npj | artificial intelligence

Review

https://doi.org/10.1038/s44387-025-00011-z

Large language models for disease diagnosis: a scoping review

Shuang Zhou[1,13], Zidu Xu[2,13], Mian Zhang[3,13], Chunpu Xu[4,13], Yawen Guo[5], Zaifu Zhan[6], Yi Fang[7], Sirui Ding[8], Jiashuo Wang[4], Kaishuai Xu[4], Liqiao Xia[9], Jeremy Yeung[1], Daochen Zha[10], Dongming Cai[11], Genevieve B. Melton[12], Mingquan Lin[1] & Rui Zhang[1] ✉

Automatic disease diagnosis has become increasingly valuable in clinical practice. The advent of large language models (LLMs) has catalyzed a paradigm shift in artificial intelligence, with growing evidence supporting the efficacy of LLMs in diagnostic tasks. Despite the increasing attention in this field, a holistic view is still lacking. Many critical aspects remain unclear, such as the diseases and clinical data to which LLMs have been applied, the LLM techniques employed, and the evaluation methods used. In this article, we perform a comprehensive review of LLM-based methods for disease diagnosis. Our review examines the existing literature across various dimensions, including disease types and associated clinical specialties, clinical data, LLM techniques, and evaluation methods. Additionally, we offer recommendations for applying and evaluating LLMs for diagnostic tasks. Furthermore, we assess the limitations of current research and discuss future directions. To our knowledge, this is the first comprehensive review for LLM-based disease diagnosis.

Automatic disease diagnosis is pivotal in clinical practice, leveraging clinical data to generate potential diagnoses with minimal human input[1]. It enhances diagnostic accuracy, supports clinical decision-making, and addresses healthcare disparities by providing high-quality diagnostic services[2]. Additionally, it boosts efficiency, especially for clinicians managing aging populations with multiple comorbidities[3–5]. For example, DXplain[6] analyzes patient data to generate diagnoses with justifications. Online services also promote early diagnosis and large-scale screening for diseases like mental health disorders, raising awareness and mitigating risks[4,7–10].

Advances in artificial intelligence (AI) have driven two waves of automated diagnostic systems[11–14]. Early approaches utilized machine learning techniques like support vector machines and decision trees[15,16]. With larger datasets and computational power, deep learning (DL) models, such as convolutional, recurrent, and generative adversarial networks, became predominant[1,2,17–20]. However, these models require extensive labeled data and are task-specific, limiting their flexibility[1,19,21]. The rise of generative large language models (LLMs), like GPT[22] and LLaMA[23], pre-trained on extensive corpora, has demonstrated significant potential in various clinical applications, such as question answering[24,25] and information retrieval[26,27]. These models are increasingly applied to diagnostics. For example, PathChat[28], a vision-language LLM fine-tuned with comprehensive instructions, set new benchmarks in pathology. Similarly, Kim et al.[29] reported that GPT-4 outperformed mental health professionals in diagnosing obsessive-compulsive disorder, underscoring its potential in mental health diagnostics.

Despite growing interest, several key questions remain unresolved: Which diseases and medical data have been explored for LLM-based diagnostics (Q1)? What LLM techniques are most effective for diagnostic tasks (see Box 1), and how should they be selected (Q2)? What evaluation methods best assess performance of various diagnostic tasks (Q3)? Many reviews have explored the use of LLMs in medicine[30–37], but they typically provide broad overviews of diverse clinical applications rather than focusing specifically on disease diagnosis. For instance, Pressman et al.[38] highlighted

[1]Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, USA. [2]School of Nursing, Columbia University, New York, New York, USA. [3]Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA. [4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR. [5]Department of Informatics, University of California, Irvine, Irvine, CA, USA. [6]Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. [7]Department of Computer Science, New York University (Shanghai), Shanghai, China. [8]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA. [9]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR. [10]Independent Researcher, San Francisco, CA, USA. [11]Department of Neurology, University of Minnesota, Minneapolis, MN, USA. [12]Institute for Health Informatics and Division of Colon and Rectal Surgery, Department of Surgery, University of Minnesota, Minneapolis, MN, USA. [13]These authors contributed equally: Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu. ✉e-mail: ruizhang@umn.edu

introducing various clinical applications of LLMs, e.g., pre-consultation, treatment, and patient education. These reviews tend to overlook the nuanced development of LLMs for diagnostic tasks and do not analyze the distinct merits and challenges in this area, revealing a critical research gap. Some reviews[39,40] have focused on specific specialties,such as digestive or infectious diseases, but failed to offer a comprehensive perspective that spans multiple specialties, data types, LLM techniques, and diagnostic tasks to fully address the critical questions at hand.

This review addresses the gap by offering a comprehensive examination of LLMs in disease diagnosis through in-depth analyses. First, we systematically investigated a wide range of disease types, corresponding clinical specialties, medical data, data modalities, LLM techniques, and evaluation methods utilized in existing diagnostic studies. Second, we critically evaluated the strengths and limitations of prevalent LLM techniques and evaluation strategies, providing recommendations for data preparation, technique selection, and evaluation approaches tailored to different contexts. Additionally, we identify the shortcomings of current studies and outline future challenges and directions. To the best of our knowledge, this is the first review dedicated exclusively to LLM-based disease diagnosis, presenting a holistic perspective and a blueprint for future research in this domain.

Results

Overview of the scope

This section outlines the scope of our review and key findings. Figure 1 provides an overview of disease types, clinical specialties, data types, and modalities (Q1), and introduces the applied LLM techniques (Q2) and evaluation methods (Q3), addressing the key questions. Our analysis spans 19 clinical specialties and over 15 types of clinical data in diagnostic tasks, covering modalities such as text, image, video, audio, time series, and multimodal data. We categorized existing works based on LLM techniques, which fall into four categories: prompting, retrieval-augmented generation (RAG), fine-tuning, and pre-training, with the latter three further subdivided. Table 1 summarizes the taxonomy of mainstream LLM techniques. Figure 2 illustrates the associations between clinical specialties, modalities of utilized data, and LLM techniques in the included papers. Additionally, Fig. 3 presents a meta-analysis, covering publication trends, widely-used LLMs for training and inference, and statistics on data sources, evaluation methods, data privacy, and data sizes. Collectively, these analyses comprehensively depict the development of LLM-based disease diagnosis.

Study characteristics

As shown in Fig. 2, the included studies span all 19 clinical specialties, and some specialties receive particular attention, such as pulmonology and neurology. While most studies leveraged text modality, multi-modal data, such as text-image[41] and text-tabular data[42], are widely adopted for diagnostic tasks. Another observation is that various LLM techniques have been applied to diagnostic tasks, and all have been used with multi-modal data (Table 1). Additionally, we find an increasing number of LLM-based diagnostic studies all over the world, reflecting the field's growing significance (Fig. 3a). Among these studies, GPT[22] and LLaMA[23] families dominate inference tasks, while LLaMA and ChatGLM[43] are commonly adopted for model training (Fig. 3b). Figure 3c shows that most datasets originate from North America (50.6%) and Asia (33.9%), and 50.4% of the studies used public datasets (Fig. 3e). Evaluation methods vary: 66.8% rely on automated evaluation, 28.1% on human assessment, and 5.1% on LLM-based evaluation (Fig. 3d). Figure 3f reveals that the included studies employed large datasets (e.g., 5×10^5 samples) for pre-training diagnostic models, surpassing those primarily using fine-tuning or RAG. This phenomenon aligns with another observation that over half of pre-training models used data from multiple specialties.

Prompt-based disease diagnosis

A customized prompt typically includes four components: instruction (task specification), context (scenario or domain), input data (data to process), and output indicators (desired style or role). In this review, over 60% (N = 278) of studies employed prompt-based techniques, categorized as hard prompts and soft prompts. Hard prompts are static, interpretable, and written in natural language. The most common methods included zero-shot (N = 194), Chain-of-Thought (CoT) (N = 37), and few-shot prompting (N = 35). Among them, CoT prompting excels in thoroughly digesting input clinical cues in manageable steps to make a coherent diagnosis decision. Particularly, in differential diagnosis tasks, CoT reasoning allows the LLM to sequentially analyze medical images, radiology reports, and clinical history, generating intermediate outputs that lead to a holistic decision, with an accuracy of 64%[44]. Self-consistency prompting was used in a few studies (N = 4). For instance, a study combined self-consistency with CoT prompting to improve depression prediction by synthesizing diverse data sources through multiple reasoning paths. This hybrid approach reduced the mean absolute error by nearly 50% compared to standard CoT methods[45].

Box 1 | Terms and Concepts

Disease diagnosis: receiving clinical data, such as patient symptoms, medical history, and diagnostic tests, as input and identifying which disease explains the symptoms and signs.
Diagnostic tasks: a type of tasks that generate disease diagnoses or probability estimates for specific conditions, such as differential diagnosis and conversational diagnosis.
Large language models: a type of AI models using deep neural networks to learn the relationships between words in natural language, using large datasets of text to train.
Hallucination: an AI-generated output that is plausible but factually incorrect or unrelated to the input, arising from limitations in training or reasoning.
Prompt: an input or instruction provided to an AI model to guide its response, often designed to elicit specific or task-relevant outputs.
Chain-of-thought: a technique enabling AI to generate multi-step reasoning by breaking down complex tasks into sequential steps for improved accuracy.
Self-consistency prompt: a method that samples diverse reasoning paths and selects the most consistent solution to enhance the reliability of outputs in reasoning tasks.
Soft prompt: a learnable embedding added to the input space of a pre-trained model to guide its behavior without modifying the model's parameters.
Retrieval-augmented generation: integrates retrieved data into LLMs, enhancing responses by leveraging external information for improved context and accuracy in content generation.
Fine-tuning: the process of adapting a pre-trained model to a specific task by training it further on a smaller, task-specific dataset.
Supervised fine-tuning: refining a pre-trained model for a task using labeled data to enhance task-specific performance.
Parameter-efficient fine-tuning: adapting pre-trained models to new tasks by updating limited parameters (e.g., adapters), reducing computational costs while preserving performance.
Reinforcement learning from human feedback: a method where models improve outputs by learning from human-provided feedback, aligning behavior with human goals through reinforcement learning.
Pre-training: the foundational training phase of a model on a large, general dataset to learn broad patterns, features, and representations, which can later be adapted to specific tasks through fine-tuning.

