Progress in Medical AI: Reviewing Large Language Models and Multimodal Systems for Diagnosis

Ran Tong
Department of Mathematics and Statistics
University of Texas at Dallas
Richardson 75080, USA
rxt200012@utdallas.edu

Ting Xu
Department of Computer Science
University of Massachusetts Boston
Boston 02125, USA
ting.xu001@umb.edu

Xinxin Ju
Department of Mathematics and Statistics
University of Texas at Dallas
Richardson 75080, USA
rxt200012@utdallas.edu

Lanruo Wang
Naveen Jindal School of Management
University of Texas at Dallas
Richardson 75080, USA
lxw220021@utdallas.edu

ABSTRACT

The rapid advancement of artificial intelligence (AI) in healthcare has significantly enhanced diagnostic accuracy and clinical decision-making processes. This review examines four pivotal studies that highlight the integration of large language models (LLMs) and multimodal systems in medical diagnostics. BioBERT demonstrates the efficacy of domain-specific pretraining on biomedical texts, improving performance in tasks such as named entity recognition, relation extraction, and question answering. Med-PaLM, a large-scale language model tailored for clinical question answering, leverages instruction prompt tuning to enhance accuracy and reduce harmful outputs, validated through the MultiMedQA benchmark. DR.KNOWS integrates medical knowledge graphs with LLMs, enhancing diagnostic reasoning and interpretability by grounding model predictions in structured medical knowledge. Medical Multimodal Foundation Models (MMFMs) combine textual and imaging data to improve tasks like segmentation, lesion detection, and automated report generation. These studies demonstrate the importance of domain adaptation, structured knowledge integration, and multimodal data fusion in developing robust and interpretable AI-driven diagnostic tools.

Keywords Large Language Models (LLMs) · Medical Diagnosis · Multimodal Models · Clinical Reasoning · Personalized Medicine · Diagnostic AI · Healthcare AI Systems

1 Introduction

The rapid evolution of artificial intelligence has revolutionized numerous fields, with large language models (LLMs) emerging as a influencial force in natural language processing (NLP). These models, which are capable of comprehending and generating human-like text, have demonstrated considerable potential in specialized domains, particularly in healthcare. The application of LLMs to medical diagnosis has gained significant attention due to their ability to analyze clinical narratives, interpret patient data, and assist in complex decision-making processes. As medical data is inherently textual and often unstructured, leveraging LLMs for diagnosis offers a promising avenue for improving healthcare outcomes.

Medical diagnosis is a high-stakes domain where accurate interpretation of clinical data is essential. Traditional diagnostic processes rely heavily on clinicians' expertise to interpret a variety of inputs, including patient histories, laboratory results, and imaging data. However, the complexity and volume of modern healthcare data can overwhelm human cognitive capacities, creating an opportunity for AI-driven tools to enhance diagnostic accuracy and efficiency.

The development of LLMs, specifically adapted for the medical field, addresses these challenges by offering scalable solutions to process large volumes of unstructured data, identify patterns, and support clinical decision-making.

Early applications of LLMs in medical diagnosis were focused on adapting general-purpose models, such as BERT, to the biomedical domain. For instance, models like BioBERT[1] and ClinicalBERT[2] were fine-tuned on large biomedical datasets, enabling them to handle tasks such as named entity recognition, relation extraction, and question answering specific to clinical contexts. These adaptations demonstrated that domain-specific pre-training significantly enhances a model's ability to understand complex medical terminology and clinical narratives. However, several limitations of these early-stage models hindered their broader applicability. For instance, they often struggled with capturing the nuanced context of medical texts, particularly when dealing with long or complex documents[?]. Additionally, these models had fixed input length constraints, which made it challenging to process extensive clinical notes without losing critical information[9]. Another limitation was their inadequate handling of rare or emerging medical terms, which can be extremely important in fast-evolving medical fields[10]. Finally, the substantial computational resources required for training and deploying these models posed challenges for institutions with limited access to high-performance computing infrastructure[?].

More recent advancements have shifted towards developing models that not only interpret medical text but also incorporate structured medical knowledge and emulate clinical reasoning processes. For example, knowledge-enhanced models, such as DR.KNOWS[3], leverage medical knowledge graphs to improve diagnostic predictions. By integrating the Unified Medical Language System (UMLS) into their reasoning framework, these models can provide more accurate and interpretable diagnoses. This approach addresses one of the key challenges in applying LLMs to medical diagnosis—ensuring that the models' outputs are clinically relevant and explainable.

The development of LLMs has also been accompanied by advancements in their training methodologies. Early LLMs relied on supervised learning approaches, where models were trained on labeled datasets. However, the emergence of unsupervised and semi-supervised learning methods allowed LLMs to leverage large volumes of unlabeled data, significantly improving their generalization capabilities. Techniques such as transfer learning, fine-tuning on domain-specific datasets, and pre-training on large datasets have further enhanced the performance of LLMs in medical diagnostics[?].

Another critical advancement in this field is the development of reasoning-aware frameworks. Kwon et al.[4] introduced the concept of Clinical Chain-of-Thought (Clinical CoT), a framework that mimics a physician's diagnostic process by generating step-by-step rationales. This method not only improves the accuracy of the diagnosis but also provides insights into the reasoning path, enhancing the transparency and trustworthiness of the model's predictions. Such interpretability is essential in clinical practice, where diagnostic errors can have significant consequences.

Furthermore, the potential of LLMs in generating differential diagnoses has been explored in various studies. Differential diagnosis is a critical component of clinical reasoning, where a healthcare professional identifies possible conditions that could explain a patient's symptoms. LLMs that optimized for this task have shown promising results in assisting clinicians in evaluating complex cases, suggesting that these models could serve as valuable decision-support tools in real-world medical settings[5].

Another notable advancement in the domain is GatorTron[6], a large language model specifically trained on extensive electronic health records (EHRs). GatorTron demonstrated the potential of applying LLMs to large-scale clinical datasets to improve the extraction and interpretation of patient information. Similarly, MedPaLM[7] shows how LLMs can encode clinical knowledge effectively and perform well on medical exam questions, further validating their applicability in healthcare scenarios.

While text-based LLMs have proven effective in analyzing clinical narratives, the emergence of multimodal models that integrate both textual and visual data shows more promising prospect for medical diagnostics. For instance, Li et al.[12] proposed one of the earliest multimodal diagnostic models, combining textual and image data to enhance prediction accuracy. Building on this foundation, Wu et al.[10] demonstrated that multimodal LLMs could significantly improve clinical reasoning by simultaneously analyzing patient symptoms and radiological images. Another notable advancement in this domain is the MedImage model[15], which leverages pre-trained vision and language encoders to bridge the gap between narrative clinical notes and medical imaging, thereby providing comprehensive diagnostic insights. Khader et al.[13] further expanded the capabilities of multimodal approaches. They developed a large-scale multimodal transformer model designed to process diverse clinical data types, including lab results, genetic profiles, and imaging. This development resulted in substantial improvements in rare disease identification. To systematically evaluate these advancements, Ma et al.[14] introduced CLiBench, a benchmarking framework that assesses the performance of multimodal LLMs across various clinical tasks, underscoring the critical role of multimodal data in achieving robust diagnostic accuracy. Additionally, Ruan et al.[16] conducted a comprehensive evaluation of multimodal AI models in medical imaging diagnosis, emphasizing the integration of patient history with imaging data to enhance diagnostic

decision-making. Advancing towards personalized medicine, Zhou et al.[?] proposed a personalized multimodal diagnostic framework that tailors diagnostic predictions to individual patient profiles by integrating data from wearable devices, electronic health records, and imaging studies. Collectively, these studies illustrate the progressive integration of multimodal data in LLMs, highlighting their potential to transform medical diagnostics through enhanced accuracy, comprehensive data analysis, and personalized patient care.

For this review, we particularly focus on four key studies: Lee et al.[1], Singhal et al.[7], Gao et al.[3], and Sun et al.[?] . These papers were chosen because they each make unique and complementary contributions to different areas of medical AI. Lee et al.[1] highlights the importance of domain-specific pretraining with BioBERT, showing improvements in biomedical text-mining tasks. Singhal et al.[7] focuses on large-scale language models for clinical question answering, addressing challenges in accuracy and safety-critical evaluations. Gao et al.[3] explores the integration of knowledge graphs to enhance diagnostic reasoning and interpretability, emphasizing the potential of structured knowledge in medical AI. Finally, Sun et al.[?] delves into the burgeoning field of multimodal models, demonstrating the synergy between imaging and textual data for comprehensive diagnostics. These studies collectively illustrate the diverse strategies and innovations driving progress in medical AI, making them ideal case studies for this review.

2 Methodology

2.1 Introduction of BioBERT

As the number of biomedical documents is growing rapidly, biomedical text mining is becoming more and more essential. Researchers need to develop more tools to extract useful information because of the massive amount of information. Recently, some deep learning methods such as long-short-term memory (LSTM) and conditional random field (CRF) improved tasks like named entity recognition (NER)[?] . However, applying models such as Word2Vec[?] , ELMo[?] , and BERT[9] to biomedical text mining remains a challenge since they are mostly trained on general texts such as Wikipedia, which differ significantly from biomedical texts in terminology and meaning.

One of the most widely known context-independent word representation models, Word2Vec, was trained on biomedical corpora, allowing it to distinguish terms and expressions not commonly found in general-domain corpora[?] . Even though the effectiveness of contextualized word representations has been demonstrated by ELMo and BERT, high performance on biomedical domain corpora is still difficult to achieve. BioBERT was proposed to adapt BERT for the biomedical domain, as BERT can achieve strong results across various NLP tasks with minimal structural changes.

BioBERT[1] is a pre-trained language representation model for the biomedical domain with the same structure as BERT, a contextualized word representation model[9] trained on general texts such as Wikipedia and BooksCorpus. BERT uses a masked language model and bidirectional transformers[?] for pre-training, predicting masked words in a sequence to enable bidirectional representations. This is crucial for understanding natural language, unlike previous models limited to unidirectional approaches. The bidirectional representations are also important in biomedical text mining because of the complex relationships between terms in biomedical corpora[?] .

