ChatCAD+: Towards a Universal and Reliable Interactive CAD using LLMs

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu,
Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, Dinggang Shen, *Fellow, IEEE*

Abstract— The integration of Computer-Aided Diagnosis (CAD) with Large Language Models (LLMs) presents a promising frontier in clinical applications, notably in automating diagnostic processes akin to those performed by radiologists and providing consultations similar to a virtual family doctor. Despite the promising potential of this integration, current works face at least two limitations: (1) From the perspective of a radiologist, existing studies typically have a restricted scope of applicable imaging domains, failing to meet the diagnostic needs of different patients. Also, the insufficient diagnostic capability of LLMs further undermine the quality and reliability of the generated medical reports. (2) Current LLMs lack the requisite depth in medical expertise, rendering them less effective as virtual family doctors due to the potential unreliability of the advice provided during patient consultations. To address these limitations, we introduce ChatCAD+, to be universal and reliable. Specifically, it is featured by two main modules: (1) Reliable Report Generation and (2) Reliable Interaction. The Reliable Report Generation module is capable of interpreting medical images from diverse domains and generate high-quality medical reports via our proposed hierarchical in-context learning. Concurrently, the interaction module leverages up-to-date information from reputable medical websites to provide reliable medical advice. Together, these designed modules synergize to closely align with the expertise of human medical professionals, offering enhanced consistency and reliability for interpretation and advice. The source code is available at GitHub.

Index Terms— Large Language Models, Multi-modality System, Medical Dialogue, Computer-Assisted Diagnosis
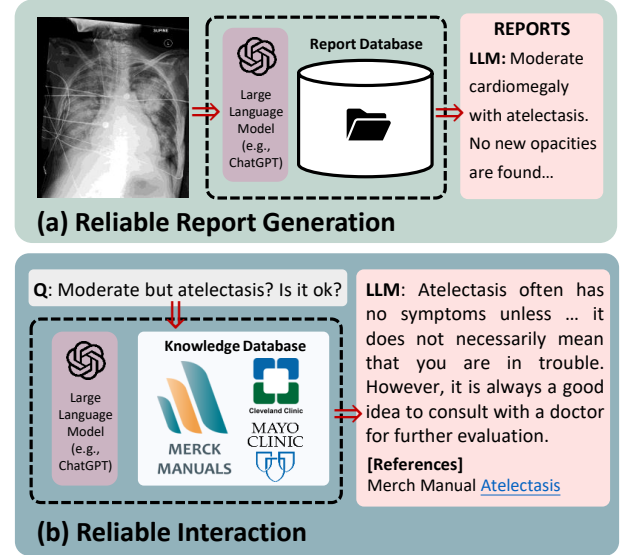
Fig. 1. Overview of our proposed ChatCAD+ system. (a) For patients seeking a diagnosis, ChatCAD+ generates reliable medical reports based on the input medical image(s) by referring to local report database. (b) Additionally, for any inquiry from patients, ChatCAD+ retrieves related knowledge from online database and lets large language model generate reliable response.

I. INTRODUCTION

Zihao Zhao, Sheng Wang, Jinchen Gu and Yitao Zhu contributed equally and are listed as first authors.

Corresponding authors: Qian Wang and Dinggang Shen.

Zihao Zhao, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, and Zhiming Cui are with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China (email: zihaozhao10@gmail.com, {gujch12022,zhuyt, meilzhj, cuizm}@shanghaitech.edu.cn)

Sheng Wang and Zixu Zhuang are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China, Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China, and also ShanghaiTech University, Shanghai 201210, China. (e-mail: {wsheng, zixuzhuang}@sjtu.edu.cn)

Qian Wang and Dinggang Shen are with the School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China, and Shanghai Clinical Research and Trial Center, Shanghai 201210, China. Dinggang Shen is also with Shanghai United Imaging Intelligence Co. Ltd., Shanghai 200230, China. (e-mail: {qianwang, dgshen}@shanghaitech.edu.cn)

LARGE Language Models (LLMs) have emerged as promising tools in various domains. It refers to advanced artificial intelligence that has been extensively trained on text data. Drawing on the combined power of sophisticated deep learning techniques, large-scale datasets, and increased model sizes, LLMs demonstrate extraordinary capabilities in understanding and generating human-like text. This is substantiated by significant projects like ChatGPT [1] and LLaMA [2]. These techniques are highly suitable for a wide range of scenarios, such as customer service, marketing, education, and healthcare consultation. Notably, ChatGPT has passed part of the US medical licensing exams, highlighting the potential of LLMs in the medical domain [3], [4].

So far, several studies have been conducted to integrate LLMs into Computer-Assisted Diagnosis (CAD) systems of medical images. Conventional CAD typically operates following pure computer vision [5], [6] or vision-language paradigms [7], [8]. LLMs have shown ability to effectively interpret findings from medical images, thus mitigating limitations of only visual interpretation. For example, in our

pilot study, we leverage the intermediate results obtained from image CAD networks and then utilize LLMs to generate final diagnostic reports [9].

Although some efforts have been made to combine CAD networks and LLMs [9]–[11], it should be noted that these studies are limited in their scopes, which often focus on specific image domains. That is, such a system may only support a single image modality, organ, or application (such as chest X-ray), which greatly limits generalizability in the real clinical workflow. The primary reason for this limitation comes from notable topologic and semantic variations observed among medical image data, which present distinct challenges when attempting to encode various images with a unified model. Furthermore, although LLMs have demonstrated the capability to articulate medical concepts and clinical interpretations [12], [13], there still exists a significant gap between LLMs and radiologists. This gap undoubtedly undermines patient confidence in a LLM-generated report. Consequently, these issues hinder practical utilization of LLMs in automated generation of medical reports. In addition, it is also articulable that the integration of LLMs and CADs may lead to medical dialogue systems [14]–[16]. In this way, patients will be able to interact through LLMs and acquire more medical advice and explanation, while this functionality is often missing in conventional CAD systems. However, existing studies show that the general LLMs typically produce medical advice based solely on their encoded knowledge, without considering specific knowledge in the medical domain [15]. As a result, patients may receive unreliable responses, dwarfing trust in CADs and thereby hindering the use of such systems in real medical scenarios.

In order to tackle the challenges mentioned above, we propose ChatCAD+ in this paper. And our contributions are made in the following aspects. **(1) Universal image interpretation.** Due to the difficulty in obtaining a unified CAD network tackling various images currently, ChatCAD+ incorporates a domain identification module to work with a variety of CAD models (c.f. Fig. 2(a)). ChatCAD+ can adaptively select a corresponding model given the input medical image. The tentative output of the CAD network is converted into text description to reflect image features, making it applicable for diagnostic reporting subsequently. **(2) Hierarchical in-context learning for enhanced report generation.** Top-k reports that are semantically similar to the LLM-generated report are retrieved from a clinic database via the proposed retrieval module (c.f. Fig. 3. The retrieved k reports then serve as in-context examples to refine the LLM-generated report. **(3) Knowledge-based reliable interaction.** As illustrated in Fig. 1(b), ChatCAD+ does not directly provide medical advice. Instead, it first seeks help via our proposed knowledge retrieval module for obtaining relevant knowledge from professional sources, e.g. Merck Manuals, Mayo Clinic, and Cleveland Clinic. Then, the LLM considers the retrieved knowledge as a reference to provide reliable medical advice.

In summary, our ChatCAD+ for the first time builds a universal and reliable medical dialogue system. The improved quality of answers and diagnostic reports of the chatbot reveals potential of LLMs in interactive medical consultation.

II. RELATED WORKS

A. Large Language Models in Healthcare

Following the impressive potential demonstrated by Chat-GPT with its 100B-scale parameters, researchers have further expanded the application of language models in health-care unprecedentedly, yielding highly promising results. Med-PaLM [17] was developed using meticulously curated biomedical corpora and human feedback, showcasing its potential with notable achievements, including a remarkable 67.6% accuracy on the MedQA exam. Notably, ChatGPT, which did not receive specific medical training, successfully passed all three parts of the USMLE and achieved an overall accuracy of over 50% across all exams, with several exams surpassing 60% accuracy [18]. ChatDoctor [14], fine-tuned on the LLaMA model using synthetic clinical QA data generated by Chat-GPT, also incorporated medical domain knowledge to provide reliable responses. However, our ChatCAD+ distinguishes its reliable interaction from ChatDoctor in several aspects. First, ChatDoctor relies on Wikipedia, which is editable for all users and may compromise the rigor and reliability required in the medical domain. Second, ChatDoctor simply divides the knowledge of each medical topic into equal sections and relies on rule-based ranking to select relevant information. In contrast, our reliable interaction leverages the Merck Manual [19], a comprehensive and trusted medical knowledge resource. It can organize medical knowledge into a hierarchical structure, thus enabling us to fully exploit the reasoning capabilities of the LLM in recursively retrieving the most pertinent and valuable knowledge.

B. Multi-modality Large Models

In present times, multi-modality large models are typically constructed based on pre-trained models rather than training them from scratch. Frozen [20] performed fine-tuning on an image encoder, utilizing its outputs as soft prompts for the language model. Flamingo [21] introduced cross-attention layers into the LLM to incorporate visual features, exclusively pre-training these new layers. BLIP-2 [22] leveraged both frozen image encoders and frozen LLMs, connecting them through their proposed Q-Former. Following the architecture of BLIP-2, XrayGPT [23] was specifically developed on chest X-ray. While the training cost for XrayGPT is demanding, our proposed ChatCAD+ can seamlessly integrate existing pre-trained CAD models without requiring further training. This highlights the inherent advantage of ChatCAD+ in continual learning scenarios, where additional CAD models can be effortlessly incorporated.

