Medical large language model for diagnostic reasoning across specialties

We developed a medical large language model with 176 billion parameters and fine-tuned it to learn physicians' inferential diagnosis. We showed that the model accurately diagnoses common and rare diseases across specialties, aligns with medical standards, and can be integrated into clinical workflows to effectively enhance physician diagnostic performance.

This is a summary of:
Liu, X. et al. A generalist medical language model for disease diagnosis assistance. *Nature* https://doi.org/10.1038/s41591-024-03416-6 (2025).

Published online: 31 January 2025

The problem

The delivery of accurate diagnoses is crucial in healthcare, especially when diagnosing rare diseases or complex conditions[1], because errors can affect patient outcomes and healthcare quality. For decades, substantial efforts have been made to improve diagnostic accuracy, such as rule-based systems or machine learning techniques, but their reliance on structured data or specialized training has restricted their use. Recently, large language models (LLMs)[2] have shown potential in addressing the 'last-mile' challenge in medical predictive analytics, with their emerging few-shot and zero-shot capabilities enabling them to perform effectively with limited or no task-specific training data in real-world applications. However, well-designed, publicly available LLMs specifically tailored for medical contexts are lacking, and their real-world application is hindered by concerns such as hallucination risks and misalignment with clinical requirements. We sought to develop a medical LLM capable of 'understanding' diverse biomedical knowledge, while further fine-tuning its diagnostic assistance to learn physicians' inferential reasoning across a broad range of tasks and specialties.

The solution

We curated a large-scale medical corpus, MedCorpus, comprising diverse medical literature and 8.7 million real-world electronic health records (EHRs). To develop MedFound, an open-source medical LLM with 176 billion parameters, we pretrained a LLM using MedCorpus to encode extensive medical knowledge and practical experience (Fig. 1a, left). To enable MedFound to learn the step-by-step diagnostic reasoning processes of physicians as a diagnostic generalist —meaning it can be widely applied across specialties, largely replacing task-specific models — we introduced chain-of-thought fine-tuning with a self-bootstrapping approach, leveraging a seed set of manual demonstrations (which are examples of human-crafted diagnostic reasoning steps for EHRs) (Fig. 1a, middle). To align the behavior of the LLM with human values and clinical decision-making preferences, we optimized the diagnostic generalist model's clinical utility using a unified preference alignment framework based on direct preference optimization[3]. This framework incorporates two types of preference: (1) diagnostic hierarchy preference, guided by International Classification of Diseases, 10th Revision (ICD-10)-based hierarchy; and (2) helpfulness preference, based on expert feedback that assessed how helpful the given diagnostic rationales were in supporting decision-making (Fig. 1a, right).

We found that our diagnostic generalist demonstrated superior diagnostic benchmarking performance across specialties and conditions (Fig. 1b, left), including common diseases (in-distribution, out-of-distribution) and rare diseases (long-tailed distribution), outperforming open-access LLMs (such as Llama 3-70B) and closed-source LLMs (such as GPT-4o). Extensive LLM ablation studies demonstrated that pretraining and chain-of-thought fine-tuning enhanced the performance of our diagnostic generalist across various medical tasks.

Our model outperformed human junior and intermediate endocrinology and pulmonology physicians and performed comparably to senior physicians (Fig. 1b, right). Furthermore, we demonstrated physicians with access to our model in their clinical workflows had enhanced performance (including EHR summarization and interpretation, diagnostic reasoning, and formulation of final diagnoses). To further evaluate the diagnostic reasoning quality of the model, we developed a human evaluation framework, CLEVER, which provided insights into the system's strengths and alignment with medical standards. A comparison of before and after preference alignment suggested that MedFound can be optimized by alignment with clinical practice standards, thus enhancing its trustworthiness and applicability.

The implications

Our LLM-based diagnostic generalist model has demonstrated the potential to assist physicians across stages in clinical workflows, offering broad applicability in clinical practice that requires extensive medical knowledge. For example, in complex cases, it can provide diagnostic reasoning for diseases across specialties, offering multidisciplinary support, compared to task-specific tools.

Although our model has undergone an initial alignment step, a limitation is that further refinement through human-in-the-loop methods will be needed to enable its continuous evolution in alignment with clinical feedback.

Our MedFound model focuses on language interaction, but future work could integrate multimodal inputs via vision-language models[4] to enable more comprehensive AI-assisted healthcare scenarios.

Guangyu Wang & Xiaohong Liu
Beijing University of Posts and Telecommunications, Beijing, China.

