Can AI models reason like clinicians?

A study evaluating frontier LLMs across the clinical workflow reveals a critical gap in the kind of reasoning that matters most for patient safety.

Peter Attia
By: Taylor Yeater, Lauren Fritsch, Peter Attia

June 20, 2026

READ TIME**8 MINUTES**

UNDERSTANDING SCIENCE

If you've spent any time recently working with a generative AI model—like ChatGPT, Claude, or Gemini, —you've likely been blown away by just how good they have gotten. The most recent iterations of any of these tools are performing so well, countless industries are having to evolve practically overnight to keep up. Just a few years ago, several general-purpose large language models (LLMs, the technology underlying these AI tools) were nearing the passing threshold of medical board examinations.[1] Newer top models are routinely exceeding 90% accuracy on those tests (where medical school students average 59.3%).[2] These developments have been so rapid, we are quickly losing the

ability to evaluate them effectively. In several cases, we have reached *benchmark saturation*: we are no longer able to design evaluations sufficiently complex to meaningfully differentiate model performance.

Understandably, the rapid improvement in these models has generated substantial enthusiasm for integrating these tools into clinical settings. But as most of us know, performing well on a test is not always the same as performing well in the real world—and when it comes to clinical decision making, real world performance is far more important than acing the exam. High scores on medical boards exams may make these models seem more clinically credible than they actually are. It is far more important that these models show excellent clinical *reasoning*. There are many ways these models could fail in practice: misreading context, anchoring on incomplete information, failing to ask the right follow-up questions, or projecting certainty where uncertainty is warranted.

This leaves two distinct but related questions: How do these models perform across the full range of clinical reasoning tasks? And where, if anywhere, do LLMs belong in actual clinical workflow—as decision support tools, as triage aids, or as something approaching autonomous agents?

Just as the development of these tools is moving at a dizzying pace, so too is progress in the space of evaluating them—new frameworks and benchmarks are appearing nearly as fast as the models themselves. In an attempt to create a lasting benchmark, one recent study proposed a new composite score designed to test LLMs across the full clinical workflow—not just final diagnosis, but the entire chain of reasoning that clinical medicine requires.[3]

Testing LLMs on the full clinical workflow

The authors tested 21 frontier LLMs—including models from OpenAI, Anthropic, xAI, DeepSeek and Google DeepMind—across 29 standardized clinical vignettes drawn from the MSD Manual (Merck Sharp & Dohme), a widely used clinical reference that publishes peer-reviewed, structured case presentations developed by independent clinical experts. Each vignette presents a full clinical picture—history of present illness, review of systems, physical examination findings, and laboratory results—and walks through the clinical encounter sequentially, from differential diagnosis through testing and management. Rather than asking models a single question and scoring the answer, the authors preserved clinical context across each step, presenting the case the way it would actually unfold. To account for variability in model outputs, each vignette was run in triplicate and scored by medical student evaluators against the MSD Manual answer keys.

The authors also introduced a new scoring system—PrIME-LLM (Proportional Index of Medical Evaluation for LLMs)—built around a simple premise: A model performing well in *one* domain while failing in *others* should not score the same as a model performing *consistently* across all five. Rather than summarizing performance as a single accuracy score, PrIME-LLM calculates a normalized

polygonal area across five domains: differential diagnosis, diagnostic testing, final diagnosis, management, and miscellaneous clinical reasoning. Think of it as a radar plot: each domain occupies one axis, and a model's score in each domain determines the shape of its polygon (**Fig 1.**). A model that excels in one domain but fails in others produces a lopsided shape with a small total area; a model that performs consistently across all five produces a larger, more balanced polygon. The final PrIME-LLM score represents that area as a proportion of a perfect score. This structure penalizes uneven performance that a simple accuracy average would obscure.