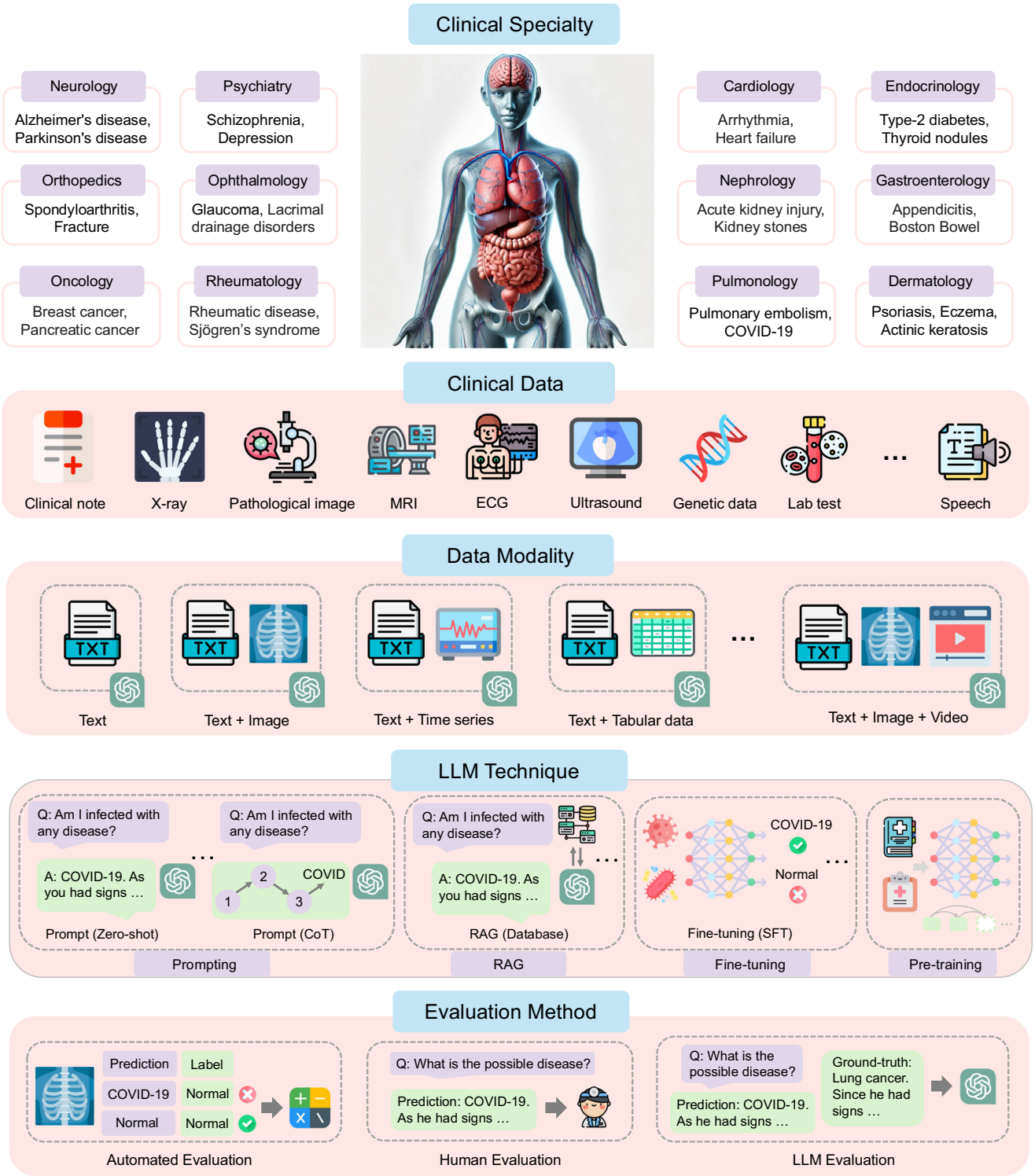

Fig. 1 | Overview of the investigated scope. It illustrated disease types and the associated clinical specialties, clinical data types, modalities of the utilized data, the applied LLM techniques, and evaluation methods. We only presented part of the clinical specialties, some representative diseases, and partial LLM techniques.

In contrast, soft prompts ($N = 6$) are continuous vector embeddings trained to adapt the behavior of LLMs for specific tasks[46]. These prompts effectively integrate external knowledge, such as medical concept embeddings and clinical profiles, making them well-suited for complex diagnostic tasks requiring nuanced analysis. This knowledge-enhanced approach achieved F1 scores exceeding 0.94 for diagnosing common diseases like hypertension and coronary artery disease and demonstrated superiority in rare disease diagnosis[47].

Most prompt-based studies ($N = 221$) focused on unimodal data, predominantly text ($N = 171$). Clinical text sources like clinical notes[48], imaging reports[49–51], and case reports[52,53] were commonly used. These studies often prompted LLMs with clinical notes or case reports to predict potential diagnoses[54–57]. A smaller subset ($N = 19$) applied prompt engineering to medical image data, analyzing CT scans[58], X-rays[59,60], MRI scans[58,61], and pathological images[62,63] to detect abnormalities and provide evidence for differential diagnoses[62,64–66].

Table 1 | Overview of LLM techniques for diagnostic tasks

Techniques	Types	Representative studies
Prompting	Zero-shot	Text[196,197], image[65,198], audio[70,72], text-image[52], text-time series[73,199], text-tabular[200]
	Few-shot	Text[25,187], image[58], text-image[41,201], text-image-tabular[153]
	CoT	Text[51,202], audio[203], time series[155], text-image[44,204]
	Self-consistency	Text[89], audio[205], text-image-tabular-time series[45]
	Soft prompt	Text[206], image[207], tabular-time series[47,208], text-image-graph[59]
RAG	Knowledge graph	Text[81], text-time series[94]
	Corpus	Text[85,87], text-image[64,86], text-time series[83]
	Database	Text[80,93], text-image[90]
Fine-tuning	SFT	Text[98,209,210], text-image[133,211,212], text-video[102,112], text-audio[111,213], text-tabular[42,200]
	RLHF	Text[116,117,214], text-image[115]
	PEFT	Text[98,124,215], text-image[104]
Pre-training	-	Text[124,129,131], text-image[109,133,137], text-tabular[135,200], text-video[213], text-omics[109]

SFT supervised fine-tuning, *RLHF* reinforcement learning from human feedback, *PEFT* parameter-efficient fine-tuning.

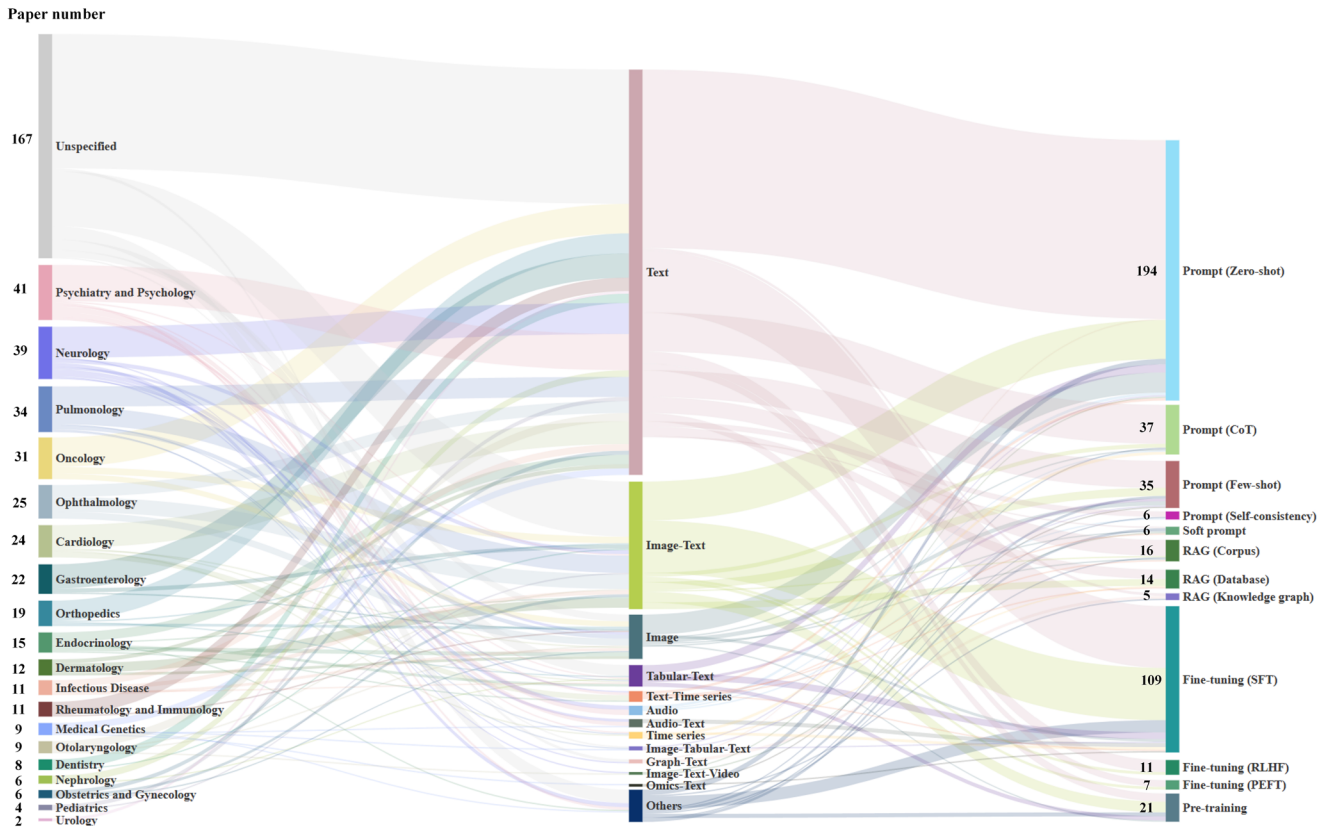

Fig. 2 | Summary of the association between clinical specialties (left), data modalities (middle), and LLM techniques (right) across the included studies on disease diagnosis.

With the advancement of multimodal LLMs, 57 studies explored their application in disease diagnosis through prompt engineering. Visual-language models (VLMs) like GPT-4V, LLaVA, and Flamingo (N = 37) integrated medical images (e.g., radiology scans) with textual descriptions (e.g., clinical notes)[67–69]. For example, incorporating ophthalmologist feedback and contextual details with eye movement images significantly improved GPT-4V's diagnostic accuracy for amblyopia[64].

Beyond image-text data, more advanced multimodal LLMs (e.g., GPT-4o and Gemini-1.5 Pro) have also integrated other data types to support disease diagnosis in complex clinical scenarios. Audio and video data have been used to diagnose neurological and neurodegenerative disorders, such as autism[70,71] and dementia[59,72]. Time-series data, such as ECG signals and wearable sensor outputs, were used to support arrhythmia detection[73,74]. With the integration of tabular data such as user demographics[75,76], and lab test results[47,77], the applications have been extended to depression and anxiety screening[45]. Omics data has been integrated to aid in identifying rare genetic disorders[78] and diagnose Alzheimer's disease[76]. Some studies further enhanced diagnostic capabilities by integrating medical concept graphs to provide a richer context for conditions such as neurological disorders[59].

Retrieval-augmented LLMs for diagnosis

To enhance the accuracy and credibility of the diagnosis, alleviate hallucination issues, and update LLMs' stored medical knowledge without needing re-training, recent studies[79–81] have incorporated external medical

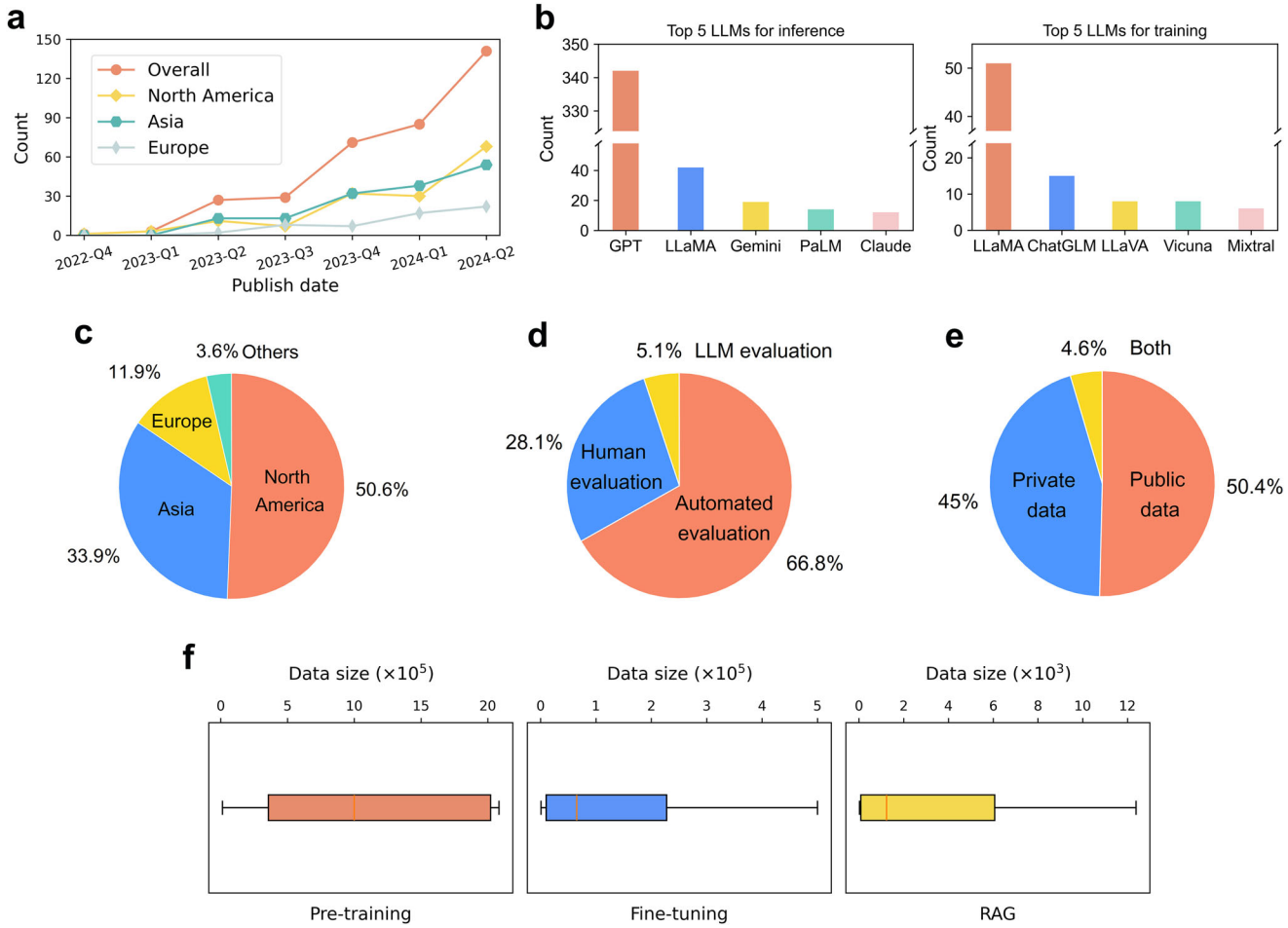

Fig. 3 | Metadata of information from LLM-based diagnostic studies in the scoping review. a Quarterly breakdown of LLM-based diagnostic studies. Since the information for 2024-Q3 is incomplete, our statistics only cover up to 2024-Q2. **b** The top 5 widely-used LLMs for inference and training. **c** Breakdown of the data source by regions. **d** Breakdown of evaluation methods (note that some papers utilized multiple evaluation methods). **e** Breakdown of the employed datasets by privacy status. **f** Distribution of data size used for LLM techniques. The red line indicates the median value, while the box limits represent the interquartile range (IQR) from the first to third quartiles. Notably, pre-trained diagnostic models were often followed by other LLM techniques (e.g., fine-tuning), yet this figure only includes studies that primarily used fine-tuning or RAG. Statistics for prompting methods are not included because: (*i*) hard prompts generally utilize zero or very few demonstration samples, and (*ii*) although soft prompts require more training data, the number of relevant studies is insufficient for meaningful distribution analysis.

knowledge into diagnostic tasks. The external knowledge primarily comes from corpus[64,79,82–88], databases[74,80,89–93], and knowledge graph[81,94], in the included papers. Based on the data modality, these RAG-based studies can be roughly categorized into text-based, text-image-based, and time-series-based augmentations.

In text-based RAG, most studies[80,82,84,85,91–93] utilized basic retrieval methods where external knowledge was encoded as vector representations using sentence transformers, such as OpenAI's text-embedding-ada-002. Queries were similarly encoded, and relevant knowledge was retrieved based on vector similarities. The retrieved data was then input into LLMs with specific prompts to produce diagnostic outcomes. In contrast, Li et al.[88] developed guideline-based GPT agents for retrieving and summarizing content related to diagnosing traumatic brain injury. They found that these guideline-based GPT-4 agents significantly outperformed the off-the-shelf GPT-4 in terms of accuracy, explainability, and empathy evaluation. Similarly, Thompson et al.[79] employed regular expressions to extract relevant knowledge for diagnosing pulmonary hypertension, achieving about a 20% improvement compared to structured methods. Additionally, Wen et al.[81] integrated knowledge graph retrieval with LLMs to enable diagnostic inference by combining implicit and external knowledge, achieving an F1 score of 0.79.

In text-image data processing, a common approach[87,91] involved extracting image features and text features and aligning them within a shared semantic space. For instance, Ferber et al.[91] used GPT-4V to extract crucial image data for oncology diagnostics, achieving a 94% completeness rate and an 89.2% helpfulness rate. Similarly, Ranjit et al.[87] utilized multi-modal models to compute image-text similarities for chest X-ray analysis, leading to a 5% absolute improvement in the BERTScore metric. Notably, one study fine-tuned LLMs with retrieved documents to enhance X-ray diagnostics[86], attaining an average accuracy of 0.86 across three datasets.

For time-series RAG, most studies focused on the electrocardiogram (ECG) analysis[74,83]. For example, Yu et al.[83] transformed fundamental ECG conditions into enhanced text descriptions by utilizing relevant information for ECG analysis, resulting in an average AUC of 0.96 across two arrhythmia detection datasets. Additionally, Chen et al.[95] integrated retrieved disease records with ECG data to facilitate the diagnosis of hypertension and myocardial infarction.

Fine-tuning LLMs for diagnosis

Fine-tuning an LLM typically encompasses two pivotal stages: supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). SFT trains models on task-specific instruction-response pairs, enabling it to interpret instructions and generate outputs across diverse modalities. This phase establishes a foundational understanding, ensuring the model processes inputs effectively. RLHF further refines the model by aligning its behavior with human preferences. Using reinforcement

learning, the model is optimized to produce responses that are helpful, truthful, and aligned with societal and ethical standards[96].

In medical applications, SFT enhances in-context learning, reasoning, planning, and role-playing capabilities, improving diagnostic performance. This process integrates inputs from various data modalities into the LLM's word embedding space. For example, following the LLaVA approach[97], visual data is converted into token embeddings using an image encoder and projector, then fed into the LLM for end-to-end training. In this review, 49 studies focused on SFT using medical texts, such as clinical notes[98], medical dialogs[99–101], or reports[102–104]. Additionally, 43 studies combined medical texts with images, including X-rays[102,105–107], MRIs[104,107,108], or pathology images[109–111]. A few studies explored disease detection from medical videos[102,112], where video frames were sampled and converted into visual token embeddings. Generally, effective SFT requires collecting high-quality, diverse responses to task-specific instructions to ensure comprehensive training.

RLHF methods are categorized as online or offline. Online RLHF, integral to ChatGPT's success[113], involves training a reward model on datasets of prompts and human preferences and using reinforcement learning algorithms like Proximal Policy Optimization (PPO)[114] to optimize the LLM. Studies have shown its potential in improving medical LLMs for diagnostic tasks[115–117]. For instance, Zhang et al.[117] aligned their model with physician characteristics, achieving strong performance in disease diagnosis and etiological analysis; the diagnostic performance of their model, HuatuoGPT, surpassed GPT-3.5 in over 60% of cases of Meddialog[118]. However, online RLHF's effectiveness depends heavily on the reward model's quality, which may suffer from over-optimization[119] and data distribution shifts[120]. Additionally, reinforcement learning often faces instability and control challenges[121]. Offline RLHF, such as Direct Preference Optimization (DPO)[122], frames RLHF as optimizing a classification loss, bypassing the need for a reward model. This approach is more stable and computationally efficient, proving valuable for aligning medical LLMs[123,124]. Yang et al.[124] reported significant performance drops on pediatric benchmarks when the offline RLHF phase was omitted. A high-quality dataset of prompts and human preferences is essential for online RLHF reward model calibration[125] or the convergence of offline methods like DPO[126], whether sourced from experts[113] or advanced AI models[127].

Since full training of LLMs is challenging due to high GPU demands, parameter-efficient fine-tuning (PEFT) reduces the number of tunable parameters. The most common PEFT method, Low-Rank Adaptation (LoRA)[128], introduces trainable rank decomposition matrices into each layer without altering the model architecture or adding inference latency. In this review, all PEFT-based studies (N = 7) used LoRA to reduce training costs[98,104,124].

Pre-training LLMs for diagnosis

Pre-training medical LLMs involves training on large-scale, unlabeled medical corpora to develop a comprehensive understanding of the structure, semantics, and context of medical language. Unlike fine-tuning, pre-training enables the acquisition of extensive medical knowledge, enhancing generalization to unseen cases and improving robustness across diverse diagnostic tasks. In this review, five studies performed text-only pretraining on the LLMs from different sources[129–132], such as clinical notes, medical QA texts, dialogs, and Wikipedia. Moreover, eight studies injected medical visual knowledge into multimodal LLMs via pretraining[109,133–137]. For instance, Chen et al.[137] employed an off-the-shelf multimodal LLM to reformat image-text pairs from PubMed into VQA data points for training their diagnostic model. To improve the quality of the image encoder, pre-training tasks like reconstructing images at tile-level or slide-level[109], and aligning similar images or image-text pairs[133] are common choices.

Performance evaluation

Evaluation methods for diagnostic tasks generally fall into three categories (Table 2): automated evaluation[138], human evaluation[138], and LLM evaluation[139], each with distinct advantages and limitations (Fig. 4).

In this review, most studies (N = 266) relied on automated evaluation, which is efficient, scalable, and well-suited for large datasets. These metrics can be grouped into three types. (1) Classification-based metrics, such as accuracy, precision, and recall, are commonly used for disease diagnosis. For instance, Liu et al.[133] evaluated COVID-19 diagnostic performance using AUC, accuracy, and F1 score. (2) Differential diagnosis metrics, including top-k precision, assess ranked diagnosis lists. Tu et al.[140] employed top-k accuracy to evaluate the correctness of differential diagnosis predictions. (3) Regression-based metrics, such as mean squared error (MSE)[141], quantify deviations between predicted and actual values[142]. Despite their efficiency, automated metrics rely on ground-truth diagnoses[143], which may be unavailable, and cannot understand contexts, such as the readability of diagnostic explanations or their clinical utility[144]. They also struggle with complex tasks, such as evaluating the medical correctness of diagnostic reasoning[145].

Human evaluation (N = 112), conducted by medical experts[24,138], does not require ground-truth labels and integrates expert judgment, making it suitable for complex, nuanced assessments. However, it is costly, time-consuming, and prone to subjectivity, limiting its feasibility for large-scale evaluation. Recent studies have explored using LLM evaluation (N = 20), a.k.a. LLM-as-Judges[139], to replace human experts in evaluation and combine the interpretative depth of LLM judgment with the efficiency of automated evaluation. While ground-truth accessibility is not strictly necessary[99,116], its inclusion improves reliability[143]. Popular LLMs used for this purpose include GPT-3.5, GPT-4, and LLaMA-3. However, this approach remains constrained by LLM limitations, including susceptibility to hallucinations[99] and difficulties in handling complex diagnostic reasoning[146]. In summary, each evaluation approach has distinct advantages and limitations, with the choice dependent on the specific requirements of the task. Figure 4 guides the selection of suitable evaluation approaches for different scenarios.

