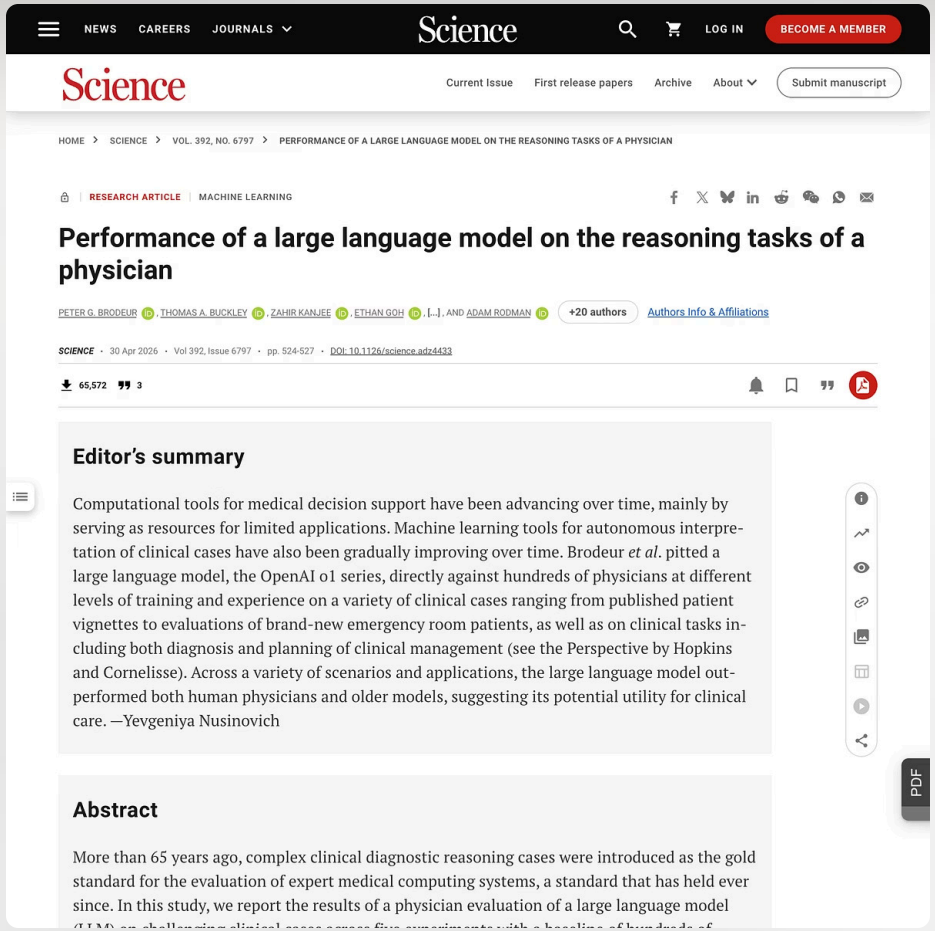


Performance of an LLM on the Reasoning Tasks of a Physician

Brodeur PG, Buckley TA, et al. *Science* 392, 524 (2026). DOI: 10.1126/science.adz4433

Background & Motivation

Historical Context

Case-based diagnostic vignettes – especially the NEJM clinicopathological conferences (CPCs) – have served as the standard benchmark for evaluating computer-aided diagnostic systems since Ledley & Lusted (1959). Every generation of differential-diagnosis tool has been tested against them. Large language models have recently surpassed older systems on these cases.

Goals

Recent LLM-in-medicine studies have largely relied on narrow or curated tasks. More critically, they have frequently **lacked human physician baselines**. With rapid LLM capability gains and growing concern about benchmark saturation, human baselines and clinically grounded tasks have become essential for meaningful evaluation.

Study Aim

Evaluate the diagnostic and management reasoning of an advanced LLM (OpenAI o1 series) across **six experiments** against baselines from hundreds of physicians, including a blinded AI-versus-expert second-opinion arm using randomly selected real emergency room patients.

Methods at a Glance

Models

o1-preview (o1-preview-2024-09-12, Oct 2023 knowledge cutoff) for the five case-based experiments (run September 2024; Grey Matters run August 2025). **Production o1 + GPT-4o** used for the ER study (run February 2025).

Comparators

GPT-4 and physician baselines are **historical controls** from prior independent studies in every experiment **except the ER study**, where models and physician second opinions were collected together in a contemporaneous design.

Analysis & Reporting

Statistical analysis in **R 4.4.2**. The ER arm was IRB-approved (BIDMC 2024P000525) with fully deidentified patient data.

EXPERIMENT OVERVIEW

Six Experiments: Study Architecture

Clinicopathological (CPC) Cases

DDx & Next Best Test

NEJM Cases

Grey Matters

Landmark Dx

Probabilistic Reasoning

ER Second Opinions

The study organizes its work as six experiments: five case-based tasks (CPC differential diagnosis and next test selection, NEJM Healer cases, Grey Matters management reasoning, Landmark diagnostic cases, and diagnostic probabilistic reasoning) plus one real-world ER second-opinion study. Only the ER arm collected human and model data contemporaneously – all other physician baselines are historical controls from prior publications.

EXPERIMENT 1

CPC – Differential Diagnosis Quality

Source

143 NEJM CPCs (2021-September 2024). Comparators are historical – GPT-4 (70 cases, Kanjee 2023) and physicians (302 responses with search, 101-case overlap, McDuff/AMIE 2025).

Task

Produce the most-likely diagnosis plus a ranked differential.

