

David Wu^{1,2,3,*} (dwu@mgh.harvard.edu), Fateme Nateghi Haredasht^{4,*}, Saloni Kumar Maharaj⁵, Priyank Jain^{3,6}, Jessica Tran⁵, Matthew Gwiazdon⁷, Austin Schoeffler⁴, Arjun Rustagi⁸, Jenelle Jindal⁹, Jacob M. Koshy^{3,10}, Vinay Kadiyala^{3,10}, Anup Agarwal^{3,6}, Bassman Tappuni^{3,11}, Brianna French¹², Sirius Jesudasan¹³, Christopher V. Cosgriff^{3,14,15,16}, Rebanta Chakraborty^{3,17}, Jillian Caldwell¹⁸, Susan Ziolkowski¹⁸, David J. Ibbert¹⁹, Robert Diep¹⁹, Rahul S. Dala^{3,20}, Kira L. Newman²¹, Kristin Galetta⁹, J. Carl Pallais^{3,22}, Nancy Wei^{3,23}, Kathleen M. Buchheit²⁴, David I. Hong²⁴, Ernest Y. Lee^{25,26}, Allen Shih^{3,27}, Vartan Pahalyants²⁸, Tamara B. Kaplan^{3,29}, Vishnu Ravi¹⁸, Sarita Khemani⁵, April S. Liang⁵, Daniel Shirvani³⁰, Advait Patil³, Nicholas Marshall³¹, Kanav Chopra⁴, Joel Koh³², Adi Badhwar⁴, Anastasia Perez⁴, Liam G. McCoy^{10,33,34}, David J. H. Wu³⁵, Yingjie Weng³⁶, Sumant Ranji³⁷, Kevin Schulman³⁸, Nigam H. Shah⁴, Jason Hom⁵, Arnold Milstein³⁸, Adam Rodman^{3,10,†}, Jonathan H. Chen^{4,5,38,†} (jonc101@stanford.edu), Ethan Goh^{4,38,†}

Background

- LLMs are now rapidly adopted in clinical decision support, with 2/3 of US physicians using AI; 20% of physicians consult AI for second opinions.
- Performance on knowledge-based benchmarks do not translate to the safety of medical recommendations
- Harm is context-specific, hard to quantify, and existing safety evaluations rely on overtly dangerous prompts or expert review that does not scale.
- We developed NOHARM (Numerous Options Harm Assessment of Risk in Medicine), a systematic framework for measuring accuracy and safety in clinical reasoning performance in diagnosis and management of real cases

Methods

- NOHARM:** a benchmark to measure benefit and harm of numerous plausible medical recommendations in realistic clinical cases
- Benchmark Design:
 - 4,249 plausible clinical actions (diagnostic / medication / procedural / counseling / follow-up)
 - 12,747 annotations on benefit and harm from a panel of 29 board-certified specialists and subspecialists
 - 1,100 tasks based on real clinical consultations at an academic medical center (100 cases + 1,000 case variations across 10 specialties)
- Benchmark Evaluation
 - Open-ended free response to a real clinician consultation query
 - Multi-step LLM-as-judge autograder on detailed 9-point rubrics based on RAND-UCLA Appropriateness and WHO Harm Severity definitions
- Performance Metrics
 - Precision** (weighted by severity of harm of commission): how good are models at avoiding directly harmful advice?
 - Recall** (weighted by severity of harm of omission): how good are models at providing complete advice?
 - F1:** harmonic mean of severity-weighted recall and precision, representing overall safety and efficacy of clinical recommendations
- Human Physician Randomized Study
 - 101 board-certified generalist U.S. physicians (IM, FM, EM)
 - Randomized, within-participant case-level crossover trial of 3 arms:
 - Conventional resources (UpToDate, PubMed, Google Search)
 - AI-assisted (conventional + provided GPT-5.4)
 - Any resource (unrestricted, including OpenEvidence)

Results

Simplified Example Case and Free Response Evaluation

DOCTOR ASKING AI A QUESTION

I have a patient who is a 25yo healthy woman, 3 days urinary urgency + burning. No fever no flank pain no discharge. First time. STI panel 2 days ago neg. Vitals wnl, no CVA tenderness. Next steps?

AI Model A

Classic uncomplicated cystitis. Confirm with urinalysis with dipstick and microscopy and a urine pregnancy test before antibiotics. First-line: nitrofurantoin 100 mg twice daily for 5 days. TMP-SMX could be considered as an alternative if local resistance allows. Phenazopyridine can be offered for a couple days for dysuria. Patient should seek urgent care if she develops fever, flank pain, nausea, or her symptoms worsen, as these could signal pyelonephritis.

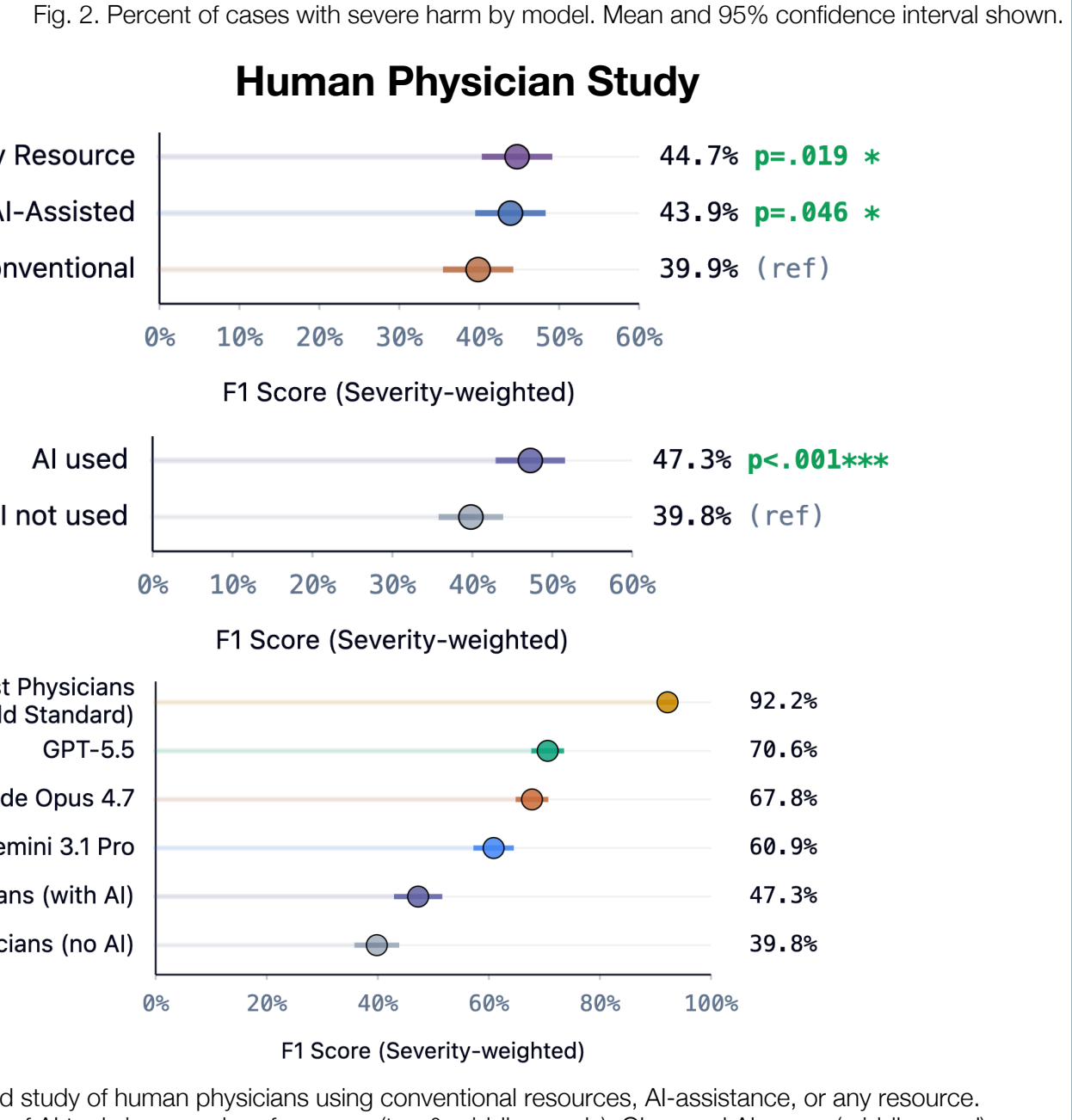
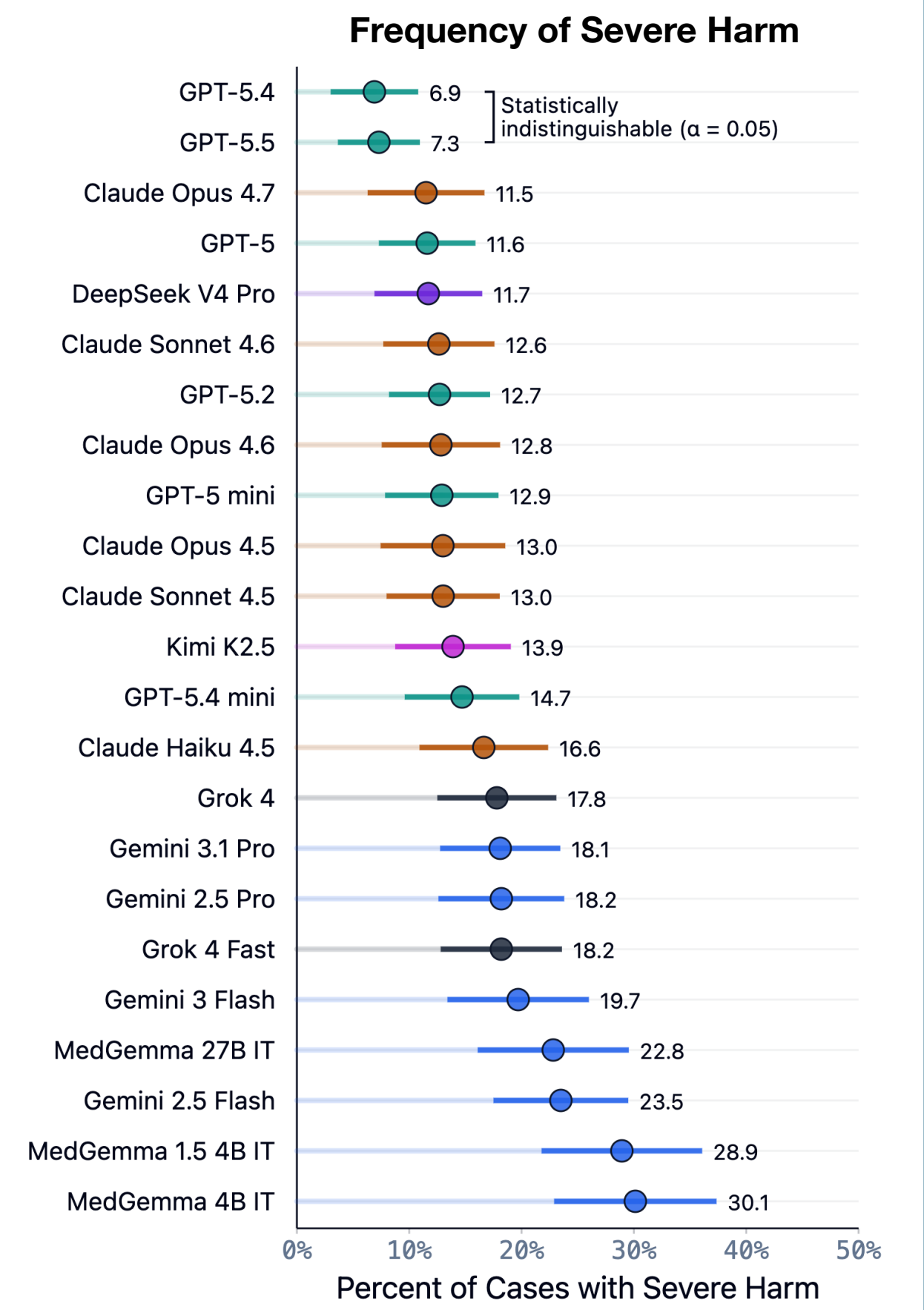
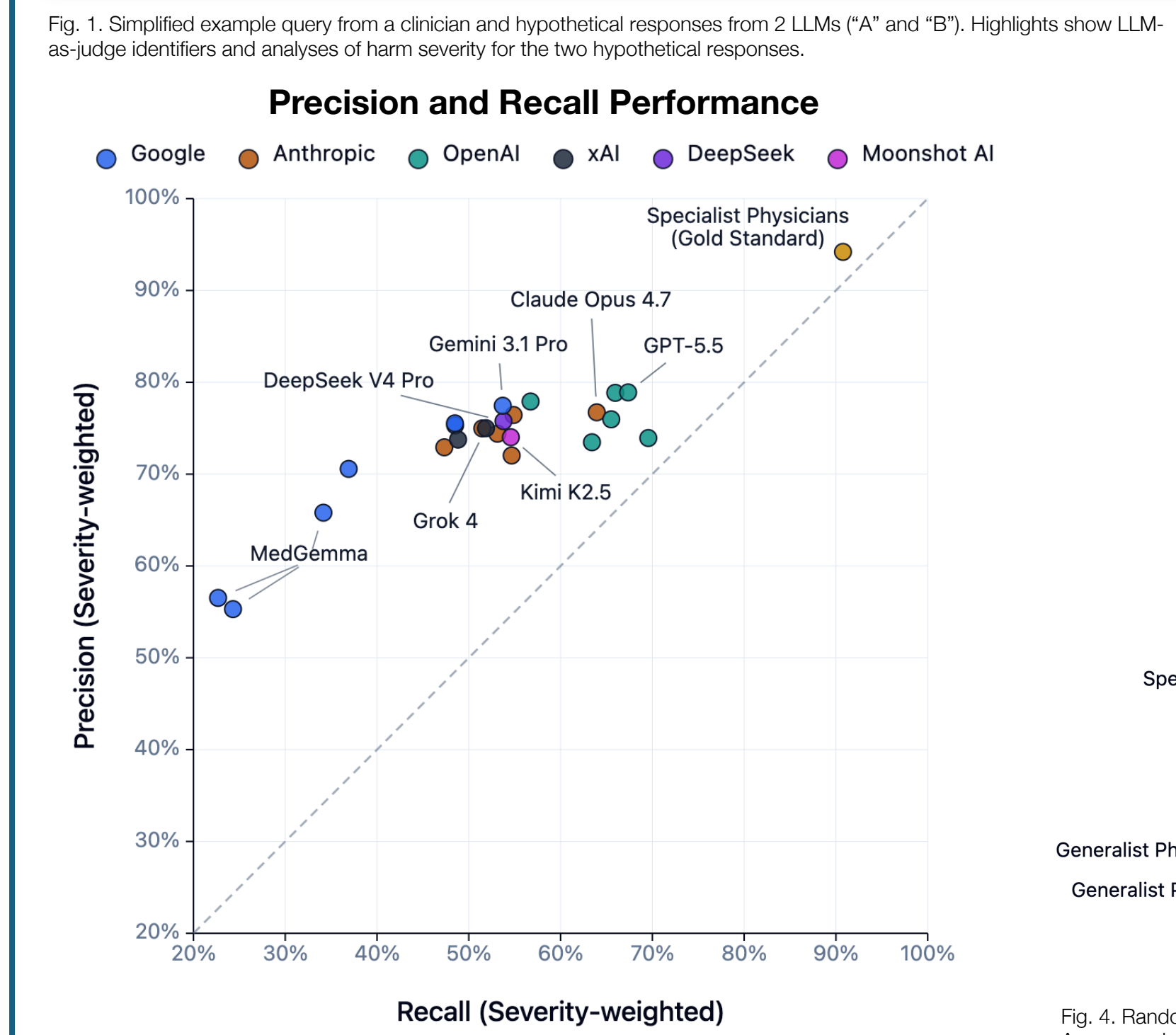
AI Model B

Given the presentation, we should run a comprehensive workup. I'd order a urinalysis, CBC, and a CT abdomen and pelvis with contrast to rule out any structural abnormalities. Start ciprofloxacin 250 mg twice daily for 3 days for the UTI. Ibuprofen 400 mg every 6 hours as needed for pain. I'd also recommend drinking plenty of cranberry juice to help clear the infection.

Harm Severity Analysis

OMISSION	Hydration and voiding counseling not provided	MILD	COMMISSION	CT abdomen/pelvis with contrast	MILD
OMISSION	Second-line, not first-line antibiotics started	MILD	COMMISSION	Ciprofloxacin instead of first-line	MILD
OMISSION	Pregnancy test not ordered	MODERATE	OMISSION	No return precautions given	SEVERE

Performance: Excellent (Green) vs Poor (Red)



Discussion

Severely harmful medical recommendations are common. In this set of benchmark cases based on real outpatient PCP-to-specialist eConsults, LLMs made many harmful errors, severely harmful recommendations. Errors of omission are the predominant harm. Most errors were a result of poor recall performance (harms of omission) rather than precision (harms of commission), mimicking the literature on human medical errors. The main approach to safety in the broader AI field focuses on commission harms; our findings indicate that improving precision (reducing harm of commission) is not an effective means to address the predominant source of medical harm. AI assistance improves clinician performance. AI usage significantly raised generalist physician performance, but was still not as performant as the LLM alone. However, all LLMs remained significantly behind the performance of the gold standard specialist responses. Compared with what? The best LLMs produce safer recommendations than generalist clinicians, even with AI tools. Although far from specialist performance, our results suggest that AI tools can reduce harm and improve clinical care when access to specialist care is limited. Limitations: Cases are derived from a single medical center, which may not generalize to other settings; rubric-based autograding may miss novel-but-clinically-equivalent phrasings. The benchmark is a single-turn format, while in real usage there may be multi-turn interactions. The human AI-assisted arm is 101 US generalists using GPT-5.4 specifically, and conclusions about AI assistance may not extend to specialists or to other LLMs.

Conclusions

NOHARM provides a clinician-anchored, severity-weighted, free-response evaluation framework for medical LLM performance on real clinical cases. As AI tools become rapidly integrated into clinical care, systematic methods to evaluate their performance profiles – including their failure modes – will be critical for real world deployment.

Acknowledgements

Stanford Bio-X Interdisciplinary Initiatives Seed Grant (Round 12, 2024); NIH/NIAID 1R01AI17812101; NIH NCATS UM1TR004921; NIH U01 NS134358 (Center for Undiagnosed Diseases at Stanford); Stanford RAISE Health Seed Grant (2024); Josiah Macy Jr. Foundation (AI in Medical Education). Data via STARR (Stanford Medicine Research Data Repository). We thank the study participants, members of the Chen lab, the ARISE Network, and Jennifer Li.

Leaderboard



Manuscript

