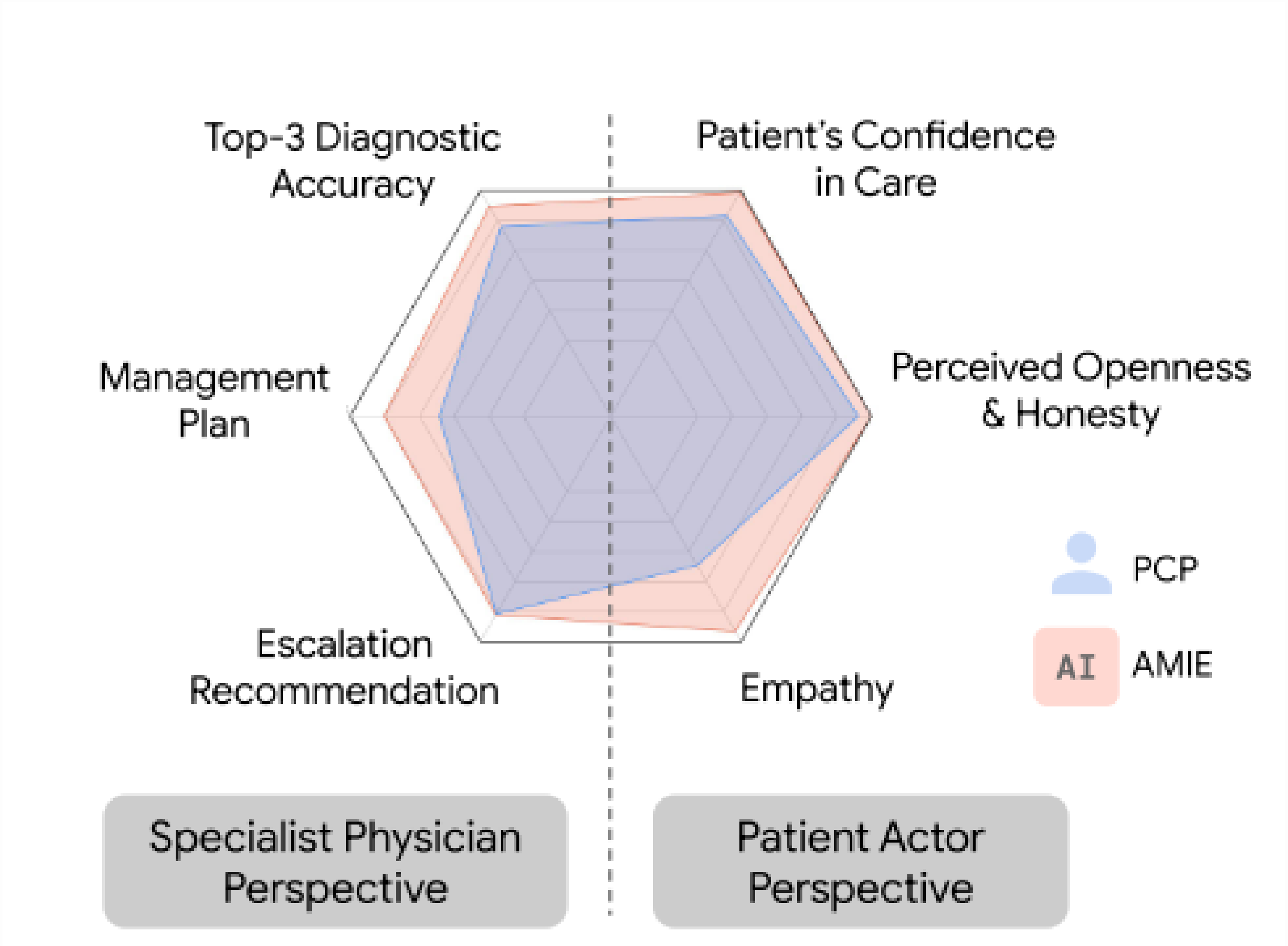
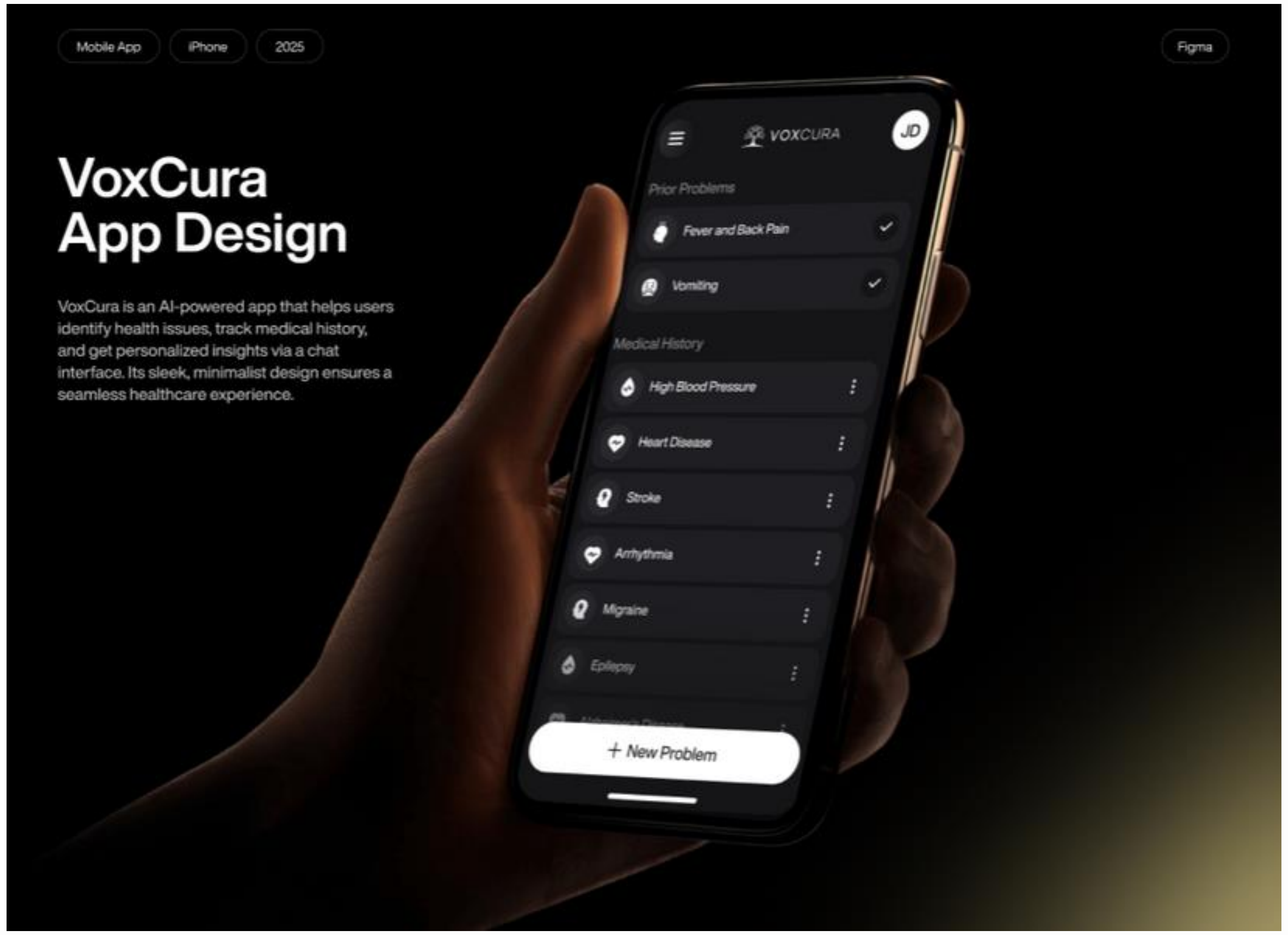