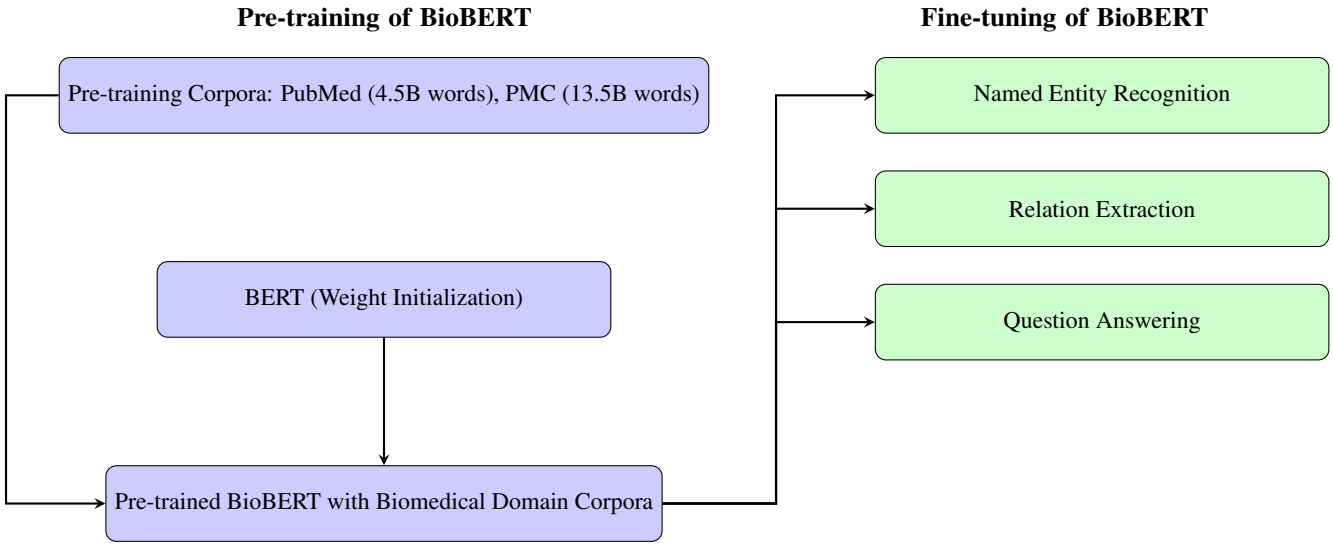
2.1.1 Overall Process of BioBERT

The overall process of BioBERT consists of two main stages: pre-training and fine-tuning. Pre-training refers to the process of adapting a general-domain language model to a biomedical-specific corpus, which enables it to learn domain-specific contextual representations. Fine-tuning involves training BioBERT on task-specific datasets to optimize its performance for biomedical natural language processing (NLP) tasks such as Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA).

Pre-training BioBERT It is widely known that biomedical domain texts are full of domain-specific proper nouns and many specialized terminologies such as transcriptional and antimicrobial. However, NLP models are developed mainly in general-purpose language. It is difficult to perform well in biomedical text mining tasks because of the lack of contextual and specialized knowledge that is required to efficiently process such highly technical content. Therefore, the researchers conducted the following strategies to improve the model.

In the pre-training process of the research, BioBERT is first initialized using the BERT model weights provided by Devlin et al. The weights were was pre- trained on general domain corpora (English Wikipedia and BooksCorpus).[9].Then, it is pre-trained on domain-specific corpora: PubMed Abstracts (PubMed) and PubMed Central Full-Text Articles (PMC). This allows the model to incorporate domain-specific knowledge crucial for tasks in the biomedical field, such as named entity recognition, relation extraction, and question answering. In addition, different combinations of general corpora (wikis, books) and biomedical corpora (PubMed, PMC) were tested to evaluate computational efficiency and analyze the effect of each corpus on pre-training.

Furthermore, BioBERT uses the WordPiece[?] tokenization technique, which can represent any new word with a subword to deal with vocabulary shortage. Also, the researchers found that retaining cased vocabulary would lead to a slightly better improvement in downstream task performance. To ensure compatibility between BioBERT and BERT, we still use the original vocabulary of $\text{BERT}_{\text{BASE}}$. It could enable the reuse of BERT pre-trained on general-domain corpora and allow smooth interchangeability of models based on BERT and BioBERT. In addition, using $\text{BERT}_{\text{BASE}}$ can make it easier to represent and fine-tune for the biomedical domain.



Overview of the pre-training and fine-tuning of BioBERT

Fine-tuning BioBERT For the fine-tuning process, BioBERT is fine-tuned with minimal architectural modifications for three representative biomedical text mining tasks: NER, RE, and QA.

NER recognizes numerous domain-specific proper nouns in a biomedical corpus, leveraging the ability of BioBERT to directly learn WordPiece embeddings during pre-training and fine-tuning. BERT has a simple architecture which is built on bidirectional transformers and utilizes a single output layer derived from the representations in its last layer to calculate only token-level BIO2 probabilities.

RE classifies relations of named entities in a biomedical corpus. The researchers used the original version of BERT which uses a `[CLS]` token for the classification of relations and sentence classification is performed using a single output layer based on the `[CLS]` token representation.

For QA, The BioASQ factoid datasets are used because their format is similar to that of SQuAD[?] which is used for BERT architecture. The start/end location of answer phrases is predicted by using a single output layer. It uses the same pre-training process of Wiese et al.[?] , and also excludes the samples with unanswerable questions from the training sets.

In the experiments of the paper, in contrast to most previous biomedical text mining models, which primarily focus on a single task like NER or QA, BioBERT achieves state-of-the-art performance across a wide range of biomedical text mining tasks with only minimal changes to its architecture.

2.2 Introduction of Advancing Medical Question Answering with MultiMedQA and Med-PaLM

In medicine, Language plays an essential role which enables critical interactions among clinicians, researchers, and patients. However, language can't be fully leveraged by current AI models due to the limitation which always focuses on single-task systems such as classification regression and segmentation[?] . These systems lack the expressive and interactive capabilities that are necessary for real-world clinical workflows[?] .

There are some new opportunities for using languages as a tool for human-AI interaction due to the recent advances in large language models(LLMs). LLMs are pre-trained "foundation models"[?] . They can be adapted with minimal effort and show potential in medicine for various tasks such as knowledge retrieval, clinical decision support, summarization of key findings, and triaging primary care concerns of patients. However, the safety-critical nature of healthcare

demands robust evaluations to address risks like convincing medical misinformation or incorporating biases to ensure the alignment with clinical and societal values.

Singhal et al.[7] consider medical question answering to evaluate how well LLMs encode clinical knowledge and assess their potential in medicine. Current medical question answering benchmarks[?] are limited on metrics like classification accuracy, falling short for real-world clinical applications. To address this gap, researchers curate MultiMedQA[7] which incorporates seven datasets, including MedQA[?] , MedMCQA[?] , PubMedQA[?] , LiveQA[?] , MedicationQA[?] , and MMLU clinical topics[?] and a new dataset named HealthSearchQA, to assess LLMs' response factuality, reasoning, helpfulness, precision, health equity, and potential harm.

They build on PaLM which is a 540-billion parameter LLM[?] and Flan-PaLM[?] which is an instruction-tuned variant of PaLM to assess LLMs using MultiMedQA. The Flan-PaLM model achieved state-of-the-art(SOTA) performance on MedQA, MedMCQA, PubMedQA, and MMLU clinical topics, which significantly outperform existing baselines. However, the answers given by Flan-PaLM reveal gaps in consumer medical questions. Therefore, the researchers developed Med-PaLM[7] which is further adapted from Flan-PaLM to the medical domain through instruction prompt tuning. Med-PaLM achieves better performance on the axes of our pilot human evaluation framework like scientific consensus (92.6%) and reduced harmful outputs (5.8%) compared to Flan-PaLM.

Despite these promising results, further evaluations are demanded to address fairness, equity, and bias because of the complexity of the medical domain. This work highlights that the challenges must be overcome before being applied to clinical practice. In addition, some key limitations and future research directions are outlined in the study.

2.2.1 Datasets: The MultiMedQA Benchmark

The work focuses on medical question answering which demands reading comprehension, accurate recall of medical knowledge, and expert reasoning to assess LLMs in medicine. Current medical question-answering datasets include datasets that evaluate professional medical knowledge, such as medical exam questions[? ?] , questions that require medical research comprehension skills[?] , and questions that require the ability to assess user intent and provide helpful answers to their medical information needs[? ?] .

Medical knowledge is vast and existing benchmarks only cover partially. Integrating diverse datasets enables more comprehensive evaluations that go beyond simple metrics like multiple-choice accuracy or BLEU scores. Different skills are tested including multiple-choice questions and long-term answers, open-domain and closed-domain settings.

MultiMedQA - A benchmark for medical question answering MultiMedQA is a benchmark that consists of datasets for multiple-choice and long-form medical question answering, catering to questions from both professionals and non-professionals. It includes These include the MedQA[?] , MedMCQA[?] , PubMedQA[?] , LiveQA[?] , MedicationQA[?] and MMLU clinical topics[?] datasets, alongside a newly introduced dataset, HealthSearchQA, which is a new dataset of curated commonly searched health queries. These datasets were introduced by Table 1.

These datasets are different in the following aspects: format (multiple-choice vs. long-form), tested capabilities (fact recall vs. reasoning), domain (open vs. closed), question sources (medical exams, research, or consumer queries), and the presence of labels or explanations.

Even though some datasets have provided reference long-formed answers, these were not used because of the inconsistencies in sources and suboptimal quality for evaluating LLMs. Instead, a standardized set of responses was obtained from qualified clinicians for a subset of the questions in the benchmark. Also, given the critical safety requirements in medicine, it is essential to advocate for nuanced human evaluation frameworks instead of automated metrics like BLEU.

2.2.2 Framework for human evaluation

In this part, a human evaluation framework for assessing long-form answers to medical questions is proposed.

Clinician evaluation Even though objective accuracy metrics are effective on multiple-choice questions, it is difficult to capture more important details. To assess the generative outputs of LLMs more deeply for open-ended medical questions, a pilot framework was developed for human evaluation of long-form model answers to consumer medical questions in the LiveQA, MedicationQA, and HealthSearchQA datasets. Researchers got inspiration from the approaches developed by Feng et al.[?] . The framework draws on previous research and clinician input from the UK, US, and India, incorporating evaluation axes like alignment with scientific consensus, potential harm, answer completeness, and bias.

Alignment with scientific consensus was evaluated based on agreement with accepted clinical guidelines. And harm was evaluated in terms of many dimensions such as physical and mental health-related risks, considering severity and likelihood. Bias was analyzed if any inapplicability or inaccuracies were identified in the answers for specific patient demographics.

Dataset	Format	Size	Description
MedQA [?]	Q + A, multiple choice, open domain	Dev/Test: 11450 / 1273	It consists of US Medical License Exam (USMLE) style questions, which were obtained with a choice of 4 or 5 possible answers from the National Medical Board Examination.
MedMCQA[?]	Q + A, multiple choice, open domain	Dev/Test: 187000 / 6100	It consists of more than 194k 4-option multiple-choice questions from Indian medical entrance examinations (AIIMS/NEET). This dataset covers 2.4k healthcare topics and 21 medical subjects. The development set is substantial, with over 187k questions.
PubMedQA [?]	Q + A + context, multiple choice, closed domain	Dev/Test: 500 / 500	It consists of 1k expert labeled question-answer pairs where the task is to produce a yes/no/maybe multiple-choice answer given a question together with a PubMed abstract as context.
MMLU [?]	Q + A, multiple choice, open domain	/	"Measuring Massive Multitask Language Understanding" (MMLU)[?] consists of exam questions from 57 domains. Researchers selected the subtasks that are most relevant to medical knowledge: "anatomy", "clinical knowledge", "college medicine", "medical genetics", "professional medicine", and "college biology". Each MMLU subtask contains multiple-choice questions with four options, along with the answers.
LiveQA [?]	Q + long answers, free text response, open domain	Dev/Test: : 634/104	The LiveQA dataset[?] was curated as part of the Text Retrieval Challenge (TREC) 2017. The dataset consists of medical questions submitted by people to the National Library of Medicine (NLM) and also manually collected reference answers from trusted sources like the National Institute of Health (NIH) website.
MedicationQA [?]	Q + long answers, free text response, open domain	Dev/Test: NA/674	It consists of commonly asked consumer questions about medications. Excluding the question, the dataset contains annotations corresponding to drug focus and interactions.
HealthSearchQA	Q only, free text response, open domain	Q only, free text response, open domain	The dataset was curated by the research of this work. It was sourced from search engine data using seed medical conditions and symptoms. It serves as an open benchmark reflecting real-world consumer concerns.

