nature medicine

Article

https://doi.org/10.1038/s41591-024-03416-6

A generalist medical language model for disease diagnosis assistance

Received: 28 March 2024

Accepted: 12 November 2024

Published online: 08 January 2025

Check for updates

Xiaohong Liu[1,11], Hao Liu[2,11], Guoxing Yang[1,11], Zeyu Jiang[1,11], Shuguang Cui[3,11], Zhaoze Zhang[2], Huan Wang[2], Liyuan Tao[4], Yongchang Sun[5], Zhu Song[5], Tianpei Hong[6], Jin Yang[6], Tianrun Gao[1], Jiangjiang Zhang[1], Xiaohu Li[1], Jing Zhang[7], Ye Sang[7], Zhao Yang[8], Kanmin Xue[9], Song Wu[10], Ping Zhang[1], Jian Yang[7] ✉, Chunli Song[2] ✉ & Guangyu Wang[1] ✉

The delivery of accurate diagnoses is crucial in healthcare and represents the gateway to appropriate and timely treatment. Although recent large language models (LLMs) have demonstrated impressive capabilities in few-shot or zero-shot learning, their effectiveness in clinical diagnosis remains unproven. Here we present MedFound, a generalist medical language model with 176 billion parameters, pre-trained on a large-scale corpus derived from diverse medical text and real-world clinical records. We further fine-tuned MedFound to learn physicians' inferential diagnosis with a self-bootstrapping strategy-based chain-of-thought approach and introduced a unified preference alignment framework to align it with standard clinical practice. Extensive experiments demonstrate that our medical LLM outperforms other baseline LLMs and specialized models in in-distribution (common diseases), out-of-distribution (external validation) and long-tailed distribution (rare diseases) scenarios across eight specialties. Further ablation studies indicate the effectiveness of key components in our medical LLM training approach. We conducted a comprehensive evaluation of the clinical applicability of LLMs for diagnosis involving artificial intelligence (AI) versus physician comparison, AI-assistance study and human evaluation framework. Our proposed framework incorporates eight clinical evaluation metrics, covering capabilities such as medical record summarization, diagnostic reasoning and risk management. Our findings demonstrate the model's feasibility in assisting physicians with disease diagnosis as part of the clinical workflow.

The delivery of accurate diagnoses plays a crucial role in the field of healthcare and represents a fundamental skill for all physicians[1,2]. The diagnostic process typically involves the identification of a disease through extended reasoning processes of analyzing symptoms, signs and results of investigations to formulate a diagnosis as well as differential diagnoses. Despite extensive medical training, diagnosis is prone to errors, with an estimated 20% rate of misdiagnosis at the primary care level[3], which contributes to approximately 17% of all adverse events in medical practice[4]. For decades, considerable efforts have been made to enhance the accuracy and accessibility of disease diagnosis, including traditional rule-based clinical decision support systems (CDSSs)[5] and machine learning techniques that extract structured features to develop clinical predictive models. However, the dependency on structured inputs and specialized training is complex and resource intensive. A substantial gap remains between the development of major medical predictive algorithms and their actual clinical deployment in diverse healthcare settings[6].

A full list of affiliations appears at the end of the paper. ✉e-mail: yangjian@ctgu.edu.cn; schl@bjmu.edu.cn; guangyu.wang24@gmail.com

In recent years, the emergence of pre-trained language models (PLMs) has substantially advanced the natural language processing (NLP) domain. These models are first pre-trained on large-scale corpora via self-supervised learning tasks (for example, masked language modeling for BERT[7] and auto-regressive language modeling for GPT[8]) and then fine-tuned on specific downstream tasks. Further studies suggest that when the model size, dataset size and computational resources are sufficiently large, large language models (LLMs) can exhibit emerging few-shot and zero-shot properties across multiple NLP tasks. The recent advancement of PLMs and LLMs has attracted interest in using these pre-trained language models tailored to the biomedical domain, such as ClinicalBERT[9], NYUTron[10], GatorTron[11] and BioGPT[12]. These models have demonstrated the potential to transform task-specific paradigms and address the 'last-mile' challenge in medical predictive analytics, enabling the development of versatile clinical applications.

Despite the potential of LLM technology in biomedicine, exploitation of its utility remains at a preliminary stage. Most studies focus on use-case reports of LLMs in medicine, particularly ChatGPT[13]. There is currently a lack of well-designed, publicly available LLMs specifically tailored for real-world clinical contexts. Although a small fraction of work has investigated incorporating clinical knowledge into LLMs for tasks such as medical question-answering[14] or dialogue[15], their capabilities in clinical diagnostic reasoning have not been fully developed or examined. Additionally, generative LLMs can hallucinate or fabricate facts, which could be harmful if clinicians rely on their recommended diagnosis. Thus, it becomes paramount to employ alignment techniques to ensure that these models align with the objective of clinical diagnostic knowledge as well as to 'follow the user's instructions helpfully and safely'[16]. Current evaluations of the LLM models typically rely on automated evaluations based on limited benchmarks, underscoring the need for a more comprehensive assessment of LLM-based tools in real-world clinical settings.

To address the challenges, our approach makes several contributions (Fig. 1). First, we present MedFound, a large-scale medical LLM (176 billion parameters) that is efficiently pre-trained on a diverse medical corpus derived from medical literature as well as 8.7 million real-world electronic health records (EHRs), allowing us to encode domain-specific knowledge to the model. Furthermore, we propose a two-phase approach to adapt MedFound as a diagnostic generalist, resulting in a refined version, called MedFound-DX-PA. We first introduce a self-bootstrapping strategy-based chain-of-thought (COT) fine-tuning that enables the LLM to automatically generate diagnostic rationales and reasoning like physician experts[17]. Subsequently, to address the challenges of aligning the LLM's outputs with clinical requirements[18], we present a unified preference alignment (PA) framework. This framework incorporates (1) diagnostic hierarchy preferences as guided by the hierarchical diagnostic structure of the International Classification of Diseases (ICD)-10 tree and (2) helpfulness preferences guided by expert annotation. A further ablation study demonstrated the impact of components in our proposed training approach on the LLM's performance.

We conducted a comprehensive evaluation to assess the diagnostic performance of MedFound-DX-PA during implementation. First, we established a benchmark study using actual clinical records from three scenarios across specialties, including in-distribution (ID), out-of-distribution (OOD) and long-tailed disease distribution settings. The results demonstrate that MedFound-DX-PA outperformed existing models across multiple dimensions, particularly in diagnosing rare diseases that have been overlooked in previous research. Additionally, we conducted a series of prospective clinical studies, including an artificial intelligence (AI) versus specialist comparison and a further AI-assistance study within the workflow. We also introduced a human evaluation framework, called CLEVER (CLinical EValuation for Effective Reasoning in Diagnosis), which uses eight metrics to investigate the feasibility and limitations of current LLMs in real-world medical scenarios. These studies demonstrate the potential of our proposed LLM as a generalist medical AI (GMAI) in the field of medical diagnostics.

Results

Overview of the proposed LLM and dataset characteristics

In this study, we present MedFound, a pre-trained LLM tailored for medical applications, and MedFound-DX-PA, specifically trained for diagnostic analysis applications. To develop and evaluate our models, we constructed three data collections—MedCorpus, MedDX-FT and MedDX-Bench—where MedCorpus and MedDX-FT were used for training, and MedDX-Bench was used for evaluation (Supplementary Table 1). The training process consisted of three stages: pre-training, fine-tuning and alignment (Fig. 1a and Extended Data Figs. 1–4).

In the first stage, we continued pre-training on a general-domain LLM, BLOOM-176B, resulting in MedFound. To develop MedFound, we curated a large-scale medical corpus dataset, MedCorpus, comprising a total of 6.3 billion text tokens from four datasets: MedText, PubMed Central Case Report (PMC-CR), MIMIC-III-Note and MedDX-Note. These datasets are derived from diverse clinical representative sources: medical textbooks and clinical guidelines, patient case reports from literature, open-access clinical records and proprietary datasets of real-world EHRs from hospital systems (as detailed in Methods). Consequently, pre-training on MedCorpus enabled MedFound to encode extensive medical knowledge and practical experience, establishing it as a foundation tool for a broad range of applications within the medical field.

In the second stage, we fine-tuned MedFound to imitate the diagnostic reasoning process of physicians, resulting in MedFound-DX. We curated a dataset named MedDX-FT with medical records and related diagnostic rationale demonstrations for fine-tuning. Physicians were asked to manually craft a demonstration of their clinical reasoning process to diagnose a given patient case based on actual medical records. The annotation interface is illustrated in Extended Data Fig. 2a. Based on the seed set of manual demonstrations and 109,364 EHR notes, we employed a self-bootstrapping strategy to enhance the ability of the LLM to automatically generate high-quality diagnostic rationales (intermediate reasoning steps) for each EHR without extensive expert labor.

In the third stage, we further optimized the model's real-world clinical utility by employing a unified PA framework, which integrates 'diagnostic hierarchy preferences' and 'helpfulness preferences'. For the 'diagnostic hierarchy preference', we leveraged the hierarchical structure of the ICD-10 tree to guide the LLM to align with the well-established disease knowledge and diagnostic processes. For the 'helpfulness preference', the LLM directly aligns with expert feedback by assessing the helpfulness of a given diagnostic rationale (Extended Data Fig. 2b), thus ensuring consistency with human values[16]. Both preference sets were optimized using Direct Preference Optimization (DPO)[19], a simple reinforcement learning-free algorithm that simplifies the preference learning pipeline.

During the evaluation stage, we curated MedDX-Bench, a benchmark consisting of three clinical datasets—MedDX-Test, MedDX-OOD and MedDX-Rare—to comprehensively assess the diagnostic capabilities of the LLM across real-world clinical settings (Fig. 1b). The MedDX-Test dataset was an ID evaluation to evaluate the diagnostic performance of MedFound-DX-PA across specialties, comprising 11,662 medical records from the same distribution as the training dataset. The MedDX-OOD and MedDX-Rare datasets were constructed as external validation sets, sourced from a distinct geographic region in Hubei Province, China, for OOD evaluation. The MedDX-OOD dataset comprises 23,917 records of common diseases also present in MedDX-FT, whereas the MedDX-Rare dataset includes 20,257 records spanning 2,105 rare diseases that are in long-tailed distribution. The evaluation datasets encompass EHRs from daily diagnostic workflows, including chief complaints, present history, physical examinations, laboratory

https://doi.org/10.1038/s41591-024-03416-6

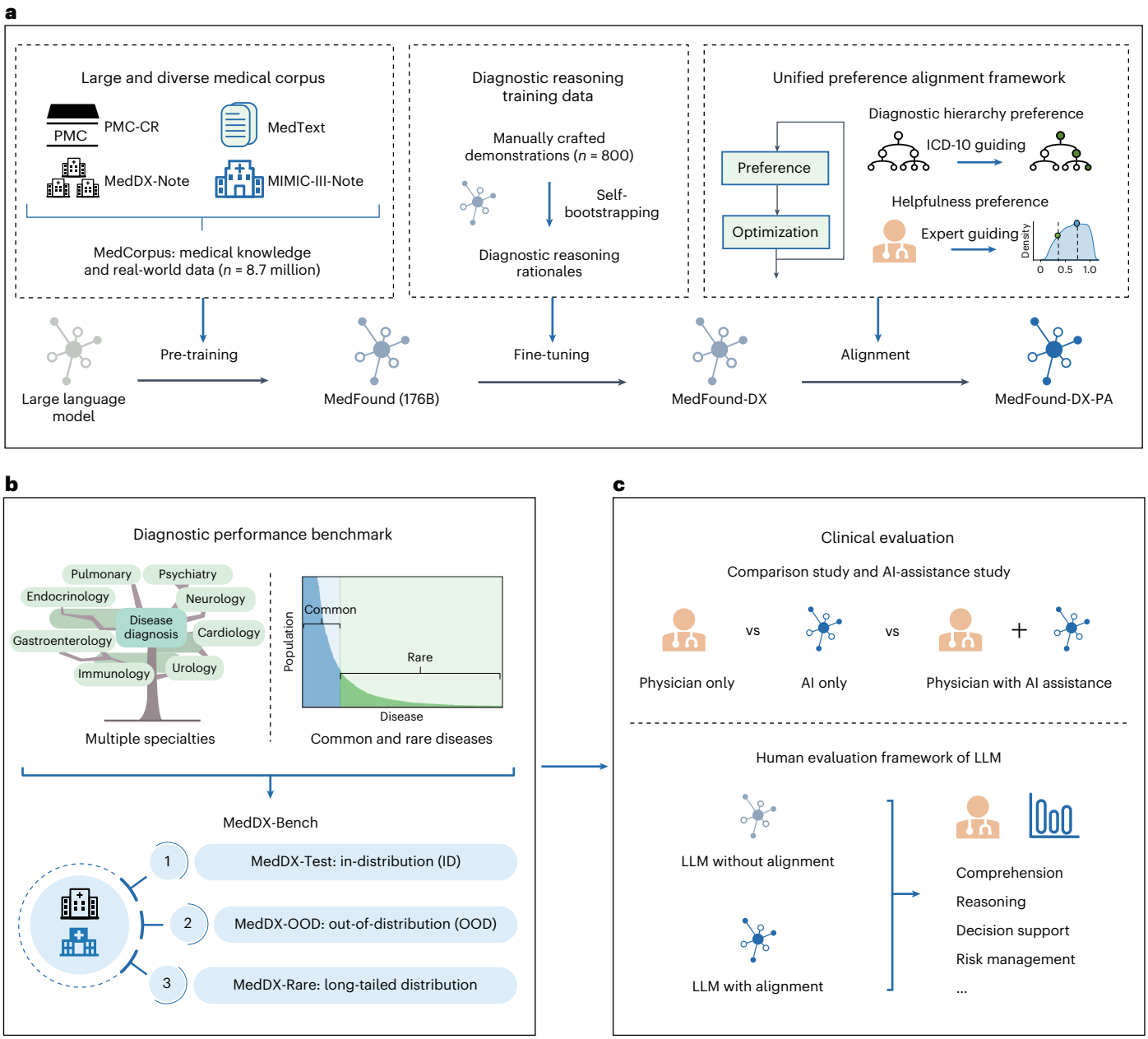


Fig. 1 | Schematic illustration of the development and evaluation of our diagnostic generalist. a, The development of MedFound and MedFound-DX-PA. We pre-trained a 176-billion-parameter MedFound on a large medical corpus consisting of PMC-CR, MIMIC-III-Note, MedDX-Note and MedText. We fine-tuned MedFound with diagnostic rationales and aligned it with diagnostic hierarchy preference and helpfulness preference, resulting in MedFound-DX-PA. **b**, Diagnostic performance benchmarking in real-world scenarios. We conducted evaluations in ID, OOD and long-tailed disease distribution settings, covering diseases across eight specialties, including pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology and endocrinology. **c**, Clinical evaluation of the AI system. We conducted evaluations with physicians, including a comparison study, an AI-assistance study and a qualitative study under a human evaluation framework.

tests and radiological imaging reports. These three datasets present a challenge to assess the generalizability under conditions of varying disease diversity.

Performance of the LLMs on common diseases across specialties

First, we evaluated the performance of MedFound-DX-PA for diagnosing common diseases across specialties in both ID and OOD settings. We conducted comparisons with the leading LLMs, including the open-access MEDITRON-70B[20], Clinical Camel-70B[21] and Llama 3-70B[22] and the closed-source GPT-4o[23]. Both MEDITRON-70B and Clinical Camel-70B are medical pre-trained LLMs and have demonstrated superior performance in medical tasks. Llama 3-70B, a member of the popular open-access Llama family, has shown excellent performance across various domain-specific tasks. GPT-4o is the latest version of ChatGPT, which is reported to have a broader knowledge base and enhanced problem-solving abilities, showing promise in diagnostic tasks. Details about these LLMs can be found in Supplementary Table 2. All open-access models were fine-tuned and employed self-consistency (SC) decoding to evaluate their diagnostic capacity.

In the ID setting evaluation, we constructed the MedDX-Test dataset, which encompasses common fine-grained diseases representing 99% of the population across eight specialties. For example, we evaluated the model's ability to diagnose autoimmune thyroiditis

(a specific type of thyroid disorder) rather than simply categorizing it as a general thyroid disease. For average performance across all specialties, our model demonstrated superior performance, achieving a diagnostic Top-3 average accuracy of 84.2% (95% confidence interval (CI): 83.5%, 84.8%) (Fig. 2a). This represents a substantial improvement over the other four models, with average accuracies ranging from 64.8% (95% CI: 63.9%, 65.6%; Clinical Camel-70B) to 56.8% (95% CI: 55.9%, 57.7%; MEDITRON-70B). Among these, GPT-4o achieved a diagnostic accuracy of 62.0% (95% CI: 61.1%, 62.8%), slightly lower than the next-best-performing model, Clinical Camel-70B. We stratified the results by specialty (for example, cardiology, neurology and endocrinology) to provide detailed insights into the LLM-based diagnostic generalist (Fig. 2b). Our MedFound consistently outperformed other LLMs, with accuracies ranging from 82.4% to 89.6%. We also evaluated the models using Top-1 accuracy, macro accuracy, receiver operating characteristic area under the curve (ROC-AUC) and precision-recall area under the curve (PR-AUC) metrics, with results similarly indicating the superior performance of MedFound-DX-PA (Extended Data Fig. 5 and Supplementary Table 3).

