numbers 2023B0303040001, 2021B0101420006).

Author contributions

S.W. and Z.Z. conceptualized the study, collected data, designed the algorithm, and performed experiments. S.W., Z.Z., and X.O. drafted the manuscript. Q.W. and D.S. contributed to the conceptual design, supervised the project, and edited the paper. T.L. provided edits to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-024-00271-8.

Correspondence and requests for materials should be addressed to Qian Wang or Dinggang Shen.

Peer review information *Communications Engineering* thanks Huang Yixing and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Or Perlman and Mengying Su and Ros Daw. A peer review file is available.