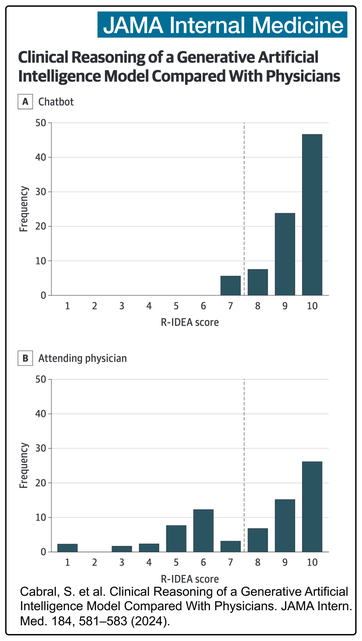
Preceptron: Towards Automated Assessment of Large Language Models in Medicine

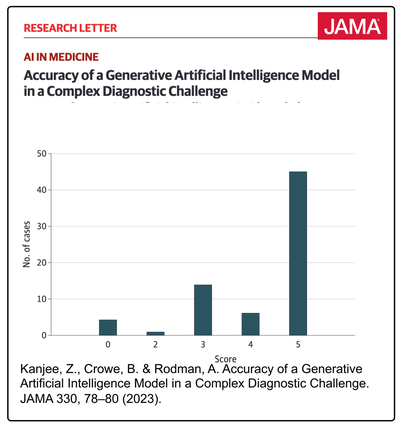
Thomas Buckley[1](thomas_buckley@g.harvard.edu), Adam Rodman, M.D.[2], and Arjun K. Manrai, Ph.D.[1*]
[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[2] Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA

Background

LLMs have been shown to be capable in clinical reasoning and diagnosis through qualitative physician evaluation.

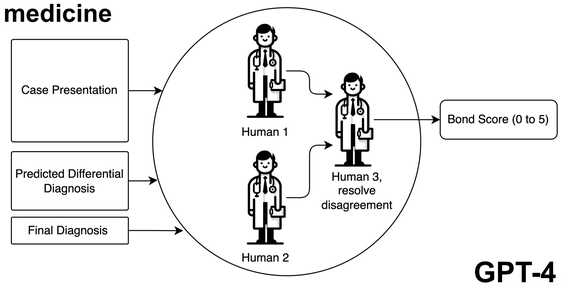


Cabral, S. et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Intern. Med. 184, 581–583 (2024).

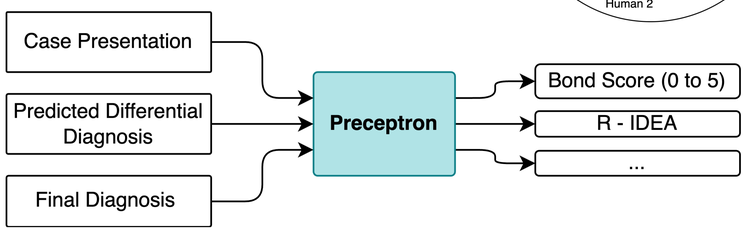


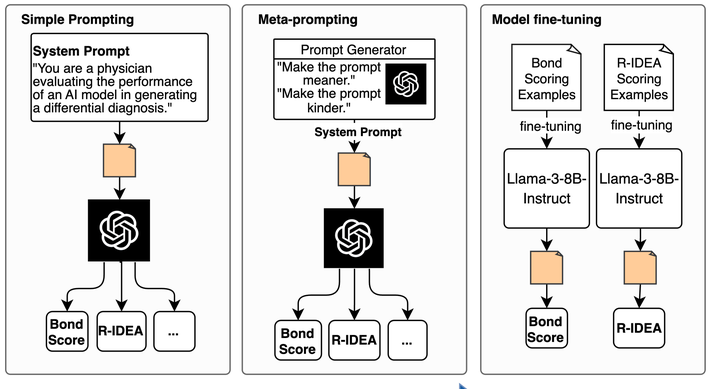
Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 330, 78–80 (2023).

