BIO Integration

Conversational AI in Medicine: Redefining Diagnostics with AMIE

Ziyang Lin[1,2], Yawei Du[3], Wu Yang[4], Yu Qian[1,*] and Junjie Li[1,*]

Effective clinical diagnosis and treatment rely on precise and empathetic doctor-patient communication. Studies have indicated that 60–80% of preliminary diagnoses can be achieved through medical history collection alone [1]. However, disparities in global healthcare resources can limit access to high-quality diagnostic services. With the recent rapid advances in natural language understanding and generation, the potential for large language models (LLM), such as GPT-4, PaLM 2, and Gemini, in medical applications is becoming increasingly evident [2–4]. These models are evolving from passive knowledge repositories into proactive conversational agents capable of guiding discussions, gathering information, and performing diagnostic reasoning as "digital physicians" [5].

In a recent study in *Nature*, Tu et al. introduced Articulate Medical Intelligence Explorer (AMIE), a conversational Artificial Intelligence (AI) system designed to perform clinical diagnostics (**Figure 1**) [6]. Building on the team's earlier work demonstrating LLMs' ability to generate differential diagnosis lists for complex cases [7], Tu et al. demonstrated that AMIE outperformed primary care physicians in simulated clinical dialogues across multiple dimensions, thus marking an important milestone in medical AI's ability to balance understanding and communication [6].

AMIE shifts from a doctor-centered model to an AI-assisted approach, enabling autonomous history-taking and iterative reasoning via multi-turn dialogue, blending fundamental process changes with efficiency gains to enhance collaboration,ultimately becoming a pivotal tool for diagnosis.

To enhance AMIE's clinical dialogue capabilities, the authors integrated three core mechanisms. First, AMIE, which was based on PaLM 2, underwent multimodal data fusion training integrating diverse datasets, including medical conversation transcripts to achieve natural dialogue and empathy; expert clinical summaries to improve structured data processing; question-answering datasets to sharpen diagnostic reasoning; and patient-focused questions to address varying needs in a relevant manner. These datasets enhanced AMIE's fluency, accuracy, and adaptability, as demonstrated by its strong performance in 159 Objective Structured Clinical Examination (OSCE) scenarios. AMIE also uses a dual-loop self-play system. An inner loop simulates doctor-patient dialogues, in which AMIE switches roles to create and refine conversations for accuracy and relevance, whereas an outer loop collects the best dialogues to fine-tune the model, thereby boosting fluency and precision in the absence of constant human oversight. Additionally, AMIE applies a step-by-step reasoning approach during inference, by analyzing dialogue history to refine responses and mimicking clinical logic to achieve robust performance.

This design surpasses basic language generation by weaving clinical reasoning and expert questioning into natural, dynamic conversations. AMIE can gather medical histories and create management plans over multiple dialogue turns, thus closely mimicking human clinical interactions. To test its performance, the authors used a robust validation framework based on the OSCE, a medical education standard. In a double-blind, crossover, randomized trial, AMIE and 20 primary care physicians each handled 159 standardized patient cases across six major specialties in Canada, the UK, and India. Simulated patients and specialists evaluated the dialogues on the basis of patient experience and clinical quality. AMIE outperformed physicians in 30 of 32 expert-rated categories and 25 of 26 patient-rated categories. Beyond achieving better diagnostic accuracy than the physicians, AMIE showed clear communication, empathy, and organized dialogue, particularly in respiratory and internal medicine, although its performance was slightly weaker in obstetrics and urology. For example, in a chest pain

[1]Department of Orthopedic Surgery, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), No. 54 Youdian Road, Hangzhou 310006, P. R. China

[2]The First School of Clinical Medicine, Zhejiang Chinese Medical University, No. 548 Binwen Road, Hangzhou 310053, P. R.China

[3]Department of Orthopaedics, Shanghai Key Laboratory for Prevention and Treatment of Bone and Joint Diseases, Shanghai Institute of Traumatology and Orthopaedics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Road, Shanghai 200025, P. R. China

[4]Pharmaceutical Sciences Laboratory, Faculty of Science and Engineering, Åbo Akademi University, Turku 20520, Finland

*Correspondence to: Yu Qian, Department of Orthopedic Surgery, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), No. 54 Youdian Road, Hangzhou 310006, P. R. China. E-mail: qianyu@zcmu.edu.cn; Junjie Li, Department of Orthopedic Surgery, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), No. 54 Youdian Road, Hangzhou 310006, P. R. China. E-mail: docljj@163.com

