

Microsoft says its MAI-DxO system diagnosed up to 85% of NEJM case challenges

Source copy kept by wheresthe.ai for the story "Microsoft says its MAI-DxO system diagnosed up to 85% of NEJM case challenges" (<https://wheresthe.ai/d/us-microsoft-mai-dxo-diagnosis-2025-06/>).

Source: <https://microsoft.ai/news/the-path-to-medical-superintelligence/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 6dbbc73abbf011c4228b8e7737e37a052bf4c6702f704a9a3d3d29ca236656c6

Copilot -->

Health

The Path to Medical Superintelligence

Dominic King & Harsha Nori

June 30, 2025

Health

Dominic King & Harsha Nori

LI

X

FB

Benchmarked against real-world case records published each week in the New England Journal of Medicine, we show that the Microsoft AI Diagnostic Orchestrator (MAI-DxO) correctly diagnoses up to 85% of NEJM case proceedings, a rate more than four times higher than a group of experienced physicians. MAI-DxO also gets to the correct diagnosis more cost-effectively than physicians.

As demand for healthcare continues to grow, costs are rising at an unsustainable pace, and billions of people face multiple barriers to better health – including inaccurate and delayed diagnoses. Increasingly, people are turning to digital tools for medical advice and support. Across Microsoft's AI consumer products like Bing and Copilot, we see over 50 million health-related sessions every day. From a first-time knee-pain query to a late-night search for an urgent-care clinic, search engines and AI companions are quickly becoming the new front line in healthcare.

We want to do more to help – and believe generative AI can be transformational. That's why, at the end of 2024, we launched a dedicated consumer health effort at Microsoft AI, led by clinicians, designers, engineers, and AI scientists. This effort complements Microsoft's broader health initiatives and builds on our longstanding commitment to partnership and innovation. Existing solutions include RAD-DINO which helps accelerate and improve radiology workflows and Microsoft Dragon Copilot, our pioneering voice-first AI assistant for clinicians.

For AI to make a difference, clinicians and patients alike must be able to trust its performance. That's where our new benchmarks and AI orchestrator come in.

Medical Case Challenges and Benchmarks

To practice medicine in the United States, physicians need to pass the United States Medical Licensing Examination (USMLE), a rigorous and standardized assessment of clinical knowledge and decision making. USMLE questions were among the earliest benchmarks used to evaluate AI systems in medicine, offering a structured way to compare model performance – both against each other and against human clinicians.

In just three years, generative AI has advanced to the point of scoring near-perfect scores on the USMLE and similar exams. But these tests primarily rely on multiple-choice questions, which favor memorization over deep understanding. By reducing medicine to one-shot answers on multiple-choice questions, such benchmarks overstate the apparent competence of AI systems and obscure their limitations.

At Microsoft AI, we're working to advance and evaluate clinical reasoning capabilities. To move beyond the

limitations of multiple-choice questions, we've focused on sequential diagnosis, a cornerstone of real-world medical decision making. In this process, a clinician begins with an initial patient presentation and then iteratively selects questions and diagnostic tests to arrive at a final diagnosis. For example, a patient presenting with cough and fever may lead the clinician to order and review blood tests and a chest X-ray before they feel confident about diagnosing pneumonia.

Each week, the New England Journal of Medicine (NEJM) – one of the world's leading medical journals – publishes a Case Record of the Massachusetts General Hospital, presenting a patient's care journey in a detailed, narrative format. These cases are among the most diagnostically complex and intellectually demanding in clinical medicine, often requiring multiple specialists and diagnostic tests to reach a definitive diagnosis.

How does AI perform? To answer this, we created interactive case challenges drawn from the NEJM case series – what we call the Sequential Diagnosis Benchmark (SD Bench). This benchmark transforms 304 recent NEJM cases into stepwise diagnostic encounters where models – or human physicians – can iteratively ask questions and order tests. As new information becomes available, the model or clinician updates their reasoning, gradually narrowing toward a final diagnosis. This diagnosis can then be compared to the gold-standard outcome published in the NEJM.