In more recent studies, researchers have taken a different approach, i.e., designing prompts to allow LLMs to utilize vision models for avoiding training altogether. For instance, Visual-ChatGPT [24] established a connection between ChatGPT and a series of visual foundation models, enabling exchange of images during conversations. ChatCAD [9] linked existing CAD models with LLMs to enhance diagnostic accuracy and improve patient care. Following a similar paradigm as ChatCAD [9], Visual Med-Alpaca [25] incorporated multiple vision models to generate text descriptions for medical images.
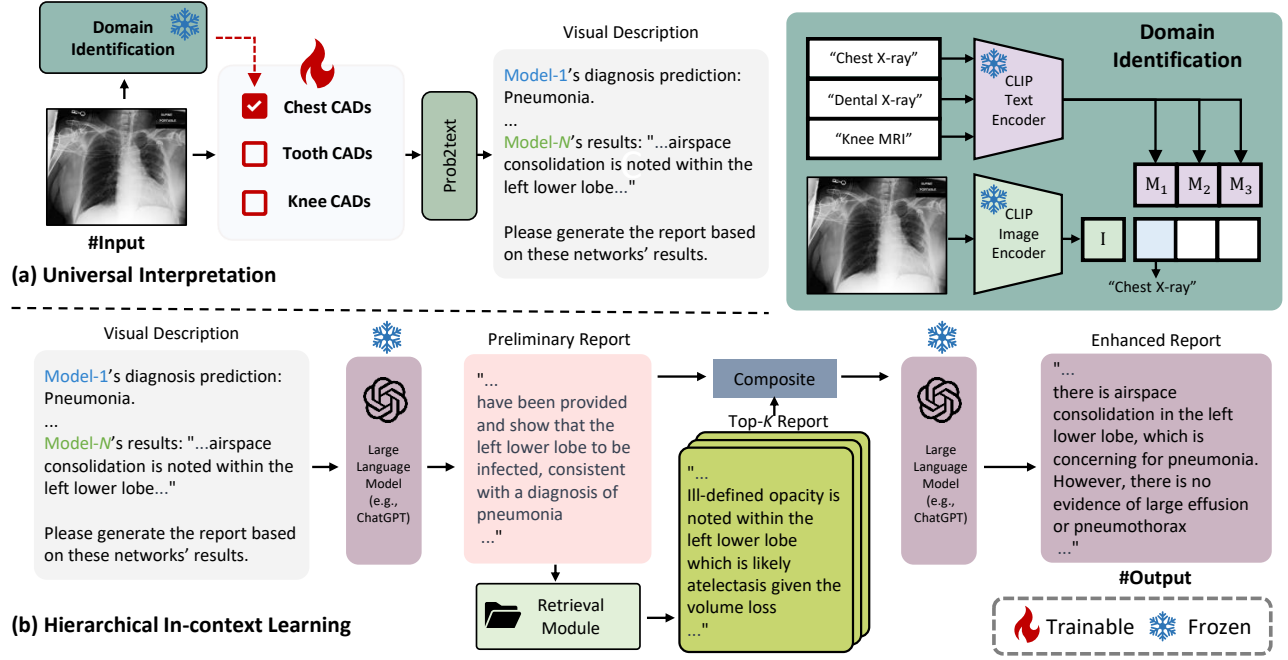
Fig. 2. **Overview of the reliable report generation.** (a) Universal interpretation: To enhance the vision capability of LLMs, multiple CAD models are incorporated. The numerical results obtained from these models are transformed into visual descriptions following the rule of **prob2text**. (b) Hierarchical in-context learning: The LLM initially generates a preliminary report based on the visual description, which is then enhanced through in-context learning using retrieved semantically similar reports. LLM and CLIP models are kept frozen, while CAD models are denoted as trainable as they can be continually trained and updated without interfering with other components.

However, when compared to our study, it is limited to a single medical imaging domain and also lacks reliability in both report generation and interactive question answering.

III. METHOD

The proposed ChatCAD+ is a multi-modality system capable of processing both image and text inputs (as illustrated in Fig. 1) through reliable report generation and reliable interaction, respectively. It is important to note that, in this paper, the term "modality" refers to language and image, which differs from the commonly used concept of medical image modality. The reliable report generation treats medical image as input, identifies its domain, and provides enhanced medical reports by referencing a local report database. Simultaneously, for the user's text query, the reliable interaction will retrieve clinically-sound knowledge from a professional knowledge database, thereby serving as a reliable reference for the LLM (e.g. ChatGPT) to enhance the reliability its response.

A. Reliable Report Generation

This section outlines our proposed reliable report generation, as depicted in Fig. 2, which encompasses two primary stages: universal interpretation and hierarchical in-context generation. Section III-A.1 illustrates the paradigm to interpret medical images from diverse domains into descriptions that can be understood by LLMs. The resulting visual description then serves as the input for subsequent hierarchical in-context learning. Following this, Section III-A.2 elucidates hierarchical in-context learning, within which the retrieval module (in Fig. 3) acts as key component.

TABLE I
THE ILLUSTRATION OF PROB2TEXT. P3 IS THE DEFAULT SETTING.

CAD Model→(**disease**: **prob**)	
P1 (direct)	"{**disease**} score: {**prob**}"
P2 (simplistic)	**prob**∈[0, 0.5): "No Finding" **prob**∈[0.5, 1]: "The prediction is {**disease**}"
P3 (illustrative)	**prob**∈[0, 0.2): "No sign of {**disease**}" **prob**∈[0.2, 0.5): "Small possibility of {**disease**}" **prob**∈[0.5, 0.9): "Patient is likely to have {**disease**}" **prob**∈[0.9, 1]: "Definitely have {**disease**}"

1) Universal Interpretation: The process of universal interpretation can be divided into three steps. Initially, the proposed domain identification module is employed to determine the specific domain of the medical image. Subsequently, the corresponding domain-specific model is activated to interpret the image. Finally, the interpretation result is converted into a text prompt for describing visual information via the rule-based prob2text module for further processing.

Domain identification chooses CAD models to interpret the input medical image, which is fundamental for subsequent operations of ChatCAD+. To achieve this, we employ a method of visual-language contrast that computes the cosine similarity between the input image and textual representations of various potential domains. This approach takes advantage of language's ability in densely describing the features of a particular type of image, as well as its ease of extensibility. In particular, we employ the pre-trained BiomedCLIP model [26] to encode both the medical image and the text associated with domains of interest. In this study, we demonstrate upon
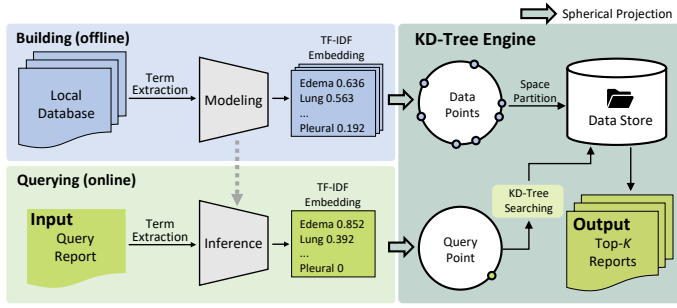
Fig. 3. **The illustration of the retrieval module within reliable report generation.** It adopts the TF-IDF algorithm to preserve the semantics of each report and converts it into a latent embedding during offline modeling and online inference. To facilitate highly efficient retrieval, we perform spherical projection on all TF-IDF embeddings, whether during building or querying. In this manner, we can utilize the KD-Tree structure to store these data and implement retrieval with a low time complexity.

three domains: chest X-ray, dental X-ray, and knee MRI. The workflow is depicted in the upper right portion of Fig. 2. Assuming there are three domains D_1, D_2, D_3, along with their textual representations M_1, M_2, M_3, and also a visual representation denoted as I for the input image, we define

$$D_{\text{pred}} = \text{argmax}_{i \in \{1,2,3\}} \frac{I \, M_i}{\|I\| \, \|M_i\|}, \quad (1)$$

where D_{pred} denotes the prediction of the medical image domain. The module thereby can call the domain-specific CAD model to analyze visual information given D_{pred}.

Since most CAD models generate outputs that can hardly be understood by language models, further processing is needed to bridge this gap. For example, an image diagnosis model typically outputs tensors representing the likelihood of certain clinical findings. To establish a link between image and text, these tensors are transformed into textual descriptions according to diagnostic-related rules, which is denoted as prob2text in Fig. 2(a). The prob2text module has been designed to present clinically relevant information in a manner that is more easily interpretable by LLMs. The details of the prompt design are illustrated in Table I. Using chest X-ray as an example, we follow the three types (P1-P3) of prompt designs in [9] and adopt **P3 (illustrative)** as the recommended setting in this study. Concretely, it employs a grading system that maps the numerical scores into clinically illustrative description of disease severity. The scores are divided into four levels based on their magnitude: "No sign", "Small possibility", "Likely", and "Definitely". The corresponding texts are then used to describe the likelihood of different observations in chest X-ray, providing a concise and informative summary of the patient's condition. The prompt design for dental X-ray and knee MRI are in similar ways. And other prompt designs, such as P1 and P2 [9], will be discussed in experiments below.

2) Hierarchical In-context Learning: The proposed strategy for hierarchical in-context learning, as shown in Fig. 2(b), consists of (1) preliminary report generation and (2) retrieval-based in-context enhancement.

In the first step, the LLM utilizes the visual description to generate a preliminary diagnostic report. This process is facilitated through prompts such as "Write a report based

on results from Network(s)." It is important to note that the generation of this preliminary report relies solely on the capability of the LLM and is still far from matching capability of human specialists. As a result, even powerful LLMs like ChatGPT still struggle to achieve satisfactory results.

