JAMA Network | Open™

Research Letter | Health Informatics

Fidelity of Medical Reasoning in Large Language Models

Suhana Bedi, BS; Yixing Jiang, BS; Philip Chung, MD; Sanmi Koyejo, PhD; Nigam Shah, MBBS, PhD

Introduction

Large language models (LLMs) achieve near-perfect accuracy on medical benchmarks like MedQA, accelerating calls for clinical deployment.[1] However, a critical question remains unaddressed: do these models reason through medical problems or exploit statistical patterns in their training data?[2]

While frameworks like MedHELM[3] have expanded evaluation to medical tasks in clinical practice, we complement this work by testing whether high performance on any medical benchmark reflects reasoning or pattern matching. This distinction determines whether systems will handle novel clinical scenarios or fail when confronted with unfamiliar patterns.[4] Our study evaluates both reasoning and standard LLMs, allowing us to test whether reasoning capabilities improve robustness.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Methods

This cross-sectional study follows Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines and was exempt from institutional review as no human participants were involved, in accordance with 45 CFR §46. We sampled 100 questions from MedQA,[5] a standard multiple-choice medical benchmark, and replaced the original correct answer choice with "None of the other answers" (NOTA). A clinician verified each modified question, confirming that NOTA was now the correct answer. Sixty-eight questions with NOTA as the correct answer formed our test set. The **Figure** illustrates our NOTA substitution approach with an example from MedQA.

We evaluated 6 models spanning different architectures and capabilities: DeepSeek-R1 (model 1), o3-mini (reasoning models) (model 2), Claude-3.5 Sonnet (model 3), Gemini-2.0-Flash (model 4), GPT-4o (model 5), and Llama-3.3-70B (model 6). For our analysis, we compared each model's performance with chain-of-thought (CoT) prompting on the 68 questions in our clinician-validated set in their original form vs their NOTA-modified versions. We used CoT to encourage explicit reasoning from all models, enabling assessment of logical reasoning vs pattern recognition. We measured accuracy as the percentage of questions answered correctly. Statistical significance was assessed using the McNemar test, and 95% CIs for the accuracy drop were calculated using bootstrapping with 1000 iterations. The McNemar test was used to calculate *P* values, and significance was set at a 2-sided $P < .05$. Python with SciPy version 1.15.2, pandas 2.1.1, and NumPy 1.26.0 (Python) were used for analyses from March to April 2025.

If models truly reason through medical questions, performance should remain consistent despite the NOTA manipulation because the underlying clinical reasoning remains unchanged. Performance degradation would suggest reliance on pattern matching rather than reasoning.

Results

All models showed decreased accuracy on the clinician-validated NOTA questions compared with their performance on the same 68 questions in their original form (**Table**). The relative accuracy drops were major: 6 of 68 questions were incorrect in model 1 (8.82%), 11 of 68 (16.18%) in model 2, 23 of 68 (33.82%) in model 3, 25 of 68 (36.76%) in model 4, 18 of 68 (26.47%) in model 5, and 26 of 68 (38.24%) in model 6.

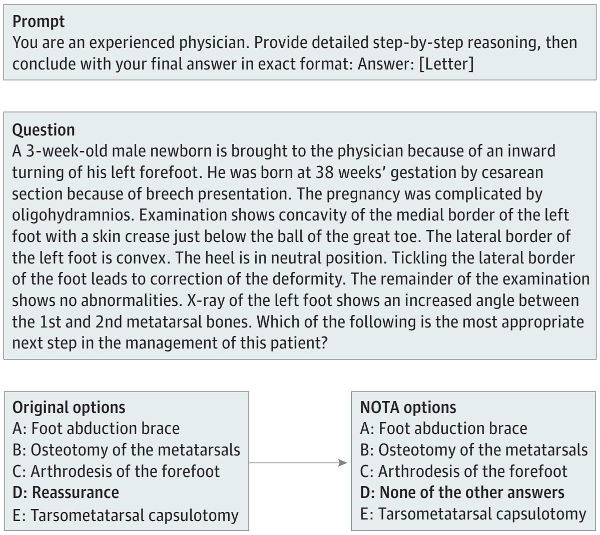
Models 1 and 2 demonstrated the most resilience to our manipulation, with the smallest relative accuracy drop. However, even these models experienced a statistically significant decline in performance.

Discussion

Our findings reveal a robustness gap for LLMs in medical reasoning, demonstrating that evaluating these systems requires looking beyond standard accuracy metrics to assess their true reasoning capabilities.[6] When forced to reason beyond familiar answer patterns, all models demonstrate declines in accuracy, challenging claims of artificial intelligence's readiness for autonomous clinical deployment.

A system dropping from 80% to 42% accuracy when confronted with a pattern disruption would be unreliable in clinical settings, where novel presentations are common. The results suggest that these systems are more brittle than their benchmark scores suggest.

