Medical AI lacks clinical reasoning, U of A researcher finds

Neurology resident Liam McCoy is looking to run clinical trials to test whether AI actually improves patient outcomes.

Sara Sunderji • January 29, 2026 3 minutes read

Lily Polenchuk

University of Alberta neurology resident Liam McCoy tested leading AI models on clinical reasoning and found they treated irrelevant details as important roughly 90 per cent of the time.

McCoy tested 10 leading AI models using 750 script-concordance questions and compared their answers with those of medical students, residents, and attending physicians. The best model scored 67.8 per cent. Overall, models performed as well as first or second-year medical students but underperformed senior residents and experts.

McCoy and colleagues evaluated leading large language models on clinical decision-making using a format different from traditional medical exams. Rather than testing AI on multiple-choice questions — where models already demonstrate superhuman performance — the research focused on script concordance testing.

“Multiple choice exams provide all the necessary information to answer the question,” McCoy said. “That’s not something that

mirrors what we do as physicians in the real world."

In real clinical practice, information arrives sequentially over time. Physicians constantly re-assess their diagnosis as new data emerges.

Script concordance testing was designed to capture this dynamic process. Each question gives a brief clinical vignette and a working hypothesis, then presents new information. Test-takers must decide whether that information makes the hypothesis more likely, less likely, or unchanged.

"It throws you right in the middle of that clinical encounter and says, okay, how do things shift and change in response to new information?" McCoy said.

To build their benchmark, McCoy's team assembled 750 script concordance testing items from 10 international datasets spanning various medical specialties.

AI performance

The results found that AI models fell short of expert clinicians. OpenAI's 03 performed the best at 67.8 per cent accuracy, followed by GPT-4o at 63.9 per cent. While models matched or exceeded medical students on some items, none reached the level of senior medical students, residents, or attending physicians.

The error patterns revealed a critical flaw. 30 per cent of test questions contained irrelevant "red herrings" that shouldn't change a diagnosis. Human clinicians recognized and ignored these distractions. AI models treated them as significant 90 per cent of the time.

"The models were really bad at that in our study, especially the more modern, powerful reasoning models," McCoy said.

This creates a dangerous dynamic, according to McCoy. Models are fine-tuned to provide confident, helpful responses — an optimization that backfires in medical reasoning.

“One of our biggest concerns is that they will explain a mistake in a way that makes you agree with them,” McCoy said.

The findings reveal a fundamental gap in AI clinical reasoning.

“What’s unique and kind of fun, but also scary about large language models is that they are simultaneously smarter than us on specific things,” McCoy said. “And then in other areas they’ll make mistakes that not even a first-year medical student would make.”

The problem is how AI processes information. Models excel at pattern matching — building associations between symptoms and diagnoses from training data. But clinical reasoning requires flexible inference.

“These more nuanced aspects of reasoning that for a human, if you understood A and C, you can infer B,” McCoy said. “The models often struggle to have enough of a world model to understand things that they weren’t explicitly trained on.”

Good physicians anticipate how test results might disconfirm a hypothesis and recognize when information is irrelevant. AI systems lack this deeper understanding.

Future directions

McCoy’s next step is running clinical trials to test whether AI actually improves patient outcomes.

“Benchmarks matter only if they translate into better patient care,” McCoy said.

With colleagues at Harvard and Stanford, McCoy is building the Medical AI Superintelligence Test: a meta-benchmark that

assesses AI behaviour, not just knowledge.

“Rather than just purely how much knowledge does it have on a zero to 100 scale, what are its tendencies?” McCoy said. “How aggressive is it in ordering tests? How likely is it to value certain different aspects of the patient’s risk of death versus the risk of disability?”

McCoy’s script concordance benchmark is now public at concor.dance. The work has attracted interest from Google and Microsoft Research.

Current models lack a critical clinical skill: knowing when information doesn’t matter. In medicine, ignoring red herrings is as important as spotting meaningful signals, McCoy explained.

“This technology is coming one way or another,” McCoy said. “So I think as physicians we need to make sure it is effective, equitable and aligned with what patients need, rather than allowing it to just be driven by external actors.”