Physician evaluations are costly, preventing larger-scale studies of AI in medicine



Methods



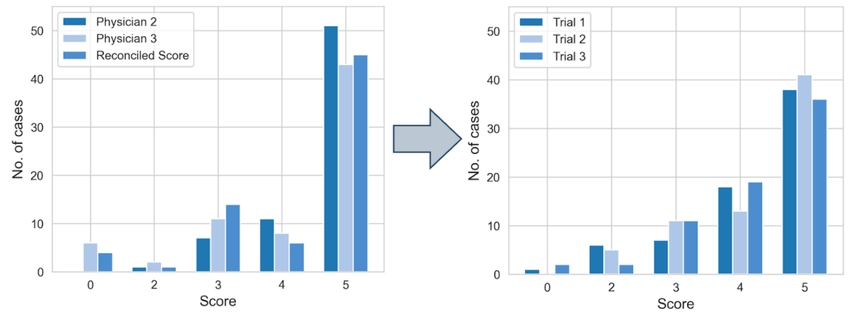


Models were evaluated on physician-labeled examples of (1) **5-point Bond Score** for predicting a differential diagnosis to an NEJM Clinicopathological Conference (CPC), and (2) **10-point R-IDEA** for clinical reasoning quality by physicians or chatbots on on 20 NEJM Healer Cases.

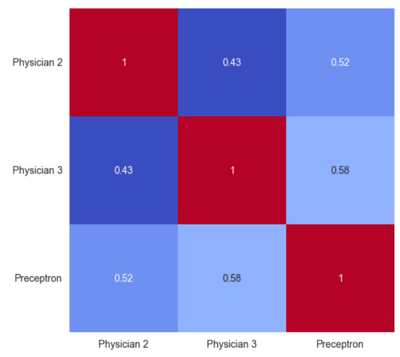
Results

A. JAMA Paper

B. Reproduced Using Preceptron



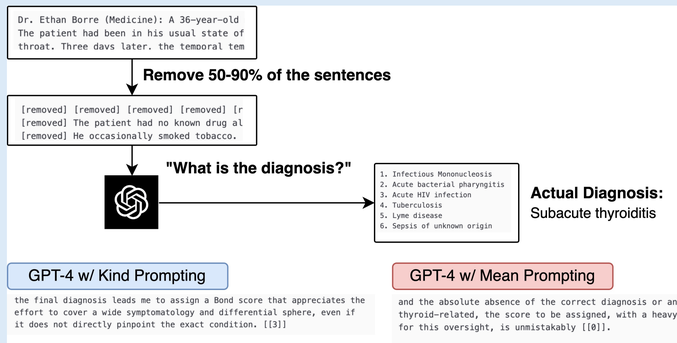
C. Linear Cohen's Kappa Between Physicians and Preceptron



GPT-4 Accurately Assigns Bond and R-IDEA Scores

	Bond Score Set (Balanced)			R-IDEA Dataset		
	Cohen's Kappa	Accuracy	Within-1 accuracy	Cohen's Kappa	Within-1 Accuracy	Within-2 accuracy
Physician 1*	0.71	0.67	0.95	0.62	0.76	0.87
Physician 2*	0.75	0.72	0.95	–	–	–
Physician 3	–	–	–	0.52	0.67	0.81
Physician 4	–	–	–	0.50	0.67	0.80

Training Llama-3-8B-Instruct to Align with Evaluator Preferences



Results in two datasets, a "mean" and "kind" dataset, each with 1860 examples. Evaluation performed with 80/20 train test split by case

Agreement of Fine-tuned Llama-3-8B-Instruct Models with Test Data

	No explanation		With explanation	
	Mean	Kind	Mean	Kind
Llama-3-8B-Instruct Baseline	0.42	0.22	0.40	0.19
Mean Llama-3-8B	**0.62**	0.42	**0.70**	0.53
Kind Llama-3-8B	0.54	**0.71**	0.59	**0.65**

Linear-weighted Cohen's Kappa on test cases

Preceptron unlocks questions that were previously impossible to answer with physician evaluations…

A. Variability of GPT-4 in Differential Diagnosis for a Complex Diagnostic Challenge

B. With 10 Attempts, Higher Temperature GPT-4 More Likely to Find Diagnosis in at Least 1 Attempt

Conclusion

- AI models can scalably evaluate other AI models, **unlocking questions that would have been impossible to answer.**
- We find that at higher temperatures, an often-overlooked parameter, the LLM can elicit a diagnosis that would have otherwise not been found.

HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS