This article in provides the information about AMIE (Articulate Medical Intelligence Explorer), an AI system which is used in medical conversations with patients and help diagnose illnesses. It is based on a Large Language Model (LLM) and focus to improve how healthcare is delivered where there is limited access to experienced doctors.

Researchers have used various real conversation to train the model to gather patient information, ask follow-up questions, and show empathy just like a real doctor. It also uses a unique learning method to practice with itself in a simulated environment to get better over time. To test AMIE, it was compared to real primary care doctors in 159 simulated cases. Patients and medical specialists rated AMIE's performance as better than the doctors in most cases both in terms of diagnostic accuracy and communication skills.

Even though this is a good step researchers accept that AMIE was tested using text-chat, not in the how doctors communicate. More research is required to implement in real-world healthcare.

The study assesses the diagnostic accuracy and communication quality of AMIE compared to actual primary care physicians (PCPs). AMIE outperformed PCPs at every top-k diagnostic guess, consistently, including during difficult encounter types that involved non-disease, such as resolved or recurrent conditions. AMIE had extremely good results in specialties such as respiratory and internal medicine, and was somewhat weaker in obstetrics/gynecology and urology. AMIE's accuracy was stable, regardless of location (India versus Canada). AMIE's advantage is coming from the reasoning capabilities and not from acquiring different information, as AMIE and PCPs had similar lengths of dialogue and then obtained similar amounts of information. IPDAM rated AMIE's accuracy to be equal to PCPs' even when AMIE was given conversations from some PCPs. Communication-wise, AMIE was rated as being better than PCPs by patient actors on all but 1 of 26 variables, and by specialists via all but 2 of 32 criteria about the consultations. The only reason AMIE did poorly was admitting to mistakes. The AI-based auto-evaluations also agreed well with the experts' scores. The training with self-play also improved the conversational quality with the same set of criteria used in the expert evaluation. Overall, AMIE is generally accurate, communicative, efficient, and robust across multiple domains and contexts.

The author compared AMIE and PCPs on two parts: diagnostic performance and conversational performance.

For diagnostic performance, AMIE showed good ability in analysing medical information and asking patients the right questions to get more important details. It did very well, similar to PCPs. This shows that AMIE has strong diagnostic reasoning skills. But the author also mentioned that the medical area is very sensitive and related to safety. So, even if AMIE sometimes performs better than PCPs, human doctors are still necessary. AMIE should always be used under the supervision of PCPs.

For conversational performance, AMIE performed better than PCPs, especially in empathy and communication. The author explained that LLM usually gives longer answers, and that helps patients feel understood. In this study, the conversation used text chat, and doctors had time pressure. So, they couldn't talk like writing emails or texts in daily life. Also, since language is the strength of LLMs, it is understandable that LLMs can communicate better. The doctors also couldn't use voice or body language in this test, so their performance was not shown fully.

Additionally, the author suggested that doctors can work together with AI. That means human can use AI's strengths to improve medical service.

Scaling of AMIE's training was done using simulation which means this was done without using real patient data. Doctor does the diagnosis, patient has a realistic medical history, moderator stops the conversation when it seems to be over and a critic who provides feedback to the doctor for improvement in terms of empathy and diagnosis quality. Despite this sophisticated a setup, there were some shortfalls like all types of patient behaviors not being captured. This was especially for those that have low literacy and moderator assumes that every conversation should ends with a diagnosis.

Testing for bias in terms of race, gender or literacy was not performed. For this aspect, authors believe there should be participatory evaluations, red-teaming and transparent reporting performed. Participatory evaluations involve engaging real people with diverse and underrepresented backgrounds in the AI's evaluation. Red-teaming means involving human or AI Agent evaluators that get deliberately try to fool the system – testing for vulnerabilities. Transparent reporting includes disclosing how the model was trained, what was the data that was used, limitations of the system and what went into the decision making of the system.

There were a few limitations in using AMIE in real world. There was no process involved wherein if the AI did not perform as expected – a human would be involved. It also needs to be updated with most recent clinical knowledge to ensure its usage is reliable for which regulatory readiness and bias auditing are necessary. Authors also emphasize not to over-rely on these systems and ensuring having appropriate guardrails to avoid that.

Authors conclude by mentioning that AMIE performs well for clinical history taking and empathetic interaction. However, they also say that the current results are only from simulation and a lot more steps need to be taken for real world usage. If this is done – tools like AMIE can be used globally for high-quality healthcare.