To address this concern, the next step involves retrieving the top-k reports that share similar semantics with the preliminary report as in-context examples. The retrieval module is designed to preserve semantic information and ensure efficient implementation. Specifically, we employ TF-IDF [27] to model the semantics of each report, and the KD-Tree [28] data structure is utilized to facilitate high-speed retrieval. Fig. 3 illustrates the detailed pipeline of the proposed retrieval module. For simplification, we use chest X-ray in MIMIC-CXR as an example in this section. To better preserve disease-related information in the report, we focus on 17 medical terms related to thoracic diseases during the modeling of TF-IDF embeddings, which form a term set $\mathcal{T}$. Assuming the training set of MIMIC-CXR as $\mathcal{D}$, the TF-IDF score for each term $t \in \mathcal{T}$ in each document $d \in \mathcal{D}$ is computed as:

$$\begin{aligned} \text{TF-IDF}(t, d) &= \text{TF}(t, d) \cdot \text{IDF}(t) \\ &= \frac{|\{w \in d : w = t\}|}{|d|} \cdot \log \frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : t \in d\}|}, \end{aligned} \quad (2)$$

where w indicates a word within the document d. In brief, $\text{TF}(t, d)$ represents the frequency of term t occurring in d, and $\text{IDF}(t)$ denotes the inverse of the frequency that document d occurs in the training set $\mathcal{D}$. Following this, we define the TF-IDF embedding (TIE) of a document d as:

$$\text{TIE}(d) = \{\text{TF-IDF}(t, d) | t \in \mathcal{T}\}. \quad (3)$$

After calculating the TIEs for all $d \in \mathcal{D}$, we organize them using the KD-Tree [28] data structure, which enables fast online querying. The KD-Tree is a space-partitioning data structure that can implement k-nearest-neighbor querying in $O(\log(n))$ time on average, making it suitable for information searching. However, since the implementation of KD-Tree relies on L_2 distance instead of cosine similarity, it cannot be directly applied to our TIEs. To address this issue, we project all TIEs onto the surface of a unit hypersphere, as shown in the right panel of Fig. 3. After spherical projection, we have:

$$L_2\left(\frac{\vec{q}}{|\vec{q}|}, \frac{\vec{v}}{|\vec{v}|}\right) = 2r \cdot \sin\left(\frac{\theta}{2}\right), \quad (4)$$

where $\vec{q}$ and $\vec{v}$ represent the TIE of the query report and a selected report, respectively. r indicates the radius of the hypersphere (radius=1 in this study), and $\theta \in [0, \pi]$ stands for the angle between $\frac{\vec{q}}{|\vec{q}|}$ and $\frac{\vec{v}}{|\vec{v}|}$. The L_2 distance between two TIEs will monotonically increase as a function of the angle between them, which favors the application of the KD-Tree data structure for efficient retrieval.

During online querying, the KD-Tree engine retrieves the top-k (i.e., k=3) reports that share the most similar semantics with the query report (i.e., preliminary report illustrated in Fig. 2). It then asks the LLM to refine its preliminary report with the retrieved k reports as references, and generate the enhanced report as the final output.
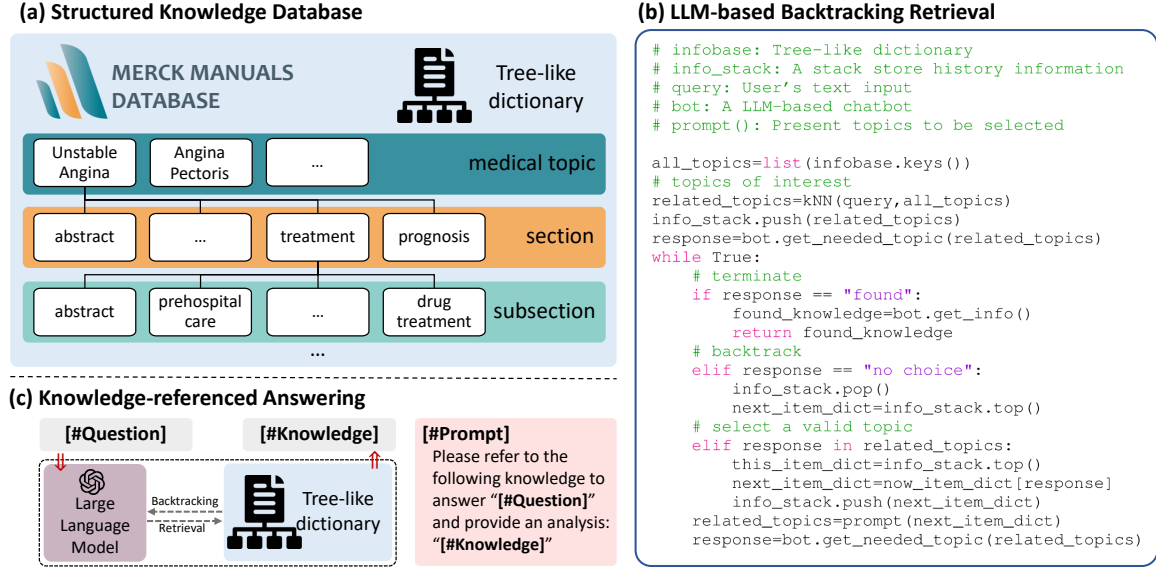
(a) Structured Knowledge Database

MERCK MANUALS DATABASE

Tree-like dictionary

| Unstable Angina | Angina Pectoris | ... | medical topic |

| abstract | ... | treatment | prognosis | section |

| abstract | prehospital care | ... | drug treatment | subsection |

...

(c) Knowledge-referenced Answering

[#Question]   [#Knowledge]

[#Prompt]
Please refer to the following knowledge to answer "[#Question]" and provide an analysis: "[#Knowledge]"

Large Language Model

Backtracking Retrieval

Tree-like dictionary

(b) LLM-based Backtracking Retrieval

```python
# infobase: Tree-like dictionary
# info_stack: A stack store history information
# query: User's text input
# bot: A LLM-based chatbot
# prompt(): Present topics to be selected

all_topics=list(infobase.keys())
# topics of interest
related_topics=kNN(query,all_topics)
info_stack.push(related_topics)
response=bot.get_needed_topic(related_topics)
while True:
    # terminate
    if response == "found":
        found_knowledge=bot.get_info()
        return found_knowledge
    # backtrack
    elif response == "no choice":
        info_stack.pop()
        next_item_dict=info_stack.top()
    # select a valid topic
    elif response in related_topics:
        this_item_dict=info_stack.top()
        next_item_dict=now_item_dict[response]
        info_stack.push(next_item_dict)
    related_topics=prompt(next_item_dict)
    response=bot.get_needed_topic(related_topics)
```

Fig. 4. Overview of the reliable interaction. (a) The illustration of structured medical knowledge database, organized as a tree-like dictionary, where each medical topic has multiple sections while sections can be further divided into subsections. (b) A LLM-based knowledge retrieval method is proposed to search relevant knowledge in a backtrack manner. (c) The LLM is prompted to answer the question based on the retrieved knowledge.

B. Reliable Interaction

This section outlines the reliable interaction, which is responsible for answering the user's text query. It serves a dual role within our proposed ChatCAD+. First, reliable interaction can function as a sequential module of reliable report generation to enhance interpretability and understandability of medical reports for patients. In this capacity, it takes the generated report from previous module as the input, and provides patients with explanations of involved clinical terms[1]. Second, it can also operate independently to answer clinical-related questions and provide reliable medical advice.

The proposed ChatCAD+ offers reliable interaction via the construction of a professional knowledge database and LLM-based knowledge retrieval. In this study, we demonstrate using the Merck Manuals. The Merck Manuals are a series of healthcare reference books that provide evidence-based information on the diagnosis and treatment of diseases and medical conditions.

The implementation of LLM-based knowledge retrieval leverages the Chain-of-Thought (CoT) technique, which is widely known for its ability to enhance the performance of LLM in problem-solving. CoT breaks down a problem into a series of sub-problems and then synthesizes the answers to these sub-problems progressively for solving the original problem. Humans typically do not read the entire knowledge base but rather skim through topic titles to find what they are looking for. Inspired by this, we have designed a prompt system that automatically guides the LLM to execute such searches. Correspondingly, we have structured the Merck Manuals database as a hierarchical dictionary, with topic titles of different levels serving as keys, as shown in Fig 4(a). Fig. 4(b) demonstrates the proposed knowledge retrieval meth-

---

[1]Clinical terms within a radiology report can be automatically extracted via tools like RadGraph [29].

---

ods. Initially, we provide the LLM with only titles of five related medical topics in the database, and then ask the LLM to select the most relevant topic to begin the retrieval. Once the LLM has made its choice, we provide it with the content of "abstract" section and present names of all other sections subject to this topic. Given that medical topics often exhibit hierarchical organization across three or four tiers, we repeat this process iteratively, empowering the LLM to efficiently navigate through all tiers of a given medical topic. Upon identifying relevant medical knowledge, the LLM is prompted to return the retrieved knowledge, terminating the process. Otherwise, in the case that the LLM finds none of the provided information as relevant, our algorithm would backtrack to the parent tier, which makes it a kind of depth-first search [30]. Finally, the LLM is prompted to provide reliable response based on the retrieved knowledge (in Fig. 4(c)).

IV. EXPERIMENTAL RESULTS

To comprehensively assess our proposed ChatCAD+, we conducted evaluations across three crucial aspects that influence its effectiveness in real-world scenarios: (1) the capability to handle medical images from different domains, (2) report generation quality, and (3) the efficacy in clinical question answering. This section commences with an introduction to the datasets used and implementation details of ChatCAD+. Then, we delve into the evaluation of these aspects in turn.

A. Dataset and implementations

Domain identification: To investigate the robustness of domain identification, which is the fundamental component of our proposed universal CAD system, we crafted a dataset covering nine medical imaging domains by randomly selecting samples from different datasets. The statistics of this dataset

	chest X-ray	dental X-ray	knee X-ray	mammography	chest CT	fundus	endoscopy	dermoscopy	blood cell
number	400	360	800	360	320	120	200	360	200
source	MedFMC [31]	Tufts Dental [32]	OAI [33]	INbreast [34]	COVID-CT [35]	DIARETDB1 [36]	MedFMC [31]	ISIC2016 [37]	PBC [38]
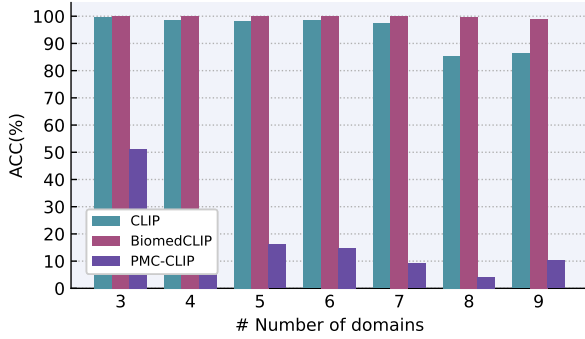
Fig. 5. Evaluation of domain identification using different CLIPs.

are presented in Table II. During the evaluation, we progressively increase the number of domains following the left-to-right order in Table II and report the accuracy (ACC) in each step. Since the pre-trained CLIP model is utilized to perform domain identification (as shown in Fig. 2), we also investigated the impact of different CLIPs in Sec. IV-B.