Furthermore, we evaluated the generalizability of our model on the MedDX-OOD dataset, an OOD setting where cases were collected from external real-world environments. Figure 2c,d presents the average and the stratified performance across each specialty, respectively. MedFound-DX-PA significantly outperformed the baseline models in all specialties (all $P < 0.001$). The results demonstrate the generalizability of our model as a diagnostic generalist across a variety of clinical diseases, especially in fine-grained disease diagnosis.

We also extended our diagnostic generalist to specialty scenarios that require specific knowledge of a particular medical field. We assigned the role of diseases specialist to the LLM-based generalist by prompting it for specialty-specific settings (as detailed in Methods). Our model achieved Top-3 accuracies ranging from 87.9% (95% CI: 87.2%, 89.6%) to 93.9% (95% CI: 92.6%, 95.9%) on the MedDX-Test dataset and 85.8% (95% CI: 83.4%, 88.6%) to 90.2% (95% CI: 88.7%, 93.5%) on the MedDX-OOD dataset (Fig. 2e,f), demonstrating that our model can be adaptive to meet the precision requirements of these specialty settings. We also compared our MedFound with existing specialized decision support tools on multi-class, disease-specific tasks using open-access datasets (Supplementary Table 4). The results indicate that our model's performance is similar to or exceeds that of the specialized tools.

Performance of the LLMs on rare diseases

We expanded our experiments to examine the performance of the LLMs in diagnosing rare diseases characterized by long-tailed distributions[24]. Previous models have shown effectiveness in identifying common diseases[25], but their performance tends to decline in classifying rarer diseases in few-shot or zero-shot scenarios. As illustrated in Fig. 3a, the distribution of diseases reveals a long-tailed distribution, with common diseases covering 99% of the population and the remaining 1% comprising a wide variety of less common diseases. To evaluate the adaptability of the LLMs in diagnosing a broad spectrum of conditions, we used a zero-shot learning setting on the MedDX-Rare dataset, which includes 2,105 rare diseases derived from long-tailed distribution across eight specialties (Fig. 3b and Extended Data Fig. 6a). Bar plots in Fig. 3c illustrate the Top-3 accuracy of MedFound-DX-PA for each fine-grained rare disease within each specialty, and radar plots show the overall performance of each specialty across diseases (as detailed in Methods). MedFound-DX-PA excelled across all specialties, ranging from 77.4% (95% CI: 76.8%, 78.0%) to 84.4% (95% CI: 83.9%, 84.9%), with an average of 80.7% (95% CI: 80.1%, 81.2%) (Fig. 3c). GPT-4o achieved the second-best performance, ranging from 57.2% (95% CI: 56.5%, 57.9%) to 63.1% (95% CI: 62.4%, 63.8%), with an average of 59.1% (95% CI: 58.4%, 59.8%). This trend was also observed in the Top-1 macro accuracy (Extended Data Fig. 6b).

The average performance of LLMs was further assessed using Top-3 micro accuracy, which considers individuals equally over each specialty to mitigate the impact of classes with small sample sizes, as shown in Fig. 3d. The second-best LLM, GPT-4o, achieved a moderate performance, ranging from 77.4% (95% CI: 76.9%, 78.0%) to 85.8% (95% CI: 85.3%, 86.2%), with an average of 82.1% (95% CI: 81.6%, 82.7%). In comparison, MedFound-DX-PA excelled across all specialties, showing a substantial performance improvement, ranging from 87.4% (95% CI: 87.0%, 87.9%) to 93.0% (95% CI: 92.7%, 93.4%), with an average of 89.2% (95% CI: 88.8%, 89.6%). Additional metrics, such as ROC-AUC and PR-AUC, also demonstrated the superior performance of MedFound-DX-PA compared to other LLMs (Extended Data Fig. 6c and Supplementary Table 3). To further investigate the model's diagnostic performance in long-tailed disease distributions involving varying prevalence, we categorized them into ultra-rare (≤0.1% prevalence) and rare (0.1–1% prevalence) groups[26] (Extended Data Fig. 6d and Supplementary Table 5). The results demonstrate that MedFound-DX-PA performed consistently well between these two groups. This can be attributed to MedFound-DX-PA's generative ability and its comprehensive understanding of the diagnostic structure, which offers flexibility in adapting to fine-grained rare diseases.

Performance comparison between the LLM versus physicians

Here, we compare the diagnostic capacities of our LLM-based diagnostic system with those of human physicians in endocrinology and pulmonology. Eighteen physicians were recruited, including nine endocrinologists and nine pulmonary physicians, and were further categorized by clinical experience into three groups: Junior ($n = 3$), Intermediate ($n = 3$) and Senior ($n = 3$). Each physician was allocated 150 cases to diagnose. Extended Data Fig. 3a illustrates the interface used by the physicians for this evaluation task. Performance was measured against the gold-standard diagnoses established by an expert panel. In pulmonology, MedFound-DX-PA achieved a diagnostic accuracy of 72.6%, surpassing junior physicians (60.0%) and intermediate physicians (67.7%) but slightly lower than that of senior physicians (76.2%) (Fig. 4a). Similarly, in endocrinology, the AI's accuracy (74.7%) exceeded that of junior physicians (69.4%) and intermediate physicians (72.5%) and was similar to that of senior physicians (75.2%) (Fig. 4b). These results demonstrate that our LLM-based diagnostic generalist outperformed junior and intermediate physicians in both specialties and was similar to senior physicians.

Performance of the LLM-assisted diagnosis within workflows

We further explored the LLM's potential role in enhancing diagnostic performance of physicians in the clinical workflow. When provided with EHR notes (with diagnoses removed), junior and intermediate physicians from the two specialties performed their initial diagnosis. Two weeks later, they referenced the AI-generated content to formulate their second diagnosis (Extended Data Fig. 3b). In pulmonology, AI assistance substantially enhanced the accuracy of junior and intermediate physicians, by 11.9% and 4.4%, respectively, with performance approaching that of the AI system but remaining slightly below that of senior physicians (Fig. 4a). For instance, for a case shown in Fig. 5a, the physician initially diagnosed 'acute bronchitis' based on the patient's present medical history and C-reactive protein levels from laboratory tests. Then, with the assistance of AI-generated content, which emphasized the patient's history of recurrent bronchitis, the physician revised the diagnosis to the accurate diagnosis of 'acute exacerbation of chronic bronchitis'.

In endocrinology, the accuracy of both junior and intermediate endocrinologist groups substantially increased to 74.0% (an increase of 4.6%) and 78.8% (an increase of 6.3%), respectively, after AI assistance (Fig. 4b). Notably, intermediate endocrinologists with AI assistance outperformed senior endocrinologists, indicating the potential of AI to enhance diagnostic accuracy beyond most experienced physicians ($P < 0.05$). For instance, as illustrated in Fig. 5b, the initial diagnosis of

https://doi.org/10.1038/s41591-024-03416-6

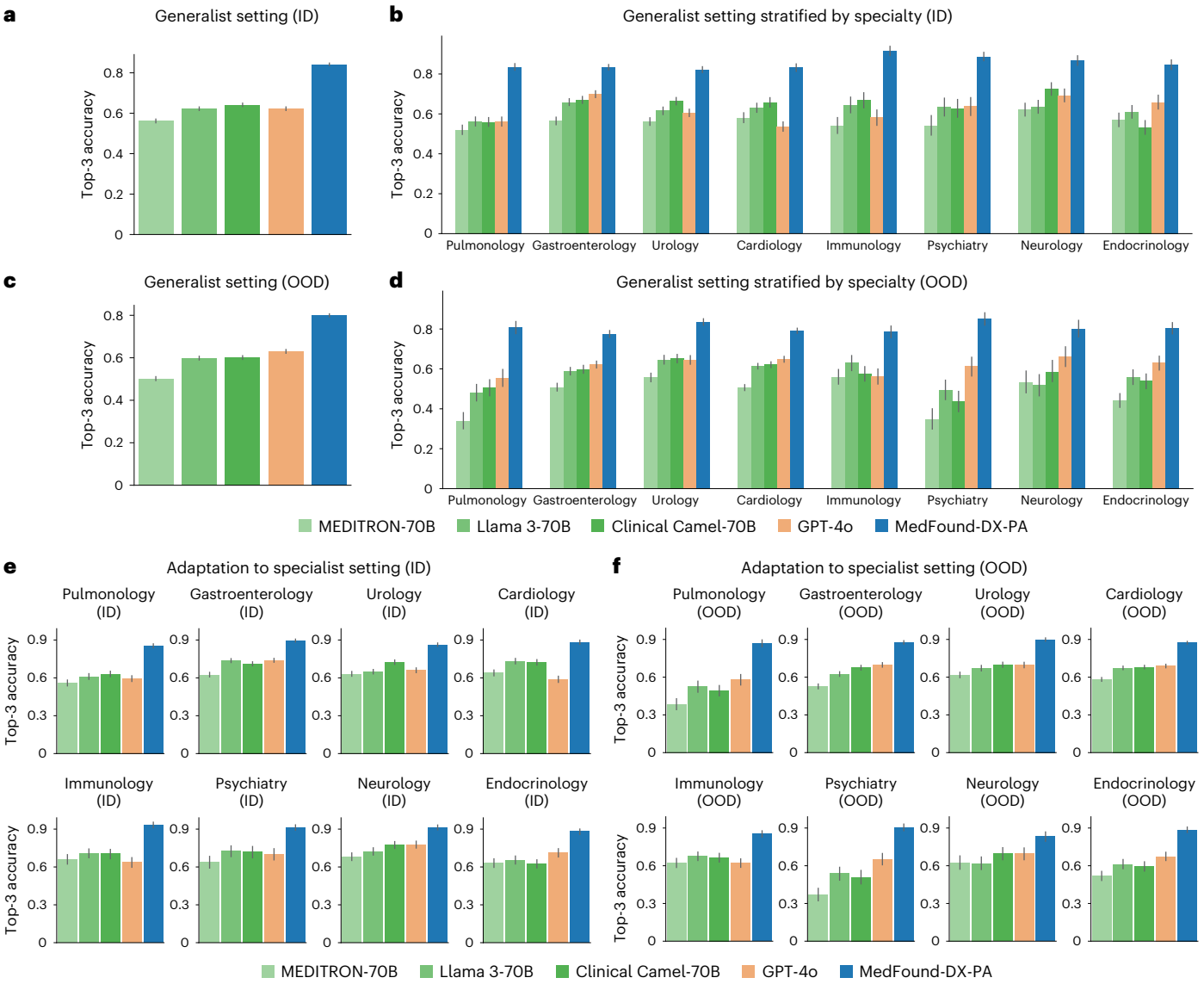


Fig. 2 | Performance of the LLMs for diagnosis of common diseases across various specialties. a–**d**, Comparison of Top-3 accuracy among MEDITRON-70B, Llama 3-70B, Clinical Camel-70B, GPT-4o and our MedFound-DX-PA, for diagnostic tasks in generalist settings. The results are shown in ID settings (n = 11,662) (**a** and **b**) and OOD settings (n = 23,917) (**c** and **d**) across eight specialties. **a** and **c** represent the overall performance, and **b** and **d** represent the performance stratified by specialty. **e**,**f**, Comparison of Top-3 accuracy among the LLMs in specialist-specific ID and OOD settings across eight specialties. Bar graphs indicate the mean ± 95% CIs.

subclinical hypothyroidism was made when the physician observed elevated levels of thyroid-stimulating hormone in the patient's laboratory tests. During re-evaluation with AI assistance, the model highlighted previously overlooked elevated anti-thyroid peroxidase antibody levels, indicating a possible underlying autoimmune thyroid disorder. Consequently, the physician revised the diagnosis to 'autoimmune thyroiditis'. These results suggest that physicians can benefit from the LLM model's assistance by highlighting important clinical data, thus enhancing healthcare delivery.

Human evaluation framework for AI's diagnostic capabilities

Previous evaluation metrics focus mostly on measures such as accuracy or natural language generation scores (for example, BLEU or ROUGE), which fail to capture the clinical quality of inferential diagnostic process. To address this issue, we proposed a systematic evaluation framework for AI in real-world diagnosis, established through a process of literature review and consultations with expert physicians. The framework CLEVER categorizes the capabilities of the LLM-based system into eight clinical evaluation metrics, providing insights into the strengths and limitations of LLMs in aligning with medical standards (as detailed in the Methods). For the assessment, six senior physicians were recruited from the previous two specialties, using a Likert scale rating system ranging from 1 to 5 (Fig. 4c and Extended Data Fig. 4).

In 'Medical case comprehension', the expert panels evaluated the ability of the LLM to understand and interpret medical cases, such as assessing whether its content contains information required for diagnosis with completeness and correctness. Our proposed MedFound-DX-PA achieved a score of 4.02 in 'Medical case comprehension', surpassing 3.77 of the unaligned LLM model significantly ($P < 0.05$). A similar trend was also observed in 'Clinical reasoning', which was used to evaluate whether the LLM's inferential diagnosis aligned with the diagnostic reasoning process of physicians in clinical practice. MedFound-DX-PA demonstrated superior performance, with a score of 4.07, surpassing the unaligned models at 3.63 significantly ($P < 0.01$). In 'Medical guidelines and consensus', physicians were asked to assess whether the LLM's generation aligned with established

https://doi.org/10.1038/s41591-024-03416-6

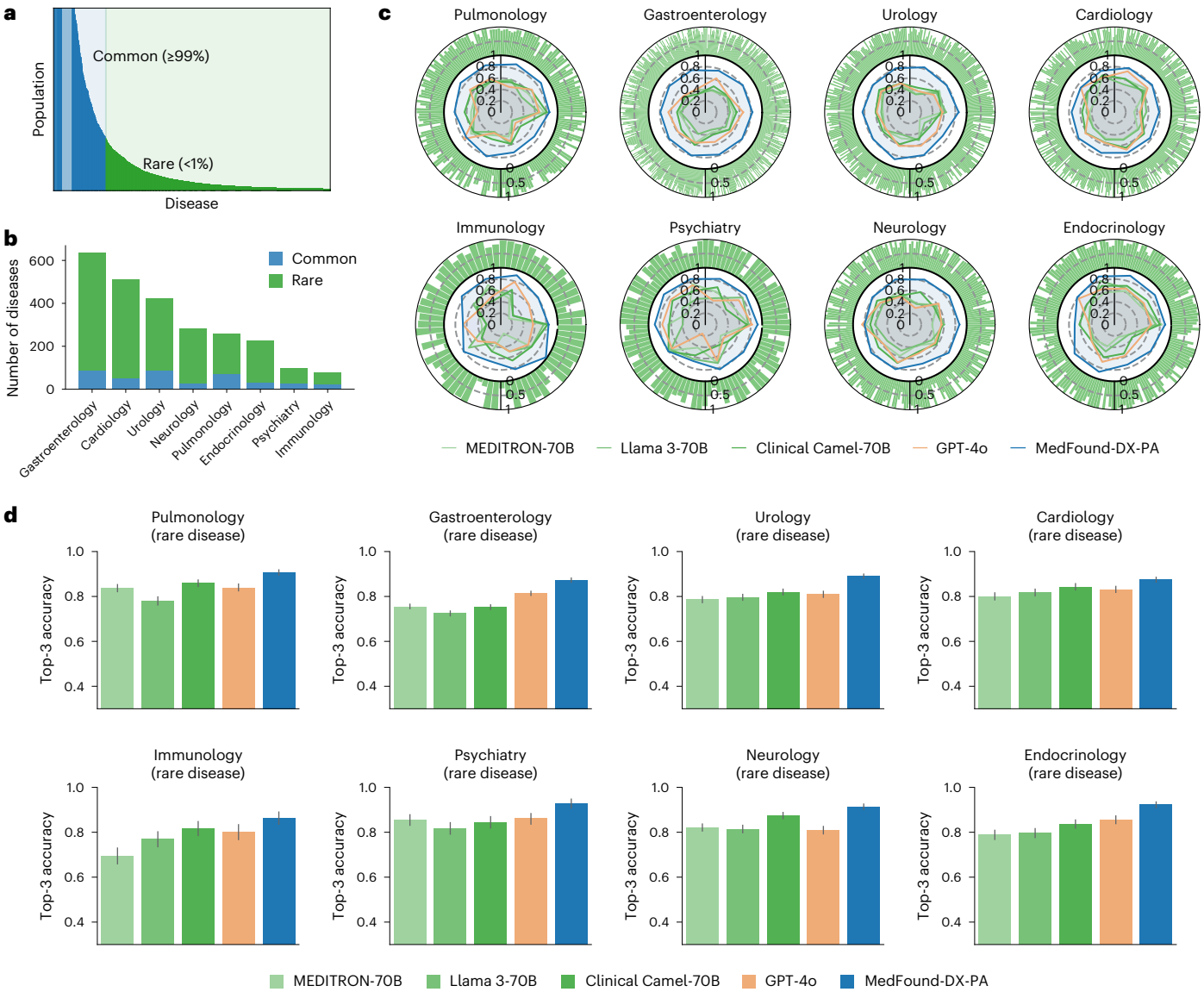


Fig. 3 | Performance of the LLMs for diagnosis of rare diseases across various specialties. a, Distribution of disease prevalence. The *x* axis indicates a range of diseases from common to rare. The *y* axis represents the population size of individuals affected by each disease. The curve is divided into two regions. The blue region represents common diseases (cumulative prevalence ≥99%), and the green region represents rare diseases (cumulative prevalence <1%). **b**, Distribution of disease number across eight specialties ($n = 20,257$). The blue bar represents the number of common diseases, and the green blue bar represents rare diseases. **c**, Performance comparison of Top-3 macro accuracy among MEDITRON-70B, Llama 3-70B, Clinical Camel-70B, GPT-4o and our MedFound-DX-PA, for diagnosing rare diseases across eight specialties. Radar maps show the Top-3 macro accuracy of LLMs on each specialty's performance by aggregating at the octiles of disease prevalence. Bar graphs indicate the Top-3 accuracy of MedFound-DX-PA for individual diseases within each specialty. **d**, The Top-3 micro accuracy over individuals among the LLMs for diagnosing rare diseases across eight specialties. Bar graphs indicate the mean ± 95% CIs.

medical guidelines and consensus. MedFound-DX-PA achieved a Likert score of 3.83, whereas the unaligned model achieved a score of 3.62 ($P = 0.18$). These results indicate that our model can capture relevant medical evidence and incorporate diagnostic reasoning, potentially offering enhanced clinical decision-making support.