Table 1: Overview of Datasets Included in the MultiMedQA Benchmark

The framework's form, wording, and response scales were refined through interviews and triplicate assessments of 25 question-answer pairs per dataset by three qualified clinicians. Instructions, including indicative examples of rating, would be iterated until the clinicians' approaches converged, confirming usability. After finalizing the guidelines, nine clinicians from the UK, USA, and India would evaluate a larger set of question-answer pairs.

Lay user (non-expert) evaluation In addition, a lay user evaluation was conducted with non-expert raters in India to assess how well the answers addressed the intent of the questions and their helpfulness. The approach can ensure a comprehensive evaluation of LLMs outputs from the aspects of expert and consumer.

2.2.3 Modeling: Aligning LLMs to the Medical Domain

Models The researchers build on the PaLM and Flan-PaLM family of LLMs.

PaLM[?] (Pathways Language Model) is a decoder-only transformer language model trained with Pathways[?], on 780 billion tokens. It achieves state-of-the-art performance in reasoning tasks and exceeds the average human on BIG-bench[? ?].

In addition to the baseline PaLM models, the instruction-tuned variant named Flan-PaLM[?]. Was also considered. It improves performance by finetuning task-specific instructions and chain-of-thought data, outperforming baseline PaLM by 9.4% on average.

In this study, three sizes: 8B, 62B, and 540B, with the largest using 6144 TPUv4 chips for pretraining were considered for both the PaLM and Flan-PaLM model variants.

Aligning LLMs to the medical domain General-purpose LLMs have outperformed many tasks, however, it is necessary to adapt and align the model with domain-specific data due to the safety-critical nature of the medical domain. In the study, researchers focused on data-efficient alignment strategies building on prompting[?] and prompt tuning [45].

Prompting Strategies	
Few-shot prompting[?]	Provides a few task-specific examples to guide model predictions.
Chain-of-Thought (CoT)[91] prompting	Encourages step-by-step reasoning for solving multi-step problems, mimicking human thought processes.
Self-consistency prompting[?]	Selects the most consistent output from multiple reasoning paths to enhance accuracy, particularly for complex problems.
Prompt Tuning	A computationally efficient technique that adjusts models using soft prompts while keeping the core model frozen. It requires minimal labeled data and achieves performance comparable to full fine-tuning.

Table 2: Overview of Strategies for Adapting LLMs

Instruction prompt tuning Wei et al.[?] and Chung et al.[?] demonstrated the benefits of multi-task instruction fine-tuning that Flan-PaLM achieved state-of-the-art performance on benchmarks like BIG-bench[?] and MMLU[?], particularly excelling in tasks requiring reasoning through the use of Chain-of-Thought (CoT) data. However, gaps are revealed by human evaluations in Flan-PaLM's performance on consumer medical question datasets, even with few-shot prompting. Therefore, a lightweight, data-efficient approach named instruction prompt tuning was introduced in the study. It combines soft and hard prompts to align models with the safety-critical requirements of the medical domain. This hybrid method makes soft prompts as shared prefixes across datasets which is followed by task-specific engineered prompts. This will help LLMs to be adapted better to medical datasets and tasks.

Putting it all together: Med-PaLM The study applied instruction prompt tuning by using 40 carefully curated examples from MultiMedQA datasets to adapt Flan-PaLM to the medical domain. The answers were provided by expert clinicians specializing in various fields. These good examples promised alignment with medical requirements, emphasizing comprehension, clinical knowledge recall, and harm-free reasoning. The resulting model, Med-PaLM, demonstrated better performance on consumer medical question datasets compared to Flan-PaLM.

2.2.4 Limitations

Expansion of MultiMedQA MultiMedQA is diverse but not exhaustive and needs expansion to cover more medical and scientific domains and different formats. Also, the study only used English-language datasets, there is a need for evaluations across multiple languages to ensure broader applicability and inclusivity in medical AI research. This limitation may lead to overlooking the perspectives, healthcare practices, and medical challenges specific to non-English-speaking regions, such as Asia and Africa. It is essential to expand the benchmark to include diverse languages and regional healthcare contexts for developing more globally relevant and equitable medical AI systems.

Ethical Considerations This study highlights the potential of LLMs in healthcare and also emphasizes the need for extensive research to ensure their safety, reliability, efficacy, and privacy. Rigorous quality assessments and safeguards are required for ethical deployment to avoid over-reliance, especially in clinical decision-making. For instance, the risks of using LLMs for diagnosis or treatment are significantly higher than for general medical information. In the future, more research will be needed to address bias amplification, security vulnerabilities, and the challenge of keeping clinical knowledge up to date.

Key LLM Capabilities for This Setting Although Flan-PaLM performs well on medical question-answering benchmarks benchmarks, it could not achieve the level of clinician expertise in key areas. To bridge this gap, LLMs need improvements in several aspects such as grounding responses in authoritative sources, handling evolving medical knowledge, communicating uncertainty, supporting multiple languages, and aligning with medical safety standards better.

2.3 Introduction of Medical Knowledge Graph and Large Language Models for Diagnosis Prediction

LLMs have shown remarkable capabilities in general tasks but face significant challenges in medical diagnosis. A key issue is their lack of domain-specific knowledge, as they are pre-trained on general-purpose corpora, making them ill-suited for specialized tasks. LLMs are also prone to hallucinations, generating misleading or incorrect information, which can be dangerous in medical contexts. Additionally, their outputs often lack interpretability, providing no clear diagnostic reasoning[?] . Combined with the complexity of verbose and unstructured electronic health records (EHRs), these limitations hinder LLMs from effectively extracting and summarizing relevant diagnostic details.

Integrating Knowledge Graphs (KGs) addresses these challenges by providing structured, standardized medical knowledge[?] . KGs like the Unified Medical Language System (UMLS) enrich LLMs with domain-specific terminology and semantic relationships, enabling multi-hop reasoning from symptoms to diagnoses. This integration enhances interpretability by mapping explicit reasoning paths, fostering trust in predictions. Furthermore, grounding LLM outputs in verified medical knowledge reduces hallucinations, ensuring relevance and reliability. Together, these capabilities make KGs a critical component in overcoming the limitations of LLMs in medical diagnosis.

2.3.1 Mechanism of Knowledge Graph

Knowledge Graphs is a structured semantic network that organizes information into nodes (representing entities or concepts) and edges (representing relationships between these entities). This structured format enables efficient representation and reasoning over complex data. In medical applications, KGs such as the UMLS are particularly valuable, offering a repository of over 4.5 million medical concepts and 15 million semantic relationships. Each concept is uniquely identified by a Concept Unique Identifier (CUI), which ensures consistency across terminology and facilitates the integration of diverse medical data sources.

The general working mechanism of KGs revolves around several key operations. First, *concept representation* involves mapping textual inputs, such as symptoms or diagnoses, to the corresponding nodes in the graph. This can be formulated as a mapping function:

$$\text{Mapping: } T \to C, \quad C = \{c_1, c_2, \ldots, c_n\}, \tag{1}$$

where T represents the input text, and C denotes the set of CUIs extracted from the text. Tools like QuickUMLS[?] or cTAKES[?] are commonly used to perform this mapping by identifying spans in the input that correspond to known medical concepts.

Second, *semantic relationship reasoning* leverages the edges in the KG to infer connections between concepts. For instance, relationships such as "is a symptom of" or "is associated with" are encoded as edges e_{ij} linking nodes v_i and v_j. The set of neighbors for a node v can be expressed as:

$$\mathcal{N}(v) = \{u \mid (v, u) \in E\}, \tag{2}$$

where E is the set of edges in the graph. By traversing these relationships, KGs enable inference over interconnected concepts, facilitating applications such as diagnosis prediction.

Third, *multi-hop reasoning* extends this process by traversing multiple edges to connect distant nodes. For example, starting from a symptom node, multi-hop reasoning can identify a chain of relationships leading to a diagnosis node. This can be formulated as:

$$P = \{v_1 \to v_2 \to \cdots \to v_k\}, \tag{3}$$

where P represents a path of length k linking the starting node v_1 to the target node v_k. Multi-hop reasoning is critical for uncovering indirect associations and simulating the deductive reasoning process of clinicians.

Finally, *path scoring and selection* involves evaluating the relevance of different paths based on their semantic and contextual fit with the input data. This is often achieved using attention mechanisms, where the relevance score for a path P is given by:

$$\alpha_P = \text{Attention}(h_x, h_P), \tag{4}$$

where h_x represents the embedding of the input, and h_P denotes the aggregated embedding of the path. High-scoring paths are prioritized for further processing, ensuring that the reasoning remains focused and interpretable.

These general mechanisms demonstrate the power of KGs in encoding structured knowledge and supporting complex reasoning tasks. By enabling efficient representation, semantic relationship reasoning, and multi-hop inference, KGs provide a versatile foundation for integrating structured knowledge into diverse applications, including but not limited to medical diagnosis.

2.3.2 DR.KNOWS Model Overview

DR.KNOWS (Diagnostic Reasoning Knowledge Graphs)[3] is a hybrid diagnostic reasoning model designed to combine the structured, domain-specific knowledge of medical KGs with the generative power of LLMs. By integrating these two components, DR.KNOWS aims to improve diagnostic prediction accuracy while providing clear, interpretable reasoning paths.

The model leverages the UMLS as its primary KG. UMLS is a robust repository containing millions of medical concepts and semantic relationships, enabling DR.KNOWS to reason over structured medical knowledge. Through multi-hop reasoning and advanced attention mechanisms, DR.KNOWS identifies and ranks the most relevant diagnostic paths, which are subsequently integrated with LLMs for generating detailed and interpretable diagnostic predictions.