Report generation: For a fair comparison, the report generation performance of ChatCAD+ is assessed on the public MIMIC-CXR dataset [39]. The MIMIC-CXR is a large public dataset of chest X-ray images, associated radiology reports, and other clinical data. The dataset consists of 377,110 de-identified chest X-ray images and associated 227,827 radiology reports. The MIMIC-CXR dataset allows us to measure the quality of diagnostic accuracy and report generation, as most report generation methods are designed specifically for chest X-ray. The quality of report generation was tested on the official test set of the MIMIC-CXR dataset, focusing on five findings (cardiomegaly, edema, consolidation, atelectasis, and pleural effusion). Since ChatGPT was adopted as the default LLM in report generation experiments, we randomly selected 200 samples for each finding of interest due to limitations on the per-hour accesses of OpenAI API-Key, resulting in a total of 1,000 test samples. To evaluate the performance of report generation, we used several widely-used Natural Language Generation (NLG) metrics, including BLEU [40], METEOR [41], and ROUGE-L [42]. Specifically, BLEU measures the similarity between generated and ground-truth reports by counting word overlap and we denote Corpus BLEU as C-BLEU by default. METEOR considers synonym substitution and evaluates performance on both sentence level and report level. Meanwhile, ROUGE-L evaluates the length of the longest common subsequence. We also measured clinical efficacy (CE) of reports using the open-source CheXbert library [43]. CheXbert takes text reports as input and generates multi-label classification labels for each report, with each label corresponding to one of the pre-defined thoracic diseases.

Based on these extracted multi-hot labels, we compute precision (PR), recall (RC), and F1-score (F1) on 1,000 test samples. These metrics provide additional insight into the performance of report generation. If not specified, we choose P3 as a default prompt design in universal interpretation and k=3 for hierarchical in-context learning.

Clinical question answering: The evaluation of reliable interaction was performed using a subset of the CMExam dataset [44]. CMExam is a collection of multiple-choice questions with official explanations, sourced from the Chinese National Medical Licensing Examination. Since CMExam covers a diversity of areas, including many clinical-unrelated questions, we focused on a subset of the CMExam test set that specifically addressed "Disease Diagnosis" and "Disease Treatment," while excluding questions belonging to the department of "Traditional Chinese Medicine". This resulted in a subset of 2190 question-answer pairs. Since the CMExam dataset is completely in Chinese, we selected ChatGLM2 [45] and ChatGLM3 [45], two 6B scale models, as the LLMs for validation. These models have demonstrated superior Chinese language capabilities compared to other open-source LLMs. We employed the official prompt design of CMExam, which required the LLM to make a choice and provide the accompanying explanation simultaneously. We acknowledge that our proposed LLM-based knowledge retrieval involves recursive requests to the LLM and will consume a large amount of tokens, making experiments using ChatGPT and GPT-4 unaffordable for us. In experimental results, we report ACC and F1 for multi-choice selection, and NLG metrics for reason explanation.

Implementation of domain-specific CADs: In Section III-A.1, we proposed to incorporate domain-specific CAD models to interpret images from different medical imaging domain. In this study, we adopted Chest CADs, Tooth CADs, and Knee CADs. Chest CADs is composed of a thoracic disease classifier [46] and a chest X-ray report generation network [7]. The former [46] was trained on the official training split of the CheXpert [47] dataset, using a learning rate (lr) of $1e^{-4}$ for 10 epochs. The report generation network [7] was trained on the training split of MIMIC-CXR [39] using a single NVIDIA A100 GPU. We used the Adam optimizer with a learning rate of $5e^{-5}$. For Tooth CAD, we utilized the periodontal diagnosis model proposed in [48], which was trained on 300 panoramic dental X-ray images collected from real-world clinics with a size of 2903x1536. The training was conducted on a NVIDIA A100 GPU with 80GB memory for a total of 200 epochs. We set the learning rate as $2e^{-5}$ using the Adam optimizer. For Knee CAD, we adopted the model proposed in [49], which was trained on 964 knee MRIs captured using a Philips Achieva 3.0T TX MRI scanner from Shanghai Sixth People's Hospital. The Adam optimizer was used with a weight decay of $1e^{-4}$.
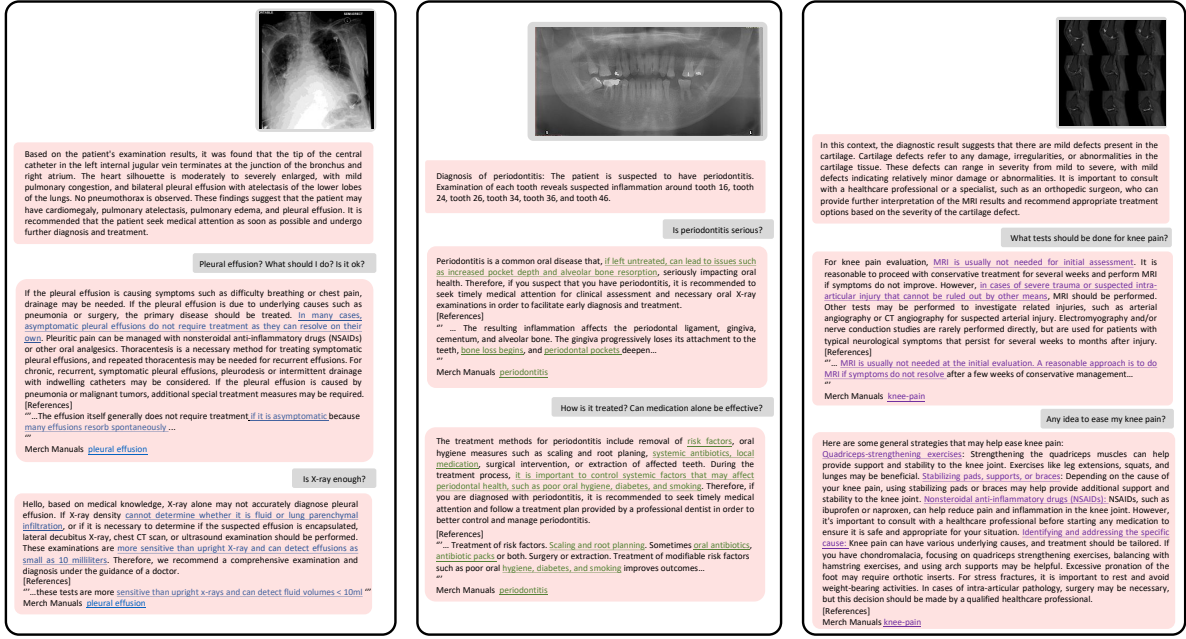
Fig. 6. Examples of universal and interactive medical consultation using ChatCAD+, with ChatGPT as the default LLM. The underlined text signifies information obtained from reputable medical websites, which is not comprehensively incorporated in ChatGPT.

The learning rate was set as $3e^{-5}$ for Transformer and graph attention network components, and $3e^{-4}$ for others.

B. Evaluation of a Universal System

To investigate the effectiveness of ChatCAD+ as a universal CAD system, we evaluated it by (1) quantitatively investigating its domain identification capability when encountering images from various domains, and (2) qualitatively investigating its ability to provide reliable response in different scenarios.

1) Robustness of domain identification: In Fig. 5, we visualize the domain identification performance of different CLIP models. Compared with the baseline CLIP model, Biomed-CLIP [26] achieves an impressive performance, showing 100% accuracy on 2560 images from 7 different domains. It can even maintain an accuracy of 98.9% when all 9 domains are involved, while the performance of CLIP is degraded to 86.3%. These observations demonstrate that BiomedCLIP has an more in-depth understanding of professional medical knowledge compared with CLIP, and thus is suitable for domain identification in our proposed ChatCAD+.

However, to our surprise, PMC-CLIP [50] performs rather poorly on this task. As shown in Fig. 5, it only achieves approximate 50% accuracy when three domains are involved. We hypothesize that this can be attributed to its lack of zero-shot capability. While PMC-CLIP has demonstrated promising quality fof fine-tuning on donwstream tasks, its lack of zero-shot inference capability may hinder its wide-speard application. In conclusion, BiomedCLIP can be a better choice than both CLIP and PMC-CLIP for domain identification.

2) Reliability of consultations: Fig. 6 demonstrates the universal application and reliability of our proposed method. The ChatCAD+ can process different medical images and provide proper diagnostic reports. Users can interact with ChatCAD+

and ask additional questions for further clarification. By leveraging external medical knowledge, the enhanced ChatCAD+ is capable of delivering medical advice in a more professional manner. In Fig. 6, each reply from ChatCAD+ includes a reference, and the pertinent knowledge obtained from the external database is underlined for emphasis. For instance, in the scenario where the question is raised about whether a chest X-ray is sufficient for diagnosing pleural effusion, ChatGPT would simply recommend a CT scan without providing a convincing explanation. In contrast, ChatCAD+ is capable of informing the patient that CT scans have the capability to detect effusions as small as 10 milliliters, thereby providing a more detailed and informative response.

C. Evaluation of Report Generation

We begin by comparing our approach to state-of-the-art (SOTA) report generation methods, while also examining the impact of various prompt designs. Subsequently, we assess the generalizability of our reliable report generation model by utilizing diverse LLMs. Following this, we explore the impact of hierarchical in-context learning by changing the number of retrieved reports. We then proceed to qualitatively showcase the advanced capabilities of ChatCAD+ in report generation. Finally, we compare our specialist-based system against others' generalist medical AI systems.