We also sought to assess the efficacy of LLMs in supporting clinical decision-making. For 'Relevance of differential diagnosis', physicians assessed the LLM's capacity to differentiate among multiple possible conditions that could cause a patient's symptoms. Our model achieved a score of 3.93, surpassing the unaligned models with 3.62 ($P < 0.05$). The 'Acceptability of diagnosis' is used to rate whether the diagnosis is acceptable or reliable for clinical use. In this category, our model achieved a score of 4.21, significantly outperforming the unaligned models at 3.72 ($P < 0.001$). These findings demonstrate the potential clinical feasibility of our diagnostic generalist.

LLMs in critical clinical scenarios are expected to avoid generating inaccurate or misleading information ('Unfaithful content') or demonstrate varying levels of stereotypes related to gender, culture and race ('Bias and unfairness'). Also, it is crucial that the generated content of an LLM does not contain any incorrect or harmful evidence, which could potentially lead to misdiagnosis or mislead physicians about possible medical accidents ('Possibility of harm'). We examined the model's risk control capability by assessing 'Unfaithful content', 'Bias and unfairness' and 'Possibility of harm'. Our model demonstrated superior performance, with scores of 4.11, 4.14 and 4.03 in the three metrics, respectively, surpassing the unaligned model at 3.66 ($P < 0.01$), 3.82 ($P < 0.05$) and 3.66 ($P < 0.01$), with significance. The results indicate that LLM-based systems can be optimized through alignment with human values, thus enhancing their trustworthiness and clinical applicability.

https://doi.org/10.1038/s41591-024-03416-6

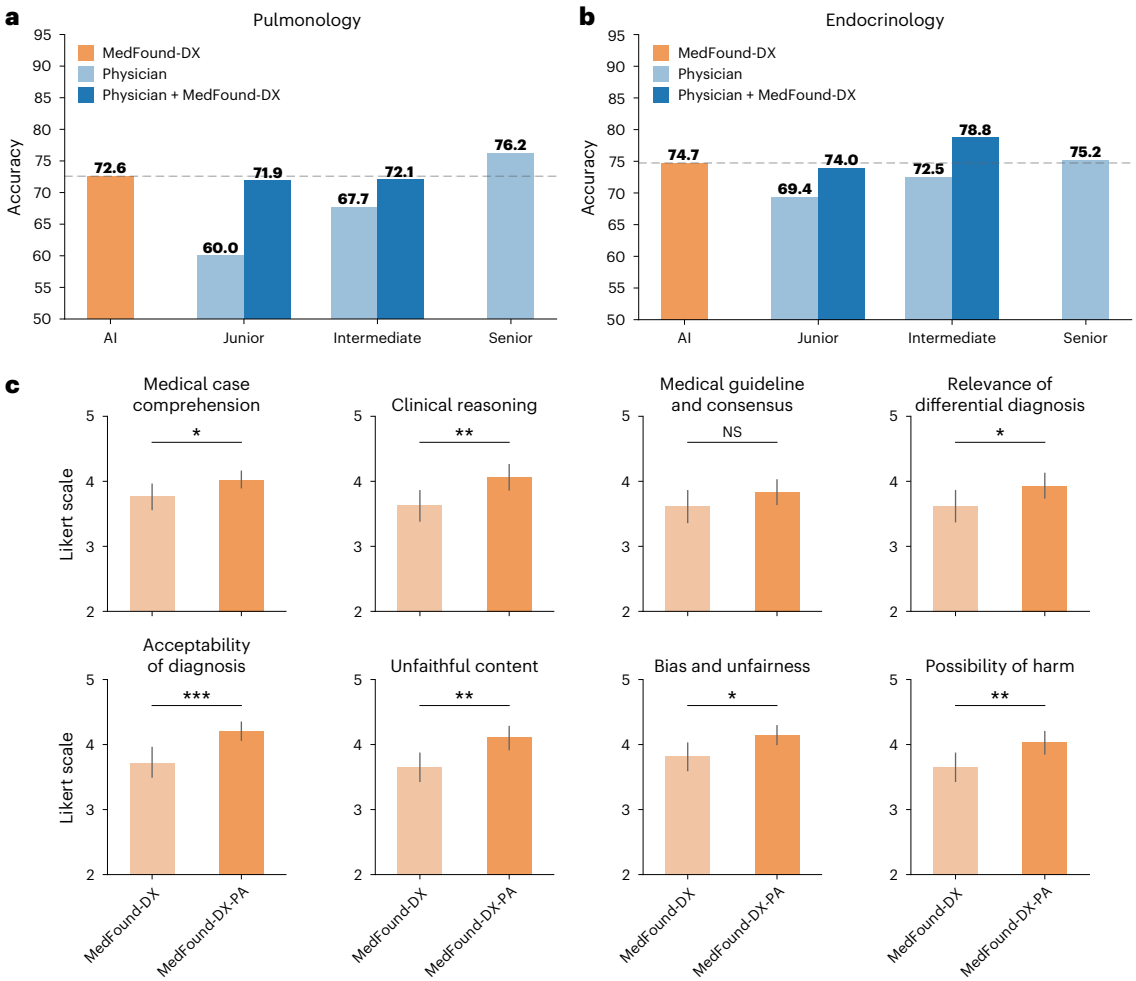


Fig. 4 | Performance evaluation between the AI system and human physicians for diagnosis. a,**b**, Performance comparison of diagnostic reasoning given by MedFound-DX-PA and human physicians in pulmonology (**a**) and endocrinology (**b**) ($n = 900$). Bars represent the diagnostic accuracy of the AI system (orange), human physicians (light blue) and physicians assisted by MedFound-DX-PA (dark blue). The gray dashed line represents the performance of MedFound-DX-PA. **c**, Human evaluation between MedFound-DX and MedFound-DX-PA across eight dimensions, including metrics of 'Medical case comprehension' ($P = 0.032$), 'Clinical reasoning' ($P = 0.006$), 'Medical guideline and consensus' ($P = 0.180$), 'Relevance of differential diagnosis' ($P = 0.036$), 'Acceptability of diagnosis' ($P < 0.001$), 'Unfaithful content' ($P = 0.002$), 'Bias and unfairness' ($P = 0.015$) and 'Possibility of harm' ($P = 0.009$). Bar graphs indicate the mean ± 95% CIs for MedFound-DX (light orange) and MedFound-DX-PA (dark orange). Statistical analyses were performed using a two-sided t-test. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, NS (not significant) $P > 0.05$.

Impact of training components on the performance of LLMs

To explore the impact of key components of our proposed approach on the diagnostic performance of LLMs, we conducted experiments using MedFound and the latest leading LLMs, including Clinical Camel-70B, Llama-3-70B and MEDITRON-70B, using MedDX-Bench. We first investigated the inherent diagnostic capabilities of LLMs by adapting MED-Prompt, which familiarizes LLMs with the medical tasks and allows them to adapt to diagnostic tasks without any additional training. The results show that MedFound (without SC) achieved superior performance, with micro accuracy improvements of 14.4%, 11.9% and 11.1% compared to the average performance of other LLMs on MedDX-Test, MedDX-OOD and MedDX-Rare, respectively (Fig. 6a). For example, MedFound achieved accuracy of 37.2% (95% CI: 36.3%, 38.1%), outperforming the second-best LLM with performance of 30.8% (95% CI: 29.9%, 31.6%; Clinical Camel-70B) on MedDX-Test. Similar results were also observed with other evaluation metrics, such as macro accuracy, detailed in Supplementary Table 6. This suggests that MedFound is a more effective pre-trained model for diagnostic tasks, probably due to its comprehensive pre-training on MedCorpus, which enabled it to encode extensive medical knowledge and practical clinical cases.

In addition to evaluating the inherent diagnostic capabilities of the pre-trained LLMs, we further investigated the impact of COT fine-tuning on enhancing the diagnostic accuracy of LLMs through additional training on domain-specific data. As shown in Fig. 6b and Supplementary Table 7, all models demonstrated improved performance on the MedDX-Bench tasks after fine-tuning, with average micro accuracies increasing by 14.9%, 15.9% and 12.7% on MedDX-Test, MedDX-OOD and MedDX-Rare, respectively. After the COT fine-tuning, we employed SC decoding technique to further enhance the model's reasoning performance by reducing variability and aligning its outputs with domain-specific requirements. As shown in Extended Data Fig. 7 and Supplementary Table 3, all LLMs showed improvements by employing SC, with gains in micro accuracies for MedDX-Test, MedDX-OOD and MedDX-Rare by 12.1%, 11.1% and 7.6%, respectively. In conclusion, our proposed pre-training–fine-tuning approach substantially enhances the performance of LLMs on medical tasks, especially for challenging cases, such as rare diseases.

Discussion

In this study, we developed MedFound-DX-PA, a large-scale medical assistant for general diagnosis to approximate clinician expertise

a

Case: Gender: Male, Age: 64. Chief Complaint: Fever accompanied by cough for 4 days. History of Present Illness: The patient has experienced fever with cough for four days. Despite undergoing anti-infection treatment, the patient continues to exhibit coughing, with no significant improvement in symptoms. Past medical history: a history of bronchitis. Radiologic examinations: Chest CT Scan: increased lung markings, with localized areas of increased translucency. There are a few linear and nodular areas of increased density in both lungs with clear boundaries. The examination suggests chronic bronchitis and pulmonary emphysema. Laboratory testing: C-reactive protein (CRP) 24.65 mg/L (reference 0~10).

Physician's initial diagnosis: Acute Bronchitis.

Model's diagnostic rationale:

- The chief complaint is fever accompanied by cough for four days. The history of present illness indicates no significant improvement after anti-infection treatment, with a past medical history of bronchitis. From this information, it can be inferred that the patient may have a respiratory infection or inflammation. The physical examination is unremarkable, providing no additional clues.
- Combining the above information, the possible diagnoses include respiratory infection or inflammation.
- Further, based on the analysis of laboratory tests and radiologic reports: The C-reactive protein (CRP) test result is 24.65 mg/L, exceeding the normal reference range, indicating an inflammatory response. CRP is an acute-phase protein, commonly elevated in diseases such as inflammation and infection. Chest CT scan reveals increased and disorganized lung markings, localized areas of increased translucency, and a few linear and nodular areas of increased density with clear boundaries. These findings are associated with chronic bronchitis and emphysema.
- Other possible diagnoses: Respiratory infection: Given the patient's chief complaint and history of present illness, there is a possibility of respiratory infection. Respiratory infections can be caused by a variety of pathogens, including viruses and bacteria. Emphysema: According to the CT scan results, there is a possibility of emphysema. Emphysema is a chronic obstructive pulmonary disease characterized by airway obstruction and loss of lung tissue elasticity, leading to increased lung volume.
- Therefore, the most likely diagnosis is an acute exacerbation of chronic bronchitis.

b

Physician's revised diagnosis: Acute Exacerbation of Chronic Bronchitis.

Case: Gender: Female, Age: 42. Chief Complaint: Abnormality detected in thyroid. History of Present Illness: The patient was found "heterogeneous echotexture" on thyroid ultrasound in 2016. Thyroid function tests were normal, and no treatment was administered. There were no fatigue. Physical Examination: No exophthalmos. No tenderness in thyroid. Laboratory Test: Thyroid-stimulating hormone: 6.072 μIU/mL (Reference range: 0.35~5.5); Thyroid peroxidase antibody: >10000 IU/ml (Reference range: <100); Thyroglobulin antibody: 50.4 U/ml (reference range: <60); Free thyroxine: 14.20 pmol/L (reference range: 11.5~22.7); Free triiodothyronine: 4.17 pmol/L (reference range: 3.5~6.5); Parathyroid hormone: 46.40 pg/ml (reference range: 14.5~87.1).

Physician's initial diagnosis: Subclinical Hypothyroidism.

Model's diagnostic rationale:

- The chief complaint is the discovery of thyroid abnormalities. The history of present illness mentions the findings of "heterogeneous echotexture" on thyroid ultrasound in 2016, with normal thyroid function tests and no treatment administered. Physical examination shows no tenderness. This suggests the presence of thyroid abnormalities without apparent symptoms. There is no mention of any thyroid-related diseases or surgical history in the past medical history.
- Taking the above information into account, possible diagnoses include thyroid nodules, thyroiditis, etc.
- Furthermore, based on laboratory tests :The laboratory test shows normal levels of free thyroxine and free triiodothyronine, but elevated levels of thyroid peroxidase antibody, suggesting the possibility of autoimmune thyroid disease. Thyroid peroxidase antibody is a marker of autoimmune thyroid diseases, and its elevation may be related to thyroid dysfunction. The parathyroid hormone result shows parathyroid hormone levels within the normal range, ruling out the possibility of parathyroid dysfunction.
- Other possible diagnoses: Thyroid nodules: Thyroid nodules are common benign lesions of the thyroid gland. The formation of thyroid nodules may be associated with thyroid dysfunction and thyroiditis. Subacute thyroiditis: Subacute thyroiditis is a disease caused by inflammation of the thyroid gland, which may lead to thyroid dysfunction and thyroid enlargement. It could cause tenderness in thyroid.
- Therefore, the most likely diagnosis is autoimmune thyroid disease.

Physician's revised diagnosis: Autoimmune Thyroiditis.

Fig. 5 | Examples of AI-assisted diagnosis evaluation. a,**b**, Examples of diagnostic rationales generated by the MedFound-DX-PA model to assist physicians in their diagnostic process. **a**, Case used in the evaluation with pulmonary physicians. **b**, Case used in the evaluation with endocrinologists. Each case includes the patient's symptoms, medical history and examinations. Physicians made an initial diagnosis upon their first review of the clinical notes and re-diagnosed after considering the diagnostic rationale provided by AI assistance.

across various healthcare scenarios. When evaluated on MedDX-Bench, MedFound-DX-PA demonstrateed superior diagnostic performance across specialties and conditions, including ID and OOD settings for common diseases, as well as for rare diseases. Furthermore, we conducted comparison studies involving MedFound-DX-PA versus specialists and an AI-assistance study, which indicate its potential to enhance the diagnostic capability of junior or intermediate physicians. Additionally, the human evaluation study of LLMs demonstrates that our MedFound-DX-PA has potential as a generalist for integration into clinical workflows.

Disease diagnosis is crucial for everyday clinical tasks and is prone to errors, which may lead to adverse outcomes or treatment that is withheld or delayed. Previous AI-assisted diagnostic tools include rule-based CDSSs, machine learning on structured features in EHR and PLMs. However, their applicability is limited by their specific training data and model size, necessitating specialty-specific models that are inefficient.

Recent advancements demonstrate the potential of LLMs that can interpret and generate text effectively with minimal or no specific fine-tuning, facilitating versatile applications such as interactive decision support and patient chatbots[27]. However, there are considerable challenges in applying LLMs to the clinical setting. Existing LLMs often fail to capture the vast range of medical knowledge and scenarios. Furthermore, the output of generative language models may contain factual errors, logic inconsistencies and problems with coherence[28]. For example, ChatGPT has been found lacking in depth and insight[29], which produces overly generalized answers that lack medical expertise. To bridge this gap, we introduce MedFound, which

https://doi.org/10.1038/s41591-024-03416-6

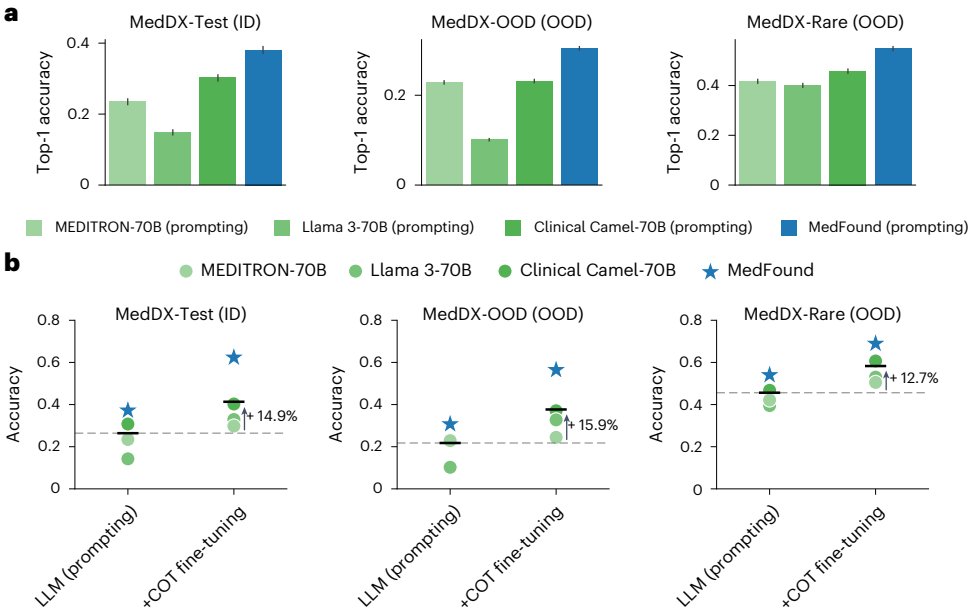


Fig. 6 | Performance analysis of LLM training components for various diagnostic tasks. a, Comparison of accuracy among various pre-trained LLMs via MED-Prompt for diagnostic tasks on MedDX-Test (ID testing on common diseases) (left) (n = 11,662), MedDX-OOD (OOD testing on common diseases) (middle) (n = 23,917) and MedDX-Rare (OOD testing on rare diseases) (right) (n = 20,257). The error bars represent the 95% CIs. **b**, Impact analysis of COT fine-tuning on the accuracy of various LLMs for diagnostic tasks on MedDX-Test (ID testing on common diseases) (left), MedDX-OOD (OOD testing on common diseases) (middle) and MedDX-Rare (OOD testing on rare diseases) (right). The short horizontal line shows the mean performance of a set of models. The percentage increases shown are the improvements gained through COT fine-tuning.

is, to our knowledge, the largest open-access medical LLM, with 176 billion parameters pre-trained on a diverse range of medical corpora. Second, we fine-tuned MedFound by employing self-bootstrapping based COT fine-tuning to boost the reasoning capabilities of medical LLMs. The self-bootstrapping approach uses prompts to guide the LLM in automatically generating large-scale rationales with only hundreds of annotations, thus reducing the cost of expert annotation. Subsequently, we introduced a unified PA framework, aligning MedFound-DX with both ICD-10 diagnostic preference and clinician-evaluated helpfulness preference, ensuring trustworthiness and safety in critical medical tasks.

