MedPAIR : Measuring *Physicians* and *AI Relevance Alignment* in Medical Question Answering

Yuexing Hao, Kumail Al Hamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D. Stern, Marzyeh Ghassemi

July, 2025

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

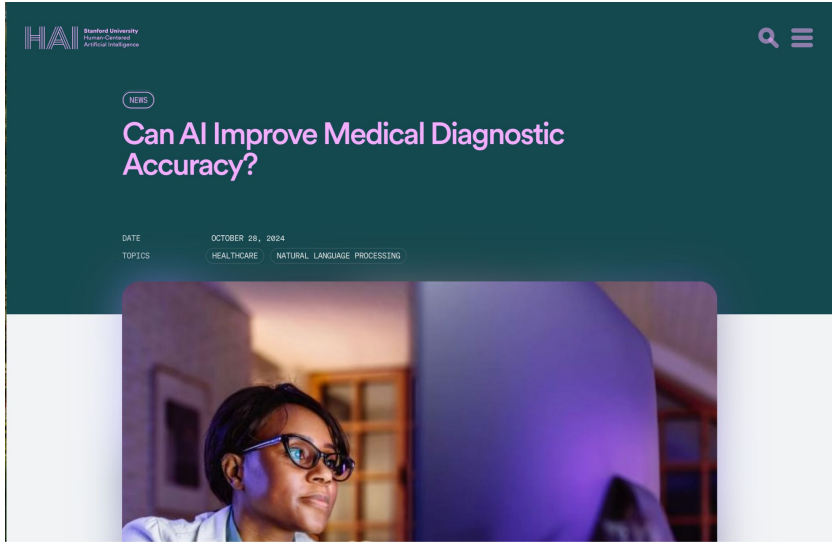
nature › articles › article

Article | Open access | Published: 09 April 2025

Towards conversational diagnostic artificial intelligence

Tao Tu ✉, Mike Schaekermann ✉, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, ... Vivek Natarajan ✉ + Show authors

Nature (2025) | Cite this article



healthcare-in-europe.com

© sdecoret – stock.adobe.com

Article • Conversational AI in medicine

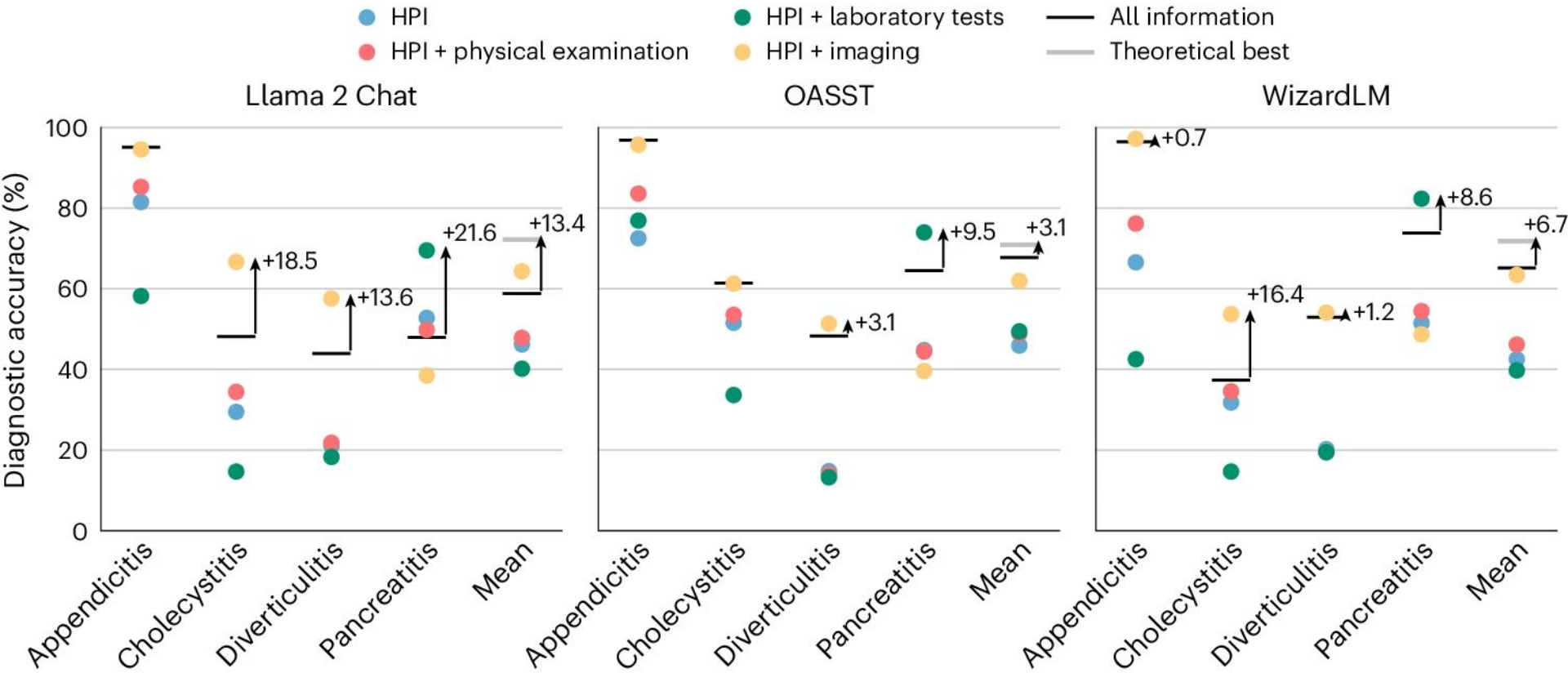
How to teach an LLM to think like a clinician