Discussion

This section analyzes key findings from the included studies, discusses the suitability of mainstream LLM techniques for varying resource constraints and data preparation, and outlines challenges and future research directions.

The rapid rise of LLM-based diagnosis studies (Fig. 3a) might partially be attributed to the increased availability of public datasets[147] and advanced off-the-shelf LLMs[57]. Besides, the top five LLMs used for training and inference differ significantly (Fig. 3b), reflecting the interplay between effectiveness and accessibility. Generally, closed-source LLMs, with their vast parameters and superior performance[143], are favored for LLM inference, while open-source LLMs are essential for developing domain-specific models due to their adaptability[148]. These factors underscore the dual influence of effectiveness and accessibility on diagnostic applications. Additionally, the regional analysis of datasets (Fig. 3c) reveals that 84.5% of datasets originate from North America and Asia, potentially introducing racial biases in this research domain[149].

Most studies employed prompting for disease diagnosis (Fig. 2), leveraging its advantages, such as minimal data requirements, ease of use, and low computational demands[150]. Meanwhile, LLMs' extensive medical knowledge allowed them to perform competitively across diverse diagnostic tasks when effectively applied[24,143]. For example, a study fed two data samples into GPT-4 for depression detection[151], and the performance significantly exceeded traditional DL-based models. In summary, prompting LLMs facilitates the development of effective diagnostic systems with minimal effort, contrasting with conventional DL-based approaches that require extensive supervised training on large datasets[2,17].

We then compare the advantages and limitations of mainstream LLM techniques to indicate their suitability for varying resource constraints, along with a discussion of data preparation. Generally, the choice of LLM technique for diagnostic systems depends on the quality and quantity of available data. Prompt engineering is particularly effective in few-data scenarios (e.g., zero or three cases with ground-truth diagnoses), requiring minimal setup[24,152]. RAG relies on a high-quality external knowledge base,

Table 2 | Overview of evaluation metrics for diagnostic tasks

Type	Evaluation metric	Purpose	Scenario	Representative task
Automated evaluation	Accuracy[216]	The ratio of all correct predictions to the total predictions	G	DD[154], DDx[217], CD[218], RP[219], DRG[105], MHD[220]
	Precision[55]	The ratio of true positives to the total number of positive predictions	G	DD[55], CD[221], MIC[44], RP[219], DRG[105]
	Recall[55]	The ratio of true positives to the total number of actual positive cases	G	DD[55], CD[221], RP[219], DRG[105]
	F1[133]	Calculated as the harmonic mean of precision and recall	G	DD[55], DDx[222], CD[221], MIC[223], RP[219], DRG[105]
	AUC[224]	The area under the Receiver Operating Characteristic curve	G	DD[59], CD[225], MIC[226], RP[219], DRG[105], MHD[227]
	AUPR[228]	The area under the precision-recall curve	G	DD[229], MIC[228], RP[230], DRG[229]
	Top-k accuracy[140]	The ratio of instances with the true label in the top k predictions to total instances	G	DD[140], DDx[168]
	Top-k precision[60]	The ratio of true positives to total positive predictions within the top k predictions	G	DD[140], DDx[222]
	Top-k recall[231]	The ratio of true positives within the top k predictions to actual positive cases	G	DD[140], DDx[222]
	Mean square error[142]	The average of the squared differences between predicted and actual values	G	DD[142], RP[141]
	Mean absolute error[141]	The average of the absolute differences between predicted and actual values	G	DD[142], RP[141]
	Cohen's κ[232]	Measure the agreement between predicted score and actual score	G	DD[232]
	BLUE[115]	Calculate precision by matching n-grams between reference and generated text	T	DD[233], CD[234], MIC[235], DRG[115]
	ROUGE[187]	Calculate F1-score by matching n-grams between reference and generated text	T	DD[233], CD[187], MIC[235], DRG[115]
	CIDEr[102]	Evaluate n-gram similarity, emphasizing alignment across multiple reference texts	T	CD[102], MIC[236], DRG[237]
	BERTScore[81]	Measure similarity by comparing embeddings of reference and generated text	T	DD[238], DDx[143], CD[187], DRG[87]
	METEOR[234]	Evaluate text similarity by considering precision, recall, word order, and synonym matches	T	DDx[143], CD[234], MIC[236], DRG[115]
Human evaluation	Necessity[187]	Whether the response or prediction assists in advancing the diagnosis	T	CD[187]
	Acceptance[239]	The degree of acceptance of the response without any revision	T	DD[54], CD[240]
	Reliability[176]	The trustworthiness of the evidence in the response or prediction	T	DD[144], CD[176]
	Explainability[88]	Whether the response or prediction is explainable	T	DDx[241], CD[218]
Human or LLM evaluation	Correctness[242]	Whether the response or prediction is medically correct	T	DD[134], DDx[217], CD[187], DRG[243], MHD[176]
	Consistency[99]	Whether the response or prediction is consistent with the ground-truth or input	T	DD[108], DDx[241], CD[99], MHD[176]
	Clarity[80]	Whether the response or prediction is clearly clarified	T	DD[149], CD[244]
	Professionality[176]	The rationality of the evidence based on domain knowledge	T	CD[149], MHD[176]
	Completeness[187]	Whether the response or prediction is sufficient and comprehensive	T	DDx[143], CD[218], DRG[243]
	Satisfaction[245]	Whether the response or prediction is satisfying	T	CD[240], DRG[237]
	Hallucination[99]	Response contains inconsistent or unmentioned information with previous context	T	DDx[222], CD[218], DRG[246]
	Relevance[80]	Whether the response or prediction is relevant to the context	T	CD[80], DRG[246]
	Coherence[247]	Assess logical consistency with the dialog history	T	CD[100], DRG[190]

Since diagnostic tasks might include explanations alongside the predicted diagnosis, existing studies also evaluated these explanatory descriptions. We categorized the metrics based on their application scenarios: G denotes that the metric requires ground-truth diagnosis for evaluation, while T indicates those applicable to textual descriptions (e.g., generated explanations). Notably, we only present a selection of representative diagnostic tasks from the included papers: disease diagnosis (DD), differential diagnosis (DDx), conversational diagnosis (CD), medical image classification (MIC), risk prediction (RP), mental health disorder detection (MHD), and diagnostic report generation (DRG).

Evaluation Approach	Accuracy & Stability	Cost-Effectiveness	Comprehension & Handling Complex Tasks	Representative Papers
Automated Evaluation	Accurate. Strong reproducibility. Quantifiable metrics.	Minimal evaluation cost. The one-time cost of ground-truth preparation. Scalable.	Inability to understand context. Inadaptable to complex tasks and scenarios.	55, 133, 140, 143, 224, 228
Human Evaluation	Accurate. Subject to individual bias.	Costly. Labor-intensive.	Remarkable context understanding. Adaptable to complex tasks and scenarios.	54, 144, 145, 176, 187, 239
LLM Evaluation	LLM-dependent accuracy. Risk hallucination. LLM-dependent stability.	Moderate cost. Scalable. Require expert design.	Moderate context understanding. Limited adaptability for complex tasks and scenarios.	80, 99, 116, 143, 176, 217

Fig. 4 | Summary of the evaluation approaches for diagnostic tasks.

such as databases[80] or corpora[82], to retrieve accurate information during inference. Fine-tuning requires well-annotated datasets with sufficient labeled diagnostic cases[133]. Pre-training, by contrast, utilizes diverse corpora, including unstructured text (e.g., clinical notes, literature) and structured data (e.g., lab results), to establish a robust knowledge foundation via unsupervised language modeling[42,153]. Although fine-tuning and pre-training facilitate high performance and reliability[133], they demand significant resources, including advanced hardware and extensive biomedical data (see Fig. 3f), which are costly and often hard to obtain[24]. In practice, not all diagnostic scenarios require expert-level accuracy. Applications such as large-scale screenings[154], mobile health risk alerts[155], or public health education[30] prioritize cost-effectiveness and scalability. Overall, balancing accuracy with resource constraints depends on the specific use case.

Despite advances in LLM-based methods for disease diagnosis, this scoping review highlighted several barriers to their clinical utility (Fig. 5). One limitation lies in information gathering. Most studies implicitly assume that the available patient information is sufficient for diagnosis, which often fails[156], especially in initial consultations or with complex diseases, increasing the risk of misdiagnosis[157]. In practice, clinical information gathering is iterative, starting with initial data (e.g., subjective symptoms), refining diagnoses, and conducting further tests or screenings[158]. This process relies heavily on experienced clinicians[140]. To reduce this dependence, recent studies have explored multi-round diagnostic dialogs to collect relevant information[159,160]. For example, AIME[140] uses LLMs for clinical history-taking and diagnostic dialog, while Sun et al.[160] utilized reinforcement learning to formulate disease screening questions. Future efforts could further embed awareness of information incompleteness into models or develop techniques for automatic diagnostic queries[161]. Another limitation arises from the reliance on single data modalities, whereas clinicians typically synthesize information from multiple modalities for accurate diagnosis[44]. Additionally, real-world health systems often operate in isolated data silos, with patient information distributed across institutions[26]. Addressing these issues will require efforts to collect and integrate multi-modal data and establish unified health systems that facilitate seamless data sharing across institutions[162].

Barriers also exist in the information integration process. Some studies utilized clinical vignettes for diagnostic tasks without fulfilling the SOAP standard[163]. While adhering to clinical guidelines is crucial[142], limited studies have incorporated this factor into diagnostic systems[164]. For example, Kresevic et al.[82] sought to enhance clinical decision support systems by accurately explaining guidelines for chronic Hepatitis C management. Besides, the integration and interpretation of lab test results pose significant value in healthcare[165]. For example, Bhasuran et al.[166] reported that incorporating lab data enhanced the diagnostic accuracy of GPT-4 by up to 30%. A future direction is the effective integration of lab test results into LLM-based diagnostic systems.

Exploring clinician-patient-diagnostic system interactions offers a promising research direction[167]. Diagnostic systems are desired to assist clinicians by providing Supplementary information to improve accuracy and efficiency[58,168], incorporating expert feedback for continuous refinement. A user-friendly interface is essential for effective human-machine interaction, enabling clinicians to input data and engage in discussions with the system. Human language interaction further enhances usability by allowing natural conversation with LLM-based diagnostic tools[168], reducing cognitive load. Additionally, LLM-aided explanations improve transparency by providing rationales for suggested diagnoses[145], fostering trust, and facilitating informed decision-making among clinicians and patients.

Most of the studies focused on diagnostic accuracy, but overlooked ethical considerations, like explainability, trustworthiness, privacy protection, and fairness[169]. Providing diagnostic predictions alone is insufficient in clinical scenarios, as the black-box nature of LLMs often undermines trust[99]. Designing diagnostic models with explainability is desired[145]. For example, Dual-Inf is a prompt-based framework that offers potential diagnoses while explaining its reasoning[143]. Besides, since LLMs suffer from hallucinations, how to enhance users' trustworthiness toward LLM-based diagnostic models is worth exploring[170]. Potential solutions include using fact-checking tools to verify the output's factuality[171]. Regarding privacy, adherence to regulations like HIPAA and GDPR, including de-identifying sensitive data, is essential[26,172]. For example, SkinGPT-4, a dermatology diagnostic system, was designed for local deployment to ensure privacy protection[173]. Fairness is another concern, as patients should not face discrimination based on gender, age, or race[169], but research on fairness in LLM-based diagnostics remains scarce[174].