MEDICAL AI

A PRIMER

It is not the strongest who survive, nor the most intelligent, but the one most responsive to change.
— Charles Darwin



DISCLOSURE

Co-Founder of VoxCura

Optimist on AI Technologies

I am not an engineer or CS expert

GOALS

I hope you leave this talk curious and excited about how AI is going to change healthcare

Speed
Things are changing fast

Magnitude
This is likely to impact all aspects of healthcare delivery

Exposure
To teach in this new world, you are going to need to be familiar with these tools and the landscape

TABLE OF CONTENTS

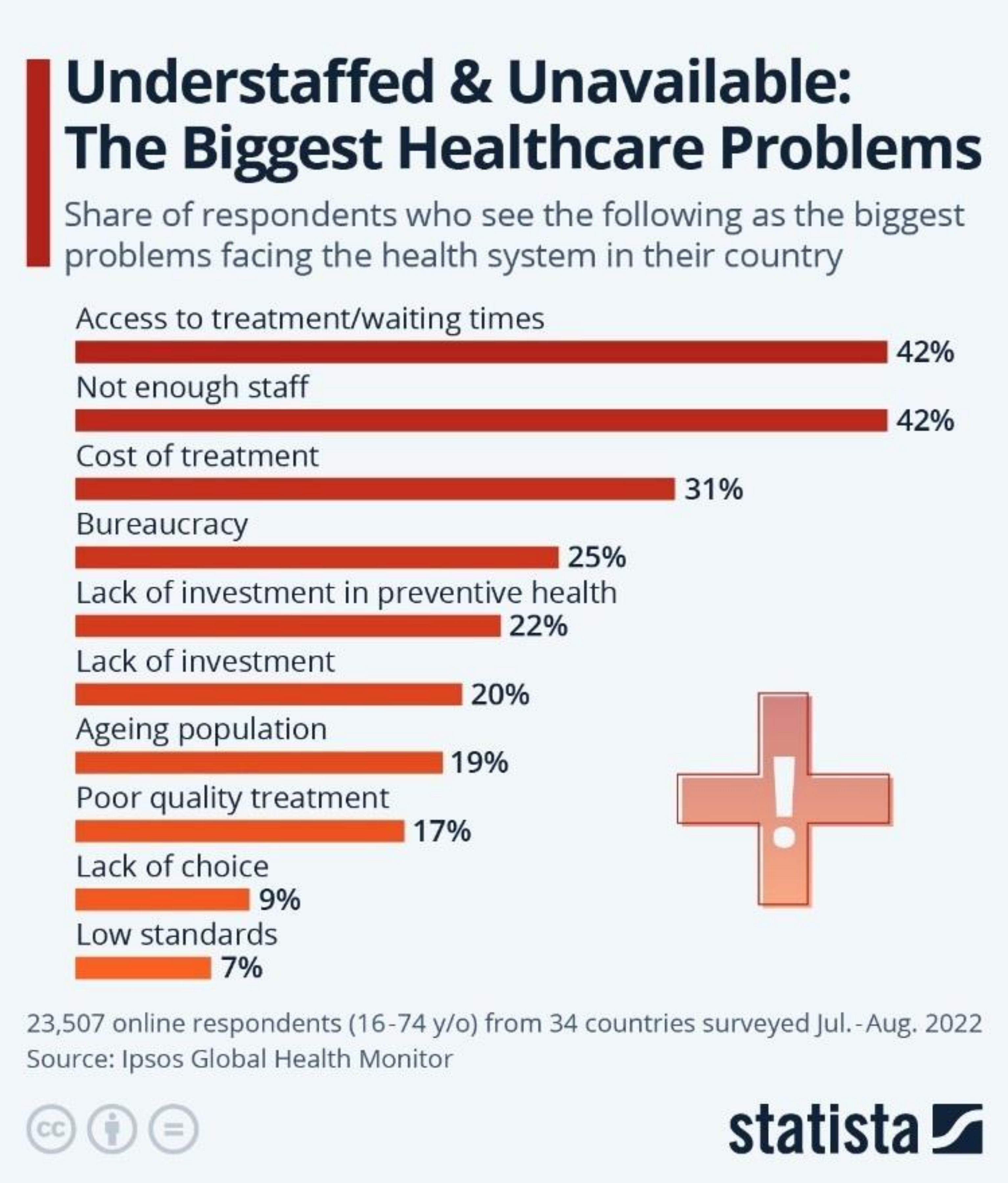
TERMINOLOGY AS KNOWLEDGE

Medical school is really learning a language to describe the form and function of the human body