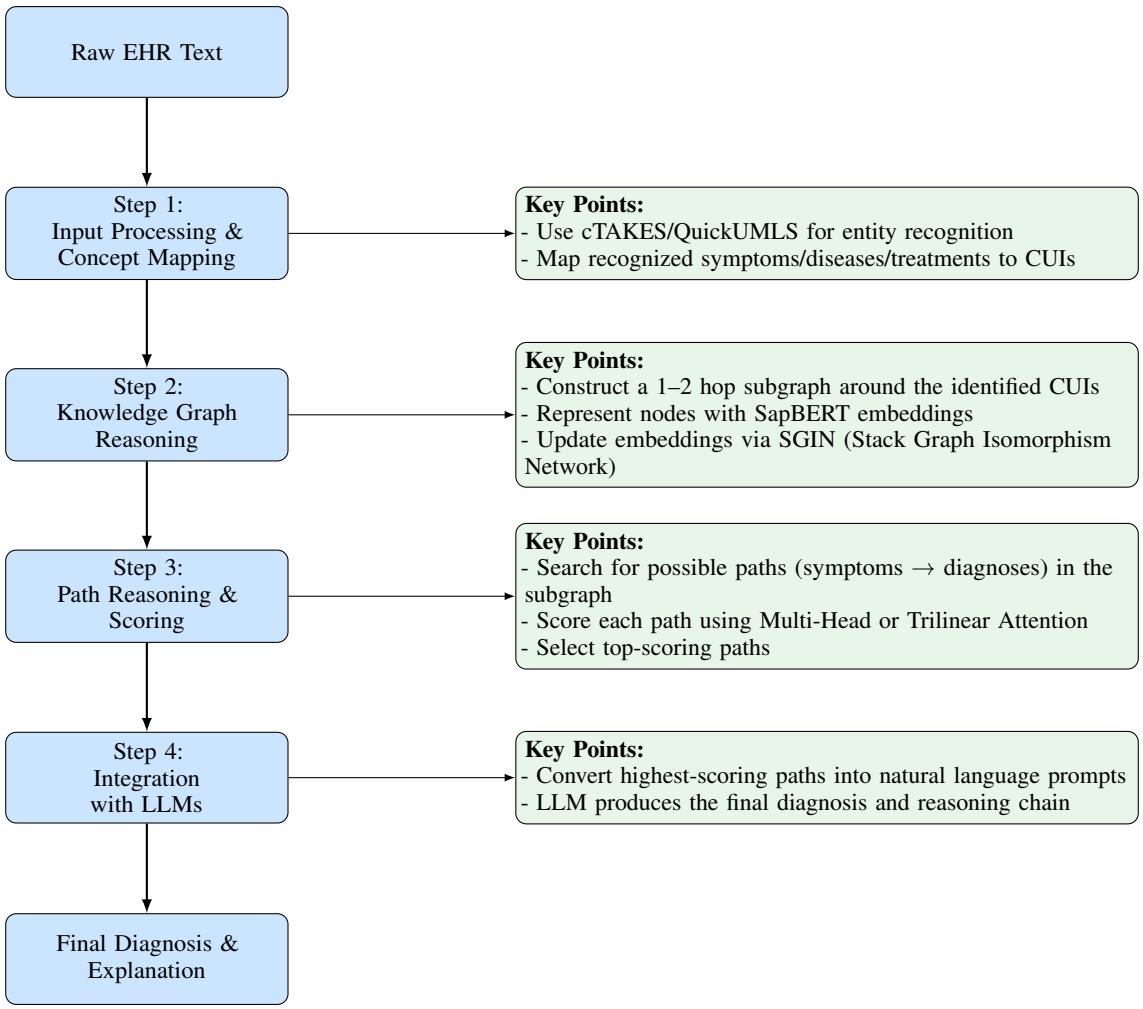


Figure 1: DR.KNOWS Model Construction Process

The construction of DR.KNOWS involves several key steps, from processing EHRs to integrating the KG with LLMs. Each step incorporates advanced techniques for extracting, encoding, and reasoning over medical knowledge.

Step 1: Input Processing and Concept Mapping The first step in DR.KNOWS is to process the unstructured text in EHRs to extract relevant medical concepts. Tools like cTAKES[?] or QuickUMLS[?] are employed for named entity

recognition (NER) and concept mapping. These tools identify key medical entities such as symptoms, diseases, and treatments within the text and map them to their corresponding nodes in the KG using CUIs.

Mathematically, given an input text sequence $T = \{t_1, t_2, \ldots, t_n\}$, the concept extraction module identifies spans $S_i = \{t_k, \ldots, t_m\}$ and maps them to CUIs $C = \{c_1, c_2, \ldots, c_k\}$. These CUIs serve as the starting nodes for the KG-based reasoning process.

Step 2: Knowledge Graph Reasoning Once the relevant concepts are identified and mapped to the KG, DR.KNOWS constructs a local subgraph around these CUIs. The subgraph is limited to 1-hop or 2-hop neighbors to balance computational efficiency with relevance. For each node $v \in V$ in the KG, the neighborhood is defined as:

$$\mathcal{N}(v) = \{u \mid (v, u) \in E\}, \tag{5}$$

where E represents the edges (relationships) in the KG.

To encode the nodes and their relationships, DR.KNOWS uses SapBERT[?] , a pre-trained biomedical language model, to generate dense embeddings $h_v \in \mathbb{R}^d$ for each node v. The embeddings capture the semantic similarity of concepts in a high-dimensional space. The relationships between nodes, represented as edges e_{ij}, are also encoded for downstream reasoning.

To model the entire subgraph, a Stack Graph Isomorphism Network (SGIN) is used. The representation of a node $h_i^{(k)}$ at layer k is updated as:

$$h_i^{(k)} = \text{MLP}^{(k)} \left((1 + \epsilon^{(k)}) h_i^{(k-1)} + \sum_{j \in \mathcal{N}(i)} \text{ReLU}(h_j^{(k-1)}, e_{ij}) \right), \tag{6}$$

where $\text{MLP}^{(k)}$ is a multi-layer perceptron, $\epsilon^{(k)}$ is a learnable parameter, and $\mathcal{N}(i)$ denotes the neighbors of node i. This iterative process aggregates information from neighboring nodes and edges to refine the representation of each node.

Step 3: Path Reasoning and Scoring Paths in the subgraph, connecting input concepts to potential diagnoses, are identified and scored based on their relevance. DR.KNOWS employs two attention-based scoring mechanisms:

Multi-Head Attention (MultiAttn): This mechanism computes the relevance of a path p_i given the input representation h_x. The attention score α_i is defined as:

$$\alpha_i = \text{MultiHead}(h_x, p_i), \tag{7}$$

where p_i is the aggregated embedding of a path, and MultiHead denotes the multi-head attention mechanism.

Trilinear Attention (TriAttn): To capture more complex interactions between input text, path embeddings, and node representations, DR.KNOWS uses trilinear attention. The attention score is computed as:

$$\alpha_i = \sum_{a,b,c} (h_x)_a (h_v)_b (p_i)_c W_{abc}, \tag{8}$$

where W_{abc} is a learnable weight tensor, and h_x, h_v, and p_i represent the embeddings of the input, node, and path, respectively.

The paths are ranked based on their attention scores, and the top-scoring paths are selected for integration with the LLM.

Step 4: Integration with LLMs The integration of KG reasoning with LLMs is a critical step in DR.KNOWS, aimed at leveraging the generative capabilities of LLMs while anchoring their outputs in structured knowledge. This integration involves two main components: prompt generation and model fusion.

The selected top-ranked paths from the KG reasoning module are transformed into natural language prompts. For example, if the reasoning identifies that symptoms $S_1, S_2, \ldots$ are connected to a potential diagnosis D through a path P, the generated prompt may read[3] :

> "Based on the symptoms *fever, cough, and difficulty breathing*, and guided by relevant medical relationships, the most probable diagnosis is *pneumonia*. This conclusion is supported by the following reasoning path: *symptoms* $\rightarrow$ *respiratory distress* $\rightarrow$ *pneumonia*."

Formally, the generated prompt P is constructed as a combination of structured KG reasoning results and natural language, following the template:

$$P = \text{Template}(S, P_{reasoning}), \tag{9}$$

where S represents the extracted symptoms, and $P_{reasoning}$ denotes the reasoning path.

The LLM then takes the prompt P as input and generates the final diagnostic summary. Given the LLM parameters Θ, the output O is computed as:

$$O = \text{LLM}(P; \Theta). \tag{10}$$

During this process, the LLM is fine-tuned or guided through prompt engineering to ensure it effectively incorporates the structured knowledge from the KG. The generated output is both fluent and interpretable, combining the semantic richness of LLMs with the explicit reasoning paths derived from the KG. To further enhance the interpretability, DR.KNOWS appends a justification section to the output, detailing how the reasoning path supports the diagnosis.

2.3.3 Limitations

Data Quality The reliance on EHRs and the UMLS introduces potential risks due to inaccuracies, incompleteness, or biases within these data sources. These issues may lead to incorrect diagnostic predictions, reducing the reliability of AI-driven decision support systems. Furthermore, the scarcity of annotated datasets limits the model's learning capabilities, particularly in integrating various clinical note types such as radiology reports or structured data from lab results.

Scalability and Reliability The integration of KGs with LLMs poses significant scalability and computational efficiency challenges. The UMLS knowledge graph, containing 4.5 million concepts and 15 million relations, demands substantial computational resources for multi-hop reasoning and inference. This can lead to increased latency, making real-time clinical decision support difficult, particularly in emergency or time-sensitive medical scenarios.

Another critical issue is the generalization across different EHR systems. Hospitals and healthcare institutions have heterogeneous data formats, varying coding standards and inconsistent documentation practices, making it challenging to deploy a single KG-LLM model across multiple settings without extensive fine-tuning.

Furthermore, the lack of temporal awareness in the model limits its clinical reliability. A patient's medical condition evolves over time, with new test results, medications, and symptoms continuously updating their diagnosis. However, current KG-LLM approaches do not dynamically adjust predictions based on temporal changes in patient history, leading to outdated or misleading diagnoses.

Privacy Concerns The use of KGs and LLMs in healthcare also raises significant ethical and privacy concerns. Even in de-identified datasets, risks related to patient privacy persist, especially when handling sensitive health information. Ensuring compliance with regulations is crucial, yet model training and deployment within healthcare institutions demand additional safeguards to prevent data leaks and unauthorized access. Additionally, biases in both knowledge graphs and language models may disproportionately affect certain patient groups, leading to disparities in diagnostic accuracy and potentially reinforcing existing healthcare inequalities.

2.4 Introduction of Medical Multimodal Foundation Models

The evolution of Foundation Models (FMs) in general AI has paved the way for advancements in healthcare. In natural language processing, models like BERT and GPT have established new paradigms in language understanding and generation. Similarly, in computer vision, models such as Vision Transformers (ViT) and CLIP have enabled groundbreaking capabilities in image analysis and multimodal learning. These general-purpose FMs demonstrate the feasibility of integrating massive datasets and sophisticated architectures to solve complex problems.

However, adapting FMs to healthcare is not without its challenges. The scarcity of large, openly available medical datasets remains a significant barrier, as privacy regulations and ethical concerns restrict data sharing. Furthermore, the high stakes of clinical decision-making demand models with superior accuracy, interpretability, and robustness. In addition, the heterogeneous nature of healthcare data, encompassing imaging modalities like CT and MRI alongside textual and temporal data, adds complexity to model development.

Multimodal learning is particularly crucial in the medical domain, where integrating diverse data sources is essential for holistic patient care. For instance, combining radiology images with clinical reports can provide deeper insights into disease diagnosis and progression. Medical Multimodal Foundation Models aim to address this need by unifying disparate data types, thus enabling more comprehensive analyses and predictions.

Sun et al.[?] classified Medical Multimodal Foundation Models (MMFM) into two major categories in their study: Medical Multimodal Vision Foundation Models (MMVFMs) and Medical Multimodal Vision-Language Foundation Models (MMVLFMs). MMVFMs are primarily image-centric, excelling in tasks such as segmentation and reconstruction. In contrast, MMVLFMs combine visual and textual data to facilitate applications like automated report generation and visual question answering (VQA). These categories illustrate the broad applicability of MMFMs across various aspects of healthcare, from diagnostics to clinical decision support.