Scoring

Bond score – A 0-5 ordinal rating of how well a differential captures the true diagnosis. **5** = exact diagnosis present; **4** = very close but not exact; **3** = closely related, possibly helpful; **2** = related but unlikely helpful; **0** = nothing close (no score of 1). Scored by 2 attending raters (κ = 0.66). Ground truth = confirmed CPC diagnosis.

Results

Metric	o1-preview	Comparator
Correct Dx in Differential	78.3% (95% CI: 70.7-84.8%)	GPT-4: 72.9% (70-case overlap, P=0.015); Physicians top-1: 24.3%, top-10: 44.5%
Leading Dx Correct	52% (95% CI: 44-61%)	Physicians top-1: 24.3%
Exact or Close (Score 4-5)	97.9% (95% CI: 94.0-99.6%)	–

Comparisons

vs GPT-4 (70-case overlap): o1-preview scored 88.6% vs GPT-4's 72.9% – favoring o1 (P = 0.015).

vs Physicians (historical baseline): o1-preview's top-1 accuracy (66.3%) more than doubled the physician top-1 rate (24.3%), and its top-10 recall (81.2%) nearly doubled the physician top-10 rate (44.5%) published in McDuff/AMIE 2025

Memorization check: Pre- vs post-Oct 2023 cutoff performance was 79.8% vs 73.5% (P = 0.59) – not significant. However, this test had only ~11.5% statistical power to detect the 6.3-point gap it observed. The non-significant result therefore cannot rule out that high CPC scores partly reflect recall of training data rather than reasoning.

EXPERIMENT 2

CPC – Next Diagnostic Test Selection

Source

A **136-case CPC subset** (130 scored; 7 deemed not applicable).

Task

Name the next diagnostic test(s) to order. **01-preview only** – no comparator arm for this experiment.

Scoring – Test-Plan Likert

A 0-2 scale comparing proposed next test(s) to the actual workup plan: **2** = essentially the same test(s); **1** = different but would have been helpful or reached the diagnosis via an alternative route; **0** = unhelpful. Scored by 2 raters (113/130, 87%, but κ = 0.26). Ground truth = the actual test plan documented in the CPC case.

Results

Score	Interpretation	% of Cases
2	Correct / Identical	87.5%
1	Helpful	11%
0	Unhelpful	1.5%

The authors conclude the model produced a correct or helpful next test in approximately **98.5% of cases**.

No physician comparator arm exists for this experiment. The 87.5% figure reflects agreement with the documented CPC workup plan, not independent expert validation. Inter-rater reliability was low (κ = 0.26).

EXPERIMENT 3

NEJM Healer Cases – Reasoning Documentation

Source

20 Healer cases × 4 sections = 80 responses per arm (o1-preview run by authors; comparators historical from Cabral 2024).

Task

Per section, respondents provided: a problem representation, a prioritized differential with justification (primary task), and cannot-miss diagnoses at triage (secondary task).

Scoring

A validated 10-point clinical reasoning *documentation* score: **Interpretive Summary** (0–4: one point each for naming key risk factors, chief complaint, illness time course, and using semantic qualifiers) + explicit prioritized **Differential** (0–2) + **Lead-Dx explanation** with objective data (0–2) + **Alternative-Dx explanation** (0–2). Cannot-miss = fraction of a pre-specified list of dangerous diagnoses appearing in the triage differential. Two raters (κ = 0.89 for R-IDEA).

Results

Arm	Perfect Scores	P vs o1-preview
o1-preview	78 / 80	–
GPT-4	47 / 80	<0.0001
Attendings	28 / 80	<0.0001
Residents	16 / 72	<0.0001

Cannot-Miss Diagnoses

o1-preview median **0.92** (IQR 0.62–1.0). Difference vs all comparator groups: **not significant**.

Cannot-miss result was not statistically significant vs any comparator group.