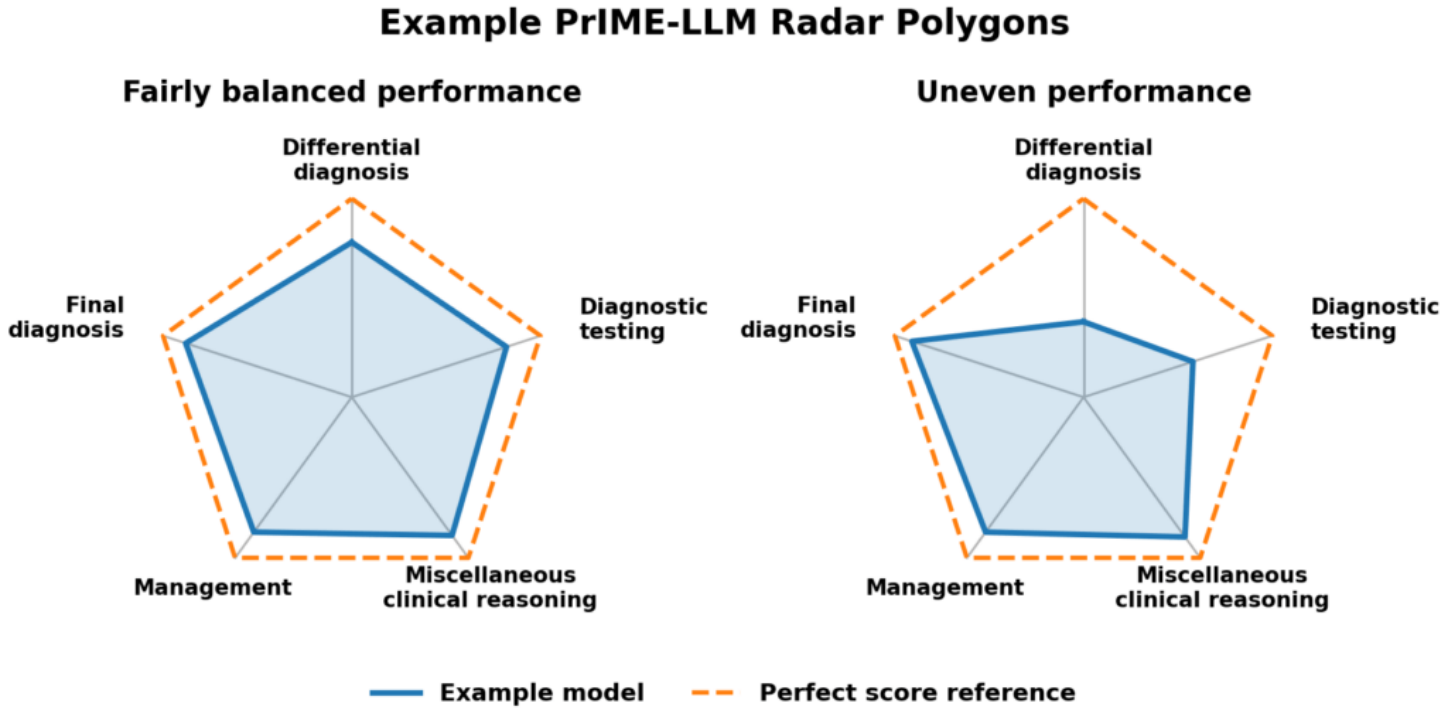

Figure: Example of well balanced (A) and unbalanced (B) PrIME-LLM polygons

When getting it right isn't good enough

On the surface, the raw accuracy results look reassuring. Across 29 vignettes and 16,254 individual responses, overall accuracy across all 21 models clustered between 81% and 90%—a familiar pattern for anyone tracking LLM performance.

But the PrIME-LLM scores told a different story. A perfect score is 100%, meaning a model performed at ceiling across all five domains with no weak links. The key distinction from raw accuracy is that a simple average can hide lopsided performance—a model scoring well on most domains, but 50% on one domain might still average out to 85% accuracy, but its PrIME-LLM score would be dragged down more by the gap. Using this metric, the models ranged from 64% to 78%, showing more separation than the narrow 81-90% accuracy band and offering a clearer picture of balanced clinical reasoning.

The domain-by-domain results start out looking reasonable. On final diagnosis and management—where the model is given a relatively complete clinical picture and asked to name the condition or outline a treatment plan—models performed well, with accuracy rates of about 85—95%. The authors also reported a stricter measure: failure rate, which counted a response as incorrect unless the model got the *full* answer right, with no partial credit. Even by this standard, final diagnosis failure rates stayed below 40%—far from flawless, but the models can pattern-match a diagnosis when most of the information is already on the table.

But differential diagnosis was where the models struggled most. Differential diagnosis is the stage at which a clinician considers not only the most likely explanation for a patient's presentation, but also the less likely and potentially more dangerous alternatives. It is where the question "what else could this be?" determines which tests get ordered, which red flags get flagged, and which diagnoses don't get missed. Accuracy rates were around 75% across models, but remember that failure rate awards no partial credit—and by that measure, failure rates exceeded 80% across all 21 models tested. Not some of them, *all* of them. Because PrIME-LLM rewards balanced performance, this single domain dragged composite scores down significantly—even for the top performers.

The authors offer a pointed interpretation of this gap. LLMs appear to collapse prematurely onto a single answer rather than preserving uncertainty and iteratively refining competing possibilities, which is precisely what good clinicians do. Strong performance on final diagnosis may actually reflect this tendency rather than contradict it. When the correct answer can be pattern-matched from a complete set of clinical information, LLMs do well. When the task requires holding multiple competing possibilities in tension, weighting them against incomplete information, and knowing which question would most efficiently narrow the field—they falter.

For anyone who interacts with these models regularly, this may be familiar: While LLMs excel at giving a relatively clear answer the first time, they struggle to give alternatives when you reject the initial idea. If you haven't played around with these models much, an analogy may be better. Imagine a cooking competition in which a contestant must identify a dish by tasting it. They detect tomato, basil, mozzarella, and bread—and they say "pizza." They might be right. But a skilled chef would also consider what else it could be—bruschetta, caprese on toast, flatbread, chicken parmesan. And a great chef would know which questions narrow the field further. Is the bread crispy or soft? Are the tomatoes fresh or cooked into a sauce? Is there pasta underneath?