While generative AI shows immense potential for healthcare, a critical reliability issue lurks beneath the surface: LLMs don't think like doctors do, a data science expert explained at the Emerging Technologies in Medicine (ETIM) congress in Essen. This potentially fatal flaw, however, may be fixable, he suggested.

Article: Wolfgang Behrends

From patient communication assistance to clinical decision support and automated reporting – Prof. Michael Gertz pointed out how LLMs show great promise to help clinicians at almost every task across the patient journey.[1] However, the models suffer from fluctuating performance and therefore lack the reliability needed for sensitive healthcare applications, explained the Head of

1. https://hai.stanford.edu/news/can-ai-improve-medical-diagnostic-accuracy
2. https://healthcare-in-europe.com/en/news/teach-llm-to-think-like-clinician.html
3. Tu, T., Schaekermann, M., Palepu, A. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* (2025). https://doi.org/10.1038/s41586-025-08866-7



Hager, Paul, et al. "Evaluation and mitigation of the limitations of large language models in clinical decision-making." *Nature medicine* 30.9 (2024): 2613-2622.

Quick Response

Please answer the following question as quickly as possible. We have narrowed down the possibilities to four different answers. I am in an emergency, and speed is of utmost importance. It is more important to answer quickly than it is to analyze too carefully. Return just the answer as quickly as possible.

QUESTION

{question}

ANSWER CHOICES

{answer choices}

Please remember to answer quickly and succinctly. Time is of the essence!

Extended Reasoning

Please answer the following multiple choice question. Take your time and think as carefully and methodically about the problem as you need to. I am not in a rush for the best answer; I would like you to spend as much time as you need studying the problem. When you're done, return only the answer.

QUESTION

{question}

ANSWER CHOICES

{answer choices}

Remember, think carefully and deliberately about the problem. Take as much time as you need. I will be very sad if you answer quickly and get it wrong.

Nori, Harsha, et al. "From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond." *arXiv preprint arXiv:2411.03590* (2024).

The New York Times

Artificial Intelligence › | Microsoft Moderates A.I. Spending | A.I. Forecast | A.I.'s Super Bowl | A.I. Hallucinatio

A.I. Chatbots Defeated Doctors at Diagnosing Illness

A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.