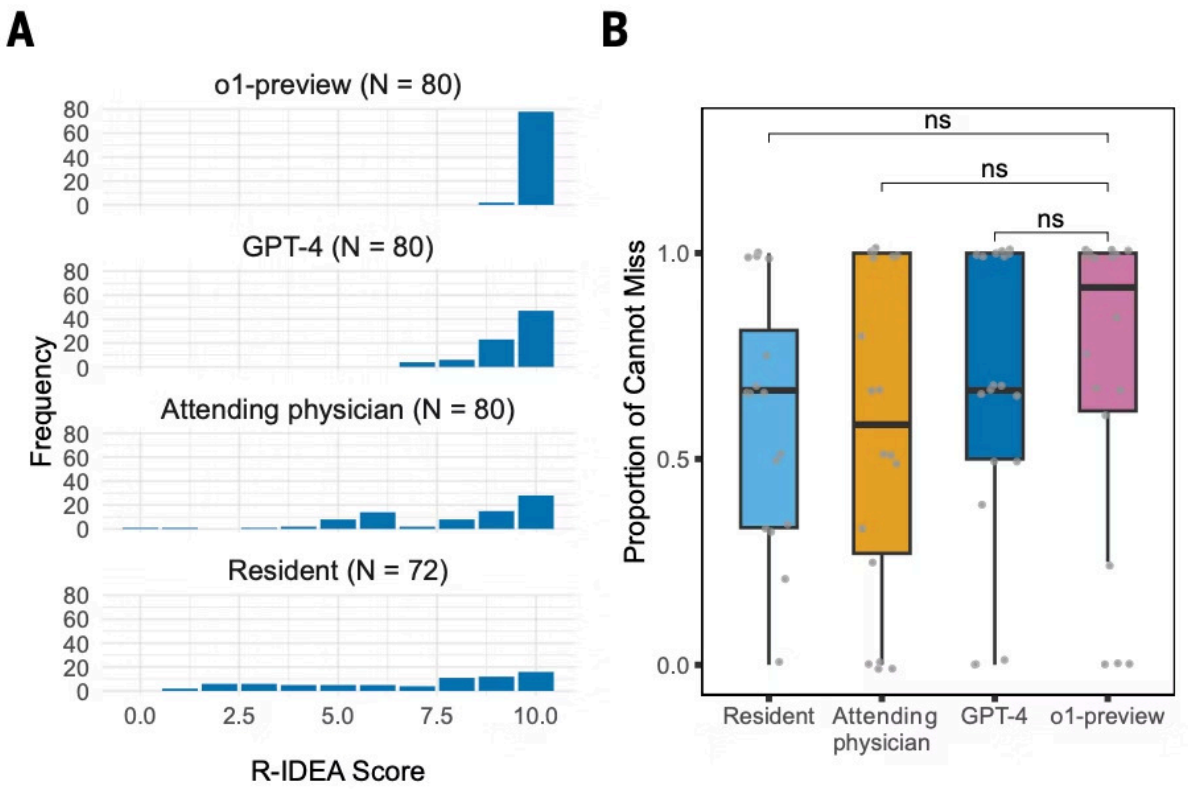


Fig. 3. Comparison of o1-preview, GPT-4, and physicians for clinical diagnostic reasoning. (**A**) Distribution of 312 R-IDEA scores stratified by respondents on 20 *NEJM* Healer cases. (**B**) Box plot of the proportion of cannot-miss diagnoses included in differential diagnosis for the initial triage presentation. The total sample size in this figure is 70, with 18 responses from attending physicians, GPT-4, and o1-preview, and 16 responses from residents. Two cases were excluded because the cannot-miss diagnoses could not be identified. ns, not significant.

EXPERIMENT 4

Grey Matters – Management Reasoning

Source

5 ACP "Core IM" management vignettes (Goh 2025). Comparators historical.

Task

o1-preview run by authors with 3 runs per case (August 2025). Answer next-step management questions per vignette.

Scoring

Expert-built, case-specific rubrics awarding points for each recommended action named (specific tests, screenings, management steps), summed and normalized to **0–100%**. Scoring is additive over a menu of acceptable items. Two raters (κ = 0.71).

ⓘ This is an uncapped additive rubric – enumerating more appropriate items yields a higher score.

Results

Arm	Median	IQR	P-value
o1-preview	89%	79 - 91%	–
GPT-4	42%	33 - 52%	<0.001
Physicians + GPT-4	41%	31 - 54%	<0.001
Physicians + resources	34%	23 - 48%	<0.001

EXPERIMENT 5

Landmark Diagnostic Cases

Source

6 vignettes (Berner 1994, accessed via Goh 2024); cases were never publicly released. Comparators historical (original study: 50 physicians).

Task

o1-preview run by authors (1 run per case). Task: provide top-3 differentials with supporting and opposing factors, a final diagnosis, and up to 3 next steps.

Scoring

Points awarded across: initial diagnoses (1 pt each, up to 3), supporting factors (0 - 2), opposing factors (0 - 2), final diagnosis (1 - 2), plus up to three next steps (0 - 2 each), normalized to 0 - 100%. Two raters (κ = 0.42).

Results

Arm	Median	IQR	P vs o1-preview
o1-preview	97%	95 - 100%	–
GPT-4	92%	82 - 97%	P = 0.70, **ns**
Physicians + GPT-4	76%	66 - 87%	P = 0.076, **ns**
Physicians + resources	74%	63 - 84%	P = 0.055, **ns**

o1-preview is numerically highest but statistically comparable to all comparators. The o1-vs-GPT-4 confidence interval spans zero in both directions (−19 to +27.7).

EXPERIMENT 6

Diagnostic Probabilistic Reasoning

Source

Primary-care vignettes from Morgan 2021. Main text describes five cases; Table S6 and Fig S2 present four (pneumonia, breast cancer, cardiac ischemia, UTI). Comparators are historical: GPT-4 (100 runs per question) and clinicians (n = 553: 290 residents, 202 attendings, 61 NP/PA).

Task

Every arm estimated disease probability before a test, after a positive test, and after a negative test – **12 estimates in all**. o1-preview was run 100 times per question.

Scoring

Error is summarized as **MAE** (mean absolute error, in percentage points) and **MAPE** (mean absolute percentage error). MAE is the interpretable metric; MAPE inflates sharply at low reference values and is unstable.

What the Numbers Represent

Each **model value** is the median of 100 runs (IQR = spread across runs). Each **clinician value** is the median across 553 practitioners, one estimate each. A median is the middle of a distribution, not the most frequent answer.

Results

Clinicians were the least calibrated. Their median sat above the reference in all 12 cells, with the largest errors after a positive test and for low-prevalence conditions: breast cancer post-positive (median 50% vs ref 3 - 9%), pneumonia post-negative (50% vs 10 - 19%), cardiac ischemia post-positive (70% vs 2 - 11%).

Both LLMs were closer to the reference. o1-preview had the lowest error in 9 of 12 cells; GPT-4 in 2. Exception: UTI post-positive (ref 0 - 8.3%), where all three groups estimated ~80 - 90% and clinicians were least far off. The one large model-to-model gap is cardiac ischemia post-positive:

Source	Median Estimate (IQR)	MAE	MAPE
o1-preview	10.9% (6 - 13.5%)	5.7	87.1
GPT-4	68.7% (65 - 70%)	56.5	869.1
Clinicians	70% (50 - 90%)	56.3	865.9

(Reference range 2 - 11%.)

Density plots of all estimates for the four conditions at the three decision points, with the literature reference range shaded. Clinician estimates are the most dispersed; model estimates cluster more tightly.