In the context of modeling, building superior models for accurate and reliable diagnosis remains an exploration. While pre-training on extensive medical datasets benefits diagnostic reasoning[175], many medical LLMs generally lag behind general-domain counterparts in parameter scale[148,176], underscoring the potential of developing large-scale generalist models for disease diagnosis. Besides, LLMs are prone to catastrophic forgetting[177], where previously acquired knowledge or skills are lost when learning new information. Addressing this issue facilitates the development of generalist diagnostic models but requires incorporating robust continuous learning capabilities[178]. One alternative approach for accurate diagnosis involves coordinating multiple specialized models, simulating interdisciplinary clinical discussions to tackle complex cases[179]. For example, Med-MoE[180] is a mixture-of-experts framework leveraging medical texts and images and achieved an accuracy of 91.4% in medical image classification. Additionally, hallucinations in LLMs undermine diagnostic reliability[170], necessitating solutions such as knowledge editing[181], external knowledge retrieval[82], and novel model architectures or pre-training strategies[175]. Another promising avenue is longitudinal data modeling, as clinicians routinely analyze EHRs

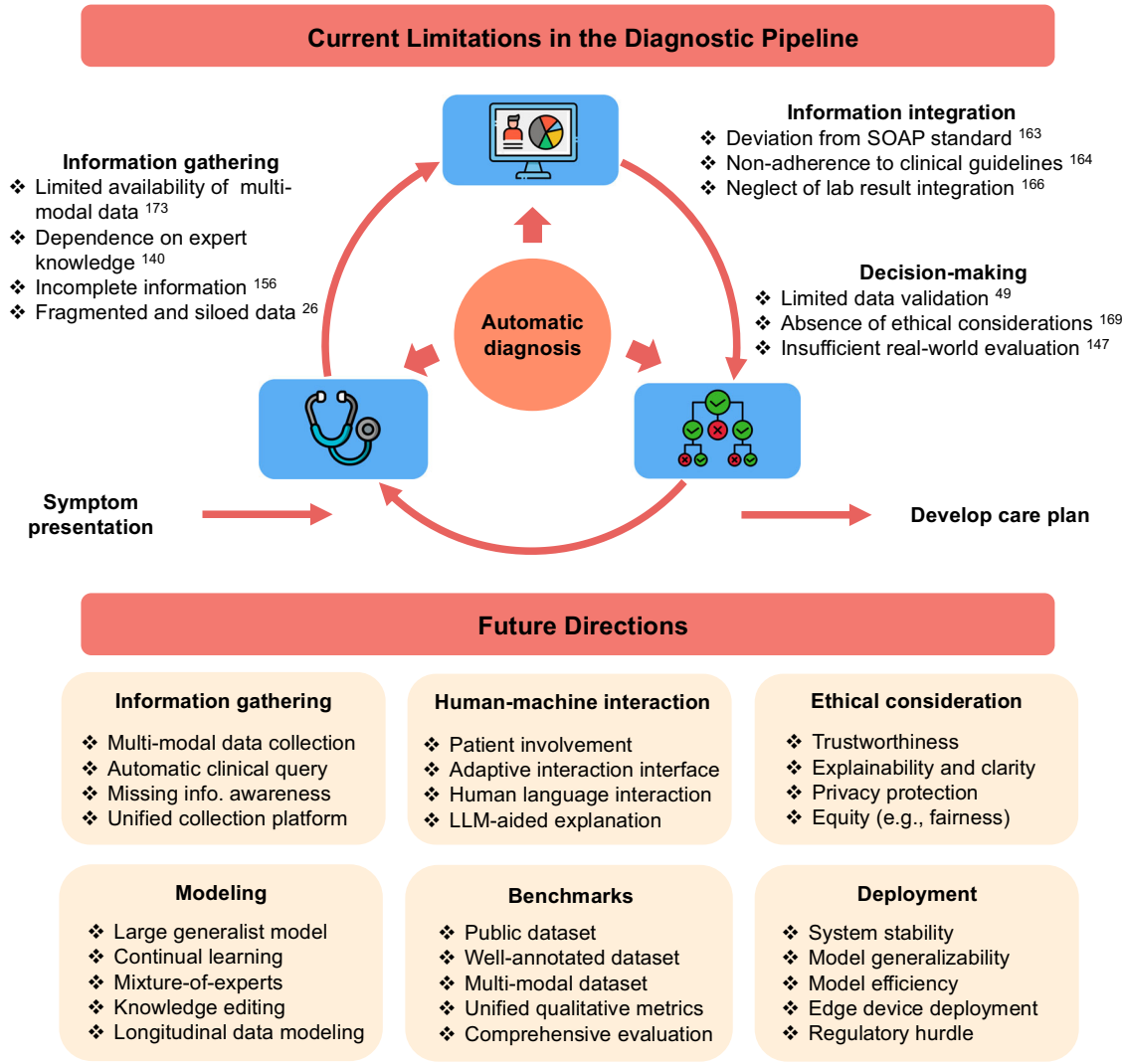

Fig. 5 | Summary of the limitations and future directions for LLM-based disease diagnosis.

spanning multiple years to inform decision-making[182,183]. Besides, modeling temporal data helps with early diagnosis[56,184] to improve patient outcomes. For example, early detection of lung adenocarcinoma might increase the 5-year survival rate to 52%[185]. However, challenges like irregular sampling intervals and missing data persist[186], necessitating advanced methodologies to effectively capture temporal dependencies[25].

Another challenge in developing diagnostic models is benchmark availability[147]. In this review, 49.6% of the included studies relied on private datasets, which were often inaccessible due to privacy concerns[82]. Additionally, the scarcity of annotated data limits progress, as well-annotated datasets with ground-truth diagnosis enable automated evaluation, reducing reliance on human assessment[143]. Hence, constructing and releasing annotated benchmark datasets would greatly support the research community[147]. Regarding performance evaluation, some studies either used small-scale data[57] or unrealistic data, such as snippets from college books[145] and LLM-generated clinical notes[147], for disease diagnosis, while large-scale real-world data can truly validate diagnostic capabilities[182]. Besides, the lack of unified qualitative metrics is another issue. For example, the evaluation of diagnostic explanation varies in different studies[143,187], including necessity[187], consistency[108], and compeleteness[143]. Unifying qualitative metrics foster a fair comparison. Additionally, many included studies failed to compare with conventional diagnostic models, while recent studies reported that traditional models, e.g., Transformer[188], might beat LLM-based counterparts in clinical prediction[189]. Therefore, future studies should compare with traditional baselines for comprehensive evaluation.

Regarding the deployment of diagnostic systems, several challenges warrant further investigation, including model stability, generalizability, and efficiency. Current studies have highlighted that LLMs often struggle with diagnosis stability[182], fail to generalize well across data from different institutions[190], and encounter efficiency limitations[191]. For instance, even minor variations in instructions, such as from asking "final diagnosis" to "primary diagnosis", can drop the accuracy 10.6% on cholecystitis diagnosis[182]. Addressing these limitations will advance the reliability and applicability of diagnostic models. Another promising avenue is deploying diagnostic algorithms on edge devices[192]. Such systems could enable the real-time collection of health data, such as ECG rhythms[19], to support continuous health monitoring[95]. However, regulatory barriers, including the stringent approval standards imposed by agencies such as the U.S. Food and Drug Administration (FDA) and the European Union's Medical Device Regulation (MDR)[193], remain a significant obstacle to clinical adoption. Overcoming these challenges will be vital to ensure the safe and effective integration of LLM-based diagnostics into clinical practice.

In conclusion, our study provided a comprehensive review of LLM-based methods for disease diagnosis. Our contributions were multifaceted. First, we summarized the disease types, the associated clinical specialties, clinical data, the employed LLM techniques, and evaluation methods within this research domain. Second, we compared the advantages and limitations

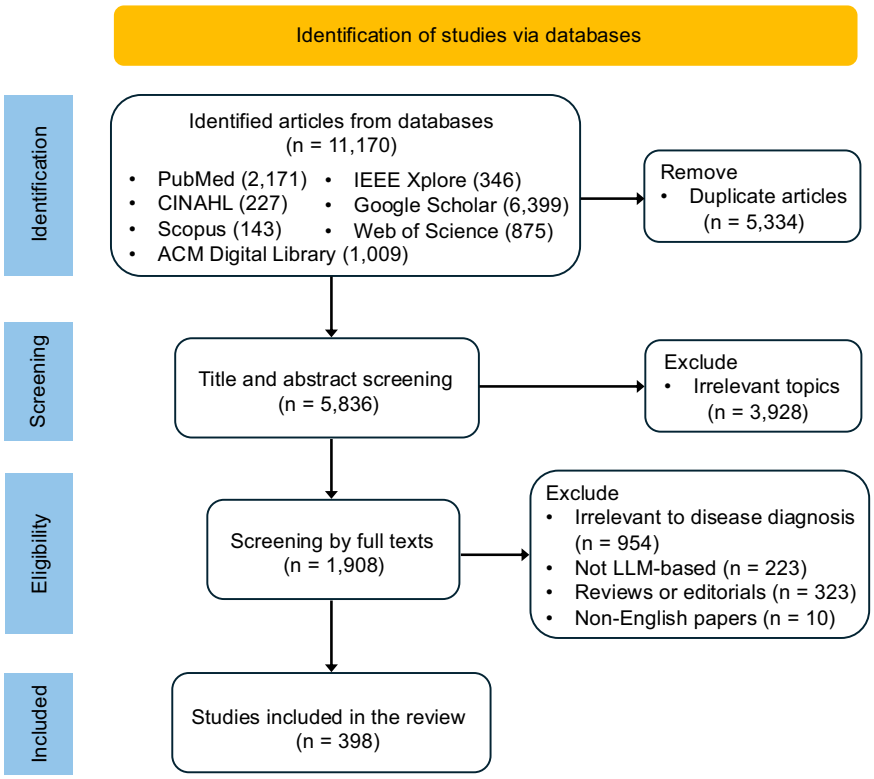

Fig. 6 | PRISMA flowchart of study records. PRISMA flowchart showing the study selection process.

of mainstream LLM techniques and evaluation methods, offering recommendations for developing diagnostic systems based on varying user demands. Third, we identified intriguing phenomena from the current studies and provided insights into their underlying causes. Lastly, we analyzed the current challenges and outlined the future directions of this research field. In summary, our review presented an in-depth analysis of LLM-based disease diagnosis, outlined its blueprint, inspired future research, and helped streamline efforts in developing diagnostic systems.

Methods

Search strategy and selection criteria

This scoping review followed the PRISMA guidelines, as shown in Fig. 6. We conducted a literature search for relevant articles published between January 1, 2019, and July 18, 2024, across seven electronic databases: PubMed, CINAHL, Scopus, Web of Science, Google Scholar, ACM Digital Library, and IEEE Xplore. Search terms were selected based on expert consensus (see Supplementary Data 1).

A two-stage screening process focused on LLMs for human disease diagnosis. The first stage involved title and abstract screening by two independent reviewers, excluding papers based on the following criteria: (a) articles unrelated to LLMs or foundation models, and (b) articles irrelevant to the health domain. The second stage was full-text screening, emphasizing language models for diagnosis-related tasks (Supplementary Data 2), excluding non-English articles, review papers, editorials, and studies not explicitly focused on disease diagnosis. The scope included studies that predicted probability values of diseases (e.g., the probability of depression) and the studies in which the foundation models involved text modalities (e.g., vision-language models) and utilized non-text data (e.g., medical images) as input. Our review excluded the foundation models without text modality, such as vision foundation models, because the scope highlighted "language" models. Following related works[194], we further excluded studies purely built on non-generative language models, like BERT[188] and RoBERTa[195], since the generative capability is a critical characteristic of LLMs to facilitate the development of the diagnostic system in the era of generative AI[30,31]. Final eligibility was determined by at least two independent reviewers, with disagreements resolved by consensus or a third reviewer.