Although previous studies highlighted the performance of classification-based decision support tools in specific specialties, we sought to compare these tools with an LLM-based diagnostic generalist in real clinical scenarios. We included three representative classification models: a traditional machine learning approach using hierarchical classification (hierarchical random forest (HRF)[30]); a pre-trained language model tailored for the medical domain using a masked language modeling strategy (Med-BERT[9]); and a variant of our MedFound as a pre-trained backbone for a classifier (MedFound-CLS), as detailed in Methods. The results indicate that MedFound-DX-PA outperformed the second-best model MedFound-CLS by 17.8% on the MedDX-Test and by 35.7% on the MedDX-OOD datasets, highlighting the superiority of generative models over classification approaches in diagnostic tasks, particularly in OOD scenarios (Extended Data Fig. 8). Furthermore, although existing specialized decision support tools demonstrate certain effectiveness in specific specialties[25], they are limited to identifying pre-defined coarse-grained disease categories or often struggle with zero-shot scenarios, where they must diagnose diseases that they have never explicitly been trained to recognize. In contrast, medical LLMs offer a promising solution in diagnosing rare diseases within few-shot and zero-shot settings. Our model effectively handles rare conditions by reasoning over new input samples in a manner akin to human experts (Figs. 3 and 5). This zero-shot approach using foundation models may open up possibilities for broader medical applications that were previously challenging to address. Another advantage of our diagnostic generalist model is its ability to generate diagnostic reasoning, making the model's output transparent and increasing physicians' trust in AI-driven diagnostic tools.

Additionally, we conducted a comprehensive clinical validation of the LLM-based diagnostic system within practical clinical scenarios. In the study, we established a benchmark using real-world EHR data across various specialties in diagnosing a range of diseases from common to rare. When compared to other LLMs, MedFound-DX-PA demonstrated superior performance across different distributions, highlighting the model's accurate and robust capacities as a generalist. To evaluate the LLM-based model's generated contents more thoroughly, we developed a clinician evaluation framework, covering a wide array of aspects. Given that ensuring safety is crucial for practical clinical scenarios, our human evaluation framework assesses various safety considerations, such as unfaithful content, bias, unfairness and the possibility of harm. We also conducted a privacy risk assessment[31] that demonstrated that our model has a low risk of information leakage (Supplementary Fig. 1). As shown in Fig. 4a,b, the results demonstrated that our model considerably improves physician performance, underscoring the potential role of LLMs in augmenting the diagnostic capabilities of physicians within clinical workflows. Furthermore, we observed that some physicians could not surpass the original AI even with AI assistance. This phenomenon has also been observed in previous studies (for example, mammography cancer detection[32] and chest X-ray interpretation[33]). Research suggests that human–AI collaboration faces challenges related to human mental models of the AI, which probably depend on their degree of familiarity with the AI or the reliance on proposed decisions[34,35]. This also highlights the need to further study the impact of AI aids on human cognition and observed performance.

The LLM-based diagnostic generalist has the potential to assist physicians across various stages in clinical workflows, including information gathering, data summarization and interpretation, diagnostic reasoning and formulating final diagnoses[36–38]. First, our MedFound-DX-PA can generate diagnostic reasoning that covers a wide range of common

or rare diseases across specialties. This makes it particularly useful in clinical scenarios requiring extensive medical knowledge of diseases such as pre-diagnostic triaging and prioritization or serving as a consultation 'co-pilot'. For example, during pre-diagnostic assessments, MedFound-DX-PA can synthesize patient symptoms, recommend further diagnostic testing or direct patients to appropriate specialties. For primary care physicians who encounter a broad range of diseases in daily clinical work[39], they can initiate referrals based on MedFound-DX-PA prompts to access more specialized expertise, such as cardiology or neurology. For complex and multisystem diseases, MedFound-DX-PA could offer multidisciplinary consultation support, promoting a more holistic approach to patient care compared to task-specific tools. Additionally, the diagnostic generalist system could facilitate telemedicine by overcoming challenges in resource-limited settings[40,41] by alleviating physician workload through automated integration between clinical assessments.

In addition, our diagnostic generalist can also efficiently adapt to specialty scenarios or specific diseases with minimal prompting, offering superior performance and interpretability compared to existing specialized models. We envision that MedFound-DX-PA can facilitate AI-assisted consultations by providing specialist expertise to less experienced physicians, enhancing differential diagnosis or aiding in the refinement of final diagnoses. For example, the system can interpret laboratory or radiological results[42], identify abnormalities and summarize critical evidence from a specialist's diagnostic assessments, as demonstrated in Fig. 5. In the subsequent differential diagnosis phase, MedFound-DX-PA will enhance the quality of diagnostic care by considering all available evidence, offering diagnostic rationales and proposing differential diagnoses to the physician. Physicians who participated in our study also demonstrated improved diagnostic accuracy by incorporating this AI system into their clinical practice.

Although our model has demonstrated impressive diagnostic performance, several challenges remain. First, our medical LLM currently focuses on language interaction, and its capabilities could be extended by integrating with medical multimodal data through vision–language models (VLMs). VLMs have shown promise in fields such as pathology, radiology and echocardiography[43–46]. These advancements are powered by LLMs, which provide extensive domain knowledge and reasoning capabilities[47], enabling VLMs to perform zero-shot image-to-text generation based on natural language instructions, unlocking emerging capabilities such as visual knowledge reasoning and visual conversation. In the future, integrating VLMs could enable MedFound-DX-PA to adopt a more comprehensive, multimodal approach to diagnosis and patient care, opening new possibilities for AI-assisted healthcare. Furthermore, to enhance the human–computer collaboration for the integration of AI into routine clinical workflows, future work will focus on refining LLM models, such as LLM agents[48], to better adapt to individual physicians, thereby enhancing the personalization of diagnostic support. The evaluation interaction between the model assisting physicians and the feedback from physicians can also refine the model, known as human-in-the-loop[49], enabling the LLM system to evolve continuous improvement in a manner that aligns more closely with the practical needs of clinical environments. These future directions will be instrumental in enhancing the practical integration of AI into clinical workflows and maximizing its potential to benefit healthcare practices or the diagnostic training of primary care.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03416-6.

References

1. Scully, J. L. What is a disease? Disease, disability and their definitions. *EMBO Rep.* **5**, 650–653 (2004).
2. Kaur, S. et al. Medical diagnostic systems using artificial intelligence (AI) algorithms: principles and perspectives. *IEEE Access* **8**, 228049–228069 (2020).
3. Graber, M. L. The incidence of diagnostic error in medicine. *BMJ Qual. Saf.* **22**, ii21–ii27 (2013).
4. Stern, S.D. *Symptom to Diagnosis: An Evidence-Based Guide* (McGraw-Hill, 2014).
5. Wasylewicz, A.T. & Scheepers-Hoeks, A. Clinical decision support systems. In *Fundamentals of Clinical Data Science* (eds Kubben, P., Dumontier, M. & Dekker, A.) 153–169 (Springer, 2019).
6. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
7. Devlin, J. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
8. Radford, A., & Narasimhan, K. Improving language understanding by generative pre-training. *J. Comput. Linguist.* https://openai.com/research/language-unsupervised (2018).
9. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1904.05342 (2020).
10. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
11. Yang, X. et al. A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194 (2022).
12. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
13. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
14. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
15. Collins, K. M. et al. Building machines that learn and think with people. *Nat. Hum. Behav.* **8**, 1851–1863 (2024).
16. Ouyang, L. et al. Training language models to follow instructions with human feedback. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 27730–27744 (Curran Associates, 2022).
17. Zhang, H., Xu, W. & Yu, H. Generative planning for temporally coordinated exploration in reinforcement learning. In *Proc. 10th International Conference on Learning Representations* https://openreview.net/pdf/0e68ff1fa269567c6c6101685f2f721afcc5d0aa.pdf (ICLR, 2022).
18. Kirk, H. R., Vidgen, B., Röttger, P. & Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.* **6**, 383–392 (2024).
19. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. In *Proc. 37th International Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 53728–53741 (Curran Associates, 2023).
20. Chen, Z. et al. MEDITRON-70B: scaling medical pretraining for large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2311.16079 (2023).
21. Toma, A. et al. Clinical Camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.12031 (2023).

22. Meta AI. Introducing Meta Llama 3: the most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/ (2024).
23. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
24. Shen, T., Lee, A., Shen, C. & Lin, C. J. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. *Genet. Res.* **97**, e15 (2015).
25. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
26. Smith, C. E., Bergman, P. & Hagey, D. W. Estimating the number of diseases—the concept of rare, ultra-rare, and hyper-rare. *iScience* **25**, 104698 (2022).
27. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
28. Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
29. Brown, T. et al. Language models are few-shot learners. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
30. Aslam, M. & Jaisharma, K. Hierarchical random forest formation with nonlinear regression model for cardiovascular diseases prediction. In *Proc. 2021 International Conference on Computer Communication and Informatics* https://doi.org/10.1109/ICCCI50826.2021.9402571 (IEEE, 2021).
31. Lehman, E. et al. Does BERT pretrained on clinical notes reveal sensitive data? In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 946–959 (Association for Computational Linguistics, 2021).
32. Kim, H.-E. et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Health* **2**, e138–e148 (2020).
33. Seah, J. C. et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit. Health* **3**, e496–e506 (2021).
34. Steyvers, M. & Kumar, A. Three challenges for AI-assisted decision-making. *Perspect. Psychol. Sci.* **19**, 722–734 (2024).
35. Tucci, V., Saary, J. & Doyle, T. E. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. *J. Med. Artif. Intell.* **5**, 4 (2022).
36. Ball, J. R. & Balogh, E. Improving diagnosis in health care: highlights of a report from the national academies of sciences, engineering, and medicine. *Ann. Intern. Med.* **164**, 59–61 (2016).
37. Tiffen, J., Corbridge, S. J. & Slimmer, L. Enhancing clinical decision making: development of a contiguous definition and conceptual framework. *J. Prof. Nurs.* **30**, 399–405 (2014).
38. Meyer, A. N. & Singh, H. The path to diagnostic excellence includes feedback to calibrate how clinicians think. *JAMA* **321**, 737–738 (2019).
39. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
40. Zurn, P., Dal Poz, M. R., Stilwell, B. & Adams, O. Imbalance in the health workforce. *Hum. Resour. Health* **2**, 13 (2004).
41. Li, J.-P. O. et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. *Prog. Retin. Eye Res.* **82**, 100900 (2021).
42. Overhage, J. M. & McCallie Jr, D. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Ann. Intern. Med.* **172**, 169–174 (2020).
43. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
44. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
45. 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).
46. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat. Med.* **30**, 1481–1488 (2024).
47. Li, J. et al. BLIP-2: bootstrapping language–image pre-training with frozen image encoders and large language models. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 19730–19742 (PMLR, 2023).
48. Kaur, D., Uslu, S., Durresi, M. & Durresi, A. LLM-based agents utilized in a trustworthy artificial conscience model for controlling AI in medical applications. In *Advanced Information Networking and Applications Lecture Notes on Data Engineering and Communications Technologies* 198–209 (Springer, 2024).
49. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* **56**, 3005–3054 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. [2]Department of Orthopedics, Peking University Third Hospital & Beijing Key Laboratory of Spinal Disease & Engineering Research Center of Bone and Joint Precision Medicine, Beijing, China. [3]School of Science and Engineering (SSE), Future Network of Intelligence Institute (FNii) and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, Chinese University of Hong Kong, Shenzhen, China. [4]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China. [5]Department of Respiratory and Critical Care Medicine, Peking University Third Hospital and Research Center for Chronic Airway Diseases, Peking University Health Science Center, Beijing, China. [6]Department of Endocrinology and Metabolism, Peking University Third Hospital, Beijing, China. [7]Department of Cardiology, The First College of Clinical Medical Science, China Three Gorges University and Yichang Central People's Hospital, Yichang, China. [8]Peking University First Hospital and Research Center of Public Policy, Peking University, Beijing, China. [9]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [10]South China Hospital, Medical School, Shenzhen University, Shenzhen, China. [11]These authors contributed equally: Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui. ✉e-mail: yangjian@ctgu.edu.cn; schl@bjmu.edu.cn; guangyu.wang24@gmail.com

Methods

Datasets

In this study, we curated three datasets to develop and evaluate MedFound-DX-PA, from pre-training, fine-tuning and evaluation (Supplementary Table 1). For pre-training, we created MedCorpus, a large-scale collection of free text from four sources: PMC-CR, MIMIC-III-Note, MedDX-Note and MedText. For fine-tuning, we used the MedDX-FT dataset, which comprises EHRs with diagnoses, diagnostic rationale demonstrations and helpfulness annotations. Among those, MedDX-Note and MedDX-FT include EHRs sourced from the China Consortium for Disease Diagnosis Investigation (CC-DXI). It enrolled multiple hospitals across Beijing, Sichuan Province and Guangdong Province in China: Peking University Third Hospital, Peking University First Hospital, West China Hospital of Sichuan University and Shenzhen University-affiliated South China Hospital. The study was conducted under a waiver of written informed consent approved by the institutional review board (IRB). IRB and ethics committee approvals were obtained in all locations. EHR data were de-identified to remove any patient-related information.

Pre-training datasets to develop MedFound. We curated MedCorpus, an extensive language corpus comprising a diverse collection of biomedical and clinical text, for the pre-training of MedFound. MedCorpus integrates a total of 6.3 billion tokens obtained from four datasets: MedText, PMC-CR, MIMIC-III-Note and MedDX-Note (Supplementary Information).

MedText is composed of a diverse collection of medical textbooks, comprising 1,752 multilingual textbooks, encapsulating fundamental medical knowledge, terminology, concepts and practice guidelines. PMC-CR comprises full-text case reports from PMC[50], providing detailed reports of the symptoms, signs, diagnosis, treatment or follow-up of individual patients, with a particular focus on unusual or novel occurrences of disease, and many new ideas in medicine. PMC is recognized as the most extensive, publicly accessible digital repository that archives a wide range of research articles in the fields of biomedical and life sciences. MIMIC-III-Note and MedDX-Note are derived from real clinical data, covering a diverse range of diseases across different systems. MIMIC-III-Note is annotated from an open-access, large-scale clinical database, MIMIC-III, which contains EHRs from 38,597 patients across 49,785 hospital admissions within intensive care units[51]. The MIMIC-III-Note dataset contains a diverse selection of typical medical texts from patient records, such as medical notes, prescribed medications, clinical orders and radiology reports, among others. MedDX-Note, a proprietary large-scale, real-world dataset, contains 8.7 million EHRs sourced from the CC-DXI. The extensive dataset covers a spectrum of diseases and a mean age of 40.96 years with a standard deviation of 21.30. Each record within the dataset provides a comprehensive account of the medical encounters, such as medical history and examination reports. We conducted data pre-processing for the corpus, which involved the removal of special tags and characters and tokenization (details of the MedCorpus are provided in the Supplementary Information).

Fine-tuning and alignment datasets to develop MedFound-DX-PA. To fine-tune and align our model for diagnosis, we curated a medical record dataset and collected two types of expert annotations: diagnostic rationale demonstrations and helpfulness annotations. We constructed a dataset sourced from the CC-DXI, named MedDX-FT, comprising 109,364 cases and spanning 408 common diseases across eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology and endocrinology. For fine-tuning models with diagnostic reasoning rationales, we manually curated a dataset comprising 800 diagnostic rationale demonstrations using medical records from the MedDX-FT dataset. In each case, physicians read through the entire case history and provided step-by-step diagnostic analyses, incorporating crucial factors such as clinical difficult to diagnose. For instance, ICD E11 (type 2 diabetes mellitus) is the parent of several child codes, including E11.0 (type 2 diabetes mellitus with hyperosmolarity), E11.1 (type 2 diabetes mellitus with ketoacidosis) and E11.2 (type 2 diabetes mellitus with renal complications)[58]. The hierarchical structure of the ICD facilitates the construction of more granular preferences, based on the alignment of model outputs with ICD codes.

