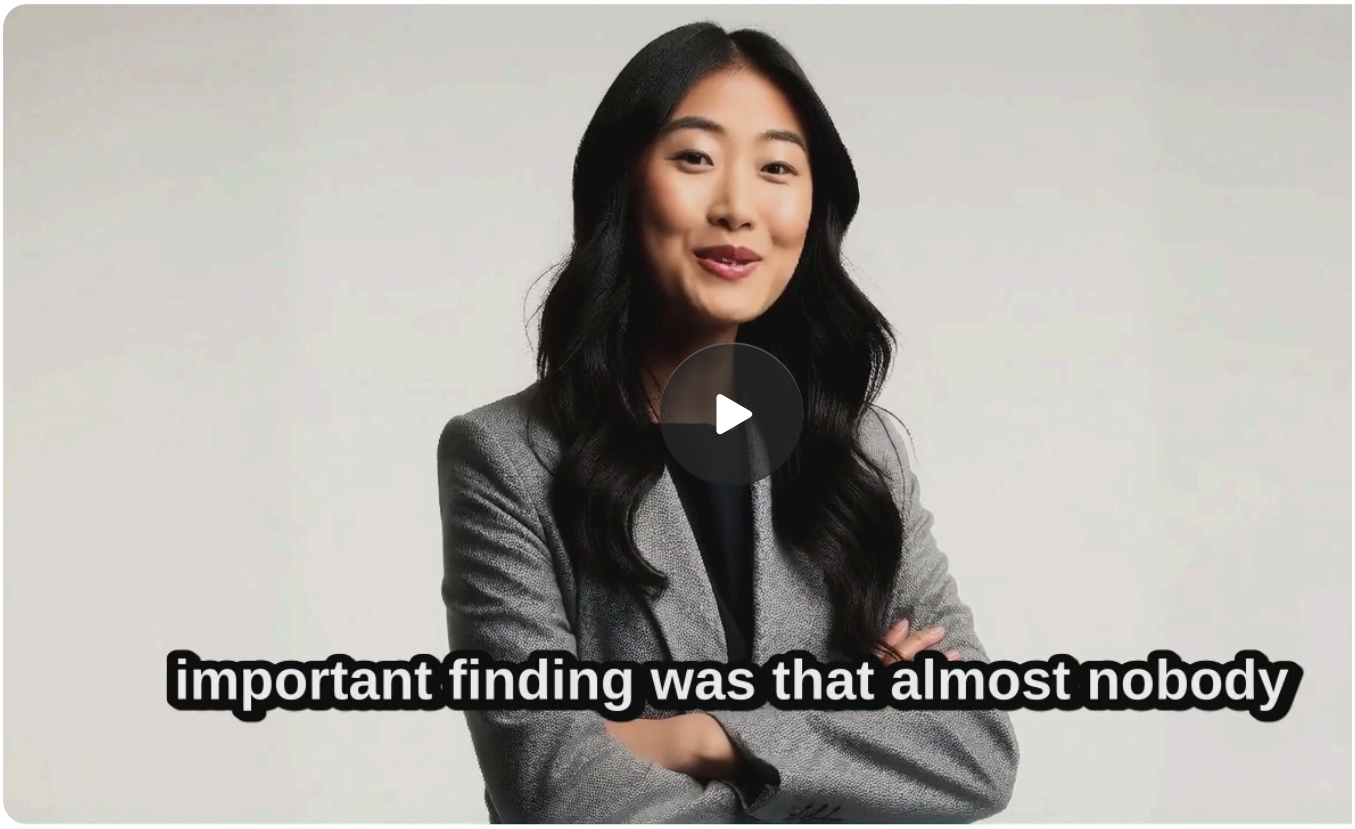
CLINICAL AI & PATIENT CARE

What the Harvard ER Study Says Abo o1 Beating Doctors at Diagnosis, Why Means Differential Diagnosis Just Stopped Being a Scarce Cognitive Asset, and Where the Money Goes Ne

MAY 05, 2026 • PAID

2

Video Preview



🎧 Listen on Spotify, Apple Podcasts, or the Web Player

0:00 -

Full episode for paid subscribers available below paywall.

Abstract

A Harvard-led team published in Science evaluating OpenAI's o1 reasoning mod against board-certified physicians on real cases pulled from a Boston emergency department. The headline numbers got picked up by Vox, the Guardian, the SF Chronicle, STAT, and roughly every healthcare Twitter account that has ever qu tweeted Eric Topol. The popular takeaway is "AI beat doctors at diagnosis." The popular takeaway is wrong, or at least incomplete enough to be useless for anyor making product, policy, or capital allocation decisions.

What follows is a closer read of what the study actually showed, what the most overlooked finding really implies for clinical workflow design, and why the inves thesis here has very little to do with "AI diagnostic startups."