Each requested investigation also incurs a (virtual) cost, reflecting real-world healthcare expenditures. This allows us to evaluate performance across two key dimensions: diagnostic accuracy and resource expenditure. You can watch how an AI system progresses through one of these challenges in this short video.

Getting to a Correct Diagnosis

We evaluated a comprehensive suite of frontier generative AI models against the 304 NEJM cases. The foundation models tested included GPT, Llama, Claude, Gemini, Grok, and DeepSeek.

Beyond baseline benchmarking, we also developed the Microsoft AI Diagnostic Orchestrator (MAI-DxO), a system designed to emulate a virtual panel of physicians with diverse diagnostic approaches collaborating to solve diagnostic cases. We believe that orchestrating multiple language models will be critical to managing complex clinical workflows. Orchestrators can integrate diverse data sources more effectively than individual models, while also enhancing safety, transparency, and adaptability in response to evolving medical needs. This model-agnostic approach promotes auditability and resilience, key attributes in high-stakes, fast-evolving clinical environments.

Fig 1.

The MAI-Dx Orchestrator turns any language model into a virtual panel of clinicians: it can ask follow-up questions, order tests, or deliver a diagnosis, then run a cost check and verify its own reasoning before deciding whether to proceed.

MAI-DxO boosted the diagnostic performance of every model we tested. The best performing setup was MAI-DxO paired with OpenAI's o3, which correctly solved 85.5% of the NEJM benchmark cases. For comparison, we also evaluated 21 practicing physicians from the US and UK, each with 5-20 years of clinical experience. On the same tasks, these experts achieved a mean accuracy of 20% across completed cases.

MAI-DxO is configurable, enabling it to operate within defined cost constraints. This allows for explicit exploration of the cost-value trade-offs inherent in diagnostic decision making. Without such constraints, an AI system might otherwise default to ordering every possible test – regardless of cost, patient discomfort, or delays in care. Importantly, we found that MAI-DxO delivered both higher diagnostic accuracy and lower overall testing costs than physicians or any individual foundation model tested.

Comparison of AI powered diagnostic agents by accuracy and average diagnostic test cost per case. Top performing agents appear toward the top left quadrant, reflecting higher accuracy and lower cost. The lower dotted line represents the performance range of the best individual foundation models. The purple line traces the performance of MAI-DxO across different configurations. The red cross indicates the average performance of 21 practicing physicians.

What's Next?

Physicians are typically characterized by the breadth or depth of their expertise. Generalists, like family physicians, manage a wide array of conditions across ages and organ systems. Specialists, such as rheumatologists, focus deeply on a single system, disease area or even condition. No single physician, however,

can span the full complexity of the NEJM case series. AI, on the other hand, doesn't face this trade-off. It can blend both breadth and depth of expertise, demonstrating clinical reasoning capabilities that, across many aspects of clinical reasoning, exceed those of any individual physician.

This kind of reasoning has the potential to reshape healthcare. AI could empower patients to self-manage routine aspects of care and equip clinicians with advanced decision support for complex cases. Our findings also suggest that AI reduce unnecessary healthcare costs. U.S. health spending is nearing 20% of US GDP, with up to 25% of that estimated to be wasted – per having little influence on patient outcomes.

Of course, our research has important limitations. Although MAI-DxO excels at tackling the most complex diagnostic challenges, further testing is needed to assess its performance on more common, everyday presentations. Clinicians in our study worked without access to colleagues, textbooks, or even generative AI, which may feature in their normal clinical practice. This was done to enable a fair comparison to raw human performance.

A novel aspect of this work is its attention to cost. While real-world health costs vary across geographies and systems, and include many downstream factors that we don't account for, we apply a consistent methodology across all agents and physicians evaluated to help quantify high level trade-offs between diagnostic accuracy and resource use.

For us, this is just the first step. We're energized by the opportunities ahead. Important challenges remain before generative AI can be safely and responsibly deployed across healthcare. We need evidence drawn from real clinical environments, alongside appropriate governance and regulatory frameworks to ensure reliability, safety, and efficacy. That's why we're partnering with leading health organizations to rigorously test and validate these approaches—an essential step before any broader roll out.

Together with our partners, we strongly believe that the future of healthcare will be shaped by augmenting human expertise and empathy with the power of machine intelligence. We are excited to take the next steps in making that vision a reality.

Further information

SD Bench and MAI-DxO are research demonstrations only and are not currently available as public benchmarks or orchestrators. You can find more detail on the underlying methodology and results in a pre-print paper published alongside this blog. We are in the process of submitting this work for external peer review and are actively working with partners to explore the potential to release SDBench as a public benchmark.

Acknowledgments

We are grateful to NEJM Group for permission to use the NEJM cases in the research reported in this blog post. The research described here has benefited from the insights of many people. We are grateful to the authors named on the arXiv paper and the wider team at MAI. We also thank further colleagues both inside and outside of Microsoft for sharing their insights including Bryan Bunning, Nando de Freitas, Andrija Milicevic, Hoifung Poon, David Rhew, Karén Simonyan, Eric Topol, and Jim Weinstein. Gianluca Fontana and Kevin

Hawkins (Prova Health) provided support on the health economics and outcomes section.

View Publication

Q&A

Is this AI safe to use for healthcare?

The work presented here is not yet approved for clinical use and would only be approved after rigorous safety testing, clinical validation, and regulatory reviews. For now, this represents exciting initial research. At the heart of any plans to deploy this technology in the real world is our commitment to safety, trust, and quality ensuring that any healthcare solutions are clinically grounded, ethically designed, and transparently communicated.

Will AI replace doctors?

While AI is becoming a powerful tool in healthcare, our team of practicing clinicians believes AI represents a complement to doctors and other health professionals. While this technology is advancing rapidly, their clinical roles are much broader than simply making a diagnosis. They need to navigate ambiguity and build trust with patients and their families in a way that AI isn't set up to do. Clinical roles will, we believe, evolve with AI giving

clinicians the ability to automate routine tasks, identify diseases earlier, personalize treatment plans, and potentially prevent some diseases altogether. For consumers, they will provide better tools for self-management and shared decision making.

What is an AI orchestrator?

In the context of generative AI, an orchestrator is like a digital conductor helping to coordinate multiple steps in achieving a complex task. In healthcare, the role of orchestration is crucial given the high stakes of each decision. Our orchestrator sits above underlying language models making sure each point in getting a diagnosis is handled systematically, reducing the risk in future of errors and offering the necessary stability, consistency and transparency to ultimately build trust from users.

Why have you looked at costs?

We initially wanted to understand whether the AI was simply requesting excessive diagnostic workups to reach the right diagnosis. What we found was that our Orchestrator was able to reach the correct answer with much less money spent on testing. In some ways this is not a surprise as diagnostic over-testing is recognized as being a widespread challenge, accounting for millions of unnecessary tests annually in the US. This work suggests AI creates an opportunity for clinicians – and consumers – to reach a faster, more accurate diagnosis while reducing costs.

Build the Future With Us

We're a lean, fast-moving lab made up of some of the world's most talented minds. We have an exciting roadmap of compute at MAI, with our next-generation GB200 cluster now operational. And we have an ambitious mission we truly believe in. We're also fortunate to partner with incredible product teams giving our models the chance to reach billions of users and create immense positive impact. If you're a brilliant, highly-ambitious and low ego individual, you'll fit right in—come and join us as we work on our next generation of models!

Explore all jobs

Related Stories

An AI companion for everyone

copilot

AI companions will change our lives

time100-voices

Two in-house models in support of our mission

models