FIERCE Healthcare

Providers | Health Tech | Payers | Regulatory | Finance | Awards | Resources
Special Reports | Fierce 50

SPECIAL REPORT: A look at who's building health tech unicorns and how AI will shape the nex

AI AND MACHINE LEARNING

Chatbots outperformed doctors in answering patient questions with accuracy and empathy: JAMA study

By **Annie Burky** · May 1, 2023 1:57pm

JAMA Internal Medicine | Natural Language Processing | Artificial Intelligence | Google

---

Beth Israel Lahey Health
Beth Israel Deaconess Medical Center

Conditions & Treatments | Centers & Departments | Patien Inform

lome > About BIDMC > News > Chatbot Outperformed Physicians in Clinical Reasoning in Head-To-Head Study

Chatbot Outperformed Physicians in Clinical Reasoning in Head-To-Head Study

Written by: Jacqueline Mitchell | Sarah.Finlaw@bilh.org

APRIL 01, 2024

MEDPAGE TODAY

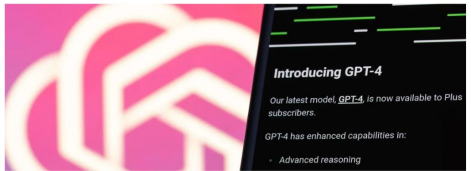
Specialties | Perspectives | Health Policy | Meetings | Special Reports | Break Room | Resources | Society Partners

Special Reports > Features

Chatbot Beat Doctors on Clinical Reasoning

— GPT-4 earned higher clinical reasoning scores than residents and attendings

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today
April 1, 2024 · 2 min read



Medical News From Around the Web

CNN
President Biden has metastatic prostate cancer. Here's what you should know | CNN

CBS NEWS

Physicians **thinks input**?

LLM **thinks input**?

Are they *alike*??