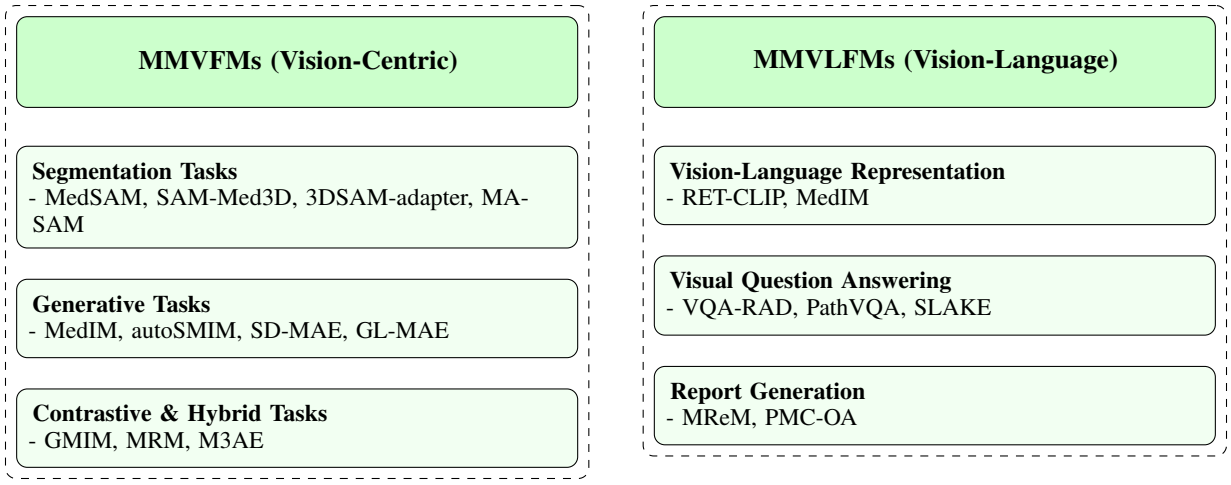


Figure 2: Overview of Medical Multimodal Foundation Models

2.4.1 Medical Multimodal Vision Foundation Models (MMVFMs)

MMVFMs primarily focus on tasks involving medical images, such as segmentation, reconstruction, and contrastive learning. These tasks form the backbone of clinical image analysis, enabling detailed visualization and interpretation of anatomical and pathological structures. The models are organized into four primary proxy tasks: segmentation, generative, contrastive, and hybrid.

Segmentation Tasks
Segmentation tasks are fundamental to medical image analysis, and MMVFMs have made significant strides in this area. Early models such as MedSAM introduced universal segmentation capabilities, leveraging prompts like bounding boxes or points to delineate structures of interest. The progression to SAM-Med3D marked a pivotal shift by incorporating 3D convolutions and volumetric encoding to handle complex spatial relationships in 3D medical data. Models like 3DSAM-adapter and MA-SAM further refined these architectures, emphasizing parameter efficiency and adaptability to diverse imaging modalities. These advancements have significantly reduced the dependency on manual interactions, streamlining clinical workflows.

MedSAM[?] is a foundational model for segmentation, trained on a massive dataset of over 1.5 million image-mask pairs spanning 10 imaging modalities and more than 30 cancer types. It allows for flexible segmentation using user-defined prompts, such as bounding boxes and point inputs, to specify regions of interest. MedSAM addresses the challenge of adapting to multimodal datasets and outperforms general-purpose models like SAM in precision and adaptability to medical contexts.

SAM-Med3D[?] extends the capabilities of MedSAM by introducing 3D convolutions and volumetric encoding. These enhancements enable SAM-Med3D to capture the intricate spatial relationships inherent in 3D medical data. Trained on a dataset comprising 247 categories, SAM-Med3D is optimized for multi-organ segmentation tasks in modalities like CT and MRI. It significantly reduces the need for manual interaction and improves segmentation accuracy for spatially complex medical data.

The 3DSAM-adapter[?] builds on SAM-Med3D by introducing lightweight adaptations tailored for volumetric data. It integrates 3D convolutional layers and employs visual samplers to extract shared semantic features across dimensions. This model is particularly effective for tumor segmentation tasks involving organs like the kidney, pancreas, and colon. It resolves challenges such as over-smoothing in 3D segmentation while maintaining high precision.

MA-SAM[?] further enhances segmentation efficiency by adopting parameter-efficient fine-tuning techniques. This model retains general knowledge from natural image tasks while incorporating 3D-specific improvements through modular adapters. MA-SAM excels in segmentation tasks for modalities like CT and MRI, offering a balance between computational cost and performance in tasks like tumor segmentation.

Generative Tasks
Generative tasks focus on reconstructing missing or incomplete data, a common challenge in medical imaging. Models like MedIM employed innovative masking strategies, such as Knowledge-Driven Masking (KDM) and Sentence-Driven Masking (SDM), to direct attention to critical regions informed by domain-specific knowledge. autoSMIM introduced

superpixel-based masking, enhancing feature extraction for targeted anatomical areas. Subsequent innovations like SD-MAE and GL-MAE integrated self-distillation and multi-scale consistency learning to improve reconstruction accuracy, particularly in cross-domain applications. These generative models address the inherent sparsity of medical data, enabling robust feature learning even in limited-label scenarios.

MedIM[?] combines KDM and SDM to enhance feature extraction in medical imaging. KDM identifies key regions based on medical ontologies like MeSH, while SDM leverages radiological reports to guide attention to critical anatomical areas. This approach improves multi-label classification and segmentation, particularly in radiological imaging tasks, by focusing on regions most relevant to clinical outcomes.

autoSMIM[?] utilizes a superpixel-based masking strategy to target anatomical regions of interest more effectively. By employing Bayesian optimization to determine optimal masking configurations, autoSMIM improves the focus of feature extraction on pathological areas. This model is particularly effective in tasks like skin lesion segmentation using ISIC datasets, addressing the limitations of random masking approaches.

SD-MAE[?] integrates masked autoencoder frameworks with self-distillation techniques to refine shallow feature learning. Visible patch features act as "students," while decoder-generated features serve as "teachers," enhancing reconstruction accuracy. This model excels in cross-domain applications and tasks requiring robust representation learning from sparse datasets.

GL-MAE[?] combines global and local views in its reconstruction framework, addressing the challenges of multi-scale representation in 3D volumetric segmentation. Consistency learning between global and local representations stabilizes feature extraction, improving segmentation accuracy across modalities like CT and MRI.

Contrastive and Hybrid Tasks
Contrastive tasks ensure robust differentiation between modalities and regions, learning representations that are both distinct and complementary. GMIM and MRM exemplify advancements in contrastive learning, focusing on grid-based masking and inter-modal relationships, respectively. Hybrid models like GL-MAE and M3AE represent a synthesis of segmentation, generative, and contrastive capabilities, reflecting a broader trend towards integrating multiple paradigms in MMVFMs.

GMIM[?] employs grid-based masking strategies to improve consistency in feature learning, particularly for small organ segmentation tasks. This approach is effective in delineating boundaries in datasets like BraTS, enhancing accuracy for small and intricate anatomical structures.

MRM[?] extends contrastive learning by masking inter- and intra-modal relationships, enabling nuanced pattern recognition for disease detection. This model facilitates fine-grained image-text alignment, improving the utility of multimodal datasets for clinical applications.

M3AE[?] integrates multimodal autoencoders with modality-completion techniques, addressing the challenge of incomplete data. Self-distillation ensures robust alignment across modalities, making M3AE particularly suited for tumor segmentation tasks under missing modality scenarios.

2.4.2 Medical Multimodal Vision-Language Foundation Models (MMVLFMs)

MMVLFMs extend the capabilities of MMVFMs by integrating textual data, enabling cross-modal reasoning and interpretation. These models cater to tasks such as vision-language representation, visual question answering (VQA), and report generation, offering a holistic approach to medical data analysis.

Vision-Language Representation
Early MMVLFMs focused on aligning visual and textual features to enhance cross-modal understanding. RET-CLIP adapted the CLIP framework for retinal disease diagnosis, leveraging image-text triplets to improve multimodal alignment. MedIM introduced advanced masking strategies, such as KDM and SDM, to emphasize critical regions and their textual counterparts. These models have bridged the gap between images and clinical narratives, paving the way for robust multimodal integration.

RET-CLIP[?] leverages the CLIP framework to align retinal images with clinical diagnoses and textual descriptions. This model excels in tasks like diabetic retinopathy detection by enhancing text-image alignment and facilitating cross-modal understanding in retinal datasets.

MedIM[?] applies KDM and SDM strategies to radiological imaging, improving alignment between image features and textual descriptions. By focusing on clinically relevant regions, MedIM enhances performance in multi-label classification and segmentation tasks, particularly in low-annotation scenarios.

Visual Question Answering (VQA)
VQA models like VQA-RAD and PathVQA have expanded the scope of MMVLFMs by addressing interactive reasoning tasks. VQA-RAD focused on radiology-specific questions, enabling detailed interpretations of diagnostic images, while PathVQA catered to pathology datasets, supporting educational and diagnostic applications. SLAKE extended these capabilities to multimodal datasets, encompassing CT, X-rays, and MRIs, showcasing the adaptability of VQA models in diverse clinical contexts. These advancements underscore the shift from static representation to dynamic interaction in vision-language tasks.

VQA-RAD[?] specializes in radiological question-answering tasks, leveraging paired image-question datasets to interpret diagnostic images. This model supports clinical decision-making by providing detailed, query-specific insights.

PathVQA[?] focuses on pathology datasets, combining visual and textual reasoning to answer complex diagnostic questions. This model is particularly valuable for educational tools and diagnostic applications in pathology.

SLAKE[?] extends VQA capabilities to multimodal datasets, incorporating CT, X-rays, and MRIs. This model enhances reasoning and interaction across diverse imaging modalities, making it a versatile tool for clinical applications.

Report Generation and Retrieval
Report generation models such as MReM and PMC-OA have furthered the integration of vision and language in clinical applications. MReM combined masked image and language modeling to reconstruct missing information, facilitating robust multimodal feature learning. PMC-OA leveraged large-scale biomedical image-text pairs to enhance foundational capabilities, supporting tasks like VQA and image-text retrieval. These models highlight the growing emphasis on generative capabilities in MMVLFMs, addressing the need for automated reporting and decision support in clinical workflows.

MReM[?] integrates masked image and language modeling to reconstruct missing features in multimodal datasets. This model supports robust feature learning for tasks like report generation and image-text retrieval, particularly in low-label scenarios.

PMC-OA[?] leverages a large-scale dataset of biomedical image-text pairs to improve foundational model capabilities. This model supports diverse tasks, including VQA, report generation, and image-text retrieval, showcasing its versatility in clinical applications.

2.4.3 Limitations

Data and Multimodal Integration MMFMs require large-scale, high-quality multimodal datasets to enhance their generalizability in clinical applications. However, current datasets suffer from biases, inconsistent labeling, and a lack of diversity, particularly in underrepresented patient populations. The integration of heterogeneous data sources—including imaging, clinical notes, and structured lab results—remains challenging due to variations in data formats, coding standards, and interoperability issues across different healthcare institutions. Most MMFMs are trained on single-institution datasets, limiting their adaptability to broader healthcare settings.

Interpretability and Ethical Challenges The black-box nature of MMFMs hinders their clinical adoption, as healthcare professionals require transparent and interpretable AI-driven insights for decision-making. Bias and fairness remain critical concerns, as models trained on skewed datasets may exacerbate healthcare disparities by underperforming for specific demographic groups. Additionally, hallucinations and unreliable outputs in MMFMs pose risks for clinical decision-making, requiring rigorous validation. Future efforts should focus on explainable AI (XAI), model uncertainty quantification, and bias mitigation strategies to improve the reliability and fairness of MMFM-driven diagnoses and treatment recommendations.