Data extraction

Information from the articles was categorized into four groups: (1) Basic information: title, publication venue, publication date (year and month), and region of correspondence. (2) Data-related information: data sources (continents), dataset type, modality (e.g., text, image, video, text-image), clinical specialty, disease name, data availability (private or public), and data size. (3) Model-related information: base LLM type, parameter size, and technique type. (4) Evaluation: evaluation scheme (e.g., automated or human) and evaluation metrics (e.g., accuracy, precision). See Supplementary Table 1 for the data extraction form.

Data synthesis

We synthesized insights from the data extraction to highlight key themes in LLM-based disease diagnosis. First, we presented the review scope, covering disease-associated clinical specialties, clinical data, data modalities, and LLM techniques. We also analyzed meta-information, including development trends, the most widely used LLMs, and data source distribution. Next, we summarized various LLM-based techniques and evaluation strategies, discussing their strengths and weaknesses and offering targeted recommendations. We categorized modeling approaches into four areas (prompt-based methods, RAG, fine-tuning, and pre-training), with detailed subtypes. Additionally, we examined challenges in current research and outlined potential future directions. In summary, our synthesis covered data, LLM techniques, performance evaluation, and application scenarios, in line with established reporting standards.

Data availability

The analyzed data are included in this article. Aggregate data analyzed in this study is available at https://github.com/betterzhou/Awesome-LLM-Disease-Diagnosis

Received: 22 December 2024; Accepted: 17 May 2025;
Published online: 09 June 2025

References

1. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908 (2020).
2. Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. *Nat. Med.* **26**, 1224–1228 (2020).
3. Li, X. et al. Artificial intelligence-assisted reduction in patients' waiting time for outpatient process: a retrospective cohort study. *BMC health Serv. Res.* **21**, 1–11 (2021).
4. Li, B. et al. The performance of a deep learning system in assisting junior ophthalmologists in diagnosing 13 major fundus diseases: a prospective multi-center clinical trial. *npj Digit. Med.* **7**, 8 (2024).
5. Qiu, S. et al. Development and validation of an interpretable deep learning framework for alzheimer's disease classification. *Brain* **143**, 1920–1933 (2020).
6. Barnett, G. O., Cimino, J. J., Hupp, J. A. & Hoffer, E. P. Dxplain: an evolving diagnostic decision-support system. *JAMA* **258**, 67–74 (1987).
7. Su, C., Xu, Z., Pathak, J. & Wang, F. Deep learning in mental health outcome research: a scoping review. *Transl. Psychiatry* **10**, 116 (2020).
8. Gkotsis, G. et al. Characterisation of mental health conditions in social media using informed deep learning. *Sci. Rep.* **7**, 1–11 (2017).
9. Du, J. et al. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Med. Inform. Decis. Mak.* **18**, 77–87 (2018).
10. Caraballo, P. J. et al. Trustworthiness of a machine learning early warning model in medical and surgical inpatients. *JAMIA Open* **8**, ooae156 (2025).
11. Sajda, P. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.* **8**, 537–565 (2006).
12. Stafford, I. S. et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit. Med.* **3**, 30 (2020).
13. Kline, A. et al. Multimodal machine learning in precision health: A scoping review. *npj Digit. Med.* **5**, 171 (2022).
14. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit. Med.* **4**, 65 (2021).
15. Myszczynska, M. A. et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat. Rev. Neurol.* **16**, 440–456 (2020).
16. Fatima, M. & Pasha, M. Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **9**, 1–16 (2017).
17. Choy, S. P. et al. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. *NPJ Digit. Med.* **6**, 180 (2023).
18. Mei, X. et al. Interstitial lung disease diagnosis and prognosis using an AI system integrating longitudinal data. *Nature communications* **14.1**, 2272 (2023).
19. Zhou, S. et al. Open-world electrocardiogram classification via domain knowledge-driven contrastive learning. *Neural Netw* **179**, 106551 (2024).
20. Zhou, Q. et al. A machine and human reader study on ai diagnosis model safety under attacks of adversarial images. *Nat. Commun.* **12**, 7281 (2021).
21. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
22. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).
23. Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
24. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
25. Yang, Z., Mitra, A., Kwon, S. & Yu, H. ClinicalMamba: A generative clinical language model on longitudinal clinical notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 54–63 (Association for Computational Linguistics, 2024).
26. Peng, L. et al. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *NPJ Digit. Med.* **7**, 127 (2024).
27. Zhan, Z., Zhou, S., Li, M. & Zhang, R. Ramie: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association* ocaf002 (2025).
28. Lu, M. Y. et al. A multimodal generative AI copilot for human pathology. *Nature* 1–3 (2024).
29. Kim, J. et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digit. Med.* **7**, 193 (2024).
30. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
31. Zhou, H. et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).
32. Meng, X. et al. The application of large language models in medicine: A scoping review. *Iscience* **27**, (2024).
33. Zhang, Y. et al. Data-centric foundation models in computational healthcare: A survey.*arXiv preprint arXiv:2401.02458* (2024).
34. Du, X. et al. Generative large language models in electronic health records for patient care since 2023: A systematic review.*medRxiv* 2024–08 (2024).
35. Wang, C. et al. A survey for large language models in biomedicine. *arXiv preprint arXiv:2409.00133* (2024).
36. Li, L. et al. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs). *arXiv preprint arXiv:2405.03066* (2024).
37. He, Kai, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Information Fusion (2025): 102963.
38. Pressman, S. M. et al. Clinical and surgical applications of large language models: A systematic review. *J. Clin. Med.* **13**, 3041 (2024).
39. Omar, M., Brin, D., Glicksberg, B. & Klang, E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *Am J Infect Control* **52**, 992–1001 (2024).
40. Giuffrè, M. et al. Systematic review: The use of large language models as medical chatbots in digestive diseases. *Alimentary pharmacology & therapeutics* **60.2**, 144–166 (2024).
41. Mai, A. S., Adnan, K. & Mohammad, Y. Medpromptx: Grounded multimodal prompting for chest x-ray diagnosis. *ArXiv* abs/2403.15585 (2024).
42. Kraljevic, Z. et al. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digital Health* **6**, e281–e290 (2024).
43. GLM, T. et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).
44. Busch, F. et al. Integrating text and image analysis: Exploring GPT-4v's capabilities in advanced radiological applications across subspecialties. *J. Med Internet Res.* **26**, e54948 (2024).
45. Kim, Y., Xu, X., McDuff, D., Breazeal, C. & Park, H. W. Health-llm: Large language models for health prediction via wearable sensor data. *Conference on Health, Inference, and Learning* (2024).
46. Gao, Z., Hu, Y., Tan, C. & Li, S. Z. Prefixmol: Target- and chemistry-aware molecule design via prefix embedding. *ArXiv preprint* abs/2302.07120 (2023).

47. Niu, S. et al. Ehr-knowgen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Inf. Fusion* **102**, 102069 (2024).
48. Chung, P. et al. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA surgery* **159.8**, 928–937 (2024).
49. Delsoz, M. et al. Performance of ChatGPT in diagnosis of corneal eye diseases. *Cornea* **43.5**, 664–670 (2024).
50. Fink, M. A. et al. Potential of chatgpt and gpt-4 for data mining of free-text ct reports on lung cancer. *Radiology* **308**, e231362 (2023).
51. Moallem, G., Gonzalez, A. D. L. M., Desai, A. & Rusu, M. Automated labeling of spondylolisthesis cases through spinal mri radiology report interpretation using ChatGPT. In *Medical Imaging 2024: Computer-Aided Diagnosis*, vol. 12927, 702–706 (SPIE, 2024).
52. Benary, M. et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **6**, e2343689–e2343689 (2023).
53. Reese, J. T. et al. On the limitations of large language models in clinical diagnosis. *medRxiv* 2023-07 (2024).
54. Sarangi, P. K., Irodi, A., Panda, S., Nayak, D. S. K. & Mondal, H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. *Indian J. Radiol. Imaging* **34**, 269–275 (2024).
55. Wang, J. et al. Augmented risk prediction for the onset of alzheimer's disease from electronic health records with large language models. *arXiv preprint arXiv:2405.16413* (2024).
56. Du, X. et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *eBioMedicine* **109**, 105401 (2024).
57. Haider, S. A. et al. Evaluating large language model (LLM) performance on established breast classification systems. *Diagnostics* **14**, 1491 (2024).
58. Siepmann, R. et al. The virtual reference radiologist: comprehensive AI assistance for clinical image reading and interpretation. *European Radiology* **34**, 6652–6666 (2024).
59. Peng, L. et al. Mmgpl: Multimodal medical data analysis with graph prompt learning. *Med. Image Anal.* **97**, 103225 (2024).
60. Xu, S. et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317* (2023).
61. Gertz, R. J. et al. Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology* **311**, e232714 (2024).
62. Ono, D., Dickson, D. W. & Koga, S. Evaluating the efficacy of few-shot learning for GPT-4vision in neurodegenerative disease histopathology: A comparative analysis with convolutional neural network model. *Neuropathol. Appl Neurobiol.* **50**, e12997 (2024).
63. Dai, Y., Gao, Y. & Liu, F. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* **11**, 1384 (2021).
64. Upadhyaya, D. P. et al. A 360° View for Large Language Models: Early Detection of Amblyopia in Children Using Multi-view Eye Movement Recordings. In *International Conference on Artificial Intelligence in Medicine* (pp. 165-175). (Cham: Springer Nature, Switzerland, 2024).
65. Noda, M. et al. Feasibility of multimodal artificial intelligence using GPT-4 vision for the classification of middle ear disease: Qualitative study and validation. *JMIR AI* **3**, e58342 (2024).
66. Antaki, F., Chopra, R. & Keane, P. A. Vision-language models for feature detection of macular diseases on optical coherence tomography. *JAMA Ophthalmol* **142**, 573–576 (2024).
67. Peng, Z. et al. Development and evaluation of multimodal AI for diagnosis and triage of ophthalmic diseases using ChatGPT and anterior segment images: protocol for a two-stage cross-sectional study. *Front. Artif. Intell.* **6**, 1323924 (2023).
68. Suh, P. S. et al. Comparing diagnostic accuracy of radiologists versus GPT-4v and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology* **312**, e240273 (2024).
69. Pugliese, G. et al. Are artificial intelligence large language models a reliable tool for difficult differential diagnosis? An a posteriori analysis of a peculiar case of necrotizing otitis externa. *Clin. Case Rep.* **11**, e7933 (2023).
70. Hu, C. et al. Exploiting ChatGPT for diagnosing autism-associated language disorders and identifying distinct features. *arXiv preprint arXiv:2405.01799* (2024).
71. Deng, S. et al. Hear me, see me, understand me: Audio-visual autism behavior recognition. (IEEE Transactions on Multimedia, 2024).
72. Rezaii, N. et al. Artificial intelligence classifies primary progressive aphasia from connected speech. *Brain* **147**, 3070–3082 (2024).
73. Liu, C., Ma, Y., Kothur, K., Nikpour, A. & Kavehei, O. Biosignal copilot: Leveraging the power of LLMs in drafting reports for biomedical signals. *medRxiv* 2023.06.28.23291916 (2023).
74. Yu, H., Guo, P. & Sano, A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *ML4H@NeurIPS* (2023).
75. Wu, D. et al. Multimodal machine learning combining facial images and clinical texts improves the diagnosis of rare genetic diseases. *arXiv preprint arXiv:2312.15320* (2023).
76. Feng, Y., Xu, X., Zhuang, Y. & Zhang, M. Large language models improve alzheimer's disease diagnosis using multi-modality data. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 61–66 (IEEE, 2023).
77. Ma, M. D. et al. Clibench: Multifaceted evaluation of large language models in clinical decisions on diagnoses, procedures, lab tests orders and prescriptions. *arXiv preprint arXiv:2406.09923* (2024).
78. Liang, L. et al. Genetic transformer: An innovative large language model driven approach for rapid and accurate identification of causative variants in rare genetic diseases. *medRxiv* 2024-07 (2024).
79. Thompson, W. et al. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. In *Deep Generative Models for Health Workshop NeurIPS 2023* (2023).
80. Shi, W. et al. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 1–10 (2023).
81. Wen, Y., Wang, Z. & Sun, J. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10370–10388 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
82. Kresevic, S. et al. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit. Med.* **7**, 102 (2024).
83. Yu, H., Guo, P. & Sano, A. ECG semantic integrator (ESI): A foundation ECG model pretrained with LLM-enhanced cardiological text. *Trans. Mach. Learn. Res*. (2024).
84. Ghersin, I. et al. Comparative evaluation of a language model and human specialists in the application of European guidelines for the management of inflammatory bowel diseases and malignancies. *Endoscopy* **56**, 706–709 (2024).
85. Ge, J. et al. Development of a liver disease–specific large language model chat interface using retrieval-augmented generation. *Hepatology* **80**, 1158–1168 (2024).
86. Xia, P. et al. RULE: Reliable multimodal RAG for factuality in medical vision language models. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1081–1093 (Association for Computational Linguistics, Miami, Florida, USA, 2024).