EXPERT OPINION

"The paper proposes MedFound, a medical LLM with 176 billion parameters pretrained on a large-scale medical corpus derived from diverse medical texts, and then fine-tuned with chain-of-thought, to learn physicians' inferential diagnosis rationale. From the clinical perspective, the authors have constructed the paper in a very complete way, including all the essential components for evaluation of the LLMs' clinical applicability (for example, comparing AI versus physician performance and including an AI-assistance study)." **Weidi Xie, Shanghai Jiao Tong University, Shanghai, China.**

FIGURE

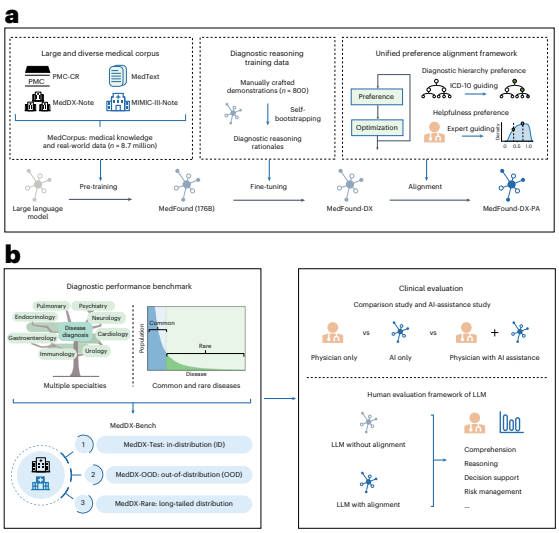

Fig. 1 | Development and evaluation of MedFound as a diagnostic generalist model. a, We pretrained a 176-billion-parameter LLM using a large biomedical corpus (MedText, PMC-CR, MIMIC-III-Note and MedDX-Note), which resulted in MedFound. We then applied chain-of-thought fine-tuning with human crafted demonstrations for it to learn diagnostic reasoning (MedFound-DX). Last, we aligned the model with diagnostic hierarchy and helpfulness preferences, guided by ICD-10 standards and expert feedback to create a diagnostic generalist (MedFound-DX-PA). **b**, Our LLM-based diagnostic generalist demonstrated superior diagnostic benchmarking performance across specialties (left), including common diseases (in-distribution and out-of-distribution) and rare diseases (long-tailed distribution). It outperformed less experienced physicians, and an AI-assistance study and a qualitative study under a human evaluation framework demonstrated its utility to physicians of any experience (right). © 2025, Liu, X. et al.

BEHIND THE PAPER

Building a clinically applicable, cross-specialty diagnostic system has been a focus of our work for years. We began by using machine learning for AI-assisted diagnostic tools. However, applying these tools to a new specialty often required extracting expert-curated clinical features and training from scratch, making it labor-intensive and time-consuming. Inspired by the impact of LLMs in other fields, we set out to develop MedFound-176B, a LLM tailored for biomedicine, and fine-tuned it to learn physicians' inferential diagnosis. Compared with classification-based decision support tools in specific specialties, which are limited to identifying predefined coarse-grained diseases, MedFound demonstrates superior performance across specialties, especially in diagnosing rare diseases in zero-shot scenarios. In addition, it generates diagnostic rationales rather than just diagnoses, which enhances physicians' trust in the system. Our interdisciplinary team, comprising computer scientists, physicians from various departments, and software engineers, has collaborated closely to advance MedFound's diagnostic capabilities in both its development and evaluation. **G.W. & X.L.**

REFERENCES

1. Singh, H. et al. Improving diagnosis in health care—the next imperative for patient safety. *N. Engl. J. Med.* **373**, 2493–2495 (2015).
   This perspective discusses diagnostic errors as a critical and underappreciated patient safety issue.
2. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
   This paper reports a LLM that can be used for clinical question-answering.
3. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. In *Proc. 37th Conference on Neural Information Processing Systems* 53728–53741 (NIPS, 2023).
   This paper introduces the direct preference optimization algorithm, a reinforcement learning-free algorithm with stable training and computational efficiency, designed to align language models with human preferences.
4. Zhang, K. et al. A generalist vision-language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
   This paper introduces a vision-language model for diverse biomedical tasks, such as clinical question answering and summarization tasks.

FROM THE EDITOR

"This work is important for several reasons. First, it develops a medical LLM that is shown to perform well across a broad range of applications and medical specialties. Second, it does so in a completely open-source framework, which is relevant in a landscape dominated mainly by commercial endeavors, and shows competitive performance with respect to its counterparts." **Editorial Team, *Nature Medicine*.**