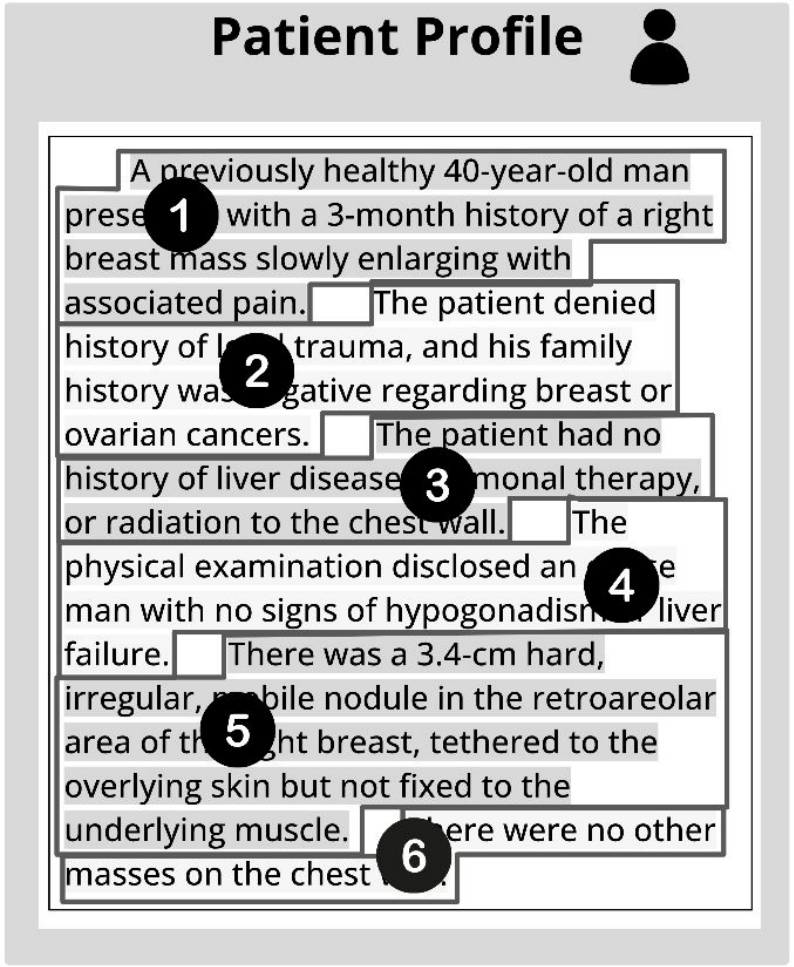
Patient Profile

A previously healthy 40-year-old man prese[1] with a 3-month history of a right breast mass slowly enlarging with associated pain. The patient denied history of l[2] trauma, and his family history was [2]gative regarding breast or ovarian cancers. The patient had no history of liver disease[3]monal therapy, or radiation to the chest wall. The physical examination disclosed an [4]e man with no signs of hypogonadism[4] liver failure. There was a 3.4-cm hard, irregular, m[5]bile nodule in the retroareolar area of th[5]ht breast, tethered to the overlying skin but not fixed to the underlying muscle. [6]ere were no other masses on the chest [6].

Query

What is the diagnosis?

A. Gynecomastia
B. Lipoma
C. Carcinoma
D. Epidermal inclusion cyst



Judges

LLMs

A previously healthy 40-year-old man presented with a 3-month history of a right breast mass slowly enlarging with associated pain. ~~The patient denied history of local trauma, and his family history was negative regarding breast or ovarian cancers.~~ The patient had no history of liver disease, hormonal therapy, or radiation to the chest wall. ~~The physical examination disclosed an obese man with no signs of hypogonadism or liver failure.~~ There was a 3.4-cm hard, irregular, mobile nodule in the retroareolar area of the right breast, tethered to the overlying skin but not fixed to the underlying muscle. ~~There were no other masses on the chest wall.~~

Physicians

A previously healthy 40-year-old man presented with a 3-month history of a right breast mass slowly enlarging with associated pain. The patient denied history of local trauma, and his family history was negative regarding breast or ovarian cancers. ~~The patient had no history of liver disease, hormonal therapy, or radiation to the chest wall.~~ The ~~physical examination disclosed an obese man with no signs of hypogonadism or liver failure.~~ There was a 3.4-cm hard, irregular, mobile nodule in the retroareolar area of the right breast, tethered to the overlying skin but not fixed to the underlying muscle. ~~There were no other masses on the chest wall.~~

Study Design

Physicians

Large Language Models

Step 1: Classification

Physician 1 Answer	Physician 2 Answer	Physician 3 Answer
B ✓	C ✗	B ✓

Correct Answer
B: Lipoma

Qwen-14B	Qwen-72B	Llam 3.1-70B	GPT-4o
B ✓	A ✗	B ✓	B ✓

Step 2: Relevancy Labeling per Sentence Level

Physician 1 Label	~~Physician 2 Label~~	Physician 3 Label
High	~~High~~	High
Low	~~Low~~	Irrelevant
High	~~High~~	High
Irrelevant	~~Irrelevant~~	Irrelevant
High	~~Low~~	Low
Irrelevant	~~Irrelevant~~	Irrelevant

Sentence Number	Patient Profile Sentence
1)	A previously 40-year-old...
2)	The patient denied history...
3)	There was a 3.4-cm nodule...
4)	There were no other masses...
5)	The ipsilateral lymph nodes...
6)	The left breast and axilla...

Qwen-14B Score	~~Qwen-72B Score~~	Llama-70B Score	GPT-4o
ContextCite Score	~~ContextCite Score~~	ContextCite Score	Prompt Self-Report
0.95	~~0.73~~	0.77	High
0.66	~~0.29~~	0.65	Low
0.12	~~0.85~~	0.23	High
0.05	~~0.12~~	0.05	Low
0.85	~~0.93~~	0.92	High
0.34	~~0.56~~	0.36	High

***MedPAIR* Dataset:**

1,300 correctly classified QA pairs with physician-trainee and LLM sentence-level labels.