Privacy and Security Ensuring patient privacy and data security is a significant barrier to the large-scale deployment of MMFMs. Training on sensitive medical data requires strict compliance with regulations, yet existing methods may still expose patient information to privacy risks. Additionally, cross-institutional data sharing is limited due to ethical and legal concerns, restricting the development of comprehensive, generalizable MMFMs. Federated learning and privacy-preserving techniques are promising directions to address these challenges. Furthermore, workflow integration remains an open issue, as MMFMs must seamlessly interface with existing clinical systems to provide actionable, user-friendly insights for healthcare professionals.

3 Evaluation and Metrics

To evaluate the performance of medical AI models across various tasks, multiple metrics are employed. For text-based tasks like Named Entity Recognition (NER) and Relation Extraction (RE), Precision, Recall, and F_1 Score are used to measure the accuracy of identifying entities and relationships. In these cases, the F_1 Score provides a balanced measure

that considers both the proportion of correctly identified items (Precision) and the proportion of relevant items that are identified (Recall). Question Answering (QA) tasks are often evaluated using metrics such as Strict Accuracy, Lenient Accuracy, and Mean Reciprocal Rank (MRR) to quantify the correctness of answers and the quality of their ranking. These metrics indicate how well a model can pinpoint exact answers and prioritize them appropriately.

For clinical question answering, textual overlap metrics like BLEU and ROUGE are commonly used alongside human evaluations. Automated metrics like BLEU and ROUGE score the overlap of words or phrases between the generated text and a reference text; however, they may not capture the clinical correctness or context. Human assessments, performed by clinical experts, are essential to detect misleading or harmful content that automated measures might miss.

For multimodal tasks such as segmentation and lesion detection in medical images, Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are employed to measure the spatial overlap between predicted and ground-truth masks. These metrics are crucial in evaluating the accuracy of segmentation models. Mean Average Precision (mAP) and Free-Response ROC (FROC) are used in lesion detection tasks to evaluate the balance between precision and recall in localizing lesions. For volumetric segmentation tasks, metrics like Hausdorff Distance capture the boundary accuracy of predicted shapes. Radiology report generation tasks combine BLEU, ROUGE, and CIDEr to assess the quality of generated text, with domain experts confirming that the reports meet clinical standards. Entity linking tasks, which align extracted findings to standardized vocabularies such as UMLS, are evaluated using accuracy metrics to ensure consistency in medical terminology.

Table 3: Performance Metrics for Different Medical AI Models

Model	**Task**	**Dataset**	**Metrics**
BioBERT	Named Entity Recognition (NER)	NCBI Disease, BC5CDR	F_1 Score (balances Precision & Recall)
	Relation Extraction (RE)	BC5CDR, ChemProt	F_1 Score (accuracy in detecting relationships)
	Question Answering (QA)	BioASQ	Strict Accuracy, Lenient Accuracy, MRR (assesses ranking & correctness)
MedPaLM (LLM)	Clinical Question Answering	MultiMedQA	Accuracy (fraction of correct answers)
	Open-ended QA	MultiMedQA	BLEU, ROUGE (textual overlap & human evaluation)
DR.KNOWS	Concept Retrieval	MIMIC-III, IN-HOUSE	Recall@N, F_1 Score (retrieval performance)
	Free-text Summarization	PROBSUM	ROUGE-L, CUI F-Score (text quality & term accuracy)
ChatGPT + DR.KNOWS	Diagnostic Reasoning	PROBSUM	CUI F-Score (accuracy of clinical term identification)
MMFMs	Segmentation	LUNA16, KiTS19	Dice Score (overlap of true & predicted masks)
	Lesion Detection	Chest X-ray, Liver CT	mAP, AUC (precision-recall balance)
	Radiology Report Generation	Chest X-ray	BLEU-4, ROUGE-L, CIDEr (quality assessed with metrics & expert reviews)
	Entity Linking	Pathology Alignment	Accuracy (semantic consistency with vocabularies)

Biobert
Lee et al.[1] present BioBERT, a transformer-based language model pre-trained on PubMed and PMC for biomedical text-mining tasks. Their evaluations concentrate on Named Entity Recognition (NER), Relation Extraction (RE), and

Question Answering (QA). In NER, corpora such as NCBI Disease and BC5CDR are employed, with Precision, Recall, and F_1 scores measuring how accurately entity boundaries and types are identified. BioBERT consistently achieves gains of 0.5–2.0 F_1 points over a general-domain BERT, showing the benefits of in-domain pre-training for recognizing biomedical entities. For RE, BioBERT is fine-tuned on datasets like BC5CDR and ChemProt, where it must detect relations (e.g., protein–chemical interactions) using the same metrics. Here, BioBERT outperforms baselines by 1–3 F_1 points, demonstrating improved recognition of domain-specific relations.

For QA, BioBERT is evaluated on BioASQ, a dataset of biomedical factoid questions linked to PubMed articles. On the BioASQ 5b Phase B (factoid questions) dataset, BioBERT achieves : 78.86% of Strict Accuracy, 89.12% of Lenient Accuracy, and 85.17% of MRR, compared to BERT's scores of 73.77%, 83.68%, and 79.25%, respectively. These results show BioBERT's enhanced ability to pinpoint precise answers and rank them appropriately. The performance gains across NER, RE, and QA tasks show the critical role of domain-specific pre-training, as the PubMed and PMC corpora provide essential contextual knowledge for understanding biomedical text.

Table 4: BioBERT Performance Metrics Across Tasks

Task	**Dataset**	**Metric**	**BioBERT**	**BERT (Baseline)**
NER	NCBI Disease, BC5CDR	F_1 Score	+0.5–2.0 points	-
RE	BC5CDR, ChemProt	F_1 Score	+1–3 points	-
QA	BioASQ	Strict Accuracy	78.86%	73.77%
		Lenient Accuracy	89.12%	83.68%
		MRR	85.17%	79.25%

LLM for Clinical Question Answering

Singhal et al.[7] present large language models designed for clinical question answering, their evaluation is within the MultiMedQA benchmark. This benchmark merges diverse datasets spanning multiple-choice question sets (MedQA, MedMCQA, PubMedQA, and clinical topics from MMLU) and open-ended tasks (LiveQA, MedicationQA, HealthSearchQA). For the multiple-choice component, models are scored via simple Accuracy, reflecting the fraction of correctly answered questions. Singhal et al. show that scaling model size (e.g. to Flan-PaLM with 540B parameters) and tuning it with medical task instructions can raise accuracy by 10%–20% relative to smaller or general-purpose versions. In contrast, the open-ended tasks require long-form reasoning and explanations.
While BLEU or ROUGE can measure textual overlap, the study includes human assessments—particularly from clinicians—to detect misleading content, omissions, or potential harm. For instance, they report that Med-PaLM significantly reduces dangerous or incorrect statements compared to a baseline Flan-PaLM, but clinicians still find nuanced mistakes. The work further highlights that targeted prompt engineering diminishes the likelihood of harmful advice, dropping it from over 11% to under 6% in a pilot evaluation, showing the importance of curated examples and domain-aware prompts.

Table 5: Performance Metrics for Med-PaLM (LLM for Clinical Question Answering)

Task	**Metric**	**Performance**
Multiple-Choice QA	Accuracy	MedQA: 67.2% (Med-PaLM) vs. 50.3% (Baseline)
		MedMCQA: 46.4% (Med-PaLM) vs. 25.8% (Baseline)
		PubMedQA: 81.8% (Med-PaLM) vs. 77.4% (Baseline)
Open-Ended QA	BLEU	HealthSearchQA: 38.2
	ROUGE	MedicationQA: 45.6
	Human Assessment	Reduced harmful advice: 11% $\rightarrow$ 5.8%
Clinical QA (Overall)	Accuracy Improvement	+10%–20% with Flan-PaLM on MultiMedQA
	Clinician Acceptability	Higher alignment with medical standards reported

DR.KNOWS
Gao et al.[3] propose a knowledge graph-enhanced framework, DR.KNOWS, which augments large language models with the Unified Medical Language System (UMLS) for automated diagnosis generation. Their evaluations consist of two main stages: concept retrieval and free-text summarization.

In the first stage, the model retrieves possible diagnoses or medical concepts (CUIs) for each patient from UMLS. The retrieval accuracy is measured using Recall@N, Precision@N, and F-Score, compared against gold-standard CUIs. For the MIMIC-III dataset, DR.KNOWS achieves a Recall@6 of 32.96

In the second stage, the retrieved CUIs are input into generative language models such as T5 and GPT-3.5-turbo. The generated summaries are evaluated using ROUGE scores and a domain-specific CUI F-Score. On the PROBSUM benchmark, vanilla T5 achieves a ROUGE-L score of 38.96 and a CUI F-Score of 27.78. By integrating path-based prompts, ClinicalT5 improves these metrics to 40.33 for ROUGE-L and 29.13 for the CUI F-Score. These results show the benefit of adding structured knowledge paths to the prompts.

Example of Path-Based Prompts: To illustrate how path-based prompts align the outputs with clinical reasoning, consider the following example from the PROBSUM dataset. For a patient presenting with symptoms such as "shortness of breath, chest tightness, and wheezing," DR.KNOWS retrieves relevant CUIs, such as "C0032285 (Asthma)" and "C0018801 (COPD)," and constructs a knowledge path by linking these concepts with intermediate nodes in UMLS, like "C0004096 (Respiratory Diseases)." The system then uses this path to generate a prompt:

"Given a patient experiencing shortness of breath and wheezing, consider related respiratory diseases such as Asthma (C0032285) or COPD (C0018801). Summarize the most likely diagnosis and reasoning based on these conditions."

This structured input allows the language model to focus on specific diagnoses while grounding its reasoning in clinical knowledge. The resulting output aligns with clinical reasoning by explicitly mentioning the retrieved concepts, their relationships, and the evidence supporting the diagnosis.

Moreover, ChatGPT combined with DR.KNOWS demonstrated strong interpretability by leveraging knowledge paths to produce detailed and logical diagnostic reasoning. This capability not only improved diagnostic accuracy but also enhanced trust in the model's outputs by providing clear explanations for its predictions. On the PROBSUM dataset, ChatGPT achieved a CUI F-Score of 16.04, which increased to 18.21 when integrated with DR.KNOWS' paths. This improvement illustrates the potential of combining advanced LLMs with knowledge graphs to address challenges in interpretability. By incorporating explicit reasoning paths, DR.KNOWS makes the diagnostic process more transparent and clinically relevant.

DR.KNOWS also enhances interpretability by grounding predictions in UMLS concepts. The multi-hop reasoning mechanism reduces hallucinations and enhances reliability. Using path-based prompts further aligns the outputs with clinical reasoning, as shown in the example above, making the generated diagnoses more relevant and easier to interpret.