Most knowledge is learning how to accurately describe the world

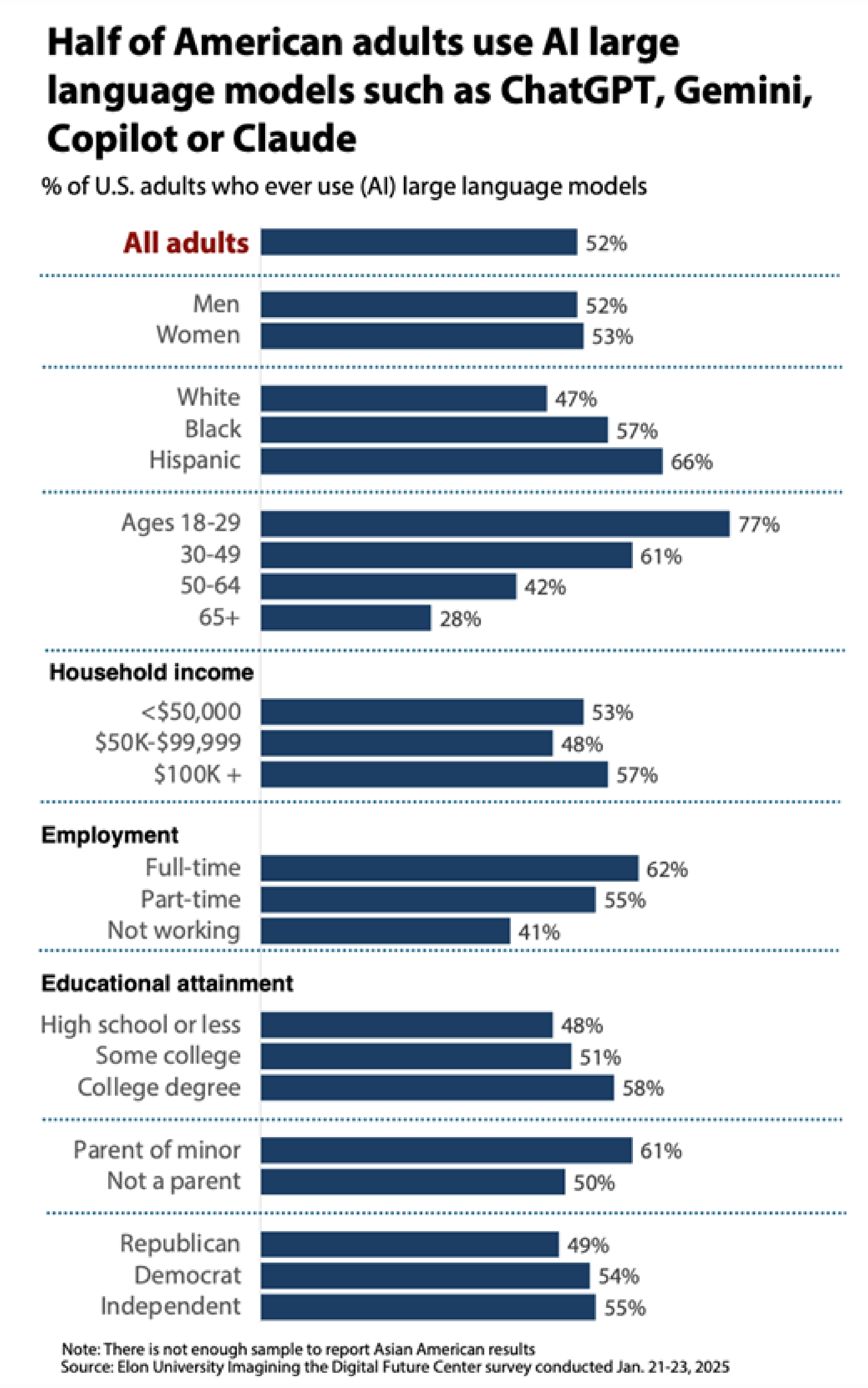
Large Language Models are complex programs that use machine learning to link semantic patterns

What LLMs do is not too dissimilar to what you are doing when you learn to be a physician