1) Comparison with SOTA methods: The numerical results of different methods was compared and presented in Table III. The advantage of ChatCAD+ lies in its superior performance across multiple evaluation metrics compared to both previous models and different prompt settings. Particularly, ChatCAD+ (P3) stands out with the highest scores in terms of C-BLEU, METEOR, REC, and F1. Despite the inferiority in specific metrics, ChatCAD+ (P3) still holds a significant advantage

Model	Size	C-BLEU	ROUGE-L	METEOR	PRE	REC	F1
R2GenCMN [7]	244M	3.525	**18.351**	21.045	0.578	0.411	0.458
VLCI [8]	357M	3.834	16.217	20.806	**0.617**	0.299	0.377
ChatCAD (P3)	256M+175B	3.594	16.791	23.146	0.526	0.603	0.553
ChatCAD+ (P1)	256M+175B	4.073	16.820	22.700	0.548	0.484	0.502
ChatCAD+ (P2)	256M+175B	4.407	17.266	23.302	0.538	0.597	0.557
ChatCAD+ (P3)	256M+175B	**4.409**	17.344	**24.337**	0.531	**0.615**	**0.564**

Model	Size	Reliable	C-BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	PRE	REC	F1
ChatGLM2 [45]	6B	✗	2.073	16.999	3.426	0.926	0.313	14.119	23.008	**0.518**	0.632	0.553
		✔	**3.236**	**25.172**	**5.268**	**1.461**	**0.566**	**15.559**	**23.397**	0.501	**0.639**	**0.555**
PMC-LLaMA [51]	7B	✗	**1.709**	**15.003**	**2.432**	**0.748**	**0.313**	**13.983**	**20.590**	**0.492**	**0.702**	**0.532**
		✔	1.059	6.5059	1.591	0.504	0.241	8.140	18.551	0.477	0.664	0.525
MedAlpaca [52]	7B	✗	1.554	30.710	6.143	1.906	0.819	12.729	11.865	**0.526**	0.464	0.484
		✔	**3.031**	**30.795**	**7.107**	**2.408**	**1.107**	**14.761**	**15.711**	0.507	**0.518**	**0.505**
Mistral [53]	7B	✗	3.261	27.806	6.068	1.463	0.458	16.132	21.999	**0.558**	0.444	0.485
		✔	**3.547**	**30.851**	**7.178**	**1.698**	**0.493**	**17.493**	**22.347**	0.543	**0.597**	**0.558**
LLaMA [2]	7B	✗	0.807	7.575	1.265	0.337	0.128	10.016	17.194	**0.495**	0.589	0.519
		✔	**0.966**	**7.790**	**1.525**	**0.423**	**0.178**	**10.237**	**18.344**	0.478	**0.625**	**0.535**
LLaMA2 [54]	7B	✗	1.298	14.929	2.717	0.531	0.132	12.016	19.958	0.481	0.588	0.519
		✔	**3.579**	**24.798**	**5.826**	**1.638**	**0.693**	**16.850**	**23.950**	**0.512**	0.561	**0.528**
LLaMA [2]	13B	✗	0.810	6.824	1.393	0.331	0.113	8.959	16.647	0.486	0.452	0.459
		✔	**0.973**	**8.295**	**1.521**	**0.439**	**0.197**	**9.813**	**18.073**	**0.498**	**0.465**	**0.463**
LLaMA2 [54]	13B	✗	2.940	23.536	5.173	1.390	0.441	15.053	21.938	0.531	0.564	0.538
		✔	**3.650**	**25.080**	**6.148**	**1.728**	**0.666**	**16.646**	**24.015**	**0.533**	**0.571**	**0.543**
ChatGPT [1]	175B	✗	3.594	28.453	6.371	1.656	0.556	16.791	23.146	0.526	0.603	0.553
		✔	**4.409**	**31.625**	**7.570**	**2.107**	**0.760**	**17.344**	**24.337**	**0.531**	**0.615**	**0.564**

across other metrics, which validates its advanced overall performance. Note that ChatCAD+ (P3) shows a significant advantage over ChatCAD (P3) on all NLG and CE metrics, which underscores the rationality of hierarchical in-context learning. These results suggest that ChatCAD+ (P3) is capable of generating high-quality responses that are not only fluent and grammatical but also clinically accurate, making it a promising model for conversational tasks.

Additionally, the impact of prompt designs can also be observed. Generally speaking, P3 achieves the best performance while P1 leads to unsatisfying results. This phenomenon may be attributed to the similarity of P3 to human language style, as it reflects the severity of the disease using rhetoric, whereas P1 directly displays the probability without post-process.

2) Investigation of generalizability: In Table IV, we explored the generalizability of reliable report generation by adopting different LLMs. From the results, several noteworthy observations can be drawn:

- Our proposed method consistently improves the quality of generated radiology reports when combined with different LLMs, except for PMC-LLaMA. This finding demonstrates the effectiveness of our approach.

- Domain-specific LLMs, such as PMC-LLaMA and MedAlpaca, do not yield competitive enough results on reliable report generation. Despite being specifically trained on medical corpus, their performance falls short. Even with the enhancement provided by reliable report generation, MedAlpaca only achieves an F1 score of 0.505. Moreover, PMC-LLaMA does not perform well with reliable report generation, leading to significant performance degradation. We hypothesize that their inferior instruction following capability contributes to these results. Previous studies have emphasized the importance of properly controlling domain-specific fine-tuning to ensure conversational capability in LLMs [55]. Unlike MedAlpaca, PMC-LLaMA fully fine-tuned its model weights during training, which could result in worse performance degradation compared to parameter-efficiently tuned MedAlpaca. We believe this is the reason why reliable report generation shows inferior performance when combined with PMC-LLaMA.

- A larger model size does not always guarantee better performance. For example, LLaMA (7B) and LLaMA (13B) exhibit inferior performance compared to Chat-
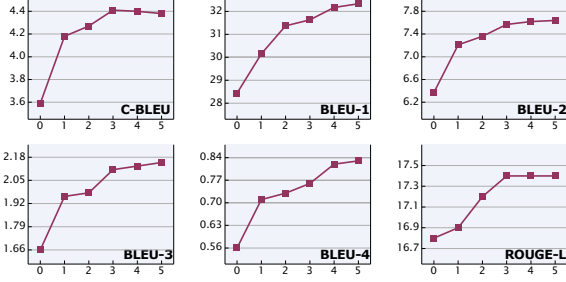
Fig. 7. Ablation study of hierarchical in-context learning w.r.t. the number of retrieved reports. A range of values, varying from 0 to 5, was chosen for k, and its influence is evaluated on several NLG metrics.
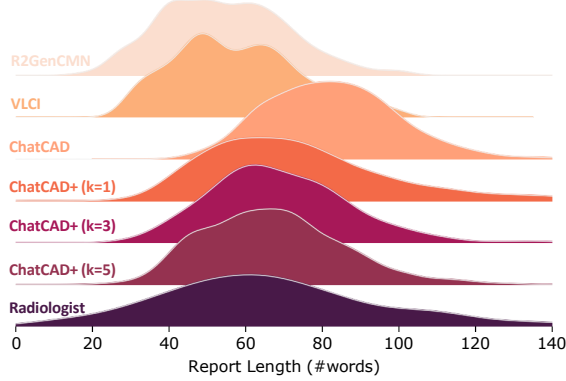
Fig. 8. Distributions of report length with respect to the changing of k.

GLM2 (6B) and other 7B-scale LLMs. Thus, blindly pursuing model size may not be the optimal solution.

- While ChatGPT unsurprisingly achieves the best overall performance, other LLMs like LLaMA2 (13B) and Mistral (7B) still demonstrate comparable performance. In particular, Mistral (7B) achieves the second-best F1 score among all the LLMs evaluated while maintaining competent NLG metrics. This highlights the potential application of consumer-level LLMs in clinical workflows.

TABLE V
PERFORMANCE COMPARISON BETWEEN CHATCAD+ AND GENERALIST MODELS IN THE MEDICAL IMAGING DOMAIN. EXCESSIVELY LOW RESULTS ARE REPRESENTED BY SLASH.

Method	C-BLEU	ROUGE-L	METEOR	PRE	REC	F1
PMC-VQA [56]	/	5.978	2.685	0.221	0.109	0.115
RadFM [57]	/	7.366	4.756	0.487	0.095	0.156
ChatCAD+ (ours)	**4.409**	**17.344**	**24.337**	**0.531**	**0.615**	**0.564**

3) Ablation study of hierarchical in-context learning: We investigate the impact of the retrieved top-k reports on reliable report generation in Fig. 7. As k increases, the overall performance shows an upward trend. Moreover, in most cases, the largest improvement in performance occurs when k changes from 0 to 1. This indicates that regardless of the quantity, as long as templates are provided as references, the quality of generated reports can be significantly enhanced. Meanwhile, the performance improvement tends to saturate around k=3, and further increasing k does not result in more significant improvement, which validates the rationality of the adopted

default setting.

The distribution of report length is also demonstrated in Fig. 8 by varying the value of k. Report length serves as an important criterion [7] to measure the similarity to ground truth radiology reports. A straightforward observation is that the distribution of ChatCAD exhibits a significant shift from that of the radiologist. In contrast, ChatCAD+ shows a more fitted curve irrespective of the value of k.

4) Qualitative analysis: In Fig. 9, each row showcases a chest X-ray along with its corresponding ground-truth annotation. To enhance clarity, we have color-coded the presence of certain findings in the generated reports. This visual representation allows for a clear illustration of the differences among the methods. For example, in the first chest X-ray image, it demonstrates the presence of four abnormalities, including cardiomegaly. However, both R2GenCMN and VLCI methods fail to acknowledge the existence of cardiomegaly in their generated reports. In the second row, the other two methods struggle to detect any abnormalities from the chest X-ray, while our method successfully identifies the possibility of cardiomegaly. The same trend can be observed in the third row. By including this qualitative analysis, we aim to highlight the advantages of our ChatCAD+ method and provide a comprehensive evaluation of the template retrieval system (hierarchical in-context learning). This comparison demonstrates the improved report quality achieved by our proposed method.

5) Comparison with generalists: While ChatCAD+ aims to integrate multiple specialist AI models incrementally, some recent studies [51], [57] have investigated generalist AI models for medical imaging. Hence, we quantitatively compare ChatCAD+ to generalists to investigate its superiority.