Notes: This experiment is descriptive, with no inferential p-values. Model values are medians of 100 runs; each clinician value is a single estimate from a different practitioner, so part of the greater clinician variability follows from how the comparison is constructed.

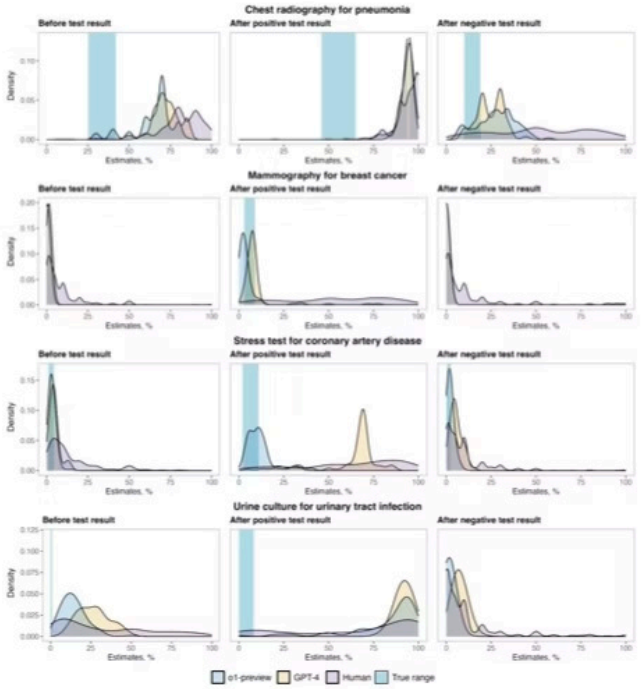


Fig. S2. Probabilistic reasoning by LLMs and humans. Shown are density plots of the distribution of responses by o1-preview, GPT-4, and humans to clinical vignettes asking for (1) the pretest probability of disease, (2) the updated probability after a positive test result, and (3) the updated probability after a negative test result. The shaded blue indicates the reference range based on a review of literature from a prior study (*21*). Human responses are from 553 medical practitioners (290 resident physicians, 202 attending physicians, and 61 nurse practitioners or physician assistants). 100 predictions were generated by GPT-4 and o1-preview for each question.

EXPERIMENT 7

Blinded ER Second Opinions

Source

80 BIDMC ED cases (2-week random sample) → 76 analyzed. Real EHR data. Three touchpoints: triage, ER physician assessment, and admission. Models and physicians were collected **together** (the only contemporaneous arm in the study). Arms: o1 vs GPT-4o vs Physician 1 vs Physician 2. 237 differentials per source; 912 total responses. Graded by 2 blinded attending raters. Blinding was successful: raters guessed source correctly 15.2% / 3.1% of the time.

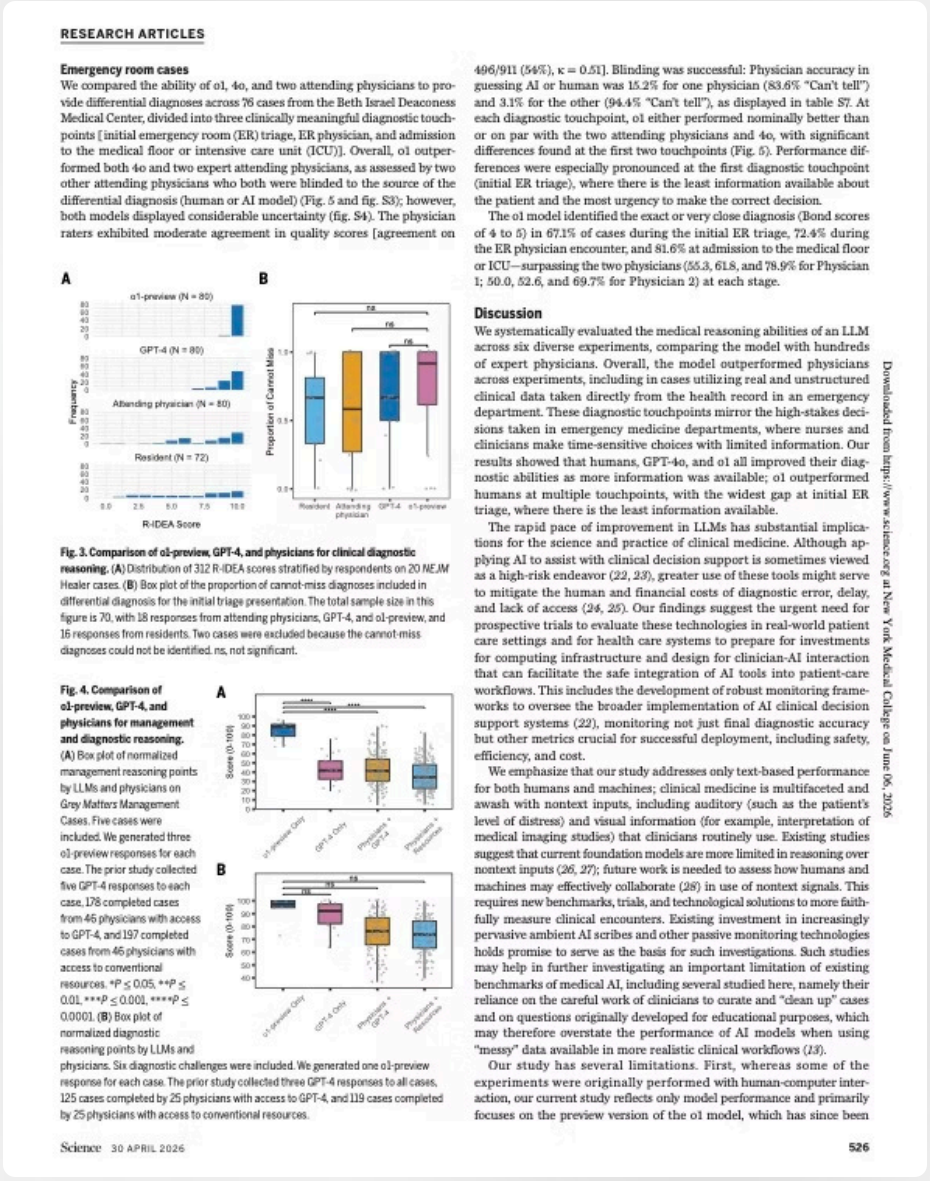
Task

Provide a prioritized second-opinion differential at each touchpoint, capped at **top 5 diagnoses**.

Results

Source	Triage	ER Physician	Admission	P vs o1
o1	67.1%	72.4%	81.6%	–
Physician 1	55.3%	61.8%	78.9%	P ⩽ 0.05 at triage and ER-physician touchpoints; ns at admission
Physician 2	50.0%	52.6%	69.7%	P ⩽ 0.05 at triage and ER-physician touchpoints; ns at admission

Differences were statistically significant at triage and ER-physician touchpoints (P ⩽ 0.05); admission comparison was not significant. o1 and GPT-4o received **different prompts**. Both models showed considerable uncertainty. All sources – human and AI – improved accuracy with more clinical information at each successive touchpoint.



RESEARCH ARTICLES

Emergency room cases

We compared the ability of o1, 4o, and two attending physicians to provide differential diagnoses across 76 cases from the Beth Israel Deaconess Medical Center, divided into three clinically meaningful diagnostic touchpoints [initial emergency room (ER) triage, ER physician, and admission to the medical floor or intensive care unit (ICU)]. Overall, o1 outperformed both 4o and two expert attending physicians, as assessed by two other attending physicians who both were blinded to the source of the differential diagnosis (human or AI model) (Fig. 5 and fig. S3); however, both models displayed considerable uncertainty (fig. S4). The physician raters exhibited moderate agreement in quality scores [agreement on 496/911 (54%), $\kappa = 0.51$]. Blinding was successful: Physician accuracy in guessing AI or human was 15.2% for one physician (83.6% "Can't tell") and 3.1% for the other (94.4% "Can't tell"), as displayed in table S7. At each diagnostic touchpoint, o1 either performed nominally better than or on par with the two attending physicians and 4o, with significant differences found at the first two touchpoints (Fig. 5). Performance differences were especially pronounced at the first diagnostic touchpoint (initial ER triage), where there is the least information available about the patient and the most urgency to make the correct decision.

The o1 model identified the exact or very close diagnosis (Bond scores of 4 to 5) in 67.1% of cases during the initial ER triage, 72.4% during the ER physician encounter, and 81.6% at admission to the medical floor or ICU—surpassing the two physicians (55.3, 61.8, and 78.9% for Physician 1; 50.0, 52.6, and 69.7% for Physician 2) at each stage.

Discussion

We systematically evaluated the medical reasoning abilities of an LLM across six diverse experiments, comparing the model with hundreds of expert physicians. Overall, the model outperformed physicians across experiments, including in cases utilizing real and unstructured clinical data taken directly from the health record in an emergency department. These diagnostic touchpoints mirror the high-stakes decisions taken in emergency medicine departments, where nurses and clinicians make time-sensitive choices with limited information. Our results showed that humans, GPT-4o, and o1 all improved their diagnostic abilities as more information was available; o1 outperformed humans at multiple touchpoints, with the widest gap at initial ER triage, where there is the least information available.

The rapid pace of improvement in LLMs has substantial implications for the science and practice of clinical medicine. Although applying AI to assist with clinical decision support is sometimes viewed as a high-risk endeavor (*22*, *23*), greater use of these tools might serve to mitigate the human and financial costs of diagnostic error, delay, and lack of access (*24*, *25*). Our findings suggest the urgent need for prospective trials to evaluate these technologies in real-world patient care settings and for health care systems to prepare for investments for computing infrastructure and design for clinician-AI interaction that can facilitate the safe integration of AI tools into patient-care workflows. This includes the development of robust monitoring frameworks to oversee the broader implementation of AI clinical decision support systems (*22*), monitoring not just final diagnostic accuracy but other metrics crucial for successful deployment, including safety, efficiency, and cost.

We emphasize that our study addresses only text-based performance for both humans and machines; clinical medicine is multifaceted and awash with nontext inputs, including auditory (such as the patient's level of distress) and visual information (for example, interpretation of medical imaging studies) that clinicians routinely use. Existing studies suggest that current foundation models are more limited in reasoning over nontext inputs (*26*, *27*); future work is needed to assess how humans and machines may effectively collaborate (*28*) in use of nontext signals. This requires new benchmarks, trials, and technological solutions to more faithfully measure clinical encounters. Existing investment in increasingly pervasive ambient AI scribes and other passive monitoring technologies holds promise to serve as the basis for such investigations. Such studies may help in further investigating an important limitation of existing benchmarks of medical AI, including several studied here, namely their reliance on the careful work of clinicians to curate and "clean up" cases and on questions originally developed for educational purposes, which may therefore overstate the performance of AI models when using "messy" data available in more realistic clinical workflows (*13*).