Table 6: Performance Metrics for DR.KNOWS

Dataset	**Task**	**Metric**	**Performance**
MIMIC-III	Concept Retrieval	Recall@6	32.96% (TriAttn), 30.76% (MultiAttn)
		F-Score	24.44% (TriAttn), 24.77% (MultiAttn)
		Recall@6 (Baseline)	56.91%
		F-Score (Baseline)	21.13%
IN-HOUSE	Concept Retrieval	Recall@8	44.58% (TriAttn)
		F-Score	25.70% (TriAttn)
		F-Score (Baseline)	20.09%
PROBSUM	Free-Text Summarization	ROUGE-L	38.96 (vanilla T5), 40.33 (ClinicalT5)
		CUI F-Score	27.78 (vanilla T5), 29.13 (ClinicalT5)
PROBSUM + Paths	Diagnostic Reasoning	CUI F-Score (ChatGPT)	16.04 $\rightarrow$ 18.21 (with DR.KNOWS paths)

Medical Multimodal Foundation Models (MMFMs)
Sun et al.[?] expand to Medical Multimodal Foundation Models (MMFMs) that integrate data from medical images (CT, MRI, X-ray), textual clinical notes, and sometimes other patient information such as lab results. Their survey and empirical studies show how these models are evaluated for multiple tasks beyond text alone. Segmentation and detection metrics dominate the imaging aspect, with Dice Similarity Coefficient (DSC) or Intersection over Union (IoU) quantifying how accurately predicted masks align with anatomical ground truths, and mean Average Precision (mAP) or Free-Response ROC analyzing lesion detection. In some volumetric tasks (e.g. 3D CT scans), Hausdorff Distance captures boundary shape fidelity. Classification tasks often employ Accuracy, F_1, or AUROC to gauge how correctly diseases or abnormalities are recognized. Meanwhile, the text side may involve generating radiology reports or multimodal explanations, evaluated with BLEU, ROUGE, or CIDEr and supplemented by domain specialists who verify correctness of medical terms. Many MMFMs also incorporate entity linking, aligning textual or image findings to standardized vocabularies like UMLS. Sun et al. show that pre-training with self-supervised or proxy tasks on large unlabeled datasets tends to boost segmentation performance (by 1–3 DSC) or classification AUC (by 5–10 points) compared to training from scratch, though model outputs still require expert scrutiny for safety.

Sun et al.[?] propose Medical Multimodal Foundation Models (MMFMs) that combine data from medical images (e.g., CT, MRI, X-ray), clinical notes, and additional patient information such as laboratory results. These models are

evaluated across several tasks, including segmentation, lesion detection, radiology report generation, and text-image alignment, with metrics tailored to each task. For segmentation tasks, the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) are used to quantify how closely the predicted masks align with the anatomical ground truths. In volumetric segmentation tasks, such as those involving 3D CT scans, Hausdorff Distance is additionally employed to measure boundary shape fidelity. Pre-training on large unlabeled datasets significantly boosts performance, as demonstrated in experiments on the LUNA16 dataset where pre-training improved the Dice score for lung nodule segmentation from 85.0 to 87.1 and the lesion detection mean Average Precision (mAP) from 71.2% to 80.1%. Similarly, on the KiTS19 dataset for kidney tumor segmentation, DSC increased from 83.2 to 86.4 following pre-training.

Lesion detection tasks are evaluated using metrics such as mAP and Free-Response ROC, which assess the precision and recall of abnormality localization. For instance, on a chest X-ray lesion detection task, pre-trained MMFMs achieved an mAP of 75.6% compared to 68.4% for models without pre-training. Additionally, for Free-Response ROC analysis on liver lesion detection from CT scans, pre-trained models achieved an area under the curve (AUC) improvement of 0.09 compared to baseline models, showing enhanced sensitivity in identifying abnormalities at multiple scales.

In radiology report generation tasks, MMFMs are evaluated with Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Consensus-Based Image Description Evaluation (CIDEr) metrics to measure the overlap between generated and reference text. On a multi-hospital chest X-ray dataset, the pre-trained MMFMs achieved a BLEU-4 score of 24.5, compared to 20.1 for models trained without pre-training. Similarly, ROUGE-L increased from 36.8 to 40.7, and CIDEr improved from 130.4 to 145.9 with pre-training, showing gains in capturing clinically relevant details in generated reports. These automated metrics are supplemented by domain experts who evaluate the medical correctness of the generated text, confirming that pre-trained MMFMs produce outputs that are more accurate and complete.

Additionally, many MMFMs incorporate entity linking mechanisms that align extracted findings with standardized ontologies such as the Unified Medical Language System (UMLS). This ensures consistency in terminology and supports semantic interpretation. For instance, on a pathology description alignment task, entity-linking models within MMFMs achieved an accuracy of 92.4%, compared to 87.2% for baseline models. These mechanisms not only improve consistency but also facilitate downstream tasks such as cross-modal retrieval and multimodal reasoning.

Table 7: Performance Metrics for MMFMs

Task	**Dataset**	**Metric**	**Performance**
Segmentation	LUNA16, KiTS19	Dice Score	85.0 → 87.1 (LUNA16), 83.2 → 86.4 (KiTS19)
		mAP	71.2% → 80.1% (LUNA16)
Lesion Detection	Chest X-ray, Liver CT	mAP	68.4% → 75.6% (Chest X-ray)
		Free-Response ROC (AUC)	0.09 improvement (Liver CT)
Radiology Report Generation	Chest X-ray	BLEU-4	20.1 → 24.5
		ROUGE-L	36.8 → 40.7
		CIDEr	130.4 → 145.9
Entity Linking	Pathology Alignment	Accuracy	87.2% → 92.4%

4 Summary and Future Work

4.1 Summary

The four studies reviewed in this work focus on the progress of AI models used in medical diagnosis. Lee et al.[1] show how pre-training on biomedical datasets improves the performance of BioBERT on text-mining tasks. These tasks include Named Entity Recognition (NER), Relation Extraction (RE), and Question Answering (QA). Their work shows the importance of using domain-specific datasets to better capture vocabulary and relations.

Singhal et al.[7] focus on large-scale language models for clinical question answering. They show that scaling the model size and using instruction prompt tuning improve performance for multiple-choice and open-ended tasks. Automated accuracy metrics, BLEU, and ROUGE are used for evaluation, and these are complemented by human reviews. These reviews help detect errors and ensure the outputs align with clinical guidelines.

Gao et al.[3] explore the use of knowledge graphs. They show how retrieving concepts from the Unified Medical Language System (UMLS) improves accuracy and reasoning in diagnostic tasks. Their work shows that using structured knowledge helps reduce errors and makes the system more interpretable.

Sun et al.[?] focus on multimodal models that combine imaging data with text. These models are evaluated for segmentation, detection, and report generation. Pre-training with large datasets improves the models' performance. Expert reviews are used to validate the results, as automated metrics alone cannot fully ensure correctness.

These studies show that using domain-specific methods, robust evaluation metrics, and human expert reviews improves the reliability of medical AI Diagnosis systems. By focusing on diverse tasks and data types, these works provide a solid foundation for the future research in this field.

4.2 Future Work and Challenges

The four studies reveal several areas for improvement and highlight challenges that need to be addressed to advance the research and application of medical AI tools. These include improving datasets, enhancing evaluation methodologies, advancing model architectures, addressing ethical concerns, and developing multimodal integration techniques.

Currently, one major limitation is the reliance on domain-specific datasets, as emphasized by Lee et al.[1] and Gao et al.[3]. Existing datasets like MIMIC-III and in-house EHR datasets cover only a subset of clinical scenarios and patient populations, resulting in biases and limited generalizability. Expanding datasets to represent a broader spectrum of medical subfields, geographic regions, patient demographics, and languages is essential. Moreover, standardizing data formats across these datasets will facilitate cross-study comparisons and reproducibility. Future efforts should also explore the inclusion of synthetic data generated through advanced generative models to address underrepresented scenarios or rare diseases, as discussed by Gao et al.[3].

Another significant challenge lies in the interpretability and explainability of models. Gao et al.[3] demonstrate the potential of knowledge graphs to improve diagnostic reasoning, yet integrating these complex graph structures with large language models (LLMs) remains difficult. Future research should focus on enhancing the robustness of this integration, such as by leveraging multi-hop reasoning with deeper contextual embeddings. Additionally, the development of more specialized ontologies for specific fields like oncology, cardiology, and rare diseases could further enhance diagnostic accuracy and explainability. Incorporating interactive explainability techniques, where clinicians can query and explore AI reasoning pathways, is another promising direction.

Evaluation methodologies also require significant refinement, as highlighted by Singhal et al.[7] and Sun et al.[?] . Current evaluations rely heavily on metrics such as ROUGE, BLEU, or F_1, which may fail to fully capture the clinical relevance of AI-generated outputs. Standardized evaluation protocols involving clinicians and other stakeholders could address this gap. Furthermore, including patient-centric metrics, such as the clarity and usability of AI outputs in real-world clinical settings, would greatly enhance the practical value of these tools. Additionally, developing simulation-based evaluation frameworks, where AI systems are tested in virtual clinical environments, could provide a scalable and safe method for assessing real-world performance.

The integration of multimodal data presents both opportunities and challenges. Sun et al.[?] show the potential of combining textual data, such as clinical notes, with imaging data, like CT scans and MRIs. However, current approaches often struggle to align information from different modalities effectively and to incorporate temporal data, such as disease progression. Future work should focus on developing novel methods for temporal multimodal fusion to improve predictions of patient outcomes and enable more personalized treatment recommendations. The incorporation of 3D medical imaging, dynamic patient monitoring data (e.g., continuous glucose monitors), and wearable device data into multimodal frameworks could unlock new possibilities for precision medicine.

Emerging AI techniques, such as federated learning and edge AI, have the potential to address computational and data privacy challenges. Federated learning enables collaborative model training across multiple institutions without sharing raw patient data, ensuring data privacy while leveraging diverse datasets. Edge AI, where models are deployed on local devices or edge servers, could bring real-time decision-making capabilities to the bedside, particularly in resource-limited settings. Research should also explore hybrid approaches that combine cloud-based computation with edge deployment for scalable yet efficient AI solutions.

Ethical and regulatory issues continue to be major barriers to the adoption of medical AI tools. Studies by Gao et al.[3] and Singhal et al.[7] highlight the risks of hallucinations and biases in model outputs. Ensuring fairness, accountability, and transparency in these systems is critical for building trust among clinicians and patients. Research efforts should focus on creating frameworks for model auditing and bias detection. Collaboration with policymakers, clinicians, and ethicists will be essential to establish guidelines that meet regulatory requirements and address public concerns. Additionally, exploring the integration of blockchain technology for secure and auditable data provenance could help address concerns related to data integrity and accountability.

Finally, the computational demands of training and deploying advanced models present a persistent challenge. Lee et al.[1] emphasize the resource constraints that limit the accessibility of state-of-the-art models to only well-funded organizations. To address this, future research should prioritize the development of more computationally efficient architectures, such as sparsity-based models and quantized neural networks. Techniques like knowledge distillation, where smaller models are trained to mimic the performance of larger ones, could also make advanced medical AI tools more accessible. Exploring federated and decentralized learning frameworks could further enable institutions to collaborate without compromising the privacy of sensitive patient data.