Published Online: July 24 2025

Available at: https://bio-integration.org/

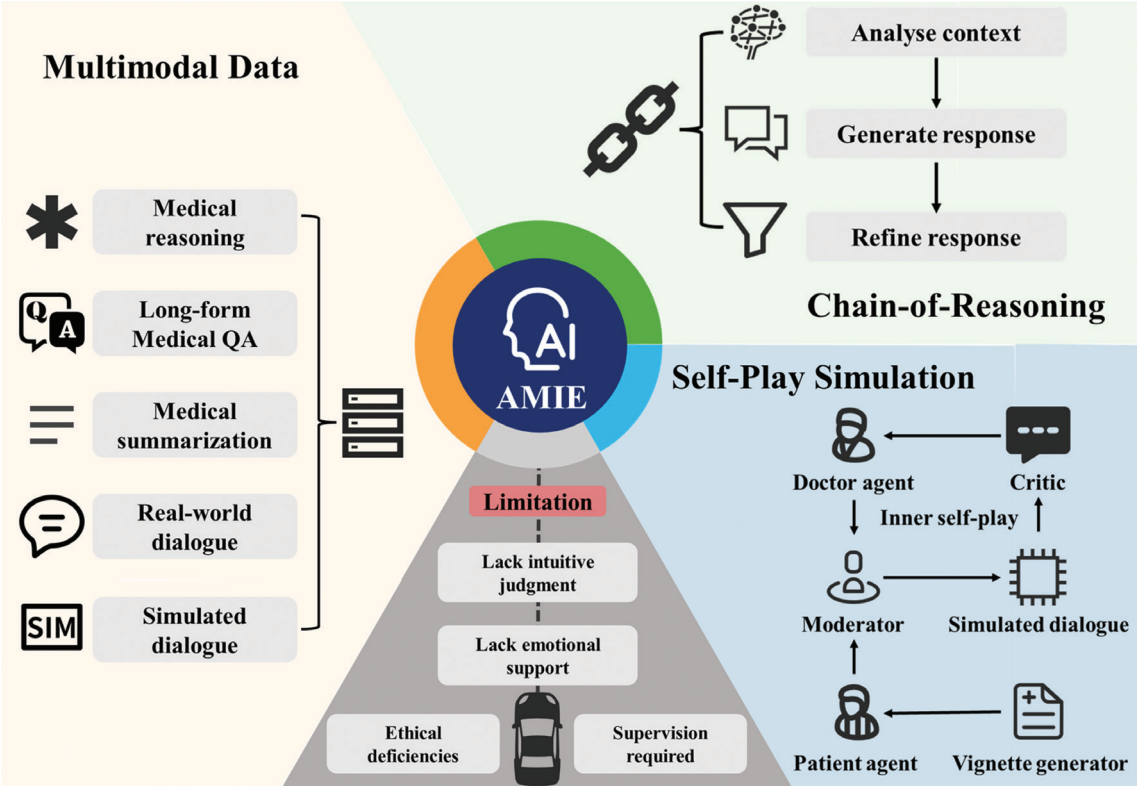

Figure 1 Advantages and limitations of the LLM-driven conversational AI diagnostic system AMIE in clinical diagnosis.

case, AMIE demonstrated clear reasoning and clinical focus by skillfully collecting details (onset, spread, and symptoms); prioritizing acute coronary syndrome and acid reflux in its diagnoses; and suggesting electrocardiogram and cardiac enzyme tests.

These empirical results underscore AMIE's advanced capabilities driven by several key innovations. Unlike other medical AI systems, such as Med-PaLM 2, GatorTron, and Clinical Camel, which focus primarily on single-turn question answering or document summarization, AMIE's conversational design enables sustained interaction, proactive questioning, and contextual integration, in a manner simulating the full course of a clinical interview. Its closed-loop self-iteration, leveraging self-play, chain-of-reasoning, and critical feedback, drives continued learning and optimization. Furthermore, instead of relying on simple accuracy metrics, the multidimensional evaluation included assessment of key clinical skills, such as listening, understanding, management recommendations, and empathy, while incorporating patient perspectives for practice-oriented assessment. Through large-scale pretraining and self-supervised learning, AMIE pioneers dialogue-driven diagnostics by delivering human-like reasoning and empathy in dynamic clinical conversations. Importantly, its design prioritizes interpretability and safety by integrating reasoning chains with expert review, thereby enhancing clinical adoptability.