Existing generalist models typically follow the architecture of DeepMind's Flamingo [21] and Salesforce's BLIP-2 [22]. For instance, RadFM [57] and PMC-VQA [56] are trained based on Flamingo and BLIP-2, respectively. We compared ChatCAD+ with PMC-VQA and Flamingo using the same test set, and conducted zero-shot inference, as shown in Table V. Overall, our method demonstrates significant advantages over the generalist models. It is worth noting that these generalist models exhibit extremely low values on C-BLEU, indicating that generalist models can only generate very short responses and lack the ability to provide comprehensive diagnostic information like ours. Generally speaking, a generalist model implies that a single image encoder needs to interpret semantic information for all possible input images. This is challenging to achieve in the medical imaging domain due to distinct morphological and clinical semantic differences among different medical imaging modalities, which require a deep understanding of professional medical knowledge. Based on these reasons, we believe that it is still not the ChatGPT time for medical imaging analysis, and specialist models will long been a preferable solution. Such preference of specialist model is also mentioned in a related survey [58].

It is also important to note that a major advantage of ChatCAD+ lies in their capability for continual learning. By incorporating the LLM, we can combine decisions from multiple CAD models, enabling fine-tuning of each CAD model individually and also their incremental ensembling. This

Chest X-ray and Ground-Truth | **R2GenCMN** | **VLCI** | **ChatCAD+**

Row 1 — Ground-truth labels: Cardiomegaly, Edema, Atelectasis, Pleural Effusion

R2GenCMN: lung volumes are low. the mediastinal contours are grossly unremarkable . there is mild pulmonary edema . small bilateral pleural effusions are noted . patchy opacities in the lung bases likely reflect areas of atelectasis . no pneumothorax is identified . there are no acute osseous abnormalities .

VLCI: ap and lateral views of the chest were obtained . the heart size is normal . the mediastinal contour is unremarkable . there are bilateral pleural effusions with overlying atelectasis . there is mild pulmonary edema . there is no pneumothorax.

ChatCAD+: Based on the information provided, the chest x-ray shows that the patient has low lung volumes and a moderately enlarged heart. The mediastinal contours appear normal, but there is mild pulmonary edema and small bilateral pleural effusions. Additionally, there are patchy opacities in the lung bases, which may indicate areas of atelectasis. No pneumothorax or acute osseous abnormalities are identified.

Row 2 — Ground-truth label: Cardiomegaly

R2GenCMN: ap upright and lateral views of the chest provided . lung volumes are low limiting assessment. there is no focal consolidation effusion or pneumothorax. the mediastinal silhouette is normal . imaged osseous structures are intact. no free air below the right hemidiaphragm is seen .

VLCI: pa and lateral views of the chest. there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . osseous structures are intact . no free air below the right hemidiaphragm is seen.

ChatCAD+: The chest x-ray reveals low lung volumes, but there is no evidence of consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette appears normal, and the imaged osseous structures are intact. No free air is seen below the right hemidiaphragm. There is a possibility of cardiomegaly.

Row 3 — Ground-truth labels: Cardiomegaly, Edema

R2GenCMN: ap portable upright view of the chest. a left chest wall port-a-cath is seen with its tip extending into the region of the right atrium . lung volumes are low limiting assessment . there is diffuse pulmonary edema with probable small bilateral pleural effusions . the heart size cannot be assessed . mediastinal contour is grossly unremarkable . bony structures are intact .

VLCI: tube is seen below the level of the diaphragm. there are low lung volumes which the bronchovascular markings . there is bibasilar atelectasis . no definite focal consolidation is seen . there is no large pleural effusion or pneumothorax . the cardiac silhouette is not enlarged . the aorta is calcified

ChatCAD+: The chest x-ray shows a left chest wall port-a-cath with its tip extending into the region of the right atrium. Lung volumes are low, which limits assessment. Diffuse pulmonary edema is present. The size of the heart remains moderately enlarged. Bony structures are intact. There is no definite focal pulmonary abnormality, and no evidence of consolidation or atelectasis. However, there is a likelihood of edema and pleural effusion. Further evaluation may be necessary to determine the underlying cause.

Fig. 9. Qualitative analysis on different report generation methods. chest X-rays and corresponding ground-truth annotations are provided. The presence of a certain finding in a report is marked in a certain color for a clear illustration.

TABLE VI
VALIDATION OF RELIABLE INTERACTION FROM PERSPECTIVES OF BOTH MULTI-CHOICE SELECTION AND REASON EXPLANATION.

Method	Multi-choice		Reason Explanation	
	ACC	F1	C-BLEU	ROUGE-L
ChatGLM2 [45]	0.435	0.427	2.430	11.396
ChatGLM2+Ours	**0.465**	**0.461**	**7.930**	**18.957**
ChatGLM3 [45]	0.458	0.455	3.270	14.915
ChatGLM3+Ours	**0.472**	**0.505**	**7.974**	**19.302**

flexibility allows us to adapt and integrate new models rapidly in response to emerging situations without the need of huge computation resources.

D. Evaluation of clinical question answering

To evaluate the efficacy of clinical question answering, we adopt a comprehensive approach that combines quantitative analysis (through testing on the CMExam dataset) and qualitative analysis (by posing potential questions in real-world scenarios). This enables a thorough examination and provides insight into how the system functions in practical context.

In Table VI, we choose ChatGLM2 and ChatGLM3 as the LLM and present the performance of multi-choice selection and reason explanation with and without our proposed knowledge retrieval method on a subset of CMExam dataset (referring to Sec. IV-A). It is important to note that the results of multi-choice selection are directly generated by the LLM via prompt design, without intervention of any text encoders. We observe that our proposed method can consistently improve the performance of question answering for both ChatGLM2 and ChatGLM3. Interestingly, we find that the generation of reason explanations is significantly enhanced compared to multi-choice selection. For ChatGLM2, the ROUGE-L score is improved from 11.396 to 18.957, and for ChatGLM3, it is improved from 14.925 to 19.302. However, the improvement in multi-choice selection is not pronounced. We hypothesize that this discrepancy arises because the LLM may occasionally select correct answer based on unrigorous or even incorrect reasoning. In such cases, the LLM's reliability is compromised. However, with our proposed method, the reasoning behind the correct answer becomes more convincing, enabling reliable interaction for patients in need.

Considering qualitative analysis, the effectiveness of our proposed LLM-based knowledge retrieval method is compared with a baseline method that does not rely on a LLM. Such methods tend to follow the paradigm of LangChain [59], which involves dividing the text into paragraphs and then utilizing sentence transformers to compare the similarity between the user query and all paragraphs. This kind of knowledge retrieval method can hardly handle the professional medical knowledge database. We select the implementation in [59] as the baseline and compare it with our proposed method in Table VII and Table VIII, where "K" and "A" indicate "retrieved knowledge" and "answer from the LLM", respectively. Note that ChatGPT is the default LLM in this qualitative analysis.

In accordance with the structure of the Merck Manual, each medical topic is divided into five main sections: "abstract", "symptoms and signs", "diagnosis", "treatment" and "progno-

Q1 Is pleural effusion serious?
K1 (Baseline) Under normal circumstances, a thin layer of pleural fluid (10-20mL) covers the surfaces of the visceral and parietal pleura, with a composition similar to plasma but with lower protein content (<1.5g/dL [<15 g/L]). Pleural effusions are usually classified into exudates, transudates, and fluid accumulation based on laboratory characteristics of the fluid (see criteria for diagnosing exudative pleural effusions in the table).Regardless of unilateral or bilateral involvement, leaky effusions usually require little investigation and can be treated, while exudative effusions require identification of the underlying cause. Various factors can cause pleural effusions (see the table for causes of pleural effusions). Leaky pleural effusions are caused by an increase in systemic circulatory hydrostatic pressure and a decrease in plasma colloid osmotic pressure.
K1 (Ours) Some pleural effusions are asymptomatic and are discovered incidentally during physical examination or on chest x-ray. Many cause dyspnea, pleuritic chest pain, or both. Pleuritic chest pain, a vague discomfort or sharp pain that worsens during inspiration, indicates inflammation of the parietal pleura. Pain is usually felt over the inflamed site, but referred pain is possible. The posterior and peripheral portions of the diaphragmatic pleura are supplied by the lower 6 intercostal nerves, and irritation there may cause pain in the lower chest wall or abdomen that may simulate intra-abdominal disease. Irritation of the central portion of the diaphragmatic pleura, innervated by the phrenic nerves, causes pain referred to the neck and shoulder. Physical examination reveals absent tactile fremitus, dullness to percussion, and decreased breath sounds on the side of the effusion. These findings can also be caused by pleural thickening.   ... Pericardial rub is best heard over the left border of the sternum in the 3rd and 4th intercostal spaces, is characteristically a to-and-fro sound synchronous with the heartbeat, and is not influenced significantly by respiration. Sensitivity and specificity of the physical examination for detecting effusion are probably low.
A1 (ChatGPT Only) Pleural effusion can be serious depending on the underlying cause and the amount of fluid accumulated in the pleural space. It can be a symptom of various medical conditions such as congestive heart failure, pneumonia, cancer, or kidney disease. If left untreated, pleural effusion can lead to breathing difficulties and complications. It's important to consult with a healthcare provider for proper diagnosis and treatment.
A1 (Ours) The severity of pleural effusion needs to be determined based on the patient's specific situation. Generally, a large amount of effusion can cause symptoms such as dyspnea and chest pain, but some patients may not have obvious symptoms. Physical examination can reveal reduced breath sounds on the effusion side, dullness to percussion, and possibly pleural friction rub. However, the sensitivity and specificity of physical examination for pleural effusion are both low, so clinical symptoms and imaging examinations need to be combined to determine the severity of the effusion and treatment plan. Common causes of pleural effusion include heart failure, pneumonia, and tuberculosis
Q2 What should I do after getting lung atelectasis?
K2 (Baseline) do not recommend oral or intravenous corticosteroids because they have not shown efficacy in metal fume fever. Patients with pre-existing lung disease (such as asthma or chronic obstructive pulmonary disease) rarely require treatment and hospitalization due to worsening of pre-existing conditions. If severe cases of either syndrome cause respiratory distress, treatment methods used for ARDS (such as mechanical ventilation and positive end-expiratory pressure [PEEP]) may also be used. ", "Atelectasis": "Abs": "Atelectasis is collapse of lung tissue with loss of volume. Patients may have dyspnea or respiratory failure if atelectasis is extensive. They may also develop pneumonia. Atelectasis is usually asymptomatic, but hypoxemia and pleuritic chest pain may be present in certain cases. Diagnosis is by chest x-ray. Treatment includes maintaining coughing and deep breathing and treating the cause."
K2 (Ours) Maximizing cough and deep breathing If obstruction by tumor or foreign body is suspected, bronchoscopy Evidence for the efficacy of most treatments for atelectasis is weak or absent. Nonetheless, commonly recommended measures include chest physiotherapy to help maintain ventilation and clearance of secretions, and encouragement of lung expansion techniques such as directed cough, deep breathing exercises, and use of an incentive spirometer. In ambulatory patients, exercise (eg, walking) is a desirable way to promote deep breathing. For patients who are not intubated and do not have excessive secretions, continuous positive airway pressure may help. For patients who are intubated and mechanically ventilated, positive end-expiratory pressure and/or higher tidal volume ventilation may help.   ... Most importantly, the cause of atelectasis (eg, mucous plug, foreign body, tumor, mass, pulmonary effusion) should be treated. For persistent mucous plugging, nebulized dornase alfa and sometimes bronchodilators are tried. N-Acetylcysteine is usually avoided because it can cause bronchoconstriction. If other measures are ineffective or if a cause of obstruction other than mucous plugging is suspected, bronchoscopy should be done.
A2 (ChatGPT Only) I'm not a doctor, but I can provide you with general information on lung atelectasis. It's important to follow your healthcare provider's advice for your specific situation. In general, treatment for lung atelectasis may involve addressing the underlying cause, such as removing any blockages in the airways, treating infections, or providing oxygen therapy. Your healthcare provider may also recommend techniques to help re-expand the collapsed lung tissue, such as deep breathing exercises, chest physiotherapy, or using a device called a spirometer. Make sure to follow up with your healthcare provider for proper diagnosis and treatment.
A2 (Ours) The treatment methods for pulmonary atelectasis include maintaining cough and deep breathing, chest physical therapy, exercise, continuous positive airway pressure, etc. However, the evidence for the effectiveness of these treatments is not sufficient, so the treatment plan needs to be developed according to the specific situation of the patient. If there is suspicion of tumor or foreign body obstruction, bronchoscopy examination is needed. Opioid drugs can be used to relieve severe pleural pain, but cough suppressant drugs should be avoided. For the causes of pulmonary atelectasis such as mucus plugs, foreign bodies, tumors, masses, pleural effusions, etc., targeted treatments should be used.

