nature medicine

Supplementary information

https://doi.org/10.1038/s41591-026-04539-8

Toward a test of medical AI superintelligence

In the format provided by the authors and unedited

Supplementary Information

Table 1: Comparative performance of physicians and AI systems on clinical reasoning tasks

Recent randomized trials and physician-adjudicated evaluations suggest that frontier LLMs can match or outperform clinicians on selected diagnostic and management reasoning tasks. However, most studies rely on simulated vignettes and structured case narratives. These findings indicate rapidly advancing clinical reasoning capability, but do not establish safety or effectiveness in real patient care.

Study	Task	Human comparator	AI	Key results	Limitation
Goh et al., 2024, JAMA Network Open	Randomized trial of diagnostic reasoning using vignettes based on real de-identified encounters	50 internal medicine, emergency medicine, and primary care physicians	OpenAI GPT-4	Diagnostic reasoning score: Human 74%; Human+AI 76%; AI 92%	Six diagnostic vignettes; information was presented as curated case narratives rather than raw, messy clinical notes
Goh et al., 2025, Nature Medicine	Randomized trial of management reasoning using vignettes based on real de-identified encounters	92 internal medicine, emergency medicine, and primary care physicians	OpenAI GPT-4	Management reasoning score: Human 35.7%; Human+AI 43.0%; AI 43.7%	Five management vignettes; information was curated and progressively revealed; cases remained more structured than routine clinical encounters
McDuff et al., 2025, Nature	Randomized study of differential diagnosis using 302 NEJM CPC cases	20 internal medicine physicians	Google PaLM 2 (fine-tuned)	Top-10 differential diagnosis accuracy: Human 44.4%; Human+AI 51.7%; AI 59.1%	NEJM CPC cases are curated diagnostic vignettes with sufficient clues for clinical reasoning; they differ from the incomplete, evolving information available at patient intake

Tu et al., 2025, Nature	Randomized, double-blind crossover study of text-based diagnostic consultations with patient actors in OSCE setting (159 scenarios)	20 primary care physicians	Google PaLM 2 (fine-tuned)	Top-1 differential diagnosis accuracy: Human 66%; AI 78% Top-3 differential diagnosis accuracy: Human 76%; AI 90%	Simulated patient-actor OSCE setting; synchronous text chat was unfamiliar to physicians
Saab et al., 2026, Nature Medicine	Randomized multimodal simulated telehealth consultations with patient actors using skin images, ECGs and clinical documents	19 primary care physicians	Google Gemini 2.0 Flash	Top-3 differential diagnosis accuracy: Human 68%; AI 80%	Synchronous text chat lacks non-verbal cues, physical examination, and usual clinician-patient interaction; modalities were limited to skin images, ECGs, and clinical documents
Liévin et al., 2026, Nature	Randomized OSCE study for longitudinal disease management across 100 multi-visit scenarios with patient actors and the use of clinical guidelines and drug formularies	21 primary care physicians	Google Gemini 1.5 Flash (fine-tuned)	Overall appropriateness of management plan for first visit: Human 72%; AI 95%	PCPs communicated with patient actors via unfamiliar text-based chat interface; cases lacked chart review that characterizes real patient care

Brodeur et al., 2026, Science	Six diagnostic and management reasoning experiments, including a real-world ER second-opinion study (76 cases)	100+ internal medicine, emergency medicine, and primary care physicians	OpenAI o1	Management reasoning: Human 34%; AI 89% Diagnostic reasoning: Human 74%; AI 97% ER triage second opinion (exact/near Dx): Human 55.3%; AI 67.1%	Text-only evaluation (no imaging or other non-text inputs) of the model alone rather than human-AI collaboration; limited to internal and emergency medicine
Ferber et al., 2026, Nature	Autonomous EHR-based emergency department simulation using 574 MIMIC-IV cases across 8 diagnoses, with head-to-head physician comparison on 311 cases	10 internal medicine, radiology and haemato-oncology physicians	OpenAI GPT-4o and o1-preview	Diagnostic accuracy: Human 78.1%; AI 87.8% Relevant procedures requested: Human 38.3%; AI 53.5%	Retrospective MIMIC-IV cases in a sandboxed EHR; patient dialogue was generated from structured discharge-summary text rather than real emergency department interactions

References

1. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw. Open https://doi.org/10.1001/jamanetworkopen.2024.40969 (2024).
2. Goh, E. et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat. Med. https://doi.org/10.1038/s41591-024-03456-y (2025).
3. McDuff, D. et al. Towards accurate differential diagnosis with large language models. Nature https://doi.org/10.1038/s41586-025-08869-4 (2025).
4. Tu, T. et al. Towards conversational diagnostic artificial intelligence. Nature https://doi.org/10.1038/s41586-025-08866-7 (2025).
5. Saab, K. et al. Advancing conversational diagnostic AI with multimodal reasoning. Nat. Med. https://doi.org/10.1038/s41591-026-04371-0 (2026).
6. Liévin, V. et al. Towards Conversational AI for Disease Management. Nature https://doi.org/10.1038/s41586-026-10764-5 (2026).
7. Brodeur, P. G. et al. Performance of a large language model on the reasoning tasks of a physician. Science https://doi.org/10.1126/science.adz4433 (2026).
8. Ferber, D. et al. Towards autonomous medical artificial intelligence agents. Nature https://doi.org/10.1038/s41586-026-10675-5 (2026).