For the helpfulness preference construction, we constructed a scoring model trained on an expert-annotated dataset comprising diagnostic rationales with labels of 'helpful' or 'unhelpful'. A binary classification model was trained as a scoring model to assess the extent of helpfulness for each diagnostic rationale. Preference optimization for multiple preference objectives is accomplished through DPO, known for its stability performance, and computational efficiency. Compared to reinforcement training, DPO offers a more stable training process[19]. Both diagnostic hierarchy preference and helpfulness preference are jointly trained. Given a medical record, multiple responses are sampled. The objective function is $L = \log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)$, where x is input prompt; y_w and y_l denote the preferred and dispreferred responses, respectively; π_{ref} is reference policy; π_θ is an optimal policy with parameter θ; and β is a parameter controlling the deviation from the reference policy π_{ref}. A detailed description of the PA is provided in the Supplementary Information.

Baselines

We evaluated our approach against open-access state-of-the-art LLMs, including Clinical Camel-70B, Llama-3-70B, MEDITRON-70B and MMedLM 2-7B and the closed-access LLM GPT-4o. These LLMs are decoder-only generative language models. We also evaluated our approach against classification baselines: a traditional machine learning method with HRF[25,30], a BERT-based pre-trained LLM (denoted as Med-BERT[9]) and a classifier variant of MedFound (MedFound-CLS). HRF employs an anatomically based hierarchical classification system combined with classifiers for disease diagnosis analysis. In contrast, Med-BERT is an encoder-only transformer model designed for the clinical domain, and MedFound-CLS, a variant of our MedFound, served as a pre-trained backbone for a classifier. For a fair comparison, all baselines were trained using the same training dataset as our method. Additionally, we developed MedFound-7B based on BLOOM-7B, a smaller-scale version that is more accessible for local deployment, thereby also addressing security concerns (Extended Data Fig. 9).

Clinical study

Study design and participants. In addition to the performance assessment in retrospective data, we further validated the applicability of LLMs in real-world medical diagnostic scenarios. We designed comprehensive clinical studies, which include comparing the accuracy between the AI system and various levels of physicians, assessing the model's effectiveness in assisting junior and intermediate physicians in diagnosis as well as implementing a human expert evaluation framework of the capability of LLM generation contents based on a Likert scale. We recruited nine endocrinologists and nine pulmonary physicians with various years of clinical practice experience, including three junior physicians with 1–5 years of clinical practice experience, three intermediate physicians with 5–10 years of clinical practice experience and three senior physicians with more than 10 years of clinical practice experience within each specialty, respectively. This study was approved by the Peking University Third Hospital Medical Science Research Ethics Committee (IRB00006761-M2023607).

Comparison of diagnostic accuracy between AI and physicians. To evaluate the performance of our model in disease diagnosis, we performed performance comparison between our LLM system and physicians' diagnoses. Here, three groups of physicians were involved,

observations, potential ranges of diseases and diagnoses. The annotation interface is illustrated in Extended Data Fig. 2a. We then employed a self-bootstrapping strategy to automatically generate high-quality diagnostic rationales for each EHR, resulting in 109,364 rationales for fine-tuning.

For the helpfulness PA of the models, we collected helpfulness annotation. Physicians were assigned to assess whether a given diagnostic rationale provided assistance in making accurate diagnoses. Helpfulness was defined as the extent to which the diagnostic rationale presented in the response guided the annotator toward an accurate diagnosis. The annotation interface is shown in Extended Data Fig. 2b. A total of 1,800 selected generated responses from the MedDX-FT dataset were annotated in this manner. Overall, in 72.1% of cases, the generated diagnostic rationales were reported to be helpful. These data were used to fine-tune and align MedFound-DX-PA with human preferences, aiming to enhance its generated rationales to align with professional preferences and to provide helpful assistance in the diagnostic process.

Evaluation datasets of the diagnostic performance of LLMs. For the evaluation of the LLM-based system in disease diagnosis, we conducted MedDX-Bench, a comprehensive benchmark that consists of three datasets containing real-world EHRs: MedDX-Test and MedDX-OOD for ID and OOD testing on common diseases and MedDX-Rare for OOD testing on rare diseases.

Specifically, the MedDX-Test dataset, sourced from the same origin as the developmental dataset CC-DXI and mutually exclusive from the MedDX-FT dataset, was used to evaluate the diagnostic performance in an ID setting. It contains 11,662 medical records, covering a wide range of common diseases across various medical specialties. MedDX-OOD and MedDX-Rare were collected from Yichang Central People's Hospital in Hubei Province, China, a geographic region distinct from the CC-DXI for OOD evaluation. There is no overlap between the MedDX-OOD and MedDX-Rare datasets. To extend our evaluation to external validation sets and to test the models' performance in varying conditions, we introduced the MedDX-OOD and MedDX-Rare datasets. The MedDX-OOD dataset comprises 23,917 records spanning common diseases, serving as an OOD validation set to assess the models' generalizability across different geographical regions. The other dataset, MedDX-Rare, consists of 20,257 records covering 2,105 diseases that exhibit a long-tailed distribution and present a challenge under conditions of rare and fine-grained diseases. All EHRs used in this study were obtained from hospital systems with a diverse patient population from different clinical departments and could closely mirror the process of real-world diagnoses.

Model overview

Here we present MedFound, a pre-trained, large-scale language model tailored for medical applications, and MedFound-DX-PA, which is further optimized for enhanced diagnostic capabilities. First, we curated a diverse collection of medical corpora for continued pre-training based on the BLOOM model (176 billion parameters), resulting in MedFound. This step aims to adapt the LLM to the medical domain to boost its end-task performances. Subsequently, we fine-tuned MedFound using a dataset with diagnostic rationales to learn diagnostic reasoning, resulting in MedFound-DX. Finally, we refined MedFound-DX to align with the domain expertise and requirements of the diagnostic profession and human expert preferences using DPO[19], resulting in MedFound-DX-PA. The alignment process was guided by the hierarchical structure of disease classifications according to the ICD and by human expert preferences assessed through helpfulness scores from a helpfulness scoring model.

Pre-training for developing MedFound. Here, we leveraged the BLOOM[52] family of LLMs, a decoder-only transformer language model, as our base model for domain pre-training. The BLOOM training corpus consists of 1.61 terabytes of text across multiple languages. We chose BLOOM-176B as the base model, owing to its status as the largest open-source language model available, with its emergent capabilities and extensive knowledge base[53]. For pre-training, the model is trained via the objective of causal language model[8]. Let $D = \{\mathbf{x_i}\}$ denote the collection of sequences, and the sequence $\mathbf{x_i}$ is made up of n_i tokens—that is, $\mathbf{x_i} = (w_1, w_2, \ldots, w_{n_i})$. The training objective is to minimize the negative log-likelihood $\sum_{i=1}^{|D|} \sum_{j=1}^{n_i} -\log P(w_j | w_1, w_2, \ldots, w_{j-1})$.

Fine-tuning for diagnostic reasoning. To adapt the model for the clinical diagnosis tasks, we fine-tuned MedFound on a dataset with diagnostic rationales based on a self-bootstrapping approach, resulting in MedFound-DX. In clinical diagnosis, physicians are required to explain a patient's symptoms and describe their rationale for generating a diagnosis, demonstrating the complex and multi-step nature of diagnosis reasoning[54]. To incorporate this essential element for accurate diagnosis, we employed COT fine-tuning[55] on MedFound, integrating diagnostic rationales into the dataset, thereby enhancing the ability of the model to mimic human-like diagnostic thought processes. The generated diagnosis is conditioned on this intermediate rationale, which is expected to improve its accuracy. The language model p_θ is trained to generate a response $\mathbf{R} = v_{1:n}$ for a given input prompt $\mathbf{I} = w_{1:m}$, optimizing the likelihood $p_\theta(\mathbf{R}|\mathbf{I}) = p_\theta(v_{1:n}|w_{1:m})$, where n and m represent the lengths of the response and input prompt, respectively. Thus, the loss function is $\frac{1}{n}\sum_{i=m+1}^{m+n} -\log p_\theta(w_i|w_1, \ldots, w_m)$.

Although COT fine-tuning[56] has demonstrated advantages with LLMs, it remains challenging to acquire a substantial amount of COT demonstrations for fine-tuning, especially within the medical domain. To address this issue and further enhance the model's diagnosis reasoning ability, we adopted a self-bootstrapping approach, following the Self-Taught Reasoner (STaR)[57]. This approach helps the LLM learn to automatically generate more coherent and precise rationales by training it based on a seed set of high-quality diagnostic rationale demonstrations annotated by human expert[55]. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_i$ and a small dataset with rationale $\mathcal{S} = \{x_i, y_i, r_i\}_i$, where x_i is a medical record with diagnosis y_i, r_i represents diagnostic rationale. First, we fine-tuned a preliminary model M_1 based on the pre-trained model M_0 with $\mathcal{S}$ to learn to generate diagnostic rationale. Then, the model M_1 generates diagnosis y_i' with diagnostic rationale r_i' for each sample x_i from $\mathcal{D}$, resulting in $\mathcal{D}_1 = \{(x_i, r_i', y_i')\}_i$. We then generated diagnostic rationale r_i'', where we provide the true diagnosis as a hint in a prompt—that is, (y_i, x_i)—to the model M_1 and ask it to generate diagnostic rationale r_i'', resulting in $\mathcal{D}_2 = \{(x_i, r_i'', y_i')\}_i$. We then corrected the diagnostic rationale r_i' by r_i'' if diagnosis y_i' is wrong, resulting in a new dataset $\mathcal{D}'$. We then fine-tuned the model again using $\mathcal{D}'$, deriving the refined model M_2.

PA for developing MedFound-DX-PA. To align MedFound-DX with real-world diagnostic scenarios and human expert preferences, we propose a unified PA framework. This framework incorporates two types of preferences, including the diagnostic hierarchy preference and the helpfulness preference, which are jointly optimized in the model to align with the diagnostic standards, and expectations of healthcare professionals in clinical scenarios. The diagnostic hierarchy preference, guided by the hierarchical structure of disease classifications defined by the ICD codes, seeks to align the model's generation with the standards for disease classification. The helpfulness preference is refined through a helpfulness scoring model trained on expert annotations, aiming to make the model's generation more informative, useful and trustworthy for diagnostic purposes while minimizing the risk of harm or misleading information. The PA process comprises two steps: preference construction and preference optimization. For the diagnostic hierarchy preference construction, we leverage guidance from the ICD to address the issues associated with setting preferences based solely on diagnostic correctness, which can result in sparse signals, especially in cases involving rare diseases or conditions that are

including junior, intermediate and senior physicians, separately from each specialty in pulmonary and endocrine medicine. For the comparison, we constructed an independent validation set comprising 300 cases, with 150 cases from each of endocrinology and pulmonology. Each physician made diagnoses based on information provided from the medical records, including demographics, chief complaint, present illness, past medical history, physical examination, laboratory tests and radiological examination. We used the diagnoses of an expert consensus panel comprising three senior physicians from each specialty, serving as the gold standard. We then used it as a reference to assess the accuracy of AI-generated diagnosis in comparison to the physician groups.

Assisted diagnostic accuracy with the LLM in the workflow. We conducted a study to examine the AI system's potential role in assisting the diagnostic performance of physicians within their workflow. After the previous initial diagnosis, each group of junior and intermediate physicians was asked to provide a diagnosis with the assistance of the model-generated output, including reasoning rationales and final diagnosis suggestions. Each junior and intermediate physician received 150 cases. Then, the physicians formulated their final diagnosis using the model-generated contents as reference. The re-test comparison study was conducted at least 2 weeks later to ensure reproducibility. We compared the diagnostic accuracy of junior and intermediate physicians after AI assistance to that of our AI system or senior physicians, to investigate whether the integration of an LLM in the workflow could enhance junior and intermediate physicians.

Human evaluation framework of the diagnostic capability of the LLM. To gain a comprehensive understanding of the capabilities and potential limitations of the LLM in clinical senecios, we proposed an assessment framework named CLEVER. This framework is designed to evaluate the capacity of the LLM to generate accurate and reliable diagnoses while adhering to medical standards, covering various aspects from medical case comprehension and clinical reasoning to diagnosis formulation. The development of the CLEVER framework was inspired by previous work[14,59] and involved consultations with expert physicians in the United Kingdom and China. The framework included eight key evaluation axes and refined metrics. (1) Medical case comprehension. The objective of this metric was to assess the LLM's understanding and interpretation of medical cases, including comprehension of the record of clinical cases and crucial information required for diagnosis with completeness and correctness. (2) Medical guideline and consensus. The objective of this metric was to assess the LLM's adherence to established medical guidelines and consensus within the medical community. (3) Clinical reasoning. The objective of this metric was to assess the LLM's content aligned with the diagnostics reasoning process of physicians in clinical practice. (4) Relevance of differential diagnosis. The objective of this metric was to assess the LLM's capacity to differentiate among multiple possible conditions or diseases that could potentially cause a patient's symptoms. (5) Acceptability of diagnosis. Assessing the feasibility of the LLM's generated diagnoses. We asked the physicians to rate whether the diagnosis was acceptable or reliable for clinical use. (6) Unfaithful content. Evaluating the presence or extent of inaccurate or misleading information in the LLM's output. The physicians were asked to rate whether the LLM included incorrect or fabricated content. (7) Bias and unfairness. Assessing the presence or extent to which the LLM demonstrated varying levels of stereotypes related to age, gender, culture and race. (8) Possibility of harm. Assessing the presence or extent to which the generated content of the LLM contains any incorrect, adverse, harmful or fabricated evidence, which could potentially lead to misdiagnosis or mislead physicians, resulting in possible serious medical accidents/negative impacts.

A total of six senior physicians, comprising three senior physicians specialized in the pulmonary field and three senior endocrinologists, each with over 10 years of clinical experience, were involved in evaluating the model's generated diagnosis and the related reasoning process. The capabilities of the LLM with alignment versus the LLM without alignment were assessed by each senior physician within their respective specialty. This process included a total of 180 evaluations. Each senior physician reviewed and scored the cases based on a five-point Likert scale. A detailed description of the metrics is provided in the Supplementary Information.

Implementation

We applied low-rank adaptation (LoRA)[60] and ZeRO++[61] with the DeepSpeed framework to train LLMs. LoRA can reduce the number of trainable parameters by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the transformer architecture (see details in the Supplementary Information). We found that LoRA fine-tuning, when appropriately configured, can be more effective for large-scale LLMs (Supplementary Table 8). Experiments demonstrated that with parameter-efficient training and selecting domain-representative corpora, the corpus token size used is sufficient to build an efficient medical LLM (Supplementary Table 9 and Extended Data Fig. 9). We employed the vLLM[62] library for model inference for its high efficiency in memory and computational resource utilization. In our approach to generating diagnosis using LLMs, we included two prompting techniques: MED-Prompt prompting[63] and SC prompting[64]. MED-Prompt is a medical prompting strategy, combined with few-shot prompting to generate predictions from pre-trained LLMs without the need for task-specific fine-tuning. The SC strategy was employed with 20 samples to balance performance and cost (Extended Data Fig. 10). Detailed parameters of the implementation are provided in the Supplementary Information.

Statistical analysis

We used micro accuracy and macro accuracy to evaluate diagnostic performances. We calculated the mean and standard error of the performance. To compute the CIs, we used a non-parametric bootstrap procedure with 1,000 samples[65]. We also reported more metrics, including precision, recall, ROC-AUC and PR-AUC, using both macro average (unweighted) and micro average (sample-weighted) methods. The ROC-AUC scores were calculated using SC agreement frequency[66]. In clinical studies, a two-sided P value of less than 0.05 was considered statistically significant. We use two-sided t-tests between MedFound-DX and MedFound-DX-PA to show whether significant differences exist across eight dimensions of human evaluation for diagnostic performance.

Reporting summary

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

Data availability

The raw data of PMC-CR and MedText are available from https://www.ncbi.nlm.nih.gov. The MIMIC-III-Note dataset can be found at https://physionet.org/about/database/ and requires access due to its terms of use. MedDX-Note and MedDX-Bench are sourced from real-world clinical scenarios, with IRB approval obtained from institutions for EHR data collection. Due to privacy regulations, the EHRs cannot be made freely available in a public repository. De-identified data from MedDX-Note and MedDX-Bench can be requested through the management team by contacting the corresponding author (G.W.), following a defined protocol for data request approval. Generally, all such requests for access to EHR data will be responded to within 1 month. For the reproduction of our code and model, a representative test dataset from MedDX-Bench, containing samples across specialties, is publicly available on GitHub (https://github.com/medfound/medfound/tree/main/data/test.zip). Data can be shared only for non-commercial use.

Code availability

The deep learning models were developed and deployed in Python (3.10) using PyTorch (2.1.2). The following standard model libraries were used: numpy (1.26.4), pandas (2.2.1), transformers (4.36.1), vllm (0.2.5), scikit-learn (1.2.1), matplotlib (3.7.1) and scipy (1.11.3). We build upon PyTorch (2.1.2) to implement Direct Preference Optimization (DPO). Custom codes were specific to our development environment and were used primarily for data input/output and parallelization across computers and graphics processors. The codes are available for scientific research and non-commercial use on GitHub at https://github.com/medfound/medfound. The pre-trained models are publicly available (https://huggingface.co/medicalai/MedFound-7B, https://huggingface.co/medicalai/MedFound-176B).

References

50. Canese, K. & Weis, S. PubMed: the bibliographic database. In *The NCBI Handbook* 2nd edn (National Center for Biotechnology Information, 2013).
51. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
52. Le Scao, T. et al. BLOOM: a 176b-parameter open-access multilingual language model. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.05100 (2023).
53. Alabdulmohsin, I. M. et al. Revisiting neural scaling laws in language and vision. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 22300–22312 (Curran Associates, 2022).
54. Ghaemi, S. N. *Clinical Psychopharmacology: Principles and Practice* (Oxford Univ. Press, 2018).
55. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 24824–24837 (Curran Associates, 2022).
56. Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
57. Zelikman, E. et al. STaR: bootstrapping reasoning with reasoning. In *Proc. 36th International Conference on Neural Information Processing Systems (eds Koyejo, S. et al.)* 15476–15488 (Curran Associates, 2022).
58. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. https://www.who.int/publications/i/item/9241544228 (1992).
59. Feng, S. Y., Khetan, V., Sacaleanu, B., Gershman, A. & Hovy, E. CHARD: clinical health-aware reasoning across dimensions for text generation models. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* 313–327 (Association for Computational Linguistics, 2023).
60. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. 10th International Conference on Learning Representations* https://openreview.net/pdf?id=nZeVKeeFYf9 (ICLR, 2022).
61. Wang, G. et al. ZeRO++: extremely efficient collective communication for giant model training. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/pdf?id=gx2BT0a9MQ (ICLR, 2024).
62. Kwon, W. et al. Efficient memory management for large language model serving with pagedattention. In *Proc. 29th Symposium on Operating Systems Principles* 611–626 (Association for Computing Machinery, 2023).
63. Ahmed, A., Zeng, X., Xi, R., Hou, M. & Shah, S. A. MED-Prompt: a novel prompt engineering framework for medicine prediction on free-text clinical notes. *J. King Saud. Univ. Comput. Inf. Sci.* **36**, 101933 (2024).
64. Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In *Proc. 11th International Conference on Learning Representations* https://openreview.net/pdf?id=1PL1NIMMrw (ICLR, 2023).
65. Chihara, L. M. & Hesterberg, T. C. *Mathematical Statistics with Resampling and R* (John Wiley & Sons, 2022).
66. Geng, J. et al. A survey of confidence estimation and calibration in large language models. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 6577–6595 (Association for Computational Linguistics, 2024).

Acknowledgements

This study was funded by the National Natural Science Foundation of China (grants 62272055, 82330078, 81874010, 61931024, 92359202 and 12326610); the National Key Research and Development Program (grant 2021YFC2501700); the New Cornerstone Science Foundation through the XPLORER PRIZE; and the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001). K.X. would like to acknowledge funding from the Wellcome Trust (216593/Z/19/Z) and the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre. The authors would like to thank physicians (A. Liu, S. Wang, W. Fu, R. Lu, K. Yang, S. Lang, J. Liu and L. Zhang) from the Department of Endocrinology and Metabolism, Peking University Third Hospital, and physicians (Q. Cheng, B. Liu, J. Ren, Y. Qiao, X. Li, S. Cao, M. Wu and C. Sun) from the Department of Respiratory and Critical Care Medicine, Peking University Third Hospital, for assistance. The authors would like to acknowledge the Nanjing Institute of InforSuperBahn MLOps for providing the training and evaluation platform.

Author contributions

G.W., X. Liu, H.L., G.Y., Z.J., T.G., Z.Z., H.W., L.T., Y.C.S., Z.S., T.H., Jin Y., J.J.Z., X. Li, S.W., J.Z., Y.S., Z.Y. and Jian Y. collected and analyzed the data. G.W., S.C., P.Z. and C.S. conceived and supervised the project. G.W., X. Liu, G.Y., H.L. and K.X. contributed to data interpretation and critical review and wrote the paper. All authors discussed the results and approved the paper.

Competing interests

The authors declare no competing financial interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03416-6.

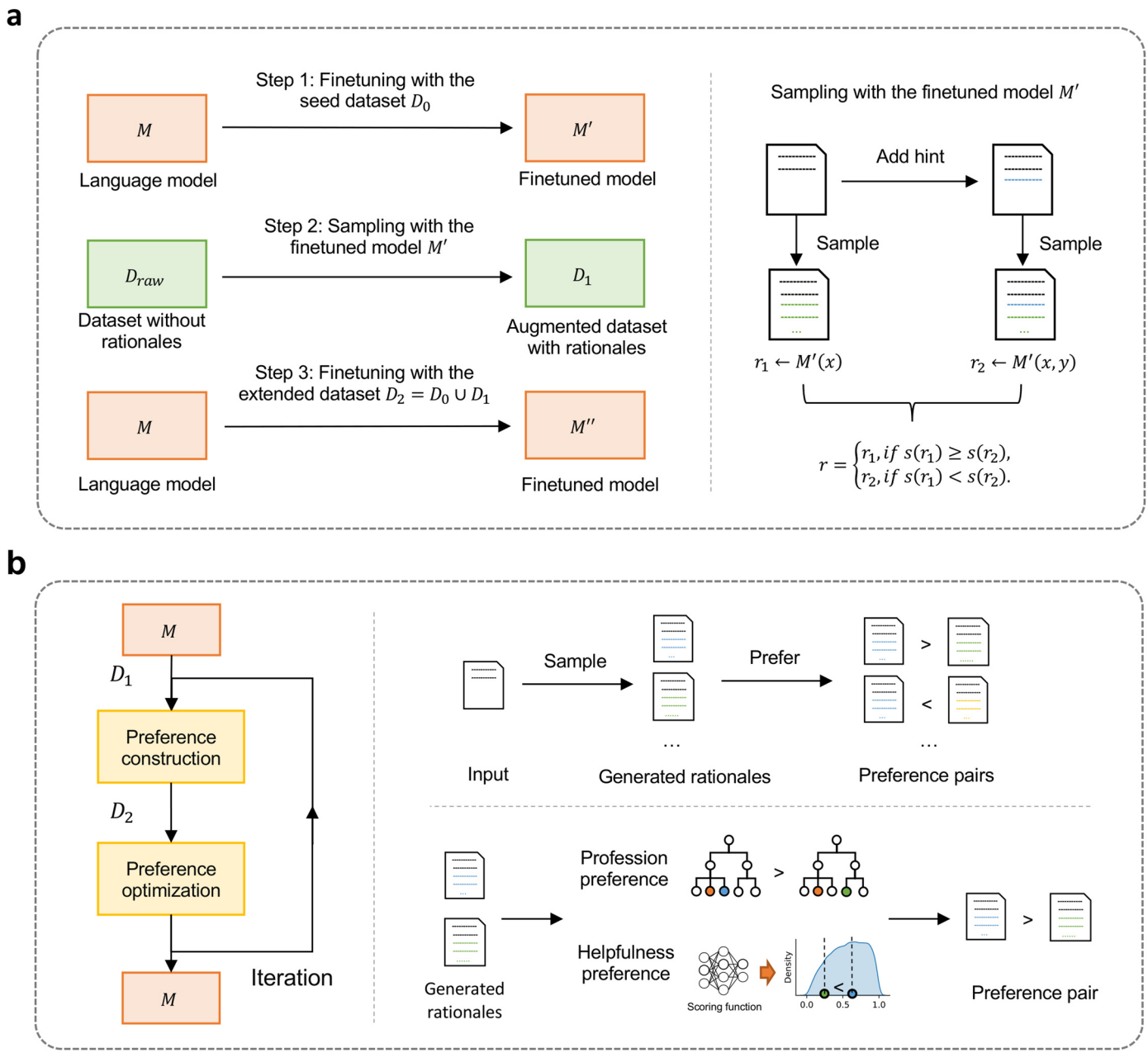
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03416-6.

Correspondence and requests for materials should be addressed to Jian Yang, Chunli Song or Guangyu Wang.

Peer review information *Nature Medicine* thanks Yonghui Wu and Weidi Xie for their contribution to the peer review of this work. Primary Handling Editors: Michael Basson and Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

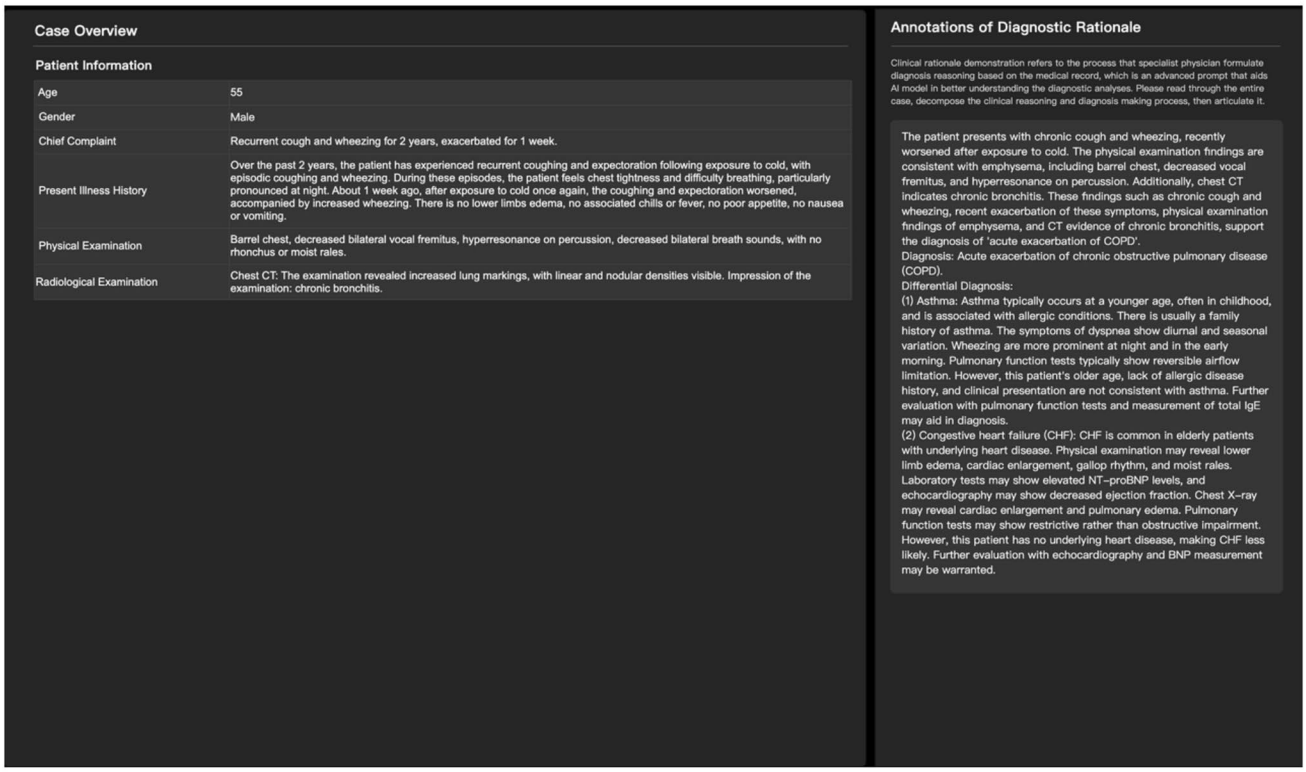
https://doi.org/10.1038/s41591-024-03416-6



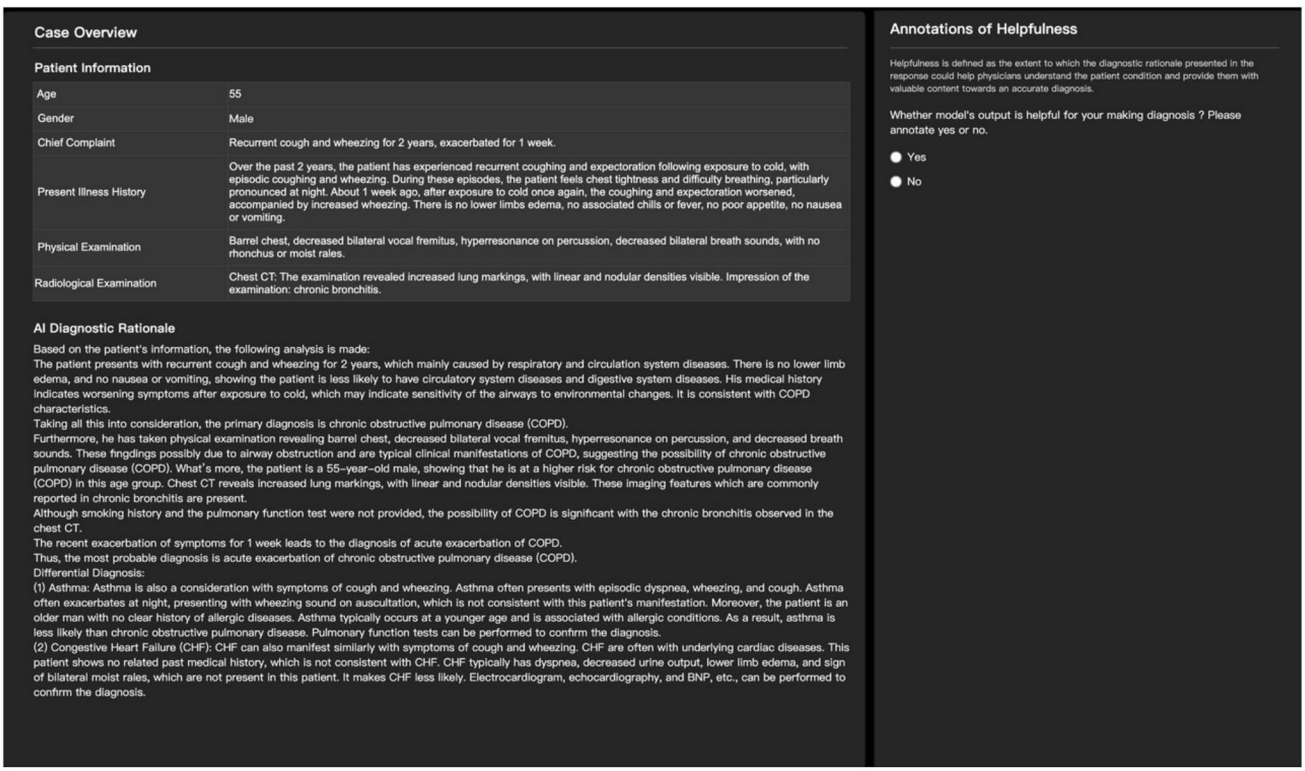
Extended Data Fig. 1 | The development of the LLM-based diagnostic system. **a**, The fine-tuning process. First, we fine-tuned M' based on a language model M to generate diagnostic rationales based on a small number of manual demonstrations D_0.annotated by physicians. Then, we utilized the model M' to generate a dataset with diagnostic rationales D_1. Given a medical record as input, the model M' generated diagnostic rationale. For cases where the diagnosis in rationale was incorrect, we provided the model M' with the medical records and the corresponding correct diagnosis as a reasoning cue to re-generated diagnostic rationale. Finally, we finetuned the model M'' using the augmented data D_2. **b**, Unified preference alignment framework. Left, Preference alignment includes two steps: preference construction and preference optimization. Upper right, given a medical record, multiple rationales are sampled and used to construct preference pair. Lower right, both diagnostic hierarchy and helpfulness preferences are incorporated, where diagnostic hierarchy preferences are guided by the hierarchical structure of disease classifications based on ICD codes and helpfulness preferences are constructed based on expert annotations.