TABLE VII
COMPARISON BETWEEN KNOWLEDGE RETRIEVAL METHODS. CHATGPT IS SELECTED AS THE DEFAULT LLM. COMPLETELY UNRELATED KNOWLEDGE IS MARKED IN RED, PARTIALLY RELATED KNOWLEDGE IS MARKED IN PURPLE. RELATED AND IMPORTANT KNOWLEDGE IS IN GREEN.

sis" as shown in Fig. 4. Therefore, to assess the effectiveness of knowledge retrieval, all the questions presented in Table VII and Table VIII are implicitly associated with certain sections. If the retrieved knowledge comes from the corresponding section, we label it as "related". If the knowledge is retrieved from a different section but still belongs to the same medical topic, we label it as "partially related". Otherwise, if the knowledge comes from a different medical topic altogether, we consider it "unrelated". For example, in the case of Q1 ("Is pleural effusion serious?"), the target section is "symptoms and signs". Unfortunately, the baseline method retrieves information from the "abstract" section, which is only "partially related" and cannot directly provide a reliable answer. In the case of Q3, we ask for the treatment of periodontitis. Our proposed method successfully retrieves the "treatment" section of periodontitis, while the baseline method mistakenly retrieves information about palmoplantar keratodermas. As a result, we labeled it as "unrelated". We hypothesize the reason behind this mistake is that palmoplantar keratodermas can lead to secondary infections of periodontitis. These results clearly

illustrate the advancement of our method. On the contrary, [59] struggles to identify relevant medical topics due to limited vocabulary size and understanding. This issue can be worse if the question of the user does not explicitly point to any medical topics or involve multiple medical entities.

We also demonstrate the answer of ChatGPT without additional knowledge. Results reveal that ChatGPT's responses are less useful and accurate. Specifically, for the first question, ChatGPT fails to elaborate on the fact that the severity of pleural effusion depends on the specific circumstances, instead merely stating that untreated pleural effusion can lead to complications. This omission prevents a thorough understanding of the condition. Furthermore, in Q4, ChatGPT provides an extensive list of possible tests for the patient, which can potentially cause unnecessary anxiety and expenses. In contrast, our reliable interaction highlights that an MRI is not imperative for initial assessment and should only be considered if knee pain does not improve after several weeks of conservative treatment. This clarification offers valuable guidance and prevents unwarranted medical interventions.

Q3 How to treat periodontitis? Is taking medication enough?
K3 (Baseline) Esophageal cancer can occur at a young age. Unna-Thost disease and Vorner disease: autosomal dominant inheritance. Papillon-Lefèvre syndrome: autosomal recessive inheritance, onset within 6 months after birth. Severe periodontitis can lead to tooth loss. Vohwinkel syndrome: autosomal dominant inheritance, patients may experience progressive toe and finger amputation and high-frequency hearing loss. Diffuse non-epidermolytic palmoplantar keratoderma: this autosomal dominant inherited form develops in infancy, causing a sharply defined symmetric keratoderma affecting the entire palm and sole. Treatment: Symptomatic treatment includes the use of emollients, keratolytic agents, and physical methods for removal.
K3 (Ours) Treatment of risk factors Scaling and root planing Sometimes oral antibiotics, antibiotic packs, or both Surgery or extraction Treatment of modifiable risk factors such as poor oral hygiene, diabetes, and smoking improves outcomes. For all forms of periodontitis, the first phase of treatment consists of thorough scaling (professional cleaning with hand or ultrasonic instruments) and root planing (removal of diseased or toxin-affected cementum and dentin followed by smoothing of the root) to remove plaque and calculus deposits. Thorough home oral hygiene is necessary and includes careful brushing and flossing to help clean. It may include chlorhexidine swabs or rinses. A therapist should help teach the patient how to do these procedures. The patient is re-evaluated after 3 weeks. If pockets are no deeper than 4 mm at this point, the only treatment needed is regular cleanings. Sometimes a flap of gum tissue is made to allow access for scaling and planing of deeper parts of the root. If deeper pockets persist, systemic antibiotics can be used. A common regimen is amoxicillin 500 mg orally 3 times a day for 10 days. In addition, a gel containing doxycycline or microspheres of minocycline can be placed into isolated recalcitrant pockets. These drugs are resorbed in 2 weeks. Another approach is to surgically eliminate the pocket and recontour the bone (pocket reduction/elimination surgery) so that the patient can clean the depth of the normal crevice (sulcus) between the tooth and gingiva. In certain patients, regenerative surgery and bone grafting are done to encourage alveolar bone growth. Splinting of loose teeth and selective reshaping of tooth surfaces to eliminate traumatic occlusion may be necessary. Extractions are often necessary in advanced disease. Contributing systemic factors should be controlled before initiating periodontal therapy. Ninety percent of patients with necrotizing ulcerative periodontitis due to HIV (HIV-associated periodontitis) respond to combined treatment with scaling and planing, irrigation of the sulcus with povidone-iodine (which the dentist applies with a syringe), regular use of chlorhexidine mouth rinses, and systemic antibiotics, usually metronidazole 250 mg orally 3 times a day for 14 days. Localized aggressive periodontitis requires periodontal surgery plus oral antibiotics (eg, amoxicillin 500 mg 4 times a day or metronidazole 250 mg 3 times a day for 14 days).
A3 (ChatGPT Only) Treating periodontitis usually involves a combination of professional dental treatments. Medication alone may not be enough to fully treat periodontitis. Professional dental treatments may include scaling and root planing to remove plaque and tartar from the teeth and roots, as well as possibly surgical procedures to correct any damage to the gums and bone. It's important to follow your dentist's recommendations closely and attend regular dental check-ups to monitor your progress and make any necessary adjustments to your treatment plan.
A3 (Ours) The treatment of periodontitis requires a comprehensive approach that considers risk factors, scaling and root planing, oral antibiotics, and local medications. For all cases of periodontitis, the first stage of treatment involves thorough cleaning, scaling and root planing, and reinforcement of daily oral hygiene. If deep periodontal pockets are present, systemic antibiotics may be used, with amoxicillin 500mg orally, three times a day for 10 days being a commonly used method. The second stage of treatment involves surgical removal of periodontal pockets and bone reshaping to allow the patient to self-clean the base of the gingival sulcus. Localized aggressive periodontitis requires periodontal surgery in combination with oral antibiotics. At the same time, systemic factors such as poor oral hygiene, diabetes, and smoking must be controlled to improve the prognosis. Therefore, drug therapy alone is not sufficient and comprehensive treatment is necessary.
Q4 What tests should be done for knee pain?
K4 (Baseline) Patellar tendon tear: the patella is palpable and displaced upwards (high-riding patella). Transverse patellar fracture: a gap between the two pieces of the patella may be palpable. However, due to significant local swelling, these symptoms may be mistaken for knee ligament injury with joint effusion. If a patient presents with knee swelling and pain after injury, the clinician will ask the patient to sit down and attempt to extend the injured leg to test knee extension actively or lie flat and elevate the injured leg. After knee swelling following trauma, remember to test the knee's ability for independent extension. Routine knee joint X-ray examination should be performed. Knee X-rays can show a high or low position of the patella. X-rays can reveal dislocation or patellar fractures, but they may also appear normal.
K4 (Ours) Not every patient requires x-rays. However, anteroposterior, lateral, and oblique x-rays are often taken to exclude fractures. The Ottawa knee rules are used to limit x-rays to patients more likely to have a fracture that requires specific treatments. X-rays should be taken only if one of the following is present: Age > 55, Isolated tenderness of the patella (with no other bone tenderness of the knee), Tenderness of fibular head, Inability to flex the knee 90°, Inability to bear weight both immediately and in the emergency department for 4 steps (with or without limping). MRI is usually not needed at the initial evaluation. A reasonable approach is to do MRI if symptoms do not resolve after a few weeks of conservative management. However, MRI is often done when severe injury or significant intra-articular injury is suspected or cannot be otherwise excluded. Other tests may be done to check for related injuries: Arteriography or CT angiography to check for suspected arterial injuries Electromyography and/or nerve conduction studies, which are rarely done immediately but more typically done when nerve symptoms persist weeks to months after the injury.
A4 (ChatGPT Only) When diagnosing knee pain, there are several tests that healthcare professionals may recommend to determine the underlying cause. These tests may include: (1) Physical examination: A healthcare provider may assess your knee by examining its range of motion, stability, and tenderness. (2) X-rays: X-rays can help identify issues such as fractures, arthritis, or bone abnormalities in the knee. (3) MRI (Magnetic Resonance Imaging): MRI scans provide detailed images of soft tissues like ligaments, tendons, and cartilage, which can help diagnose conditions such as ligament tears or cartilage damage. (4) CT scan (Computed Tomography): CT scans can provide detailed images of the bones and joints in the knee, helping to diagnose fractures or bone abnormalities. (5) Ultrasound: Ultrasound imaging can help evaluate soft tissue structures like tendons and ligaments for injuries or inflammation. (6) Arthroscopy: In some cases, a minimally invasive procedure called arthroscopy may be recommended to directly visualize and diagnose issues within the knee joint. It's important to consult with a healthcare professional to determine which tests are necessary based on your specific symptoms and medical history.
A4 (Ours) For knee pain evaluation, MRI is usually not needed for initial assessment. It is reasonable to proceed with conservative treatment for several weeks and perform MRI if symptoms do not improve. However, in cases of severe trauma or suspected intra-articular injury that cannot be ruled out by other means, MRI should be performed. Other tests may be performed to investigate related injuries, such as arterial angiography or CT angiography for suspected arterial injury. Electromyography and/or nerve conduction studies are rarely performed directly, but are used for patients with typical neurological symptoms that persist for several weeks to months after injury. Therefore, which tests are needed depends on the specific situation of the patient.