WHY AI MATTERS

Only one of these that AI cannot help with



01 Over half of americans us LLMs

02 Almost 40% of those are using it for health information.

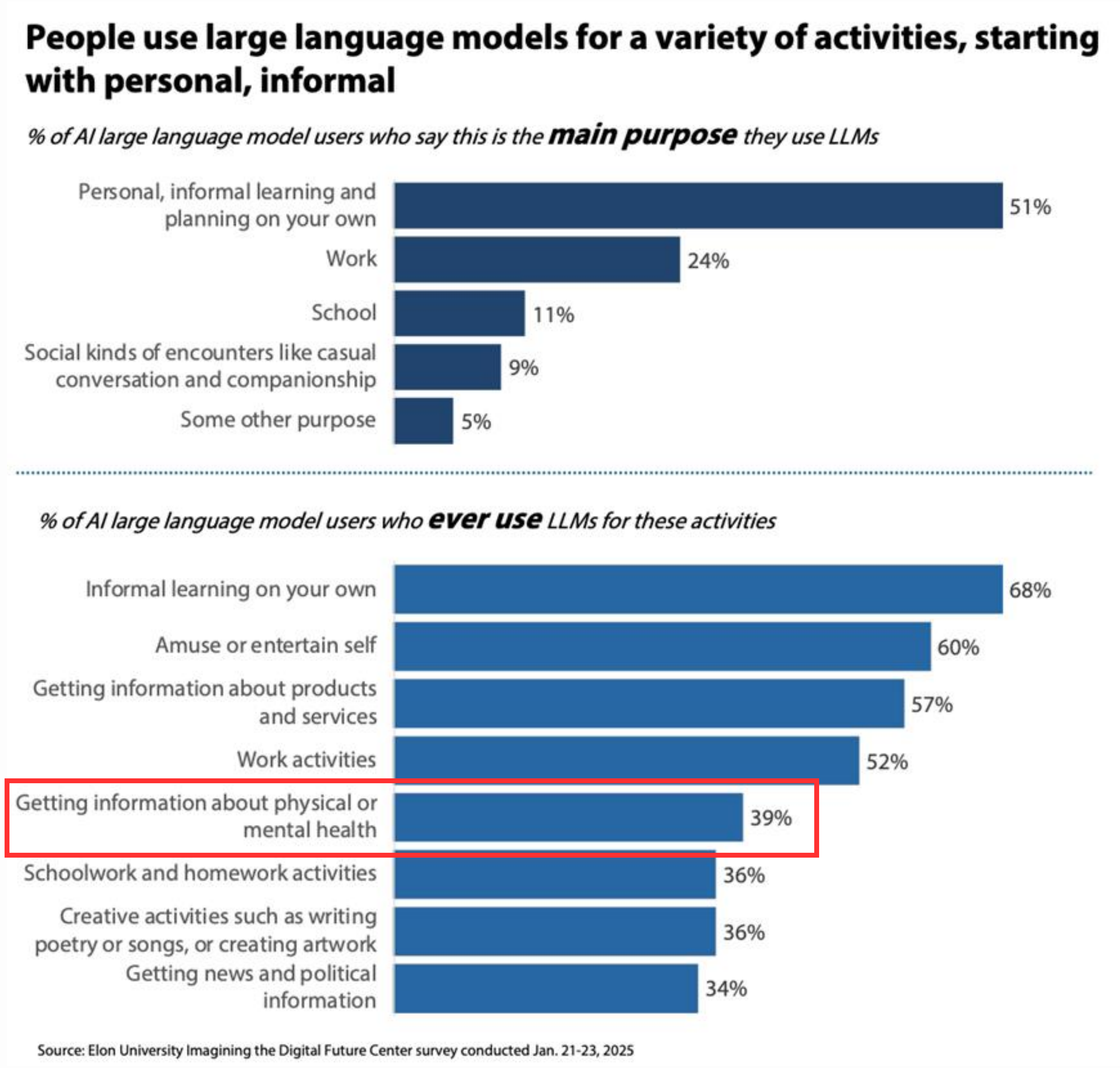
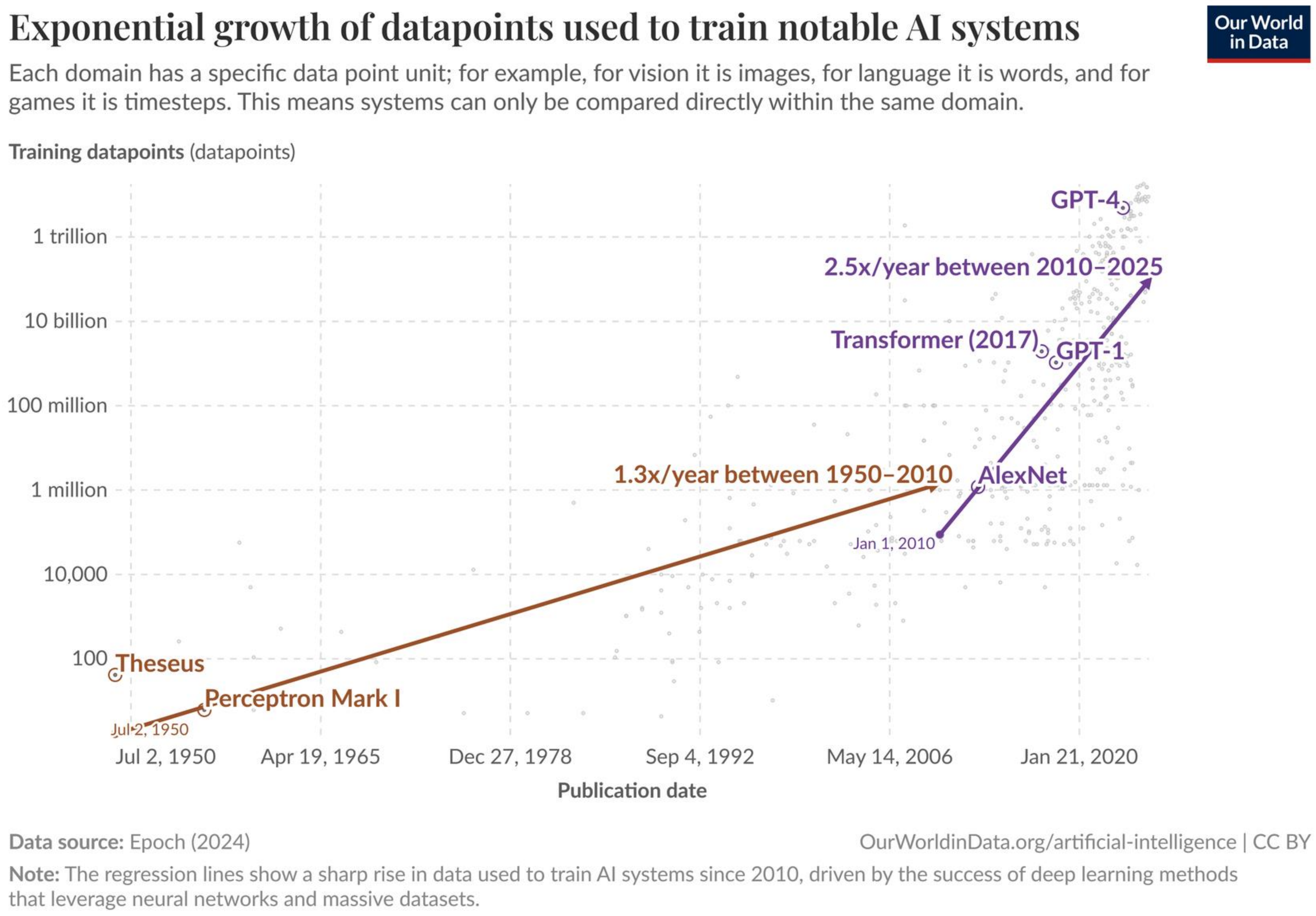


TABLE OF CONTENTS

HISTORY



1950s

Theoretical foundations for neural networks

1956 - Summer Research Project on Artificial Intelligence at Dartmouth College

2013-2016

The Rise of Embeddings

HISTORY

2017-2019

2017 - Transformer Revolution - "Attention is All You Need" - New Neural Network Training -> Pre-train then fine tune

2018 - Open AI Releases GPT (Generative Pre-trained Transformer)

2019 - GPT 2 (Q's, Short translation, small calculations)

2020-2022

2020 - OpenAIs GPT 3 with 175 billion parameters released

2022 - ChatGPT (GPT 3.5) launched bringing mainstream attention to LLMs (Complex questions, tell stories, simple software)