87. Ranjit, M., Ganapathy, G., Manuel, R. & Ganu, T. Retrieval augmented chest x-ray report generation using openai gpt models. In Deshpande, K. et al. (eds.) *Proceedings of the 8th Machine Learning for Healthcare Conference*, vol. 219 of *Proceedings of Machine Learning Research*, 650–666 (PMLR, 2023).s
88. Li, Z., Zhang, J., Zhou, W., Zheng, J. & Xia, Y. Gpt-agents based on medical guidelines can improve the responsiveness and explainability of outcomes for traumatic brain injury rehabilitation. *Sci. Rep.* **14**, 7626 (2024).
89. Abdullahi, T. et al. Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. *JMIR Med. Educ.* **10**, e51391 (2024).
90. Rifat Ahmmad Rashid, M. et al. A respiratory disease management framework by combining large language models and convolutional neural networks for effective diagnosis. *Int. J. Comput. Digit. Syst.* **16**, 189–202 (2024).
91. Ferber, D. et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *ArXiv* abs/2404.04667 (2024).
92. Soong, D. et al. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. *PLOS Digital Health* **3**, e0000568 (2024).
93. Rau, A. et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* **308**, e230970 (2023).
94. Zhu, Y. et al. Emerge: Integrating RAG for improved multimodal EHR predictive modeling. *ArXiv* abs/2406.00036 (2024).
95. Chen, C. et al. Large Language Model-Informed ECG Dual Attention Network for Heart Failure Risk Prediction. *IEEE Transactions on Big Data* **11**, 948–960 (2024).
96. Askell, A. et al. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861 (2021).
97. Liu, H. et al. Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023).
98. Toma, A. et al. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. arXiv preprint arXiv:2305.12031 (2023).
99. Wu, J., Wu, X., Zheng, Y. & Yang, J. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611* (2024).
100. He, Y., Zhang, Y., He, S. & Wan, J. BP4ER: Bootstrap Prompting for Explicit Reasoning in Medical Dialogue Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2480–2492 (ELRA and ICCL, Torino, Italia, 2024).
101. Xu, K., Cheng, Y., Hou, W ., Tan, Q. & Li, W. Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
102. Yang, L. et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162* (2024).
103. He, S. et al. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv e-prints (2024): arXiv-2404.
104. Chen, Z. et al. Dia-LLaMA: Towards large language model-driven ct report generation. arXiv preprint arXiv:2403.16386 (2024).
105. Liu, Z. et al. Radiology-llama2: Best-in-class large language model for radiology. *arXiv preprint arXiv:2309.06419* (2023).
106. Alkhaldi, A. et al. Minigpt-med: Large language model as a general interface for radiology diagnosis. arXiv preprint arXiv:2407.04106 (2024).
107. Lee, S., Youn, J., Kim, H., Kim, M. & Yoon, S. H. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. *arXiv* (2023).
108. Kwon, T. et al. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 18417–18425 (2024).
109. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
110. Zhou, J. et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).
111. Sun, Y. et al. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 5034–5042 (2024).
112. Zhang, X. et al. When LLMs Meets Acoustic Landmarks: An Efficient Approach to Integrate Speech into Large Language Models for Depression Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 146–158 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
113. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744 (2022).
114. Schulman, J. et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
115. Zhou, Z. et al. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728* (2024).
116. Wang, G., Yang, G., Du, Z., Fan, L. & Li, X. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968* (2023).
117. Zhang, H. et al. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885 (Association for Computational Linguistics, Singapore, 2023).
118. Zeng, G. et al. MedDialog: Large-scale Medical Dialogue Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. (Association for Computational Linguistics, 2020).
119. Gao, L., Schulman, J. & Hilton, J. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866 (PMLR, 2023).
120. Ye, Z. et al. Beyond Scalar Reward Model: Learning Generative Judge from Preference Data. ArXiv abs/2410.03742 (2024).
121. Henderson, P. et al. Deep reinforcement learning that matters. *Proceedings of the AAAI conference on artificial intelligence*. 32 (2018).
122. Rafailov, R. et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**, 53728–53741 (2023).
123. Ye, Q. et al. Qilin-med: Multi-stage knowledge injection advanced medical large language model. arXiv preprint arXiv:2310.09089 (2023).
124. Yang, D. et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems* **37**, 138632–138662 (2024).
125. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In Precup, D. & Teh, Y. W. (eds.) *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70 of *Proceedings of Machine Learning Research*, 1321–1330 (PMLR, 2017).
126. Tajwar, F. et al. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vol. 235, 47441–47474 (JMLR.org, 2024).
127. Bai, Y. et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022).
128. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* (OpenReview.net, 2022).

129. Rajashekar, N. C. et al. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, vol. 37, 1–20 (ACM, New York, NY, USA, 2024).s
130. Yang, X. et al. A large language model for electronic health records. *NPJ digital medicine* **5**, 194 (2022).
131. Labrak, Y. et al. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
132. Wang, J., Seng, K. P., Shen, Y., Ang, L.-M. & Huang, D. Image to label to answer: An efficient framework for enhanced clinical applications in medical visual question answering. *Electronics* **13**, 2273 (2024).
133. Liu, F. et al. A medical multimodal large language model for future pandemics. *NPJ Digit. Med.* **6**, 226 (2023).
134. Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv preprint arXiv:2308.02463 (2023).
135. Ding, J.-E. et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. *Scientific Reports* **14**, 20774 (2024).
136. Phan, V. M. H. et al. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages. 11492–11501 (2024).
137. Chen, J. et al. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
138. Lu, Z. et al. Large language models in biomedicine and health: current research landscape and future directions. *J. Am. Med. Inform. Assoc.* **31**, 1801–1811 (2024).
139. Li, H. et al. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).
140. Tu, T. et al. Towards conversational diagnostic artificial intelligence[J]. *Nature* 1–9 (2025).
141. Safranek, C. W. et al. Automated heart score determination via chatgpt: Honing a framework for iterative prompt development. *J. Am. Coll. Emerg. Phys. Open* **5**, e13133 (2024).
142. Zhang, T. et al. Incorporating Clinical Guidelines Through Adapting Multi-modal Large Language Model for Prostate Cancer PI-RADS Scoring. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Cham: Springer Nature, Switzerland, 2024).
143. Zhou, S. et al. Explainable differential diagnosis with dual-inference large language models. *npj Health Systems* **2**, 12 (2025).
144. Chen, X. et al. EyeGPT: Ophthalmic assistant with large language models. *arXiv preprint arXiv:2403.00840* (2024).
145. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med.* **7**, 20 (2024).
146. Li, S. S. et al. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. (Neural Information Processing Systems, 2024).
147. Fansi Tchango, A., Goel, R., Wen, Z., Martel, J. & Ghosn, J. Ddxplus: A new dataset for automatic medical diagnosis. *Adv. neural Inf. Process. Syst.* **35**, 31306–31318 (2022).
148. Xie, Q. et al. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digit. Med.* **8**, 141 (2025).
149. Yang, S. et al. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 17. (2024).
150. Mohammadi, S. S. & Nguyen, Q. D. A user-friendly approach for the diagnosis of diabetic retinopathy using ChatGPT and automated machine learning. *Ophthalmol. Sci.* **4**, 100495 (2024).
151. Tank, C. et al. Depression detection and analysis using large language models on textual and audio-visual modalities. *arXiv preprint arXiv:2407.06125* (2024).
152. Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat. Commun.* **15**, 2050 (2024).
153. Bae, S. et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems* **36**, 3867–3880 (2023).
154. Hu, J. et al. Designing scaffolding strategies for conversational agents in dialog task of neurocognitive disorders screening. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21 (2024).
155. Englhardt, Z. et al. From classification to clinical insights: Towards analyzing and reasoning about mobile and behavioral health data with large language models. *Proc. ACM Interact., Mob. Wearable Ubiquitous Technol.* **8**, 1–25 (2024).
156. Smith, P. C. et al. Missing clinical information during primary care visits. *JAMA* **293**, 565–571 (2005).
157. McInerney, D. et al. Towards reducing diagnostic errors with interpretable risk prediction. *Proceedings of the conference*. Association for Computational Linguistics. North American Chapter. Meeting. Vol. 2024, 2024).
158. Adler-Milstein, J., Chen, J. H. & Dhaliwal, G. Next-generation artificial intelligence for diagnosis: from predicting diagnostic labels to "wayfinding". *JAMA* **326**, 2467–2468 (2021).
159. Shi, X. et al. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics ACL 2024* (2024).
160. Sun, Z., Luo, C. & Huang, Z. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292* (2024).
161. Zou, X. et al. AI-driven diagnostic assistance in medical inquiry: Reinforcement learning algorithm development and validation. *J. Med. Internet Res.* **26**, e54616 (2024).
162. Zhang, R. et al. Making shiny objects illuminating: the promise and challenges of large language models in us health systems. *npj Health Syst* **2**, 8 (2025).
163. Cameron, S. & Turtle-Song, I. Learning to write case notes using the soap format. *J. Counsel. Dev.* **80**, 286–292 (2002).
164. Oniani, D. et al. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 694–702 (IEEE, 2024).
165. Sallam, M., Al-Salahat, K. & Al-Ajlouni, E. ChatGPT performance in diagnostic clinical microbiology laboratory-oriented case scenarios. *cureus* **15**, e50629 (2023).
166. Bhasuran, B. et al. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digit. Med.* **8**, 166 (2025).
167. Yi, Z. et al. A survey on recent advances in LM-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013* (2024).
168. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature* 1–7 (2025).
169. Haltaufderheide, J. & Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (llms). *NPJ Digit. Med.* **7**, 183 (2024).