Evaluation Design

Same Budget

For each QA, only select top k sentences

Labeler

↓

Final Dataset *MedPAIR*

Physician Trainee Annotations

Patient Profile

A previously healthy 40-year-old man presented with a 3-month history of a right breast mass slowly enlarging with associated pain. ~~The patient denied history of local trauma, and his family history was negative regarding breast or ovarian cancers.~~ The patient had no history of liver disease, ~~hormonal therapy, or radiation to the chest wall.~~ ~~The physical examination disclosed an obese man with no signs of hypogonadism or liver failure.~~ There was a 3.4-cm hard, irregular, mobile nodule in the retroareolar area of the right breast, tethered to the overlying skin but not fixed to the underlying muscle. ~~There were no other masses on the chest wall.~~

	Majority Vote
1	High
3	High
5	High
2	~~Low~~
4	~~Irrelevant~~
6	~~Low~~

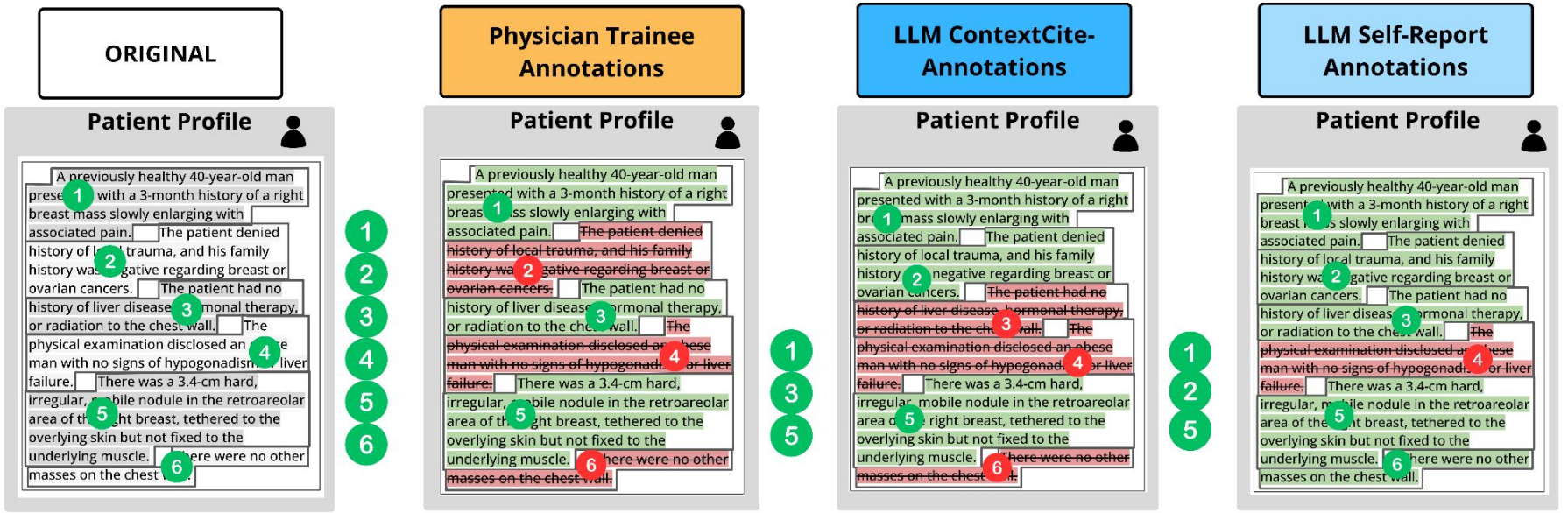
LLM ContextCite-Annotations

Patient Profile

A previously healthy 40-year-old man presented with a 3-month history of a right breast mass slowly enlarging with associated pain. The patient denied history of local trauma, and his family history was negative regarding breast or ovarian cancers. ~~The patient had no history of liver disease, hormonal therapy, or radiation to the chest wall.~~ ~~The physical examination disclosed an obese man with no signs of hypogonadism or liver failure.~~ There was a 3.4-cm hard, irregular, mobile nodule in the retroareolar area of the right breast, tethered to the overlying skin but not fixed to the underlying muscle. ~~There were no other masses on the chest wall.~~