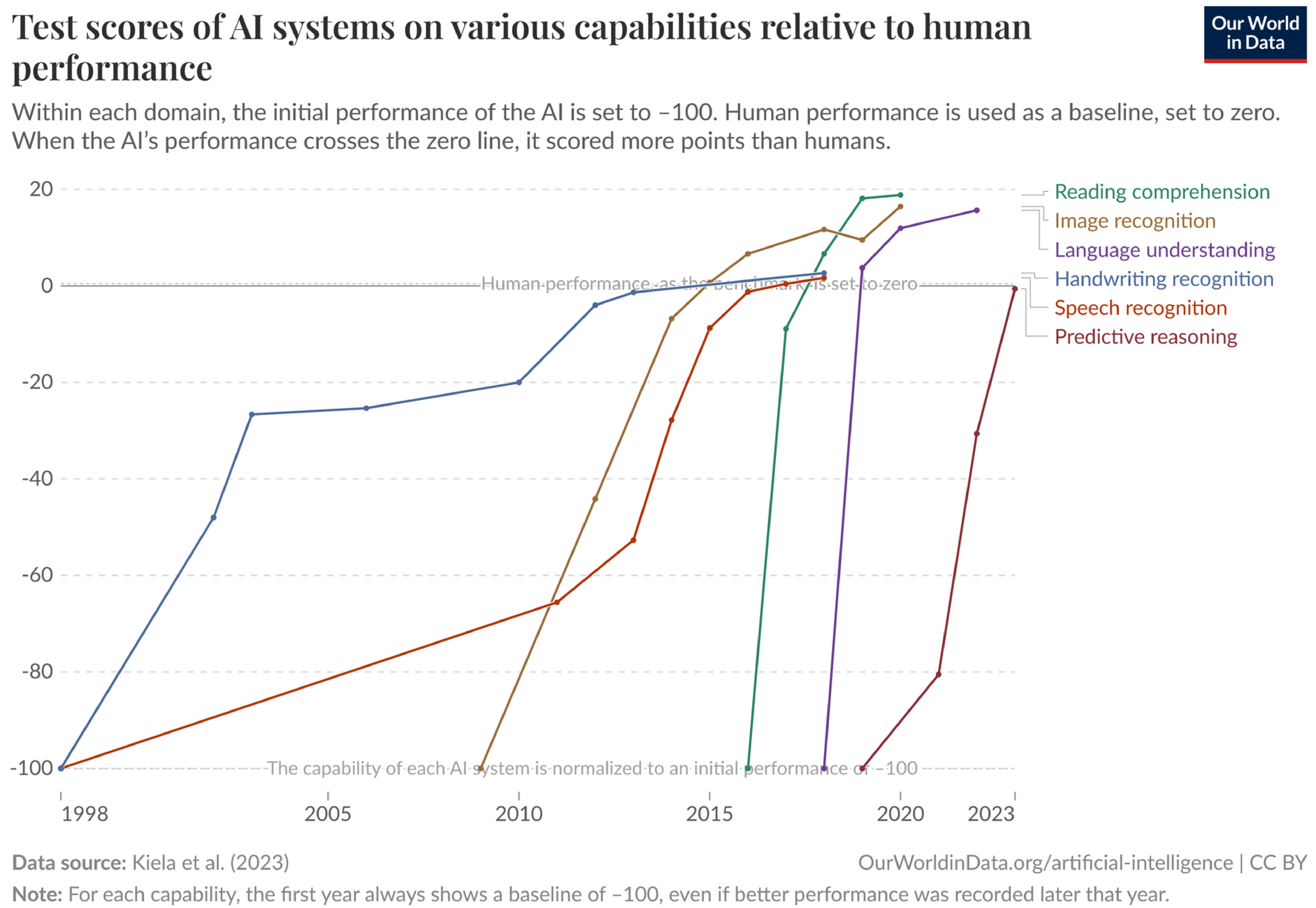
2023-2025

Mulit-agent models introduced, CoT

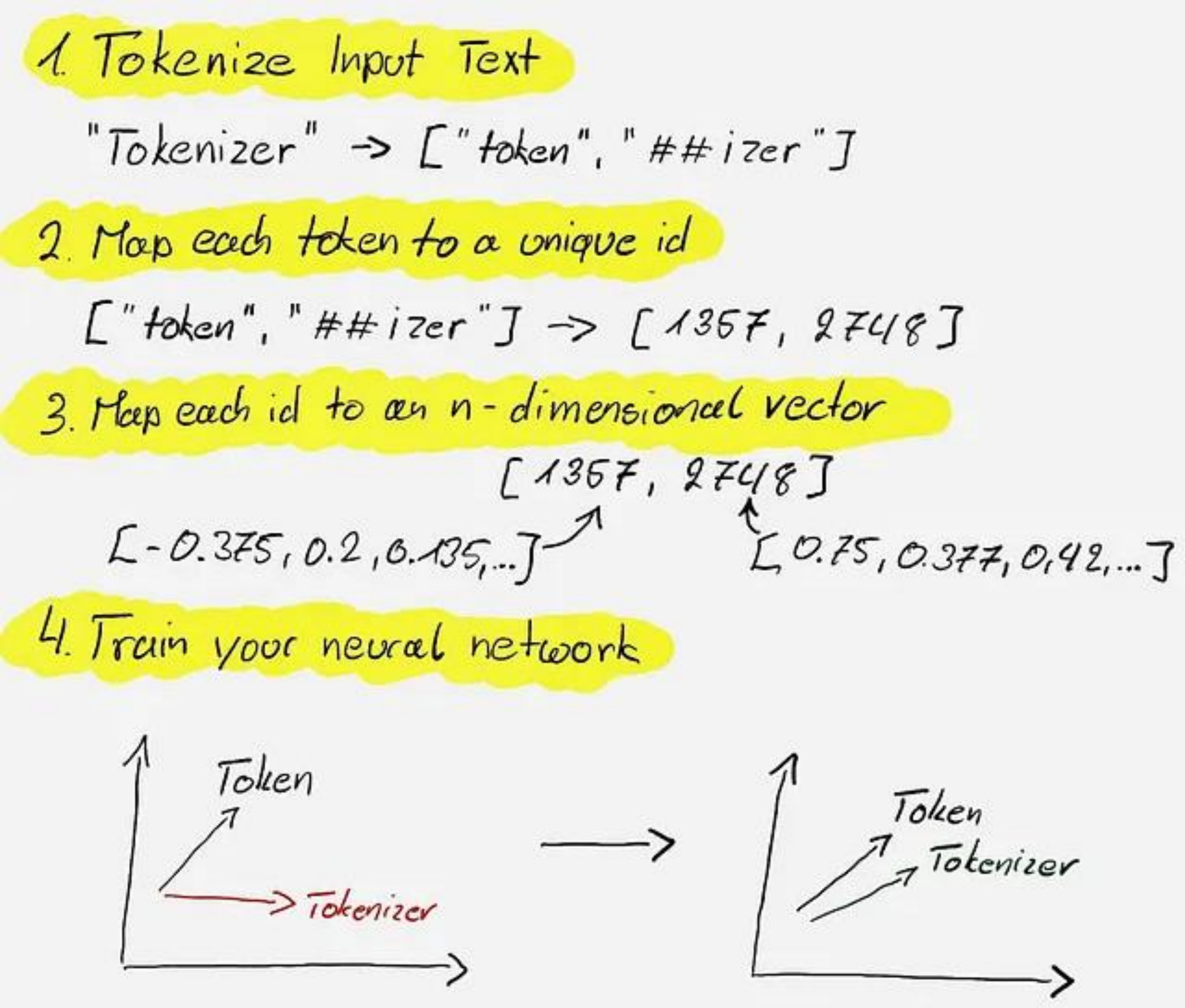
2025 - Ace PhD level exams, Code entire applications, Perfectly emulate human voices