Our study has several limitations. First, whereas some of the experiments were originally performed with human-computer interaction, our current study reflects only model performance and primarily focuses on the preview version of the o1 model, which has since been

Fig. 3. Comparison of o1-preview, GPT-4, and physicians for clinical diagnostic reasoning. (**A**) Distribution of 312 R-IDEA scores stratified by respondents on 20 *NEJM* Healer cases. (**B**) Box plot of the proportion of cannot-miss diagnoses included in differential diagnosis for the initial triage presentation. The total sample size in this figure is 70, with 18 responses from attending physicians, GPT-4, and o1-preview, and 16 responses from residents. Two cases were excluded because the cannot-miss diagnoses could not be identified. ns, not significant.

Fig. 4. Comparison of o1-preview, GPT-4, and physicians for management and diagnostic reasoning. (**A**) Box plot of normalized management reasoning points by LLMs and physicians on Grey Matters Management Cases. Five cases were included. We generated three o1-preview responses for each case. The prior study collected five GPT-4 responses to each case. 178 completed cases from 46 physicians with access to GPT-4, and 197 completed cases from 46 physicians with access to conventional resources. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$. (**B**) Box plot of normalized diagnostic reasoning points by LLMs and physicians. Six diagnostic challenges were included. We generated one o1-preview response for each case. The prior study collected three GPT-4 responses to all cases. 125 cases completed by 25 physicians with access to GPT-4, and 119 cases completed by 25 physicians with access to conventional resources.

Authors' Overall Conclusions

Outperformance of Physician Baselines

The model outperformed physician baselines across experiments, including on real unstructured ER data. Humans, GPT-4o, and o1 all improved with increasing clinical information, and o1's advantage was widest at initial ER triage – the point of least available information.

Substantial LLM Capability Across Reasoning Domains

LLMs now demonstrate substantial performance in differential diagnosis, diagnostic reasoning, and management reasoning, exceeding prior model generations and human clinicians across multiple domains tested.

Benchmarks Eclipsed, Call for Prospective Trials

The authors conclude that LLMs have eclipsed most benchmarks of clinical reasoning and call for human-computer interaction studies, prospective clinical trials, and monitoring frameworks addressing safety, efficiency, and cost.

Limitations Stated by the Authors

Model Scope & Currency

Reflects model performance alone (not human-AI interaction). Primarily the o1-preview model, which has since been supplanted by newer models (e.g., o3).

Breadth of Clinical Reasoning

Only six aspects of clinical reasoning were examined; many other relevant tasks exist. Limited to internal and emergency medicine; not representative of broader practice (e.g., surgery). Performance may vary by diagnosis, patient, or clinical setting.

Text-Only Inputs

All inputs were text-based. Clinical care also depends on non-text signals such as patient distress, physical examination findings, and imaging.

ER Study Design

The ER second-opinion task at predefined touchpoints is a proof of concept. Real ED decisions center on triage, disposition, and immediate management – not diagnostic accuracy alone.

Robustness of Improvements

Improvements over prior models were not always robust – notably for cannot-miss diagnoses and the landmark diagnostic cases, where statistical significance was not achieved.

CRITIQUE C1

Statistical Power – Three "No Difference" Findings Are Underpowered

A non-significant result from a small or noisy comparison means **"we could not detect a difference,"** not **"there is no difference."** Absence of evidence is not evidence of absence. When statistical power is low, a null result is uninformative – the test would fail to detect a real effect even if one existed.

Memorization Null

Pre- vs post-training-cutoff (Oct 2023): 79.8% (n=109) vs 73.5% (n=34), P=0.59. Estimated power to detect the observed 6.3-pt gap: **~11.5%**. Post-cutoff performance would need to fall below ~53.5% (a 26-pt drop) for 80% power. The observed gap leans in the direction of concern (pre-cutoff > post-cutoff).
Implication: Cannot rule out that headline CPC scores partly reflect recall of memorized cases rather than live reasoning.

Landmark Cases – o1 97% vs Physicians 74-76% (ns)

Minimum detectable effect at 80% power: **≈29-33 percentage points**. Observed gaps were +4.4, +18.6, and +20.2 pp. The o1-vs-GPT-4 CI (−19 to +27.7) spans zero in both directions.
Implication: The comparison is uninformative – we can conclude neither that o1 beats physicians nor that it matches them.

Cannot-Miss Diagnoses – Not Significantly Better (ns)

With n ≈ 18 per group and scores bunched near the 0.92 ceiling, the test is too weak to detect a difference even if one exists. **Implication:** The single most safety-critical comparison in the paper is inconclusive – providing no support for or against the model as a safety net for diagnoses that kill.

For contrast: the significant wins (Grey Matters, R-IDEA, CPC vs GPT-4) have effect sizes large enough that significance is secure within-study (Grey Matters z ≈ 6.5). The caveats there are scoring validity and generalization from 5-6 cases, not power.

CRITIQUE C1 (CONTINUED)

How Grading Scales Can Be Biased Toward Verbosity

Three instruments in the paper reward **length and completeness** rather than correctness per se, which structurally favors verbose LLM output over terse human clinical documentation.

R-IDEA – Presence of Documentation Elements

R-IDEA scores whether a write-up *contains* four documentation domains (interpretive summary, differential, lead-dx explanation, alternative-dx explanation). A model instructed to "document your thinking" will emit all four elements by design. Terse human admission notes frequently omit explicit alternative-dx justification – the format for which R-IDEA was originally validated. Critically, the supplement specifies only the **model's** prompt; it does not state that the historical physician comparators (Cabral 2024) received equivalent instructions to hit all four domains.

Grey Matters – Additive Uncapped Menu

The Grey Matters rubric is additive over an uncapped menu of acceptable actions: enumerating more appropriate items yields a higher score regardless of whether any single item is uniquely insightful. A model that generates a comprehensive enumeration is structurally advantaged over a clinician who identifies the highest-yield action concisely.