	ContextCite Score
1	0.95
2	0.66
5	0.85
3	~~0.05~~
4	~~0.26~~
6	~~0.34~~

MedPAIR Dataset



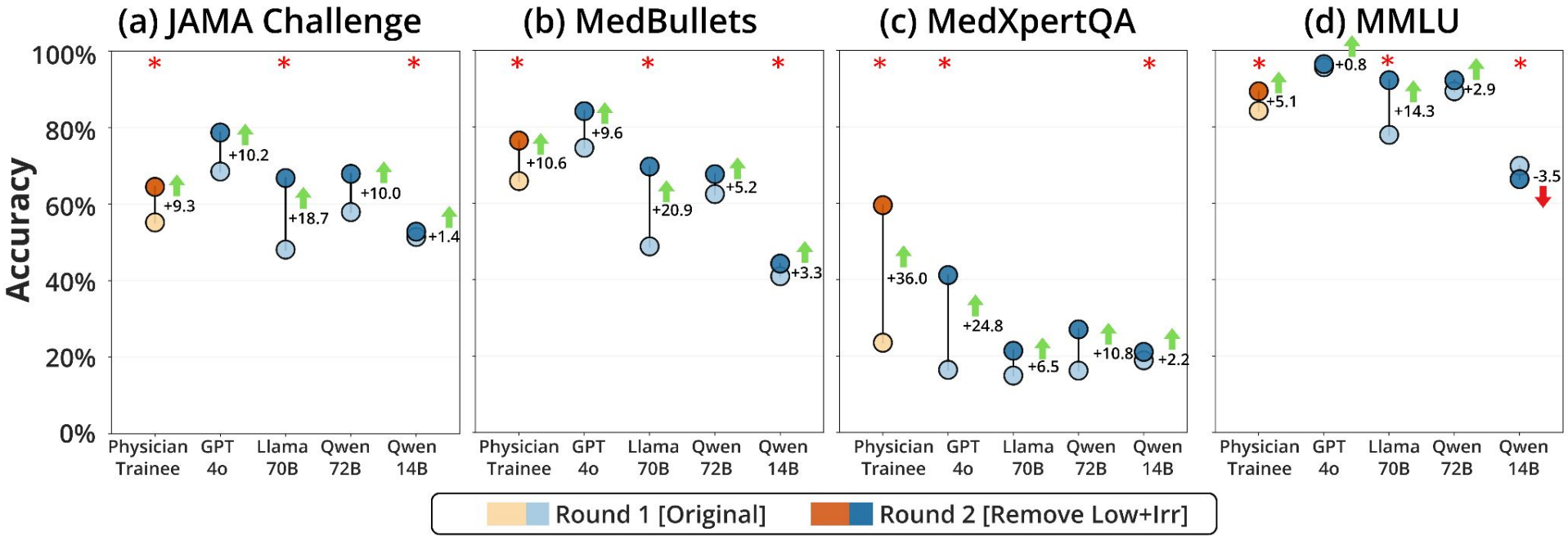
Highly relevant sentences are consistently longer and more uniform in structure

Dataset	Total QA	Total Options	Avg Sentence	Avg Words Per Sentence		Perplexity	
				High	Low/Irr	High	Low/Irr
MMLU (Precision Medicine)	193	4	15.9 (7.0)	18.7 (5.2)	12.8 (4.6)	46.4 (56.3)	58.7 (70.4)
JAMA Clinical Challenge	582	4	26.8 (8.5)	23.1 (5.6)	16.0 (5.4)	51.6 (69.3)	68.2 (92.4)
MedBullets	207	4	21.0 (4.6)	18.1 (4.2)	16.0 (4.3)	46.5 (51.1)	48.3 (65.8)
MedXpertQA	318	10	14.9 (5.6)	21.4 (6.8)	15.6 (4.9)	41.4 (43.8)	52.3 (71.0)
Overall	1300	4/10	21.3 (8.8)	21.2 (6.0)	15.4 (5.1)	48.7 (62.0)	61.0 (82.9)