TABLE VIII

COMPARISON BETWEEN KNOWLEDGE RETRIEVAL METHODS. CHATGPT IS SELECTED AS THE DEFAULT LLM. COMPLETELY UNRELATED KNOWLEDGE IS MARKED IN RED, PARTIALLY RELATED KNOWLEDGE IS MARKED IN PURPLE. RELATED AND IMPORTANT KNOWLEDGE IS IN GREEN.

We believe that this comparison could further highlight the effectiveness of our proposed reliable interaction.

V. CONCLUSION

In this work, we have developed ChatCAD+, a universal and reliable interactive CAD system that can accept images of different domains as valid input. We expect that the proposed paradigm of universal CAD networks can motivate the advancement of cross-domain model literature in medical image domain. Furthermore, the hierarchical in-context learning enhances the report generation ability of ChatCAD+ in a training-free manner and let it act more like a radiologist. The LLM-based knowledge retrieval method enables the LLM to search by itself, and utilize the external medical knowledge database to provide convincing medical treatments. In this way, our ChatCAD+ has utilized the Merck Manuals database as its primary external knowledge database, which may be considered limited in its scope. However, we are confident in the versatility of our proposed LLM-based knowledge retrieval method, as it can be readily adapted to incorporate other databases. By doing so, we anticipate that the reliability of ChatCAD+ will be greatly enhanced. There are several limitations that must be acknowledged. First, while we have succeeded in enhancing the report generation capacity of our model to closely resemble that of a radiologist, it relies on additional report database. Second, the knowledge retrieval method we employed is relatively slow, as its efficacy is closely related to the response speed of the OpenAI API.

These limitations provide an opportunity for future research to address them and improve upon the current findings.

REFERENCES

[1] OpenAI. (2023) Chatgpt: Optimizing language models for dialogue. [Online]. Available: https://openai.com/blog/chatgpt/ 1, 8

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 1, 8

[3] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao *et al.*, "Large ai models in health informatics: Applications, challenges, and the future," *arXiv preprint arXiv:2303.11568*, 2023. 1

[4] V. W. Xue, P. Lei, and W. C. Cho, "The potential impact of chatgpt in clinical and translational medicine," *Clinical and Translational Medicine*, vol. 13, no. 3, 2023. 1

[5] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, J. Huang *et al.*, "A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images," *Nature communications*, vol. 13, no. 1, p. 2096, 2022. 1

[6] S. Wang, X. Ouyang, T. Liu, Q. Wang, and D. Shen, "Follow my eye: Using gaze to supervise computer-aided diagnosis," *IEEE Transactions on Medical Imaging*, 2022. 1

[7] Z. Chen, Y. Shen, Y. Song, and X. Wan, "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*, Aug. 2021. 1, 6, 8, 9

[8] W. Chen, Y. Liu, C. Wang, G. Li, J. Zhu, and L. Lin, "Visual-linguistic causal intervention for radiology report generation," *arXiv preprint arXiv:2303.09117*, 2023. 1, 8

[9] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023. 2, 4

[10] L. Milecki, V. Kalogeiton, S. Bodard, D. Anglicheau, J.-M. Correas, M.-O. Timsit, and M. Vakalopoulou, "Medimp: Medical images and prompts for renal transplant representation learning," *arXiv preprint arXiv:2303.12445*, 2023. 2

[11] C. Niu and G. Wang, "Ct multi-task learning with a large image-text (lit) model," *bioRxiv*, pp. 2023–04, 2023. 2

[12] D. Ueda, Y. Mitsuyama, H. Takita, D. Horiuchi, S. L. Walston, H. Tatekawa, and Y. Miki, "Diagnostic performance of chatgpt from patient history and imaging findings on the diagnosis please quizzes," *Radiology*, vol. 308, no. 1, p. e231040, 2023. 2

[13] J. Liu, C. Wang, and S. Liu, "Utility of chatgpt in clinical practice," *Journal of Medical Internet Research*, vol. 25, p. e48568, 2023. 2

[14] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You, "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge," *arXiv preprint arXiv:2303.14070*, 2023. 2

[15] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," *arXiv preprint arXiv:2304.01097*, 2023. 2

[16] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu, "Huatuo: Tuning llama model with chinese medical knowledge," *arXiv preprint arXiv:2304.06975*, 2023. 2

[17] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022. 2

[18] T. H. Kung, M. Cheatham, A. Medinilla, ChatGPT, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *medRxiv*, pp. 2022–12, 2022. 2

[19] R. S. Porter and J. L. Kaplan, *The Merck manual of diagnosis and therapy*. Merck Sharp & Dohme Corp., 2011. 2

[20] 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. 2

[21] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022. 2, 9

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023. 2, 9

[23] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. S. Khan, "Xraygpt: Chest radiographs summarization using medical vision-language models," *arXiv preprint arXiv:2306.07971*, 2023. 2

[24] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023. 2

[25] S. Chang, C. Baian, L. Fangyu, F. Zihao, S. Ehsan, and C. Nigel, "Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities," github.com/cambridgeltl/visual-med-alpaca/, 2023. 2

[26] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Large-scale domain-specific pretraining for biomedical vision-language processing," *arXiv preprint arXiv:2303.00915*, vol. 2, no. 3, p. 6, 2023. 3, 7

[27] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972. 4

[28] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975. 4

[29] S. Jain, A. Agrawal, A. Saporta, S. Truong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, C. Langlotz *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 5

[30] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022. 5

[31] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen *et al.*, "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *Scientific Data*, vol. 10, no. 1, p. 574, 2023. 6

[32] K. Panetta, R. Rajendran, A. Ramesh, S. P. Rao, and S. Agaian, "Tufts dental database: a multimodal panoramic x-ray dataset for benchmarking diagnostic systems," *IEEE journal of biomedical and health informatics*, vol. 26, no. 4, pp. 1650–1659, 2021. 6

[33] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 84–92, 2019. 6

[34] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database." *Academic Radiology*, vol. 19, pp. 236–248, 2012. 6

[35] J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," *arXiv preprint arXiv:2003.13865*, 2020. 6

[36] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "The diaretdb1 diabetic retinopathy database and evaluation protocol." in *BMVC*, vol. 1, no. 1. Citeseer, 2007, p. 10. 6

[37] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016. 6

[38] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data in brief*, vol. 30, 2020. 6

[39] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019. 6

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. 6

[41] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72. 6

[42] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81. 6

[43] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert," *arXiv preprint arXiv:2004.09167*, 2020. 6

[44] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu *et al.*, "Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 6

[45] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335. 6, 8, 10

[46] W. Ye, J. Yao, H. Xue, and Y. Li, "Weakly supervised lesion localization with probabilistic-cam pooling," 2020. 6

[47] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597. 6

[48] L. Mei, Y. Fang, Z. Cui, K. Deng, N. Wang, X. He, Y. Zhan, X. Zhou, M. Tonetti, and D. Shen, "Hc-net: Hybrid classification network for automatic periodontal disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 54–63. 6

[49] Z. Zhuang, L. Si, S. Wang, K. Xuan, X. Ouyang, Y. Zhan, Z. Xue, L. Zhang, D. Shen, W. Yao *et al.*, "Knee cartilage defect assessment by graph representation and surface convolution," *IEEE Transactions on Medical Imaging*, 2022. 6

[50] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, "Pmc-clip: Contrastive language-image pre-training using biomedical documents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 525–536. 7

[51] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Pmc-llama: Further finetuning llama on medical papers," *arXiv preprint arXiv:2304.14454*, 2023. 8, 9

[52] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023. 8

[53] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023. 8

[54] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. 8

[55] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=hTEGyKf0dZ 8

[56] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "Pmc-vqa: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023. 9

[57] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," *arXiv preprint arXiv:2308.02463*, 2023. 9

[58] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, p. 102996, 2023. 9

[59] langchain ChatGLM Contributors, "Chatglm application with local knowledge implementation," https://github.com/imClumsyPanda/langchain-ChatGLM, 2023. 10, 11