List of underperformance vs humans is shrinking

HISTORY


Test scores of AI systems on various capabilities relative to human performance
Our World in Data
Within each domain, the initial performance of the AI is set to −100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.
20
0
-20
-40
-60
-80
-100
Human performance, as the benchmark, is set to zero
The capability of each AI system is normalized to an initial performance of −100
1998
2005
2010
2015
2020
2023
Reading comprehension
Image recognition
Language understanding
Handwriting recognition
Speech recognition
Predictive reasoning
Data source: Kiela et al. (2023)
OurWorldinData.org/artificial-intelligence | CC BY
Note: For each capability, the first year always shows a baseline of −100, even if better performance was recorded later that year.

STRUCTURE



Words in corpus are **tokenized**

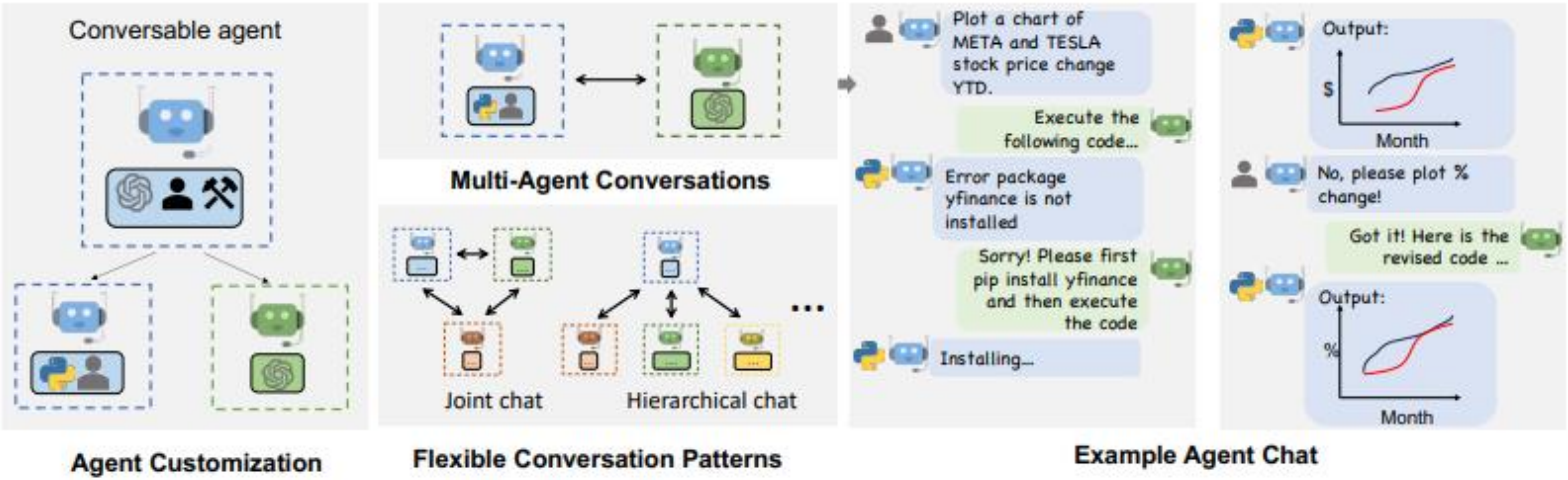
Embedded into high dimensional **vector** space

High dimensional vector space allows for ML or Neural Network

Unsupervised Learning → RLHF


Intelligence is the ability to learn, adapt, solve problems, and use reasoning to overcome challenges. It encompasses various cognitive abilities, including logical reasoning, abstract thinking,
role in
It involves cognitive processes such as reasoning, memory,
logical reasoning, abstract thinking, memory,
It involves skills such as critical