The LLMs in this study were reasonably good at identifying "pizza"—at naming the most probable diagnosis—but struggled considerably at generating the complete list of plausible alternatives and knowing which follow-up questions would distinguish among them. In a cooking competition, the cost of that failure is a lost round. In medicine, the stakes are categorically different. Clinical presentations overlap far more than dishes do, and anchoring on the most obvious answer while a more dangerous diagnosis goes unrecognized is not a lost round—it is a missed cancer, a delayed intervention, or treatment aimed at the wrong target.

A baseline, not a ceiling

Before drawing sweeping conclusions, it's worth being precise about what this study measured—and what it didn't. The authors evaluated off-the-shelf models as they come "out of the box," without any of the augmentation that a purpose-built clinical AI system might eventually have access to. No real-time search or retrieval of medical guidelines, no clinical calculators, no patient records, no structured reasoning workflows, no agentic capabilities (i.e., the ability to autonomously plan and execute tasks like ordering tests or querying databases rather than simply answering questions). If you've used these tools yourself, you know the gap: in practice, models can search the web, review attached reference documents, or use extended reasoning. None of that was enabled here. In the authors' own framing, this was a baseline evaluation of longitudinal clinical reasoning—not a ceiling test.

A more capable system would have access to diagnostic references, published guidelines, laboratory calculators, full patient history, and structured reasoning tools that could help a model generate and rank differential diagnoses more systematically. Whether those augmentations would close the differential diagnosis gap is an open question—though it is worth noting that within this study, reasoning-optimized models (those whose architecture is designed to perform internal deliberation—weighing and refining a chain of reasoning—before arriving at a final answer) already showed a meaningful edge. Reasoning models scored significantly higher on PrIME-LLM (mean: 76%) than nonreasoning models (mean: 67%), a difference that was both statistically significant and large in effect size (Cohen *d*: 2.60). That gap suggests architecture and inference design matter—and that purpose-built augmentation could push performance further still.

The study also was not designed to determine whether LLMs perform better or worse than human clinicians—a separate question, and arguably the more consequential one. Some evidence on this front does exist, and it paints a more nuanced picture. A head-to-head comparison found that GPT-4 outscored both attending physicians and residents on clinical reasoning using standardized cases.[4] A more recent study reported that OpenAI's o1-preview reasoning model matched or exceeded physician baselines across six experiments, including real emergency department cases drawn from electronic health records—with its advantage most pronounced in early-stage triage, where clinicians must make decisions with minimal information.[5] But a randomized trial found that giving physicians access to an LLM during diagnostic reasoning did not meaningfully improve their performance compared to conventional resources.[6] In other words, LLMs can outperform clinicians on structured cases when working in isolation, but the real clinical question—whether these tools actually improve physician decision-making when used as intended—remains largely unanswered. That is the scenario that matters most, and it is the one most in need of prospective evaluation.

The bottom line

These findings highlight the need for precision about which tasks, which settings, and which levels of oversight are appropriate for deploying LLMs in clinical settings. For lower-stakes, clinician-supervised work—summarizing patient information, drafting patient-friendly explanations, structuring documentation—the case for adoption is reasonable. The technology's strengths are well-matched to these tasks, and the consequences of imperfection are manageable. Autonomous diagnostic reasoning is a different proposition entirely. The concern is not just hallucination—the well-documented tendency of LLMs to generate plausible-sounding but fabricated information. It is that these models may project confidence precisely where clinical reasoning demands uncertainty.

A model that arrives at the correct final diagnosis 90% of the time sounds useful until you consider that it may fail to generate the broader differential needed to catch the other 10%. The problem isn't that the model is wrong sometimes. It's that it has no validated way of asking "what if I'm wrong?" In medicine, confidence is only warranted after the alternatives have been considered and ruled out. A model that skips that step isn't being efficient. It's being brittle in exactly the place where the cost of failure is highest.

The appeal of broader deployment is understandable—the capability gains are real, and the pace of progress makes caution feel like you are falling behind. But medicine has navigated this tension before. We do not approve drugs because they show promise in early testing; we require evidence of safety and efficacy in the populations and conditions where they will actually be used. The same standard should apply here. Until we have prospective data showing that a particular implementation of AI reliably improves patient outcomes in real clinical settings—not on simulated vignettes, but in live clinical workflows—we should continue exercising caution when using LLMs in our clinical workflows.

For a list of all previous weekly emails, click here.

podcast | website | ama

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

2. Bicknell BT, Butler D, Whalen S, et al. Critical analysis of ChatGPT 4 Omni in USMLE disciplines, clinical clerkships, and clinical skills. *JMIR Med Educ*. 2024;10:e63430. doi:10.2196/63430

3. Rao AS, Esmail KP, Lee RS, et al. Large language model performance and clinical reasoning tasks. *JAMA Netw Open*. 2026;9(4):e264003. doi:10.1001/jamanetworkopen.2026.4003

4. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern Med*. 2024;184(5):581-583. doi:10.1001/jamainternmed.2024.0295

5. Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. *Science*. 2026;392(6797):524-527. doi:10.1126/science.adz4433

6. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial: A randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969