CPC "Diagnosis Anywhere in the List" – Recall Over Unbounded Differential

The headline CPC recall metric credits any diagnosis appearing anywhere in an explicitly unbounded list. Recall rises mechanically with list length. This is addressed further in Critique C2.

CRITIQUE C2

Unbounded LLM Differentials May Inflate CPC Recall Metrics

The Core Issue

The CPC prompt explicitly instructs the model that **"there is no limit to the number of diagnoses on your differential."** Recall of the true diagnosis in an unbounded list rises mechanically with list length – the standalone headline numbers (78.3%, 97.9%) are recall-with-unlimited-guesses metrics, not precision-weighted accuracy measures.

Variation Across Experiments

This concern varies in severity across the study design:

- **Landmark cases** cap at top-3 diagnoses – list length is constrained.
- **ER arm** caps at top-5 for all sources – fair on length, contemporaneous design.
- **Table S2 human comparison** uses top-1 / top-10, which truncates list length and largely neutralizes the length advantage in the head-to-head.
- **Standalone CPC estimates** (78.3%, 97.9%) remain uncapped.

Interpretive Implication

The uncapped CPC point estimates are the most susceptible to length inflation. The head-to-head comparison with GPT-4 is model-vs-model under the same uncapped condition – and is therefore internally fair between those two arms, even if the absolute values are length-inflated for both.

The paper's overall pattern – **large performance gaps on verbosity-friendly scales, small or null gaps where lists are capped or the metric is symmetric** – is consistent with part of the apparent superiority being a scoring artifact rather than a pure reflection of reasoning quality.

No penalty is applied for over-calling on any recall metric (see Critique C3).

CRITIQUE C3

Missing Penalties and Other Methodological Biases

No Penalty for Over-Calling / False Positives

On every recall metric, a focused, correct human differential scores no better than a long LLM list that happens to contain the answer. Over-testing is rewarded, not penalized – the opposite of real-world clinical stewardship priorities.

Aggregated Model vs Individual Humans

Probabilistic estimates use the **median of 100 model runs** versus a single estimate per clinician. The conclusion that "clinicians are more variable" is partly definitional – individual model runs would show substantially more spread, just as individual clinician estimates do.

Prompt Asymmetry in the ER Arm

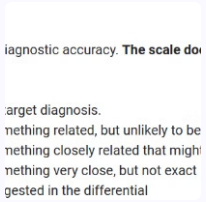
o1 received a brief second-opinion prompt; GPT-4o received an elaborate Tree-of-Thought system prompt. The o1-versus-GPT-4o contrast in the ER arm is therefore confounded by prompt design – the difference in performance cannot be cleanly attributed to model capability alone.

Scoring Rubric Reference Summary

Six distinct scoring instruments were used across the seven experiments. Understanding each rubric is essential to interpreting reported effect sizes and comparisons.

Rubric	What It Measures	Range / Scale	Used In
Bond Score	Presence of true dx in differential; how close the best answer is to ground truth	0–5 ordinal (no 1); 4–5 = exact/close	Exp. 1 (CPC), Exp. 7 (ER)
Test-Plan Likert	Match of proposed next test(s) to actual workup plan	0–2 (2=identical, 1=helpful, 0=unhelpful)	Exp. 2 (CPC next test)
R-IDEA (10-pt)	Presence of four clinical reasoning documentation elements in the write-up	0–10 (sum of four subdomains)	Exp. 3 (Healer)
Management Points	Fraction of expert-defined acceptable actions named; additive, uncapped	0–100% normalized	Exp. 4 (Grey Matters)
Diagnostic Points	Points for differential items, supporting/opposing factors, final dx, next steps	0–100% normalized	Exp. 5 (Landmark)
MAE / MAPE	Deviation of probability estimate from literature reference; symmetric penalty	Absolute / percentage error; lower = better	Exp. 6 (Probabilistic)

Rubric Source Documents


iagnostic accuracy. **The scale do**

arget diagnosis.
nething related, but unlikely to be
nething closely related that migh
nething very close, but not exact
gested in the differential

Bond Score (CPC Differential Diagnosis rubric, 0–5 scale, no score of 1)

I - Interpretive Summary

Provides a concise summary statement that uses semantic vocabulary to highlight the most important elements from history, exam, and testing and to interpret and represent the patient's main problem (s). The presence or absence of the following features is assessed:

a) Key risk factors
b) Chief complaint
c) Illness time course; and
d) Use of semantic qualifiers

SCORING

0 - No features present

Study Design Overview – Key Parameters at a Glance

Experiment	N (cases)	Models compared	Human baseline	Contemporaneous?	Key rubric
1 – CPC Differential	143	o1-preview; GPT-4 (historical)	302 physician responses, search-assisted (McDuff/AMIE 2025; 101-case overlap)	No	Bond 0-5
2 – CPC Next Test	136 (130 scored)	o1-preview only	None	N/A	Likert 0-2
3 – Healer	20 × 4 = 80	o1-preview; GPT-4 (historical)	Attendings + residents (Cabral 2024)	No	R-IDEA 0-10
4 – Grey Matters	5 vignettes	o1-preview; GPT-4 (historical)	Physicians, 46/arm: with GPT-4 and with conventional resources (Goh 2025)	No	Mgmt points 0-100%
5 – Landmark	6 vignettes	o1-preview; GPT-4 (historical)	Physicians, 25/arm: with GPT-4 and with conventional resources (Goh 2024); cases from Berner 1994	No	Diag points 0-100%
6 – Probabilistic	4-5 vignettes	o1-preview; GPT-4 (historical)	553 clinicians (Morgan 2021)	No	MAE / MAPE
7 – ER Second Opinion	76 ED cases	o1; GPT-4o (both run for this study)	2 internal-medicine physicians (BIDMC)	**Yes**	Bond 4-5 proportion

The ER arm (Experiment 7) is the only contemporaneous design – the only experiment where human and model data were collected under identical conditions at the same time. All other physician baselines are historical controls from prior independent publications.