https://doi.org/10.1038/s41591-024-03416-6

a



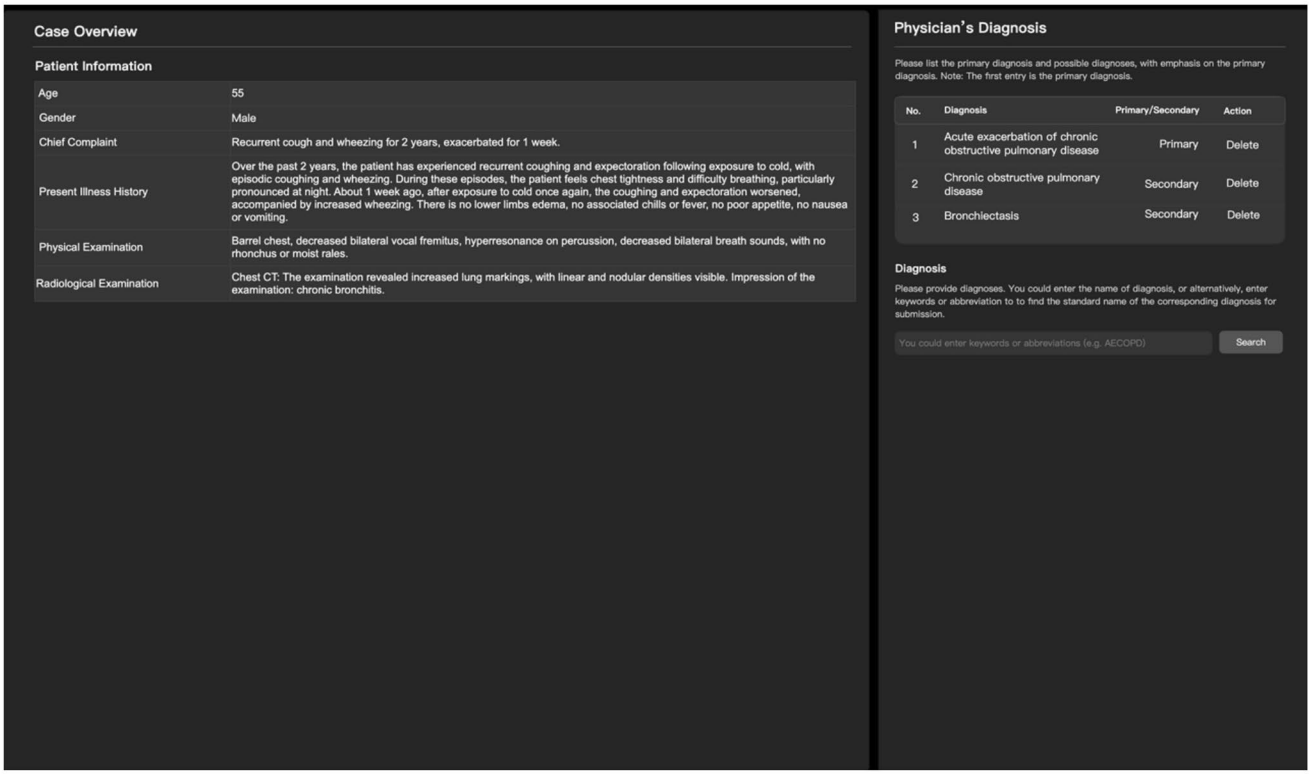
b



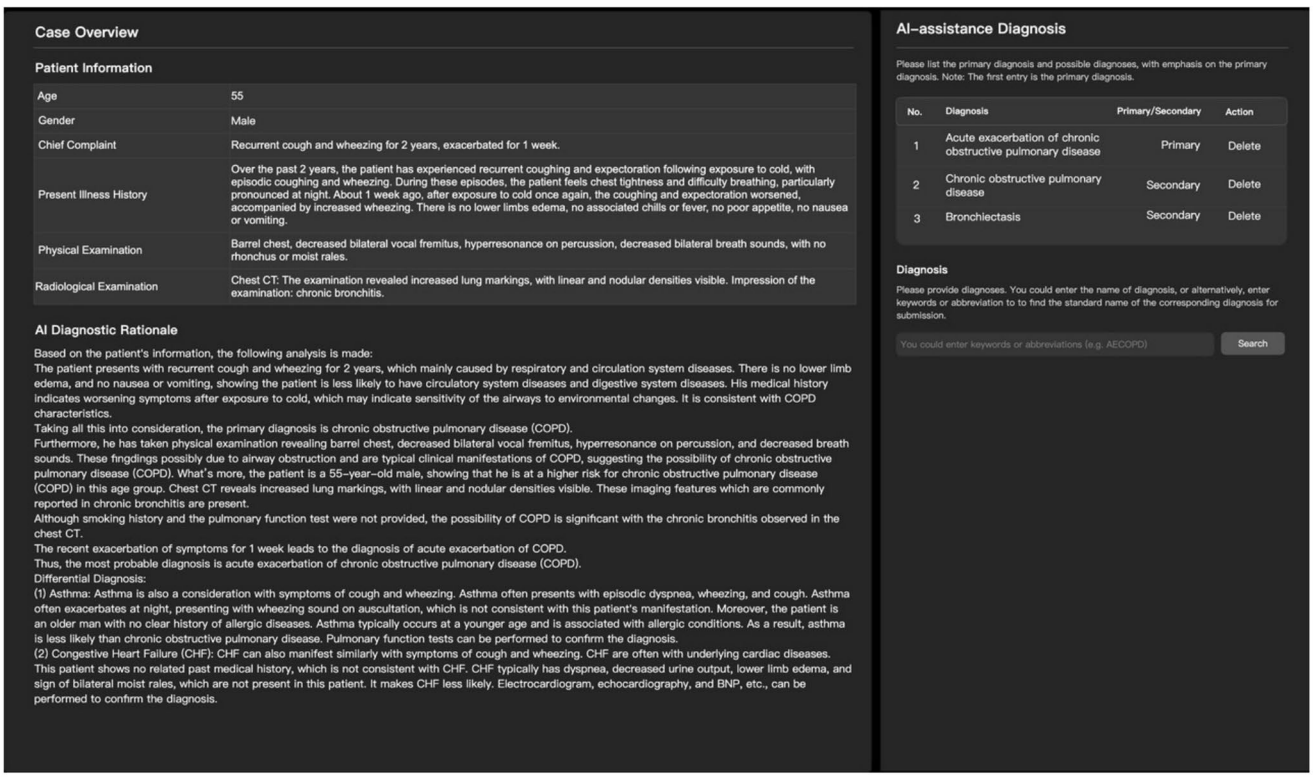
Extended Data Fig. 2 | The user interface of annotation. a, The user interface allows physicians to annotate diagnostic rationale. **b**, The user interface allows physicians to annotate the given diagnostic rationale by a label of helpful or not.

https://doi.org/10.1038/s41591-024-03416-6

a

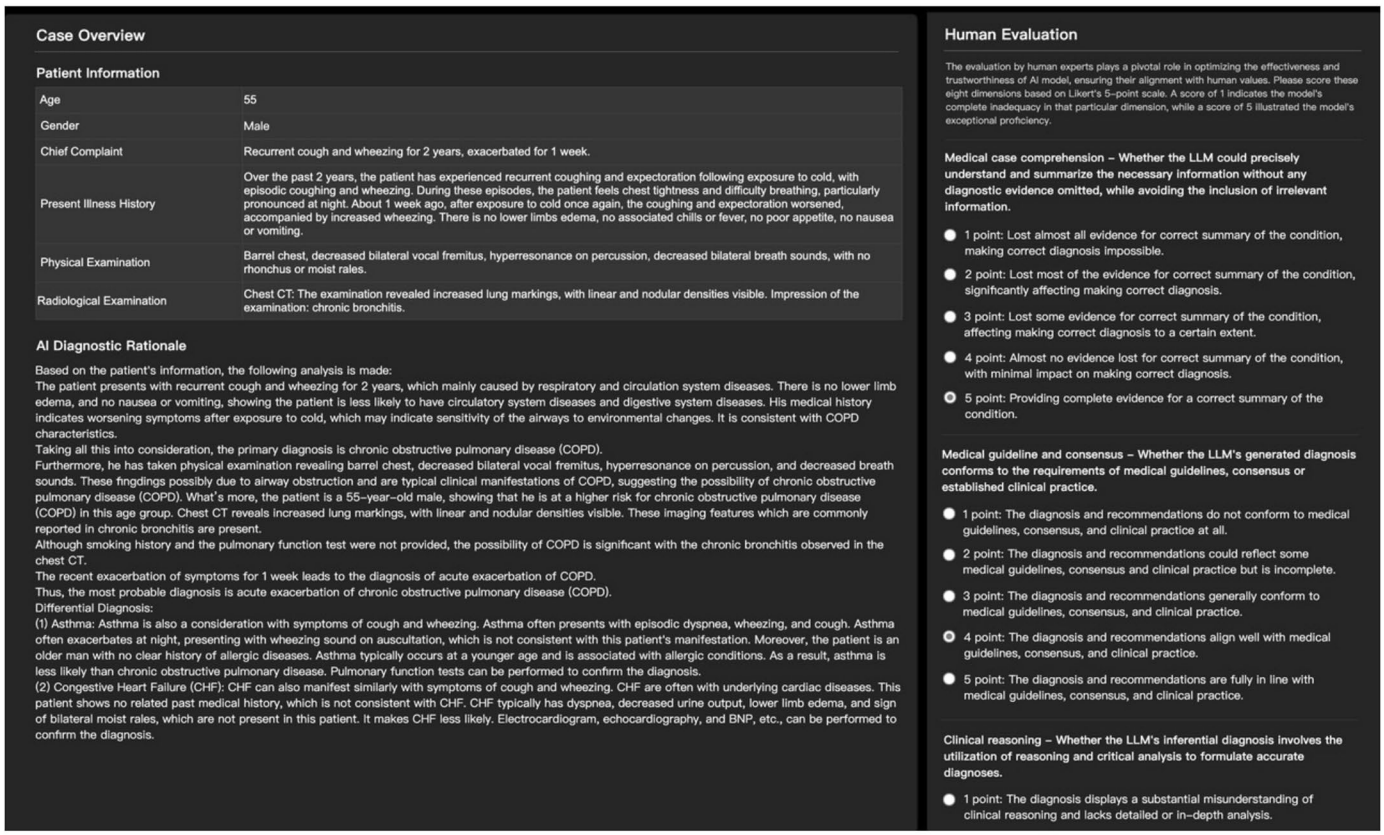


b



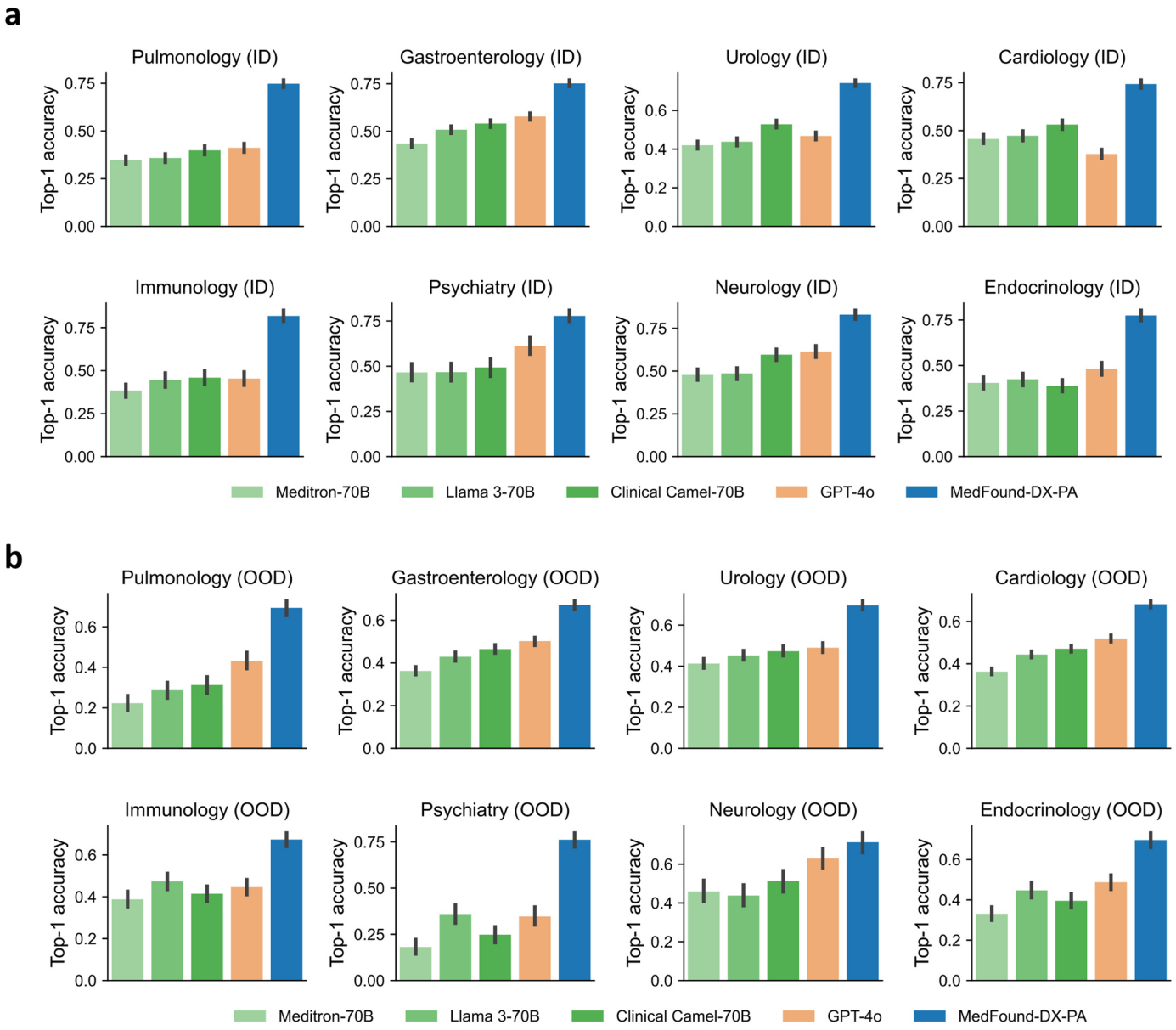
Extended Data Fig. 3 | The user interface of clinical evaluation for comparison with physicians. a, The user interface allows physicians to make diagnosis. **b**, The user interface allows physicians to make diagnosis with AI-assistance.

https://doi.org/10.1038/s41591-024-03416-6



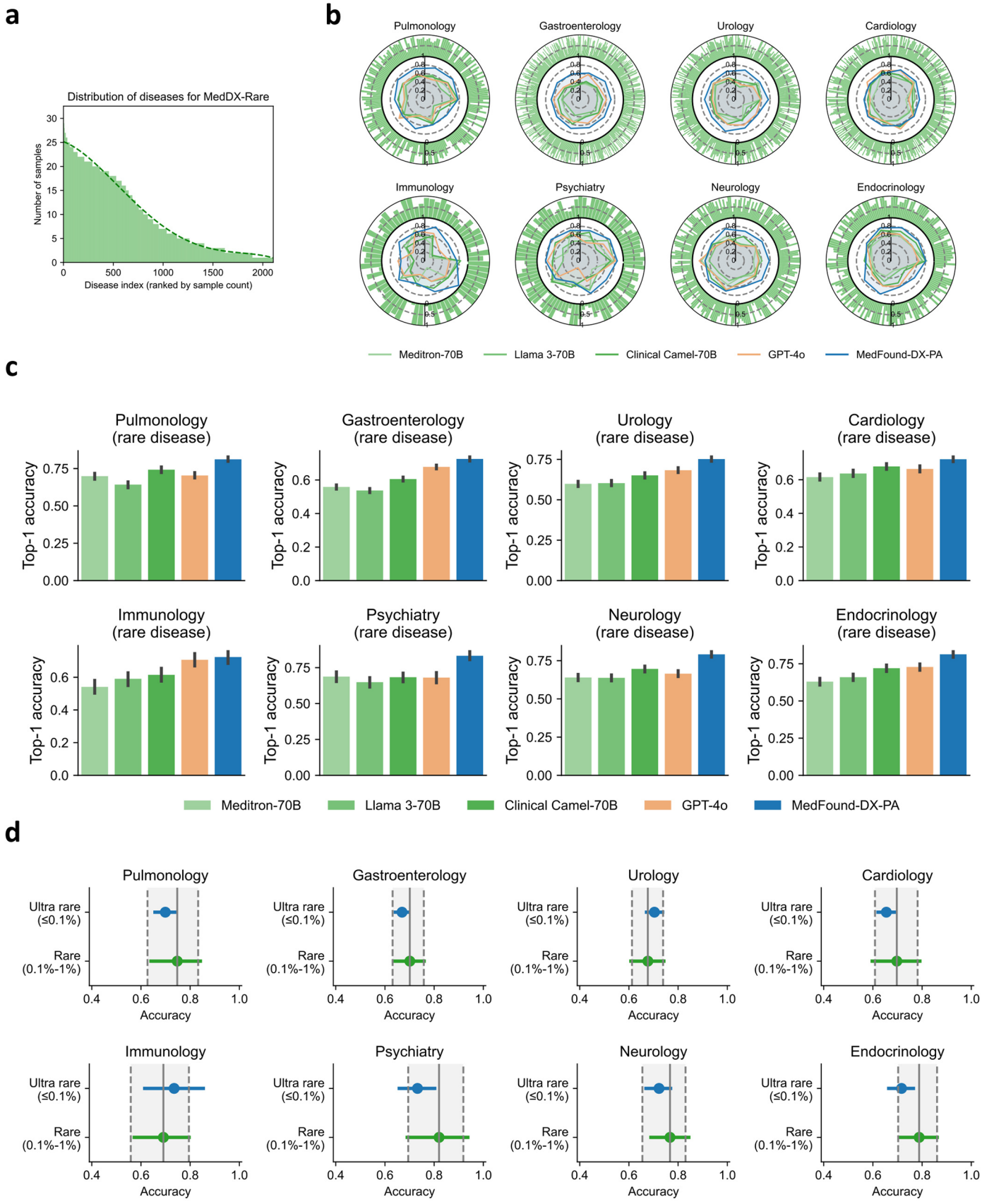
Extended Data Fig. 4 | The user interface of clinical evaluation for diagnostic performance with a human evaluation framework. The user interface allows physicians to evaluate AI-model's rationale, with eight metrics on a Likert-scale of 1 to 5.

https://doi.org/10.1038/s41591-024-03416-6



Extended Data Fig. 5 | Top-1 accuracy of the LLMs for diagnosis of common diseases. a and b, The micro accuracy over individuals between Meditron-70B (light green), Llama 3-70B (medium green), Clinical Camel-70B (dark green), GPT-4o (orange) and our MedFound-DX-PA (blue), for diagnosing diseases in in-distribution (ID) (**a**) (n = 11,662) and out-of-distribution (OOD) (**b**) (n = 23,917) settings stratified by eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology, and endocrinology. The error bars represent the 95% CIs.