Figure. None of the Other Answers (NOTA) Substitution Example in Medical Reasoning Assessment



Prompt
You are an experienced physician. Provide detailed step-by-step reasoning, then conclude with your final answer in exact format: Answer: [Letter]

Question
A 3-week-old male newborn is brought to the physician because of an inward turning of his left forefoot. He was born at 38 weeks' gestation by cesarean section because of breech presentation. The pregnancy was complicated by oligohydramnios. Examination shows concavity of the medial border of the left foot with a skin crease just below the ball of the great toe. The lateral border of the left foot is convex. The heel is in neutral position. Tickling the lateral border of the foot leads to correction of the deformity. The remainder of the examination shows no abnormalities. X-ray of the left foot shows an increased angle between the 1st and 2nd metatarsal bones. Which of the following is the most appropriate next step in the management of this patient?

Original options
A: Foot abduction brace
B: Osteotomy of the metatarsals
C: Arthrodesis of the forefoot
D: Reassurance
E: Tarsometatarsal capsulotomy

NOTA options
A: Foot abduction brace
B: Osteotomy of the metatarsals
C: Arthrodesis of the forefoot
D: None of the other answers
E: Tarsometatarsal capsulotomy

Chain-of-thought prompt, original question from MedQA with correct answer "Reassurance" (left) compared with NOTA-modified version where the correct answer is replaced with "None of the other answers" (right).

Table. Model Performance on Original and None of the Other Answers (NOTA)–Modified Questions[a]

	Accuracy, % (No./total No.)		
Model	Original	NOTA-modified	Accuracy drop, % (No./total No.) [95 % CI]
1	92.65 (63/68)	83.82 (57/68)	8.82 (6/68) [2.70-18.92]
2	95.59 (65/68)	79.41 (54/68)	16.18 (11/68) [10.81-29.73]
3	88.24 (60/68)	61.76 (42/68)	26.47 (18/68) [17.57-39.19]
4	92.65 (63/68)	58.82 (40/68)	33.82 (23/68) [24.32-47.30]
5	85.29 (58/68)	48.53 (33/68)	36.76 (25/68) [28.38-51.35]
6	80.88 (55/68)	42.65 (29/68)	38.24 (26/68) [27.03-51.35]

[a] This table compares performance on 68 clinician-validated questions. Original accuracy refers to performance on questions in their standard format, while NOTA-modified accuracy shows performance when the correct answer was replaced with "None of the other answers" (NOTA). Models are ordered by increasing accuracy drop. CIs were calculated using the McNemar test for paired nominal data.

While our study has limitations, including a small sample size and evaluation limited to 0-shot settings without exploring retrieval-augmented generation or fine-tuning techniques, our findings suggest 3 priorities for medical artificial intelligence: (1) development of benchmarks that distinguish clinical reasoning from pattern matching, (2) greater transparency about current reasoning limitations in clinical contexts, and (3) research into models that prioritize reasoning over pattern recognition. Until these systems maintain performance with novel scenarios, clinical applications should be limited to nonautonomous supportive roles with human oversight.

ARTICLE INFORMATION

Accepted for Publication: June 12, 2025.

Published: August 8, 2025. doi:10.1001/jamanetworkopen.2025.26021

Corresponding Author: Suhana Bedi, BS, Biomedical Data Science, Stanford University, 453 Quarry Rd, Palo Alto, CA 94304-1419 (suhana@stanford.edu).

Author Affiliations: Department of Biomedical Data Science, Stanford School of Medicine, Stanford, California (Bedi, Jiang, Shah); Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Stanford, California (Chung); Department of Computer Science, Stanford University, Stanford, California (Koyejo); Center for Biomedical Informatics Research, Stanford University, Stanford, California (Shah).

Author Contributions: Ms Bedi had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Bedi, Jiang, Koyejo, Shah.

Acquisition, analysis, or interpretation of data: Bedi, Chung.

Drafting of the manuscript: Bedi, Chung, Koyejo.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Bedi.

Obtained funding: Koyejo.

Administrative, technical, or material support: Jiang, Chung, Koyejo.

Supervision: Koyejo, Shah.

Conflict of Interest Disclosures: Dr Chung reported being a contractor for OpenAI outside the submitted work. Dr Koyejo reported receiving grants from the National Science Foundation, the Institute of Education Sciences, the MacArthur Foundation, Schmidt Sciences, and OpenAI during the conduct of the study. Dr Shah reported being a cofounder of Prealize Health and Atropos Health; receiving funding from the Gordon and Betty Moore Foundation and Chan Zuckerberg Institute; serving on the Board of the Coalition for Healthcare AI; and serving as a scientific advisor to Opala, Curai Health, Arsenal Capital, and JnJ Innovative Medicines. No other disclosures were reported.

Funding/Support: Ms Bedi is supported by the Stanford Graduate Fellowship. Dr Chung is supported by the Mentored Research Training Grant from the Foundation for Anesthesia Education and Research.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See the Supplement.

REFERENCES

1. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

2. Salido ES, Gonzalo J, Marco G. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice LLM evaluation benchmarks. *arXiv*. Published online February 18, 2025. http://arxiv.org/abs/2502.12896

3. Bedi S, Cui H, Fuentes M, et al. MedHELM: olistic evaluation of large language models for medical tasks. *arXiv*. Published online May 26, 2025. http://arxiv.org/abs/2505.23802

4. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-869. doi:10.1001/jama.2023.14217

5. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv*. Published online May 21, 2021. doi:10.20944/preprints202105.0498.v1

6. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. doi:10.1001/jama.2024.21700

SUPPLEMENT.

Data Sharing Statement