Nonetheless, despite its strong performance, AMIE is unlikely to replace human physicians in the near term, from the perspective of clinicians. Clinical diagnosis uses experience and subtle cues to diagnose patients holistically, whereas AMIE's reliance on clear data might limit nuance in its output. Moreover, AMIE's efficient history-taking and diagnosis suggestions might lead to decreased reliance on physicians' skills, thus potentially weakening their ability to engage patients or think iteratively. Overusing AMIE might also shift routine cases to AI-driven processes while sidelining human judgment. Although AMIE provides empathetic responses, they might be perceived as inhuman, thus risking patient trust if AMIE is seen as a replacement for physicians instead of an assistant. Decreased human contact with patients might also affect outcomes in sensitive cases. In medical training, AMIE's real-time feedback could help students; however, excessive reliance on AMIE might stunt students' ability to reason independently, if they favor AMIE's structured answers over patient-focused analysis. Finding a balance will be crucial for physicians to maintain their core clinical skills.

Additionally, strict healthcare ethical and legal frameworks require rigorous oversight of AI applications to ensure patient safety and privacy [8]. AMIE's transparent reasoning represents an advancement yet requires further refinement, given that clinical errors can have serious consequences. AMIE is best suited as a physician's assistant aiding in streamlining information collection, generating structured consultation summaries, suggesting differential diagnoses, and providing explainable recommendations, thus enabling physicians to focus on complex decisions and patient-centered care. In community hospitals, AMIE could support primary diagnostics to address limited specialist access; in tertiary hospitals, it could enhance specialists' precision in complex workflows; in resource-scarce areas, it could democratize expertise, although challenges such as internet access and training must be addressed. To create more reliable and inclusive AI systems, future efforts should focus

on three areas: improving personalized diagnostics by incorporating factors such as age, medical history, and lifestyle; enhancing handling of rare diseases and atypical cases for safe use; and adapting language models, tone, and cultural references to diverse linguistic and low-resource settings to boost global accessibility and equity.

AMIE's phased plan seeks to transform diagnostics by ensuring accuracy and accessibility. In the auxiliary diagnostics phase, AMIE can help physicians with imaging and decisions in advanced hospitals, and can aid in addressing issues such as poor data, physician distrust, and system mismatches by standardizing data, encouraging physician-AMIE teamwork, and learning the use of the system. In the enhanced diagnostics phase, AMIE can serve as a smart tool providing instant diagnoses and custom treatment plans in mid- and high-resource clinics; using secure data; working with regulators; and adapting solutions to address privacy, rules, and growth challenges. In the redefining diagnostics phase, AMIE can integrate diverse types of data and might serve as a key diagnostic tool able to be deployed in low-resource areas worldwide. AMIE can facilitate precise, tailored care by predicting health risks; ensuring fairness through better technology, training, and unbiased data; and delivering trustworthy, fair diagnostics with ethical guidance.

Advancing systems such as AMIE will require sustained multi-stakeholder collaboration involving regulatory bodies, ethical oversight committees, healthcare providers, and patients. Through such coordinated efforts, AI has the potential to transform healthcare delivery by increasing accessibility and quality. However, maintaining an overarching principle in which AI supports but does not replace human physicians will be critical, to ensure that technological innovation enhances rather than undermines patient safety, empathy, and care quality.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

[1] Levine D. History taking is a complex skill. Br Med J 2017;358:j3513. [PMID: 28729457 DOI: 10.1136/bmj.j3513]

[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, et al. Large language models in medicine. Nat Med 2023;29(8):1930-40. [PMID: 37460753 DOI: 10.1038/s41591-023-02448-8]

[3] Menezes MCS, Hoffmann AF, Tan AL, Nalbandyan M, Omenn GS, et al. The potential of Generative Pre-trained Transformer 4 (GPT-4) to analyse medical notes in three different languages: a retrospective model-evaluation study. Lancet Digit Health 2025;7(1):e35-43. [PMID: 39722251 DOI: 10.1016/S2589-7500(24)00246-2]

[4] Chen Z, Er Saw P. Integration in biomedical science 2024: emerging trends in the post-pandemic era. BIO Int 2024;5(1):998. [DOI: 10.15212/bioi-2024-1001]

[5] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31:943-50. [PMID: 39779926 DOI: 10.1038/s41591-024-03423-7]

[6] Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642:442-50. [PMID: 40205050 DOI: 10.1038/s41586-025-08866-7]

[7] McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642:451-57. [DOI: 10.1038/s41586-025-08869-4]

[8] Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health 2024;6(11):e848-56. [PMID: 39294061 DOI:10.1016/S2589-7500(24)00143-2]