Humans and LLMs Disagree on Information Relevance

Data Source	Qwen-72B	Llama-70B	Qwen-14B	GPT-4o
	CC	CC	CC	SR
MMLU	26.9 (0.2)	**70.7** (0.2)	56.9 (0.2)	50.5 (0.3)
JAMA	45.5 (0.2)	**62.1** (0.2)	59.1 (0.2)	45.2 (0.3)
MedBullets	49.8 (0.3)	**66.6** (0.2)	53.9 (0.2)	45.2 (0.3)
MedXpertQA	51.8 (0.3)	**69.3** (0.3)	51.9 (0.2)	52.1 (0.4)
Overall	44.9 (0.3)	**65.9** (0.2)	56.2 (0.2)	47.7 (0.3)

Human Relevance Improves LLM Performance



LLM Relevance Estimates Improves LLM QA's Performance

Datasets	MMLU	JAMA	MedBullets	MedXpertQA
Original	95.6	68.5	74.5	16.4
After **Physician Trainee** Labeled Low+Irr Removal	+0.8	+10.2	+9.6	+24.8
After **Qwen-72B** Low+Irr Removal	−1.8	+4.0	+2.3	+24.6
After **Llama-70B** Low+Irr Removal	−2.4	+0.7	+0.1	+22.4
After **GPT-4o** Self-Reported Low+Irr Removal	+1.8	+10.4	+8.6	+8.8

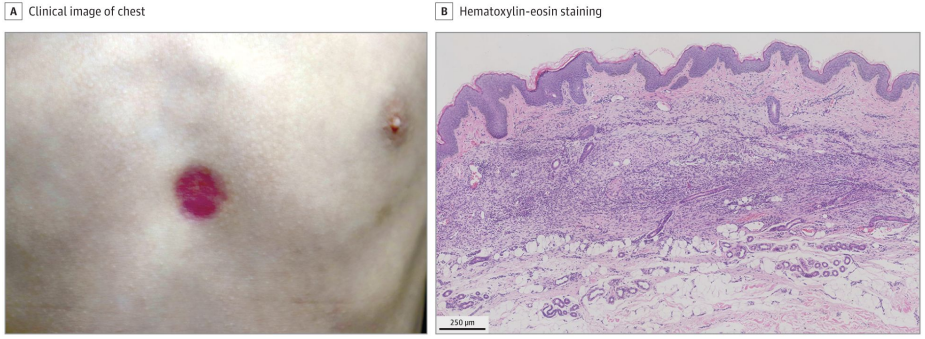
Contributions

MedPAIR is a **first benchmark** step to matching the relevancy annotated by clinical professional labelers to that estimated by LLMs. The motivation for MedPAIR is to ensure that what the LLM finds relevant in a clinical case closely matches what a physician trainee finds relevant.

Next Step?

Patient Profile

A previously healthy 40-year-old man prese[1] with a 3-month history of a right breast mass slowly enlarging with associated pain. The patient denied history of l[2] trauma, and his family history was[2] gative regarding breast or ovarian cancers. The patient had no history of liver disease[3] monal therapy, or radiation to the chest wall. The physical examination disclosed an [4]e man with no signs of hypogonadism[4] liver failure. There was a 3.4-cm hard, irregular, [5]bile nodule in the retroareolar area of th[5]ht breast, tethered to the overlying skin but not fixed to the underlying muscle. [6]here were no other masses on the chest [6].



Query

What is the diagnosis?

A. Gynecomastia
B. Lipoma
C. Carcinoma
D. Epidermal inclusion cyst