Table 2: Clinical AI benchmarks evaluating clinical reasoning, multimodality, safety, and agentic EHR task execution

Overview of recent benchmarks evaluating large language models across clinically relevant tasks, including clinical reasoning, multimodal interpretation, harm avoidance, calibration under uncertainty, and agentic EHR workflow execution. A system that saturates knowledge-based multiple-choice questions may still perform poorly on SCT-Bench, which evaluates whether a model updates its reasoning appropriately when new and conflicting clinical information appears. In NOHARM, potential for severe harm from LLM recommendations across 31 LLMs occurred in up to 22.2% of cases, and safety performance was only moderately correlated with existing AI and medical knowledge benchmarks.

Benchmark	Clinical construct	Data source	Evaluation design
ReXrank - Zhang et al., 2024	Multimodal radiology report generation from chest X-rays	10,000 chest X-ray studies from 67 U.S. medical sites; public datasets including MIMIC-CXR, IU-Xray, and CheXpert Plus	Models generate radiology findings and impressions from chest X-ray images; outputs are compared with reference radiology reports using lexical, semantic, and clinically oriented report-quality metrics
MedAgentBench - Jiang et al., 2025	Agentic EHR workflow capability	300 clinical tasks across 10 categories; 100 realistic patient profiles; 700,000+ data elements in a FHIR-compliant virtual EHR	Agents interact with FHIR APIs to retrieve data and execute clinical tasks such as ordering tests/medications
HealthBench - Arora et al., 2025	Open-ended conversational medical assistance	5,000 multi-turn health conversations; 262 physician annotations across 26 specialties; 48,562 rubric criteria	Conversation-specific physician-written rubrics; model responses graded with rubric-based evaluation
HealthBench Professional - Hicks et al., 2026	Clinician-facing medical chat tasks	525 tasks selected from 15,079 candidate clinician conversations; 190 physician contributors across 26 specialties	Example-specific rubrics authored, reviewed and adjudicated by physicians; model responses graded with rubric-based evaluation

SCT-Bench - McCoy et al., 2025	Clinical reasoning under uncertainty	750 Script Concordance Test questions from 10 international datasets	Vignettes ask how new information changes diagnostic or management likelihood; responses scored against expert-panel distributions
CPC-Bench - Buckley et al., 2025	Expert diagnostic reasoning and case synthesis	7,102 NEJM clinicopathological cases; 1,021 Image Challenges	10 text and multimodal tasks, including differential diagnosis, next-test selection, literature retrieval, and image interpretation
NOHARM - Wu et al., 2025	Harm avoidance in clinical management	100 actual primary care-to-specialist consultation cases; 10 specialties; 4,249 management options; 12,747 expert annotations	31 LLMs evaluated on clinical recommendations; outputs assessed against specialist-rated management options for benefit/harm, omission/commission, and WHO harm-severity definitions
MedHELM - Bedi et al., 2026	Broad medical-task performance	37 evaluations across 5 categories, 22 subcategories, and 121 tasks; 16 public datasets, 7 credentialed-access datasets, and 14 private/Stanford-developed datasets	Nine frontier LLMs evaluated using task-specific metrics and an automated multi-evaluator LLM-jury against expert-defined criteria
PhysicianBench - Liu et al., 2026	EHR-based clinical task execution	100 long-horizon tasks adapted from real primary care subspecialty e-consults; STARR-derived de-identified longitudinal patient records with additional privacy perturbations; 21 specialties	FHIR-based EHR environment; mean 27 tool calls per task; 670 structured checkpoints spanning retrieval, reasoning, action execution, and documentation

References

1. Zhang, X. et al. ReXrank: a public leaderboard for AI-powered radiology report generation. Preprint at https://doi.org/10.48550/arXiv.2411.15122 (2024).
2. Jiang, Y. et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. NEJM AI https://doi.org/10.1056/AIdbp2500144 (2025).
3. Arora, R. K. et al. HealthBench: evaluating large language models towards improved human health. Preprint at https://doi.org/10.48550/arXiv.2505.08775 (2025).
4. Hicks, R. S. et al. HealthBench Professional: evaluating large language models on real clinician chats. Preprint at https://doi.org/10.48550/arXiv.2604.27470 (2026).
5. McCoy, L. G. et al. Assessment of large language models in clinical reasoning: a novel script concordance test benchmark. NEJM AI https://doi.org/10.1056/AIdbp2500120 (2025).
6. Buckley, T. A. et al. Advancing medical artificial intelligence using a century of cases. Preprint at https://doi.org/10.48550/arXiv.2509.12194 (2025).
7. Wu, D. et al. First, do NOHARM: towards clinically safe large language models. Preprint at https://doi.org/10.48550/arXiv.2512.01241 (2025).
8. Bedi, S. et al. Holistic evaluation of large language models for medical tasks with MedHELM. Nat. Med https://doi.org/10.1038/s41591-025-04151-2 (2026).
9. Liu, R. et al. PhysicianBench: evaluating LLM agents in real-world EHR environments. Preprint at https://doi.org/10.48550/arXiv.2605.02240 (2026).