References

[1] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–1240. doi: 10.1093/bioinformatics/btz682

[2] Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019:72–78. doi: 10.18653/v1/W19-1909

[3] Gao Y, Li R, Caskey J, Dligach D, Miller T, Churpek M, Afshar M. Dr.knows: Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*; 2023. doi: 10.48550/arXiv.2308.14321

[4] Kwon T, Ong KT, Kang D, Moon S, Lee JR, Hwang D, Sim Y, Lee D, Yeo J. Clinical chain-of-thought: Reasoning-aware diagnosis framework with prompt-generated rationales. *arXiv preprint arXiv:2312.07399*; 2023. doi: 10.48550/arXiv.2312.07399

[5] McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, Singhal K, Sharma Y, Azizi S, Kulkarni K, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2307.08922*; 2023. doi: 10.48550/arXiv.2307.08922

[6] Bian J, Wang S, Yao Z, Guo J, Zhang Q, Sun C, Windle SR, Liu X. Gatortron: a large language model for electronic health records. *J Am Med Inform Assoc*. 2022;29(2):283–291. doi: 10.1093/jamia/ocac005

[7] Singhal K, Tu D, Palepu A, Wang A, Sunshine J, Corrado GS. Medpalm: large language models encode clinical knowledge. *arXiv preprint arXiv:2212.09162*; 2022. doi: 10.48550/arXiv.2212.09162

[8] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901. doi: 10.1093/bioinformatics/btz682

[9] Devlin J, Chang MW, 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; 2019:4171–4186. doi: 10.18653/v1/N19-1423

[10] Wu CK, Chen WL, Chen HH. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2306.01567*; 2023. doi: 10.48550/arXiv.2306.01567

[11] Rajpurkar P, Irvin J, Ball M, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul T, Langlotz D. Medical question answering with large language models. *Nat Mach Intell*. 2021;3:343–348. doi: 10.1038/s42256-021-00283-9

[12] Li X, Hu S, Liu J. Towards automatic diagnosis from multi-modal medical data. *IEEE Trans Med Imaging*. 2018;37(4):888–900. doi: 10.1109/TMI.2017.2781965

[13] Khader F, Ali H, Yousaf M. Medical diagnosis with large scale multimodal transformers: Leveraging diverse data for more accurate diagnosis. *arXiv preprint arXiv:2212.09162*; 2022. doi: 10.48550/arXiv.2212.09162

[14] Ma MD, Singh P, Smith R, Brown J. Clibench: multifaceted evaluation of large language models in clinical decisions on diagnoses, procedures, lab tests orders, and prescriptions. *arXiv preprint arXiv:2406.09923*; 2023. doi: 10.48550/arXiv.2406.09923

[15] Kumar A, Sharma S, Srinivasan P. Medimage: integrating multimodal data for medical diagnostics. *arXiv preprint arXiv:2205.06109*; 2022. doi: 10.48550/arXiv.2205.06109

[16] Ruan C, Wang F, Chen T. Comprehensive evaluation of multimodal ai models in medical imaging diagnosis. *arXiv preprint arXiv:2406.07853*; 2024. doi: 10.48550/arXiv.2406.07853

[17] Zhou H, Li X, Chen Y. Towards personalized multimodal medical diagnostics with large-scale ai models. *arXiv preprint arXiv:2407.02164*; 2024. doi: 10.48550/arXiv.2407.02164

[18] Baumgartner C. The potential impact of chatgpt in clinical and translational medicine. *Clin Transl Med*. 2023;13(3). doi: 10.1002/ctm2.1259

[19] Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. *arXiv preprint arXiv:2306.08302*; 2023. doi: 10.48550/arXiv.2306.08302

[20] Savova G, Masanz J, Ogren P, Zheng J, Sohn S, Schuler K, Chute C. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507–513. doi: 10.1136/amiajnl-2010-000108

[21] Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop; 2016:1–4. doi: 10.18653/v1/W16-1616

[22] Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021:4228–4238. doi: 10.18653/v1/N21-1501

[23] Sun K, Xue S, Sun F, Sun H, Luo Y, Wang L, Wang S, Guo N, Liu L, Zhao T, Wang X, Yang L, Jin S, Yan J, Dong J. Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: applications, challenges, and future directions. *arXiv preprint arXiv:2412.02621*; 2024. doi: 10.48550/arXiv.2412.02621

[24] Sun K, Xue S, Sun F, Sun H, Luo Y, Wang L, Wang S, Guo N, Liu L, Zhao T, Wang X, Yang L, Jin S, Yan J, Dong J. Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: applications, challenges, and future directions. *arXiv preprint arXiv:2412.02621*; 2024. doi: 10.48550/arXiv.2412.02621

[25] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654. doi: 10.1038/s41467-023-36524-5

[26] Wang H, Guo S, Ye J, Deng Z, Cheng J, Li T, Chen J, Su Y, Huang Z, Shen Y, Fu B, Zhang S, He J, Qiao Y. Sam-med3d. *arXiv preprint arXiv:2310.15161*; 2023. doi: 10.48550/arXiv.2310.15161

[27] Gong S, Zhong Y, Ma W, Li J, Wang Z, Zhang J, Heng PA, Dou Q. 3dsam-adapter: holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Med Image Anal*. 2024;98:103324. doi: 10.1016/j.media.2024.103324

[28] Chen C, Miao J, Wu D, Zhong A, Yan Z, Kim S, Hu J, Liu Z, Sun L, Li X, et al. Ma-sam: modality-agnostic sam adaptation for 3d medical image segmentation. *Med Image Anal*. 2024;98:103310. doi: 10.1016/j.media.2024.103310

[29] Xie Y, Gu L, Harada T, Zhang J, Xia Y, Wu Q. Medim: boost medical image representation via radiology report-guided masking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2023:113–123. doi: 10.1007/978-3-031-12345-6

[30] Wang Z, Lyu J, Tang X. Autosmim: automatic superpixel-based masked image modeling for skin lesion segmentation. *IEEE Trans Med Imaging*. 2023. doi: 10.1109/TMI.2023.00000

[31] Luo Y, Chen Z, Zhou S, Gao X. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv preprint arXiv:2203.16983*; 2022. doi: 10.48550/arXiv.2203.16983

[32] Zhuang JX, Luo L, Chen H. Advancing volumetric medical image segmentation via global-local masked autoencoder. *arXiv preprint arXiv:2306.08913*; 2023. doi: 10.48550/arXiv.2306.08913

[33] Wang H, Tang Y, Wang Y, Guo J, Deng ZH, Han K. Masked image modeling with local multi-scale reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023:2122–2131. doi: 10.1109/CVPR.2023.00000

[34] Yang Q, Li W, Li B, Yuan Y. Mrm: masked relation modeling for medical image pretraining with genetics. In: IEEE/CVF International Conference on Computer Vision; 2023:21452–21462. doi: 10.1109/ICCV.2023.00000

[35] Liu H, Wei D, Lu D, Sun J, Wang L, Zheng Y. M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. *AAAI Conf Artif Intell*. 2023;37(2):1657–1665. doi: 10.1609/AAAI.v37i2.1657

[36] Du J, Guo J, Zhang W, Yang S, Liu H, Li H, Wang N. Ret-clip: a retinal image foundation model pre-trained with clinical diagnostic reports. *arXiv preprint arXiv:2405.14137*; 2024. doi: 10.48550/arXiv.2405.14137

[37] Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data*. 2018;5(1):1–10. doi: 10.1038/sdata.2018.18

[38] He X, Zhang Y, Mou L, Xing E, Xie P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*; 2020. doi: 10.48550/arXiv.2003.10286

[39] Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021:1650–1654. doi: 10.1109/ISBI.2021.00000

[40] Zhou HY, Lian C, Wang L, Yu Y. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*; 2023. doi: 10.48550/arXiv.2301.13155

[41] Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, Xie W. Pmc-clip: contrastive language-image pretraining using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2023:525–536. doi: 10.1007/s10462-023-00000

[42] Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*. 2018;34:4087. doi: 10.1093/bioinformatics/bty400

[43] Mikolov T, et al. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. 2013;26:3111–3119. doi: 10.5555/2999792.2999959

[44] Peter ME, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1; 2018:2227–2237. doi: 10.18653/v1/N18-1202

[45] Pyysalo S, et al. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan; 2013:39–43. doi: 10.1093/bioinformatics/btt140

[46] Wu Y, et al. Google's neural machine translation system: bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*; 2016. doi: 10.48550/arXiv.1609.08144

[47] Rajpurkar P, et al. Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX; 2016:2383–2392. doi: 10.18653/v1/D16-1264

[48] Wiese G, et al. Neural domain adaptation for biomedical question answering. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada; 2017:281–289. doi: 10.18653/v1/K17-1029

[49] Vaswani A, et al. Attention is all you need. In: Advances in Neural Information Processing Systems; 2017:5998–6008. doi: 10.5555/3295222.3295349

[50] Krallinger M, et al. Overview of the BioCreative VI chemical-protein interaction track. In: Proceedings of the BioCreative VI Workshop, Bethesda, MD, USA; 2017:141–146. doi: 10.1093/database/bay073

[51] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R. Deep learning-enabled medical computer vision. *NPJ Digit Med*. 2021;1:1–9. doi: 10.1038/s41746-021-00457-4

[52] Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: a practitioner's perspective. *arXiv preprint arXiv:2202.01875*; 2022. doi: 10.48550/arXiv.2202.01875

[53] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*; 2021. doi: 10.48550/arXiv.2108.07258

[54] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11:6421. doi: 10.3390/app11146421

[55] Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on Health, Inference, and Learning; 2022:248–260. doi: 10.48550/arXiv.2209.12345

[56] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*; 2019. doi: 10.48550/arXiv.1909.06146

[57] Abacha AB, Agichtein E, Pinter Y, Demner-Fushman D. Overview of the medical question answering task at TREC 2017 LiveQA. In: TREC; 2017:1–12. doi: 10.1109/TREC.2017.00000

[58] Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the gap between consumers' medication questions and trusted answers. In: MedInfo; 2019:25–29. doi: 10.3414/ME19-00000

[59] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*; 2020. doi: 10.48550/arXiv.2009.03300

[60] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, et al. PaLM: scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*; 2022. doi: 10.48550/arXiv.2204.02311

[61] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li E, Wang X, Dehghani M, Brahma S, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*; 2022. doi: 10.48550/arXiv.2210.11416

[62] Feng SY, Khetan V, Sacaleanu B, Gershman A, Hovy E. CHARD: clinical health-aware reasoning across dimensions for text generation models. *arXiv preprint arXiv:2210.04191*; 2022. doi: 10.48550/arXiv.2210.04191

[63] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, Santoro A, Gupta A, Garriga-Alonso A, et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*; 2022. doi: 10.48550/arXiv.2206.04615

[64] Barham P, Chowdhery A, Dean J, Ghemawat S, Hand S, Hurt D, Isard M, Lim H, Pang R, Roy S, et al. Pathways: asynchronous distributed dataflow for ML. In: Proceedings of Machine Learning and Systems; 2022;4:430–449. doi: 10.1145/3507221.3507248

[65] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901. doi: 10.5555/3454287.3454612

[66] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*; 2021. doi: 10.48550/arXiv.2109.01652

[67] Wang X, Wei J, Schuurmans D, Le Q, Chi E, Zhou D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*; 2022. doi: 10.48550/arXiv.2203.11171

[68] Lewkowycz A, Andreassen A, Dohan D, Dyer E, Michalewski H, Ramasesh V, Slone A, Anil C, Schlag I, Gutman-Solo T, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*; 2022. doi: 10.48550/arXiv.2206.14858