https://doi.org/10.1038/s41591-024-03416-6

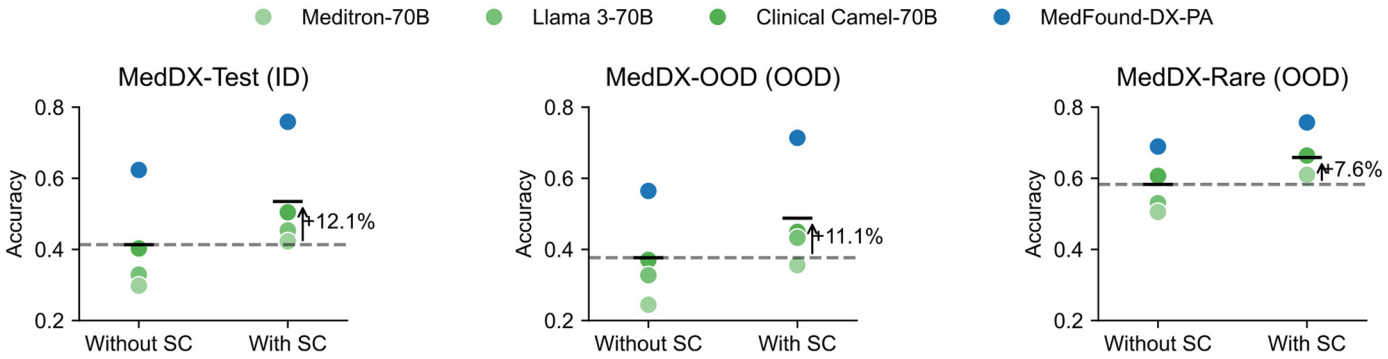


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41591-024-03416-6

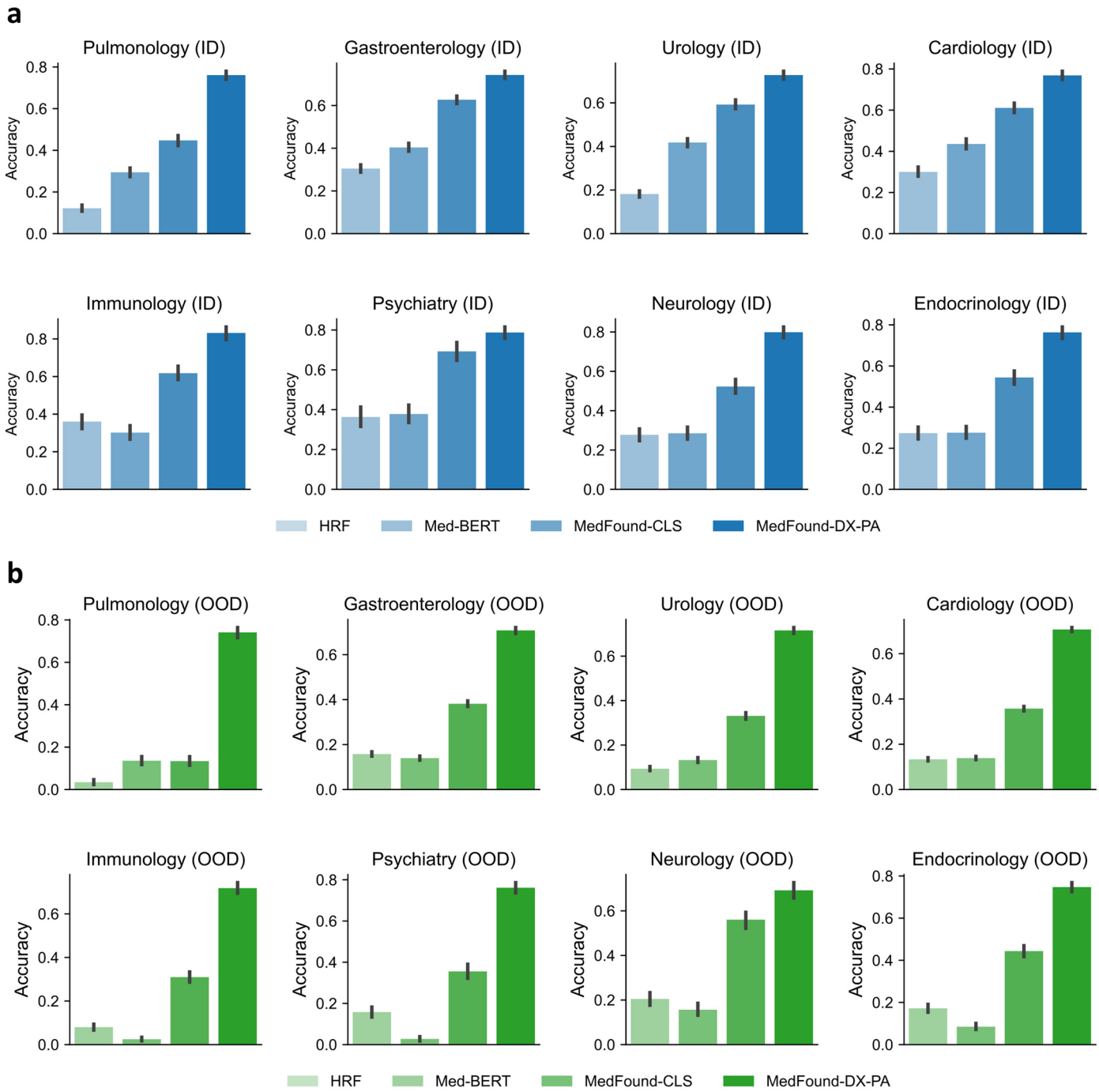
Extended Data Fig. 6 | Top-1 accuracy of the LLMs for diagnosis of rare diseases. a, The distribution of diseases in the MedDX-Rare dataset. The horizontal axis represents the diseases ranked by the number of samples illustrating the long-tail distribution, and the vertical axis represents the number of samples**. b**, Performance comparison of macro accuracy between Meditron-70B (light green), Llama 3-70B (medium green), Clinical Camel-70B (dark green), GPT-4o (orange) and our MedFound-DX-PA (blue) across eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology, and endocrinology (n = 20,257). Bar graphs indicate the MedFound-DX-PA's Top-1 accuracy for individual diseases within each specialty. Each specialty's performance on individual diseases is aggregated at the octiles of disease prevalence for averaged performance evaluation. **c**, The micro accuracy over individuals between Meditron-70B (light green), Llama 3-70B (medium green), Clinical Camel-70B (dark green), GPT-4o (orange) and our MedFound-DX-PA (blue), for diagnosing rare diseases across eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology, and endocrinology. Bar graphs indicate the mean ± 95% confidence intervals. **d**, The micro accuracy of MedFound-DX-PA for diagnosing rare diseases stratified by cumulative prevalence, including ultra-rare diseases (prevalence ≤ 0.1%) (n = 378) and rare diseases (prevalence between 0.1% and 1%) (n = 1,727). The x-axis represents the accuracy (mean ± 95% confidence intervals).

https://doi.org/10.1038/s41591-024-03416-6



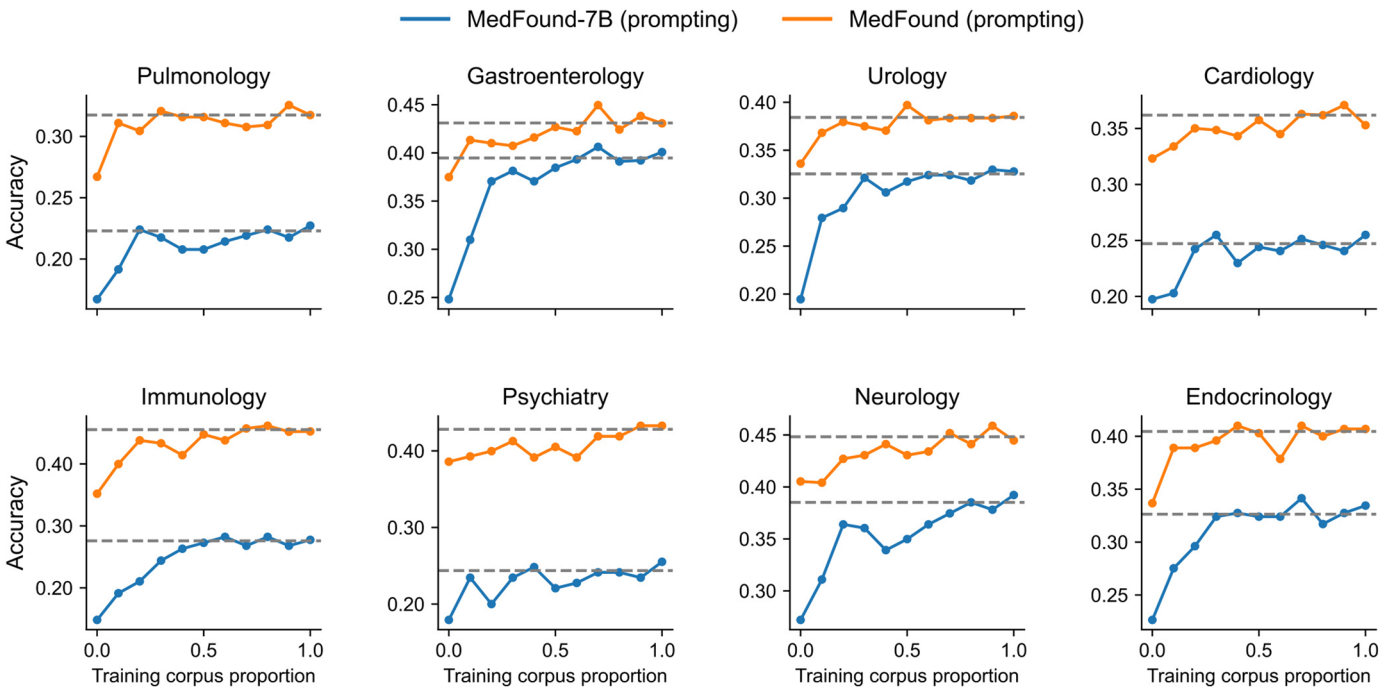
Extended Data Fig. 7 | Performance comparison of LLMs with or without self-consistency strategy for various diagnostic accuracy. Impact analysis of self-consistency strategy on the accuracy of various LLMs, Meditron-70B (light green), Llama 3-70B (medium green), Clinical Camel-70B (dark green), and our MedFound-DX-PA (blue), for diagnostic tasks on MedDX-Test (in-distribution testing on common diseases) (left), MedDX-OOD (out-of-distribution testing on common diseases) (middle), and MedDX-Rare (out-of-distribution testing on rare diseases) (right). The short horizontal line shows the mean performance of a set of models. The percentage increases shown are the improvements gained through self-consistency strategy.

https://doi.org/10.1038/s41591-024-03416-6



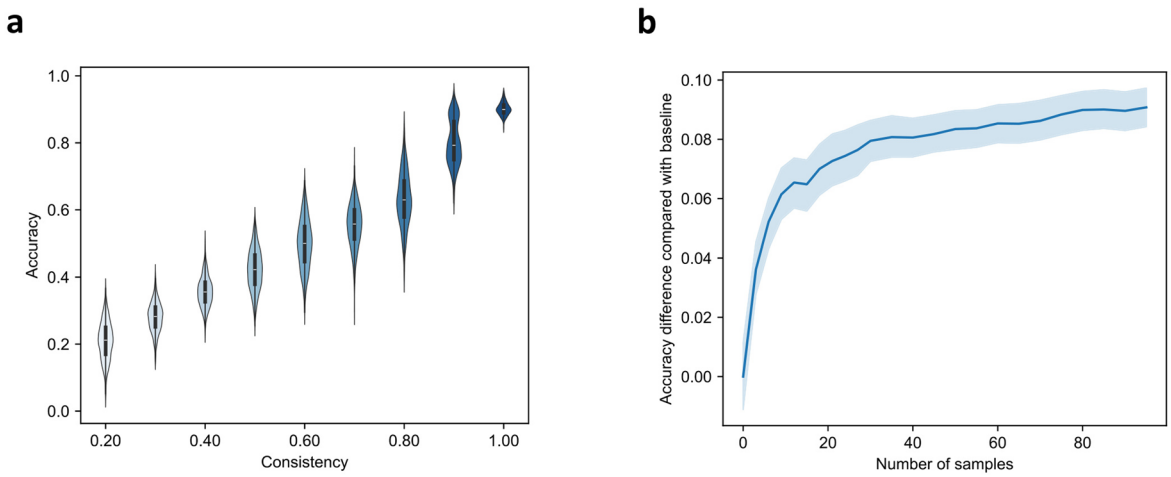
Extended Data Fig. 8 | Performance comparison of LLMs using classification and text generation for diagnosis. a and b, Comparison of accuracy between classification models (HRF, Med-BERT, MedFound-CLS) and text generation model (MedFound-DX-PA) for diagnostic tasks in in-distribution (ID) (**a**) (n = 11,662) and out-of-distribution (OOD) (**b**) (n = 20,257) settings on diseases across eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology, and endocrinology. Bar graphs indicate the mean ± 95% CI.

https://doi.org/10.1038/s41591-024-03416-6



Extended Data Fig. 9 | Performance analysis of MedFound-7B and MedFound-176B model, pre-trained on corpora of varying sizes. The performance of MedFound (orange) and MedFound-7B (blue) pre-trained on increasing proportions of the MedCorpus dataset for diagnostic tasks across eight specialties: pulmonology, gastroenterology, urology, cardiology, immunology, psychiatry, neurology, and endocrinology. The x-axis indicates the proportion of total data used for pre-training the LLM. The y-axis represents the accuracy of diagnoses. The horizontal dashed line corresponds to the mean performance of the LLMs over the last three data points. To examined the effects of corpus size for LLM pretraining, we utilized MedFound and MedFound-7B with the MED-Prompt strategy, evaluated on MedDX-Test. We observed consistent performance improvements as the training corpus ratio increased up to 70%. The improvements plateaued when further increasing the data size beyond this threshold, indicating that the current corpus meets the requirements for effective training.

https://doi.org/10.1038/s41591-024-03416-6



Extended Data Fig. 10 | The ablation results of self-consistency. a, Performance using self-consistency with various consistency level (n = 8,000). The x-axis is consistency levels. The y-axis is accuracy. Plots show the median and interquartile range. **b**, Performance using self-consistency with various sample size. The x-axis is sample size. The y-axis is accuracy change compared to baseline. The shaded area represents the 95% CI.

nature portfolio

Corresponding author(s): Guangyu Wang

Last updated by author(s): Nov 5, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scripts for data collection and preparation were written in Python (3.10) using numpy (1.26.4), pandas (2.2.1).
Data analysis	The tranining and evaluation framework used were implemented using Python (3.10), PyTorch (2.1.2), transformers (4.36.1), and others. We build upon the PyTorch to implement Direct Preference Optimization (DPO). The following Python libraries were used for analysis: scikit-learn (1.2.1), matplotlib (3.7.1), and scipy(1.11.3). The codes are available for scientific research and non-commercial use on GitHub at https://github.com/medfound/medfound. The pre-trained models are publicly available on HuggingFace (https://huggingface.co/medicalai/MedFound-7B, https://huggingface.co/medicalai/MedFound-176B).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw data of the PMC-CR and MedText is available from the website (https://www.ncbi.nlm.nih.gov). MIMIC-III-Note dataset can be found at https://physionet.org/about/database/ and requires access due to its terms of use. MedDX-Note and MedDX-Bench are sourced from real-world clinical scenarios, with institutional review board (IRB) approval obtained from institutions for EHR data collection. Due to privacy regulations, the EHRs cannot be made freely available in a public repository. De-identified data from the MedDX-Note and MedDX-Bench can be requested through the management team by contacting the corresponding author (G.W.), following a defined protocol for data request approval. Generally, all such requests for access to EHR data will be responded to within one month. For the reproduction of our code and model, a representative test dataset from MedDX-Bench, containing samples across specialties, is publicly available on GitHub (https://github.com/medfound/medfound/tree/main/data/test.zip). Data can only be shared for non-commercial use.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not perform sex and gender analysis.
Reporting on race, ethnicity, or other socially relevant groupings	We did not perform race, ethnicity, or other socially relevant groupings analysis.
Population characteristics	The population for training and internal validation has a mean age of 40.96 with a standard deviation of 21.30. The population for external validations has a mean age of 44.99 and a standard deviation of 20.98.
Recruitment	Electronic health records were sourced from MIMIC and CC-DXI (China Consortium for Disease Diagnosis Investigation). The data are representative for the generalized population with no selection biases.
Ethics oversight	Institutional review board and ethics committee approvals were obtained in all locations. The study was approved by the Ethics Committee of Peking University Third Hospital (IRB00006761-M2023607).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For pretraining, we curated a corpus comprising a total of 6.3 billion tokens obtained from medical literature as well as 8.7 million electronic health records. For fintuning and alignment, the dataset was constructed comprising 109,364 cases. Three datasets were used for evaluation, comprising 11,662, 23,917, and 20,257 cases, respectively. Sample size was determined by the data availability. No additional statistical method for sample size estimation was used.
Data exclusions	For pretraining data, no additional data exclusions were performed after data curation. For finetuning data, we excluded data that miss diagnosis label.
Replication	Replication is not relevant. We used independent validation cohorts to test the model, and the models achieved similar performances in the external validation sets.
Randomization	Samples were randomly allocated to the training and testing sets.
Blinding	During the data processing, all data was first de-identified to remove any patient related information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The study was approved by Peking University Third Hospital Medical Science Research Ethics Committee (IRB00006761-M2023607).
Study protocol	We have provided the full clinical protocol in the manuscript.
Data collection	An independent validation set was constructed comprising 300 cases, with 150 cases each from endocrinology and pulmonology.
Outcomes	The diagnoses of expert consensus panel were used as reference to assess the diagnostic accuracy.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*