Quick coverage map:

- 76-patient Boston ED dataset, multi-stage evaluation from triage through admission
- ~67% diagnostic accuracy for o1 at triage vs ~50-55% for physicians, ~80%+ v full context
- The buried lede: physicians using AI did not outperform AI alone
- Three-layer physician value stack and which layer just collapsed
- Triage inversion thesis and why front-door medicine is the highest-leverage deployment
- Search vs reasoning reframe, and what it means for product design
- The liability bottleneck nobody is pricing in yet
- Why the obvious "AI diagnostics" bet is the wrong layer to invest at

Table of Contents

1. Video Preview

2. Podcast, Part I (Free)

3. The setup and what the tweet actually means

4. The numbers, and the caveats most takes skip

5. Why physicians plus AI did not beat AI alone

6. The three-layer model of physician value

7. Diagnosis as infrastructure

8. The triage inversion thesis

9. Search vs reasoning, the real reframe

10. The AI-first differential workflow

11. Liability as the next bottleneck

12. Where the dollars actually flow

13. The punchline

14. The setup and what the tweet actually means

The study tested o1, OpenAI's reasoning-class model, against attending and resi
physicians on 76 cases drawn from the emergency department at a major Boston
academic medical center. Cases were structured into stages mirroring the actual
workflow, starting with triage-level information (chief complaint, vitals, brief hi
and progressing through ED workup and ultimately admission. At each stage, bo
the model and the physician were asked to produce a differential diagnosis and r
candidates. Physicians were blinded to the model's outputs in the comparison ar

The methodological detail that matters: this is not a full clinical interaction. The
no patient sitting in front of either the human or the model. There is no live ima

review, no live labs interpretation in the sense of looking at the raw values cycle cycle, no nursing handoff, no family history clarifying conversation in real time. inputs are text-based vignettes constructed from chart data, which is closer to a review exercise than to actual clinical care. Anyone who has worked in or aroun operations knows that the case on paper looks very different from the case in the room.

The Science Magazine writeup framed this as evidence that o1 reasons clinically above physician level. The Science Media Centre piece, which is the most useful secondary read in the entire ecosystem here, pushed back hard on that framing. STAT. So did several of the senior clinical informaticists who weighed in on Blu and LinkedIn after the paper dropped. The pushback is correct on the merits bu not actually defang what the study showed. What the study showed is still impo It is just important for reasons different from the ones being tweeted about.

The numbers, and the caveats most takes skip

At the triage stage, where the model and the physicians were given only the bare initial intake information, o1 landed correct primary diagnosis around sixty-seve percent of the time. Physicians, both attendings and residents, were closer to fif fifty-five percent. As progressively more context got added through the workup both groups improved, with the model and senior physicians converging in the e percent plus range by the time full ED workup data was on the table.

A few things to immediately notice. First, the gap is biggest where the informati sparsest. That is counterintuitive only if the prior assumption is that physicians some special intuition that gets activated at the bedside. The data suggests some closer to the opposite: physicians underperform exactly where breadth of recall most useful, because at triage the cognitive task is generating a wide net of plau diagnoses, and humans are systematically narrow-net generators. There is a deep clinical reasoning literature on premature closure, the failure mode where the diagnosing physician latches onto an early hypothesis and stops searching the sp

Singh and colleagues have published repeatedly on this; their estimates put the population-level diagnostic error rate somewhere around twelve million Americ adults per year experiencing some form of misdiagnosis, with downstream morta contributions in the forty to eighty thousand range depending on which method you trust. The Boston study is, in effect, a controlled demonstration of why thos numbers exist. Doctors are not generating wide enough differentials at the front

Second, the convergence at later stages is real and underdiscussed. Once the cha includes labs, imaging reads, and a focused history, the physician advantage at p integration largely catches up. The model does not lose ground; the human gains the right read is not "AI is better than doctors" but "AI is dramatically better th doctors at the early-information end of the workflow, and roughly comparable o enough data is on the table." That is a much more interesting finding than the tv version.

Third, this is text in, text out. There is no imaging interpretation in this study. T is no real-time signal handling. There is no conversational diagnosis where the n would have to decide what to ask next. So the result generalizes to chart-mediate reasoning, not to autonomous clinical care. STAT was right to flag this. The Scie Media Centre was right to flag it. Anyone using this paper to argue that physicia obsolete in any near-term sense is either trolling or has not read the methods sec

Why physicians plus AI did not beat AI alone

The single most important finding in the paper, and the one that has been least discussed publicly, is that the human-plus-AI condition did not outperform AI a This deserves to be sat with for a minute. Every health tech pitch deck for the pa three years has assumed an additive model: physician judgment plus machine br equals better than either alone. That is the entire premise of "copilot" framing. T is the premise of the FDA's evolving thinking on clinical decision support. That premise of the augmented intelligence positioning the AMA has been pushing.