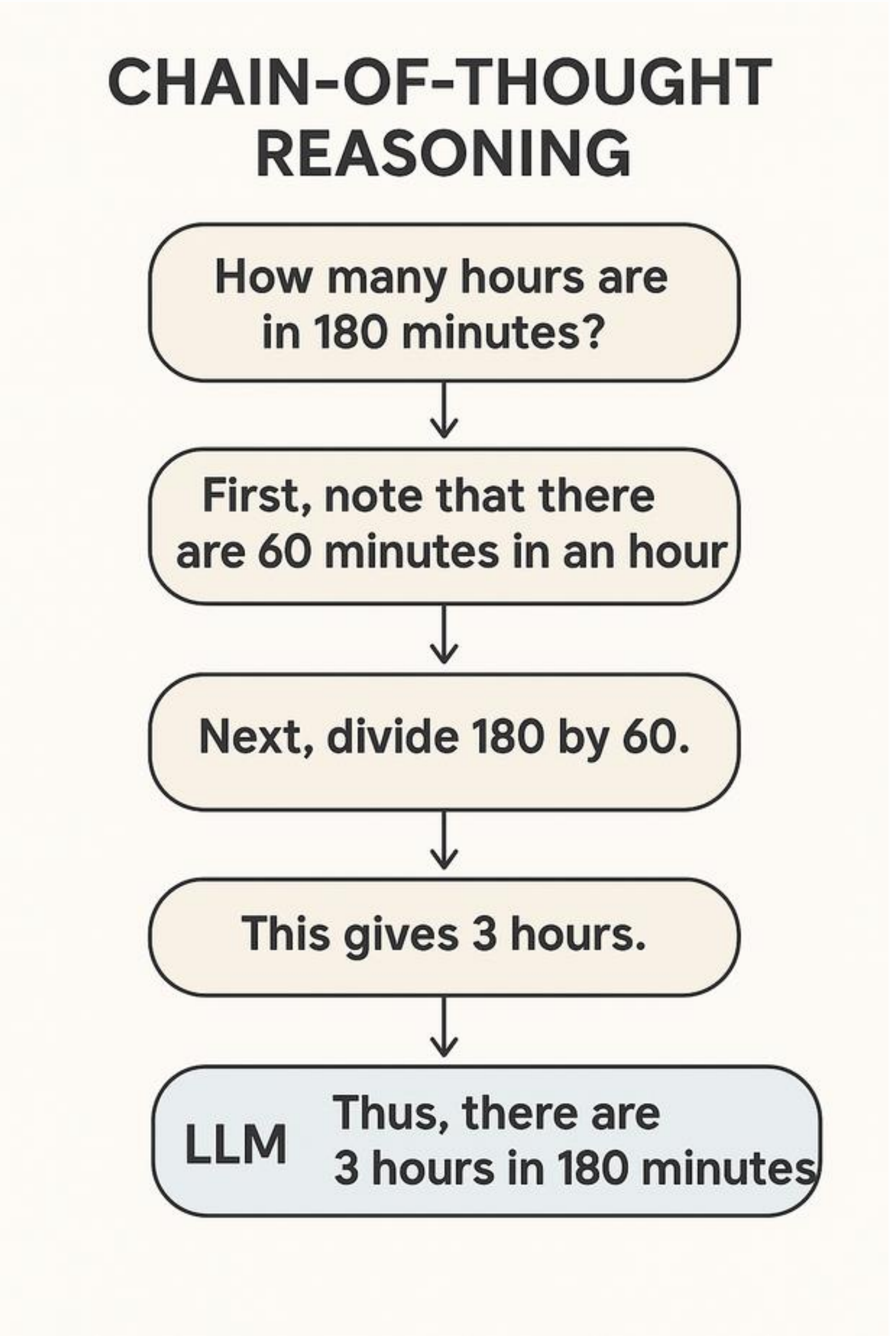
MIND MAP



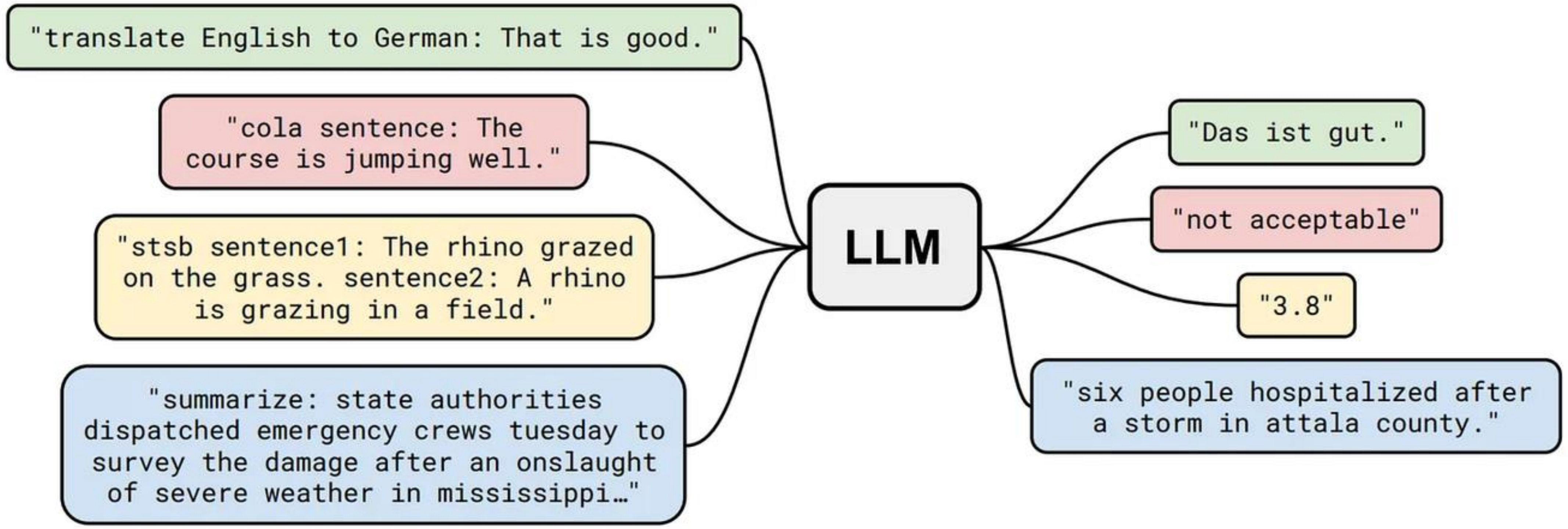
MULTI-AGENT MODELS

Agency can be given to a LLM by equipping it with tools (search engine access, a database, code base, etc.)

Multiple agents with different roles can be given a task to complete. These agents can have different goals and hierarchical structures.