TAKEAWAYS

Takeaways

Strengths

- Six experiments, each with a human physician comparator.
- A real-world emergency department component with three timed touchpoints and blinded adjudication, and the blinding was largely successful.
- Negative results were reported, including the cannot-miss and landmark nulls.
- The statistically significant results (management, R-IDEA, CPC vs GPT-4) prompt further evaluation in real world settings

Limitations

- The memorization check is underpowered and cannot rule out that the high CPC scores partly reflect recall of cases seen in training.
- The largest performance gaps occur on rubrics that reward length, structure, and completeness, and the recall metrics do not penalize over-calling.
- The non-significant findings (memorization, cannot-miss, landmark) come from underpowered comparisons and do not establish equivalence.
- Prompts were not matched in the emergency department arm, and most human baselines were historical rather than concurrent.
- The evaluated model is now deprecated.
- Differential-diagnosis quality is distinct from triage, which the authors note when they describe the emergency department task as a proof of concept.

JOURNAL CLUB TAKEAWAYS

Questions to Ask of Any Clinical-AI Evaluation

Generalized from the critique points above.

- Was the evaluation set held out from training, or could the benchmark be in the training data?
- For any claim of no difference, was the comparison adequately powered or tested for equivalence?
- Does the metric measure correctness, or does it reward form such as length, structure, or recall over an unbounded list, and is incorrect or excessive output penalized?
- Were conditions matched across arms, including prompts, output limits, and concurrent comparators?
- Does the task reflect the real clinical decision rather than a proxy?
- Does the study measure human-AI interaction or real-world outcomes, or only standalone model performance?

APPENDIX

Appendix A – NEJM Healer Prompt Template

The exact prompt presented to both the LLM and physician participants in Experiment 3 (NEJM Healer Cases). Participants responded to four sequential case sections, documenting their problem representation, differential diagnosis, and justification after each reveal.

APPENDIX

Appendix B – Grey Matters Scoring Rubrics (Questions 1-4)

Point-by-point scoring rubrics for the four Grey Matters management vignette questions used in Experiment 4. These rubrics define the maximum achievable scores and acceptable answer variants for each clinical management decision.

n 1: How would you evaluate for involuntary weight loss?

9 points)

C	1 point
tabolic Panel	1 point
/Urine Culture	1 point
Chest	1 point
Abdomen/Pelvis	1 point
ncer Biomarkers	1 point
C	½ point
rtisol Testing	½ point
d	1 point
Evaluation	1 point
/ Testing	1 point
V Testing	1 point
toimmune Serologic Testing	1 point
D	½ point
lonoscopy	½ point
o Smear	½ point
mmogram	½ point
mining Dentition	1 point
luating for Dysphagia	1 point
eening for food insecurity	1 point
eening for depression	1 point
eening for dementia	1 point

hat is your approach to evaluating her fevers?

)

d Cultures	1 point
entesis	1 point
lture	1 point
/iral Panel	1 point
ures (also acceptable: tracheal piratory culture, Bronch/BAL,	1 point max
	1 point
res	1 point
Abdomen/Pelvis, DVT Brain MRI	1 point each
ting	1 point each
arkers	1 point

Question 3: What do you do with her antibiotics and why?

(max **10** points)

Assuming no infectious source is identified after 48-72 hours and the patient is not improving, stop antibiotics (either all at once or sequentially)	10 points
Assuming no infectious source is identified after 48-72 hours, de-escalate to SBP prophylaxis	10 points
Assuming no infectious source is identified after 48-72 hours and the patient is not improving, stop vancomycin but continue meropenem	10 points
Continue empiric treatment with vanc/meropenem for 7 days	5 point
Assuming no infectious source is identified after 48-72 hours and the patient is not improving, change classes of antibiotics but maintain relatively broad-spectrum coverage	0 points
Continue empiric treatment with vanc/meropenem and add micafungin (or a similar agent)	0 points

not getting any better after two weeks in t
sures. What factors go into your decision ir

advanced care planning discussions or documer
proxy if patient wishes were unclear
s
e
ransplant candidate
functional recovery

A fifth question was also included:

Question 5: How do you explain this diagnosis to the family?

(up to **12** points)

Provide an apology AND Express Empathy	8 points max if includes both, 4 points if only includes one
Discuss the nature of drug-induced neurotoxicity, Point out how drug-induced neurotoxicity is relatively rare (and this likely led to a delay in diagnosis), Discuss that drug-induced neurotoxicity is presumed but not confirmed, Discussion that ESLD and/or renal dysfunction can increase the likelihood of drug toxicities	4 points if any of these points included

Question 5: How do you explain this diagnosis to the family? (up to 12 points)

APPENDIX

Appendix C – Landmark Diagnostic Cases Scoring Form

The complete scoring rubric for Experiment 5 (Landmark Diagnostic Cases). Points are awarded for correct diagnoses (up to 3), supporting factors (up to 2 each for 3 diagnoses = 6 total), opposing factors (up to 2 each for 3 diagnoses = 6 total), and final diagnostic decision (up to 2). Maximum possible score varies by case structure.

Diagnosis scoring: 0 or 1 point each (out of 1) for up to 3 diagnoses

Supporting factors: 0 - 2 points each (out of 2) for up to 3 diagnoses

Opposing factors: 0 - 2 points each (out of 2) for up to 3 diagnoses

Final diagnostic decision: 0 - 2 points (out of 2)