170. Dou, C. et al. Detection, diagnosis, and explanation: A benchmark for Chinese medical hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4784–4794 (2024).
171. Tran, H., Wang, J., Ting, Y., Huang, W. & Chen, T. Leaf: Learning and evaluation augmented by fact-checking to improve factualness in large language models. *arXiv preprint arXiv:2410.23526* (2024).
172. Yue, X. & Zhou, S. Phicon: Improving generalization of clinical text de-identification models via data augmentation. In *Clinical Natural Language Processing Workshop* (2020).
173. Zhou, J. et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).
174. Spitale, M., Cheong, J. & Gunes, H. Underneath the Numbers: Quantitative and Qualitative Gender Fairness in LLMs for Depression Prediction. arXiv preprint arXiv:2406.08183 (2024).
175. Chen, Z. et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
176. Yang, K. et al. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, 4489–4500 (2024).
177. Peng, J. et al. Continually evolved multimodal foundation models for cancer prognosis. *arXiv preprint arXiv:2501.18170* (2025).
178. Yi, H. et al. Towards general purpose medical AI: Continual learning medical foundation model. *arXiv preprint arXiv:2303.06580* (2023).
179. Kim, Y. et al. Adaptive collaboration strategy for LLMs in medical decision making. *NeurIPS* (2024).
180. Jiang, S. et al. Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3843–3860 (Association for Computational Linguistics, Miami, Florida, USA, 2024).
181. Xu, D. et al. Editing factual knowledge and explanatory ability of medical large language models. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (2024).
182. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med*. 1–10 (2024).
183. Kuratov, Y. et al. Babilong: Testing the limits of LLMs with long context reasoning-in-a-haystack. *Adv. Neural Inf. Process. Syst.* **37**, 106519–106554 (2024).
184. Yang, Z., Mitra, A., Liu, W., Berlowitz, D. & Yu, H. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.* **14**, 7857 (2023).
185. Huang, L. et al. Machine learning of serum metabolic patterns encodes early-stage lung adenocarcinoma. *Nat. Commun.* **11**, 3556 (2020).
186. Cui, H. et al. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *arXiv preprint arXiv:2503.04176* (2025).
187. Dou, C. et al. PlugMed: Improving Specificity in Patient-Centered Medical Dialogue Generation using In-Context Learning. *Conference on Empirical Methods in Natural Language Processing* (2023).
188. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics* (2019).
189. Chen, C. et al. Clinicalbench: Can LLMs beat traditional ML models in clinical prediction? *arXiv preprint arXiv:2411.06469* (2024).
190. Zhong, T. et al. Chatradio-valuer: A chat large language model for generalizable radiology report generation based on multi-institution and multi-system data. *arXiv preprint arXiv:2310.05242* (2023).
191. Zhan, Z., Zhou, S., Zhou, H., Liu, Z. & Zhang, R. Epee: Towards efficient and effective foundation models in biomedicine. *arXiv preprint arXiv:2503.02053* (2025).
192. Ferrara, E. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: a survey of early trends, datasets, and challenges. *Sensors* **24**, 5045 (2024).
193. Hulstaert, F. et al. Gaps in the evidence underpinning high-risk medical devices in Europe at market entry, and potential solutions. *Orphanet J. Rare Dis.* **18**, 212 (2023).
194. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit. Med.* **7**, 258 (2024).
195. Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
196. Wu, D. et al. GestaltMML: Enhancing Rare Genetic Disease Diagnosis through Multimodal Machine Learning Combining Facial Images and Clinical Texts. ArXiv (2024): arXiv-2312.
197. Mizuta, K., Hirosawa, T., Harada, Y. & Shimizu, T. Can chatgpt-4 evaluate whether a differential diagnosis list contains the correct diagnosis as accurately as a physician? *Diagnosis* **11**, 321–324 (2024).
198. Olesen, A. S. O. et al. How does ChatGPT-4 match radiologists in detecting pulmonary congestion on chest X-ray? *J Med Arti Intell* **7**, (2024).
199. Liu, X. et al. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525* (2023).
200. Slack, D. & Singh, S. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188* (2023).
201. Xia, P. et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems* **37**, 140334–140365 (2024).
202. Wada, A. et al. Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. *Diagnostics* **14**, 1541 (2024).
203. Chen, Z., Lu, Y. & Wang, W. Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304 (Association for Computational Linguistics, Singapore, 2023).
204. Vashisht, P. et al. UMass-BioNLP at MEDIQA-M3G 2024: DermPrompt - A Systematic Exploration of Prompt Engineering with GPT-4V for Dermatological Diagnosis. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 502–525 (Association for Computational Linguistics, Mexico City, Mexico, 2024).
205. Lim, S., Kim, Y., Choi, C-H., Sohn, J-Y. & Kim, B-H. ERD: A Framework for Improving LLM Reasoning for Cognitive Distortion Classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300 (Association for Computational Linguistics, Mexico City, Mexico, 2024).
206. Peng, C. et al. Improving generalizability of extracting social determinants of health using large language models through prompt-tuning. *arXiv preprint arXiv:2403.12374* (2024).
207. Zhou, W. et al. Transferring Pre-Trained Large Language-Image Model for Medical Image Captioning. In *CLEF (Working Notes),* pages 1776–1784, (2023).
208. Belyaeva, A. et al. Multimodal LLMs for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, 86–102 (Springer, 2023).
209. Ong, J. C. L. et al. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. arXiv preprint arXiv:2402.01741 (2024).

210. Vithanage, D. et al. Evaluating machine learning approaches for multi-label classification of unstructured electronic health records with a generative large language model. medRxiv (2024).
211. Liu, J. et al. Large language model locally fine-tuning (LLMLF) on Chinese medical imaging reports. In *Proceedings of the 2023 6th International Conference on Big Data Technologies* (ACM, New York, NY, USA, 2023).
212. Song, M. et al. PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis. *Med. Image Anal.* **97**, 103248 (2024).
213. Liu, W. & Zuo, Y. Stone needle: A general multimodal large-scale model framework towards healthcare. *arXiv preprint arXiv:2306.16034* (2023).
214. Dou, C. et al. Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2453–2473 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
215. Sun, M. LlamaCare: A Large Medical Language Model for Enhancing Healthcare Knowledge Sharing. arXiv preprint arXiv:2406.02350 (2024).
216. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med* **30**, 3129–3141 (2024).
217. Wu, C.-K., Chen, W.-L. & Chen, H.-H. Large language models perform diagnostic reasoning. Tiny Papers @ ICLR 2023.
218. Yang, Z. et al. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. *Journal of Medical Internet Research* **27**, e65146 (2025).
219. Chen, Z. et al. Narrative Feature or Structured Feature? A Study of Large Language Models to Identify Cancer Patients at Risk of Heart Failure. arXiv preprint arXiv:2403.11425 (2024).
220. Hayati, M. F. M., Ali, M. A. M. & Rosli, A. N. M. Depression detection on Malay dialects using GPT-3. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 360–364 (IEEE, 2022).
221. Liu, S. et al. Leveraging large language models for generating responses to patient messages-a subjective analysis. *J. Am. Med. Inform. Assoc.* **31**, 1367–1379 (2024).
222. Gao, Y. et al. Large language models and medical knowledge grounding for diagnosis prediction. medRxiv 2023-11 (2023).
223. Sushil, M. et al. A comparative study of zero-shot inference with large language models and supervised modeling in breast cancer pathology classification. Research Square (2024).
224. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).
225. Kotelanski, M., Gallo, R., Nayak, A. & Savage, T. Methods to estimate large language model confidence. *arXiv preprint arXiv:2312.03733* (2023).
226. Qu, L. et al. The rise of AI language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Adv. Neural Inf. Process. Syst.* **36**, 67551–67564 (2023).
227. Dekel, S. et al. ChatGPT Demonstrates Potential for Identifying Psychiatric Disorders: Application to Childbirth-Related Post-Traumatic Stress Disorder. Research Square (2023).
228. Du, J. et al. Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. (Cham: Springer Nature, Switzerland, 2024).
229. Blankemeier, L. et al. Merlin: A vision language foundation model for 3d computed tomography. Research Square (2024).
230. Acharya, A. et al. Clinical risk prediction using language models: benefits and considerations. *Journal of the American Medical Informatics Association* **31**, 1856–1864 (2024).
231. Chen, P.-F. et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med. Inform.* **9**, e23230 (2021).
232. Pedro, T. et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *Journal of NeuroInterventional Surgery* (2024).
233. Ren, X. et al. ChatASD: LLM-based AI therapist for ASD. In *Communications in Computer and Information Science*, Communications in computer and information science, 312–324 (Springer Nature Singapore, Singapore, 2024).
234. Weng, Y. et al. Large language models need holistically thought in medical conversational qa. *arXiv preprint arXiv:2305.05410* (2023).
235. Panagoulias, D. P., Virvou, M. & Tsihrintzis, G. A. Evaluating ILM–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730* (2024).
236. Liu, Y. et al. A systematic evaluation of GPT-4v's multimodal capability for chest x-ray image analysis. *Meta-Radiol* **2**, 100099 (2024).
237. Chen, X. et al. Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digit. Med.* **7**, 111 (2024).
238. Hill, B. L. et al. Chiron: A generative foundation model for structured sequential medical data. In *Deep Generative Models for Health Workshop NeurIPS 2023*.
239. Kottlors, J. et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology* **308**, e231167 (2023).
240. Nair, V. et al. DERA: Enhancing Large Language Model Completions with Dialog-Enabled Resolving Agents. *Clinical Natural Language Processing Workshop* (2023).
241. Umerenkov, D., Zubkova, G. & Nesterov, A. Deciphering diagnoses: how large language models explanations influence clinical decision making. *arXiv preprint arXiv:2310.01708* (2023).
242. Chen, X. et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. *Br J Ophthalmolog* **108**, 1450–1456 (2024).
243. Lyu, Q. et al. Translating radiology reports into plain language using ChatGPT and gpt-4 with prompt learning: results, limitations, and potential. *Vis. Comput. Ind. Biomed. Art.* **6**, 9 (2023).
244. Jo, E. et al. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. *JMIR Med. Educ.* **10**, e51282 (2024).
245. Guo, S. et al. Comparing ChatGPT's and Surgeon's Responses to Thyroid-related Questions From Patients. *J Clin Endocrinol Metab* **110**, e841–e850 (2025).
246. Kang, S. et al. WoLF: Wide-scope Large Language Model Framework for CXR Understanding. arXiv preprint arXiv:2403.15456 (2024).
247. He, Y. et al. BP4ER: Bootstrap Prompting for Explicit Reasoning in Medical Dialogue Generation. *International Conference on Language Resources and Evaluation* (2024).

Acknowledgements

This work was supported by the National Institutes of Health's National Center for Complementary and Integrative Health under grant number R01AT009457, National Institute on Aging under grant number R01AG078154, and National Cancer Institute under grant number R01CA287413. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health. We also acknowledge the support from the Center for Learning Health System Sciences.

Author contributions

S.Z. conceptualized the study and led the work. Z.Z., S.Z., J.Y., and M.Z. searched papers. S.Z., Z.X., M.Z., C.X., Y.G., Z.Z., S.D., J.W., K.X., Y.F., L.X., and J.Y. conducted paper screening and data extraction. S.Z., Z.X., M.Z.,

and C.X. performed data synthesis and contributed to the writing. S.Z., Z.X., M.Z., C.X., D.Z., G.M., and R.Z. revised the manuscript. R.Z. supervised the study. All authors read and approved the final version.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00011-z.

Correspondence and requests for materials should be addressed to Rui Zhang.