What the data showed is that when physicians had access to the model's outputs did not pick up the additional accuracy the model would have delivered on its ov The most plausible mechanism is automation bias combined with anchoring. Wh physicians saw the model's top-ranked diagnoses, they used those as a starting p and pruned, but they also prematurely accepted incorrect rankings and failed to model errors as often as one would hope. They sometimes also discounted correc suggestions when those suggestions conflicted with the human's prior. The net e was a wash, or close enough to a wash that the additive collaboration thesis is nc empirically on the back foot.

This finding is not unique to this paper. Radiology has been wrestling with it for better part of a decade, and there is a meaningful literature on how radiologists CAD assistance often perform comparably to or marginally worse than radiologi without it, depending on prevalence, signal quality, and trust calibration. The Be ED finding is the same shape of result but in cognitive diagnosis rather than ima read. It suggests that "human in the loop" is not automatically a quality multipli and may, under certain configurations, be a quality reducer relative to AI alone. is a deeply uncomfortable conclusion for anyone whose product strategy assume physician seat is the one paying the bill.

The three-layer model of physician valu

To make sense of where this lands economically, it helps to decompose what physicians actually get paid for. The traditional value stack has three layers, rou The first layer is diagnosis: figuring out what is wrong. The second is decision-making: choosing the appropriate treatment, weighing risk, integrating patient preference and goals. The third is execution: actually performing care, whether t procedural skill, longitudinal management, or care coordination.

Layer one is what the study just commoditized. Or, more precisely, the study adc another data point to a trend line that has been visible since GPT-4 cleared roug ninety percent on USMLE Step questions in 2023. The benchmarks have been signaling for two years that knowledge recall and pattern matching across a wid

medical literature is no longer a scarce skill. The Boston study moved the goalpo from "can pass a board exam" to "can outperform attendings on real cases at the uncertainty end of the workflow." That is a meaningful jump in evidentiary strer

Layer two and layer three are still firmly in human hands. Decision-making in th of uncertainty, particularly when patient values, social context, payer constraints risk tolerance are all in play, is not something current LLMs do credibly. They ca summarize options. They cannot sit with a sixty-eight-year-old recently widowed patient and decide whether aggressive chemo is the right call. Execution, especia procedural execution and longitudinal relationship work, is even further from b touched.

So the right read on the study is not "doctors are obsolete." The right read is "th diagnostic step, which has historically been a meaningful chunk of why physicia paid what they are paid, just got pulled out of the value stack and turned into a r zero marginal cost utility." Whatever fraction of physician comp is implicitly compensating for layer one is now exposed to repricing. Estimating that fraction hard, but if you look at how E&M coding levels work, with diagnostic complexity being a primary driver of code level and therefore reimbursement, it is not small 2021 office visit E&M revisions explicitly anchored level selection on medical decision-making complexity, of which differential diagnosis is a core input. Whe marginal cost of generating a high-quality differential goes to roughly zero, the question of why the physician is being paid at level five becomes more pointed.

Diagnosis as infrastructure

The cleaner mental model for what is happening is that diagnosis is shifting fron being an artisanal individual skill to being a piece of infrastructure. This is not a pattern. It happened to security analysis when Bloomberg terminals went mainst in the 1980s. Before Bloomberg, knowing the price of a corporate bond was a ski and the analysts who could remember the spreads got paid for it. After Bloombe that knowledge was a utility, and analyst comp moved up the stack to interpreta modeling, and origination. It happened to legal research when Westlaw and

LexisNexis got their full text databases working. Before Westlaw, finding the rel case was a skill, and junior associates billed hours for it. After, finding the case v utility, and the human work moved to argument construction and judgment. It happened to radiology, partially, with PACS and the early generation of CAD toc The radiologist's value moved from "find the lesion" toward "integrate the findi into clinical context."

Diagnosis is now in the early phase of the same shift. The right way to think abo or whatever next-generation model gets deployed at the bedside, is as the medic equivalent of a Bloomberg terminal for differential diagnosis. The value questio becomes: who owns the terminal, and who controls distribution into the workflc Not: who is the best diagnostician.

If that framing is correct, then the strategic moves that matter are the ones that determine the distribution layer. The model itself becomes a commodity over a l enough horizon. The training data, the integration into the EHR, the specific UX the order entry flow, the audit trail, and the contractual relationship with the he system are where defensibility lives. This is consistent with how Epic and Oracle Health have been positioning, and with why Microsoft's twenty-billion-dollar Nu acquisition has aged considerably better than skeptics predicted at the time. Distribution into the moment of care is the moat. Model quality is table stakes.

← Previous