CHAIN OF THOUGHT



PROMPTING

Setting the stage for a LLM alters the output

Assign a role and provide context

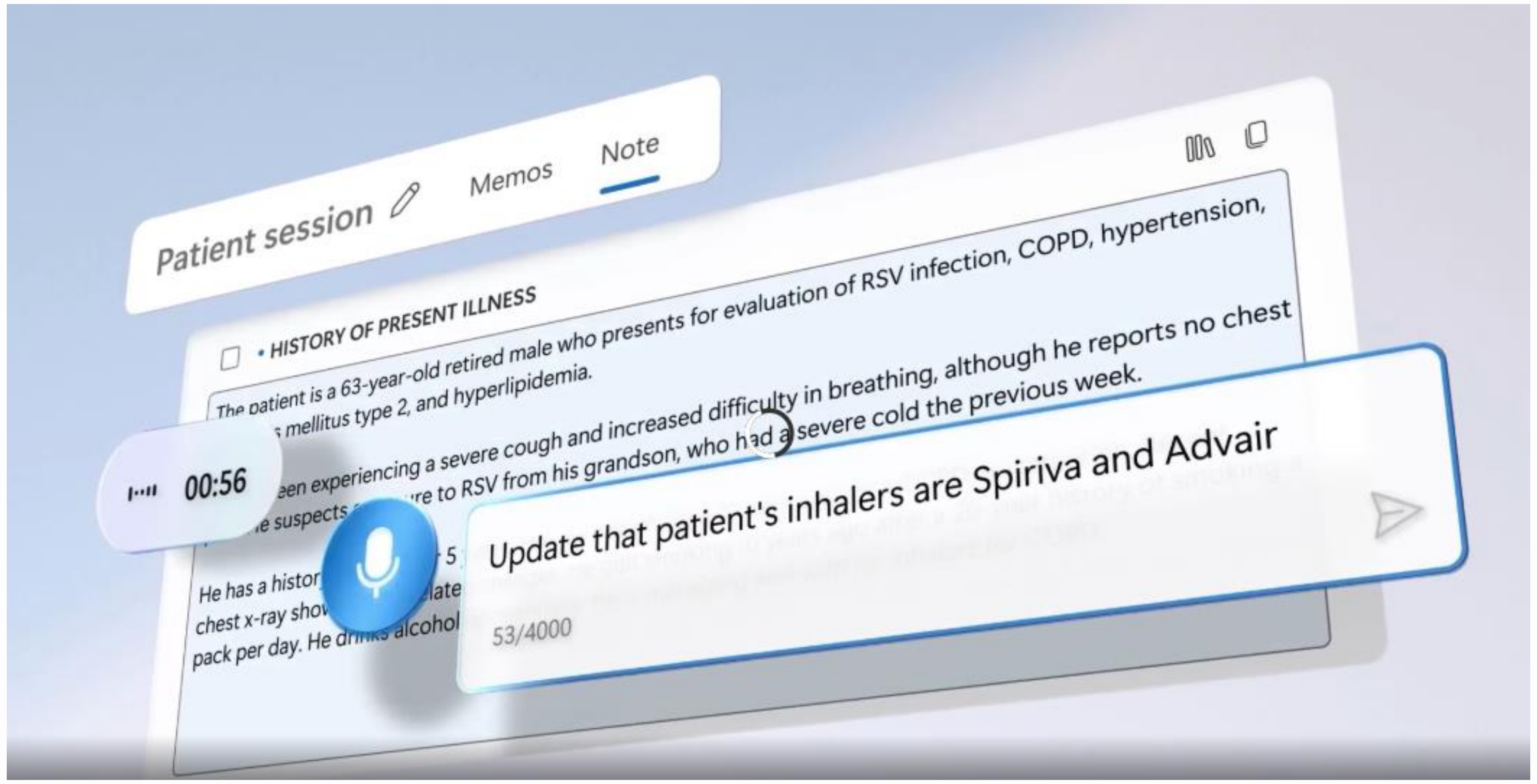
Place critical instructions at the beginning and end (bias)

Adding directions like “think step by step” can improve accuracy

Be clear and specific

AMBIENT RECORDING

Likely a multi-agent framework with text-to-speech.



Automated Notes
Drafted encounter notes

Accuracy
Enhances accuracy of the encounter and discussion points

Time Saving
Will decrease documentation burden

TABLE OF CONTENTS

CONSIDERATIONS

Bias

Representation Bias - Cultural biases in training data
Temporal Bias - Newer data more heavily weighted
Frequency Bias - Common patterns
Positional Bias - Information at beginning or end
Authority Bias - Overweight authoritative content
Coherency Bias - Coherent sounding responses

Data Use

Terms give permission to use data entered (Opt)

Privacy

While you can get HIPAA compliant agreements they are reserved for businesses.
DO NOT PUT IN IDENTIFIERS

CONSIDERATIONS

Hallucinations

Less common with newer models

Confabulations

Consent

Should be part of the informed process.

Especially Ambient recordings.

Energy Use

An hour of streaming Netflix is equivalent to 70-90,000 Llama 65B tokens

> Claude-3.5 and o1 consistently emerged as the top-performing models across this evaluation, exhibiting the lowest hallucination rates across all tasks and risk categories. Remarkably, both models achieved a 0% hallucination rate in the Diagnosis Prediction task, suggesting a high degree of reliability for diagnostic inference within this specific context. Claude-3.5 demonstrated exceptionally low hallucination rates of 0.5% (Chronological Ordering) and 0.25% (Lab Data Understanding). o1 mirrored this strong performance, with equally low or slightly superior rates of 0.25% for both Chronological Ordering and Lab Data Understanding.

CONSIDERATIONS

Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models

25 March 2025 | Publication

Download (1.2 MB)

Overview

Artificial Intelligence (AI) refers to the capability of algorithms integrated into systems and tools to learn from data so that they can perform automated tasks without explicit programming of every step by a human. Generative AI is a category of AI techniques in which algorithms are trained on data sets that can be used to generate new content, such as text, images or video. This guidance addresses one type of generative AI, large multi-modal models (LMMs), which can accept one or more type of data input and generate diverse outputs that are not limited to the type of data fed into the algorithm. It has been predicted that LMMs will have wide use and application in health care, scientific research, public health and drug development. LMMs are also known as "general-purpose foundation models", although it is not yet proven whether LMMs can accomplish a wide range of tasks and purposes.

Other language:

Korean

Read more

Emergency use of unproven clinical interventions outside clinical trials: ethical considerations

Emerging technologies

TABLE OF CONTENTS

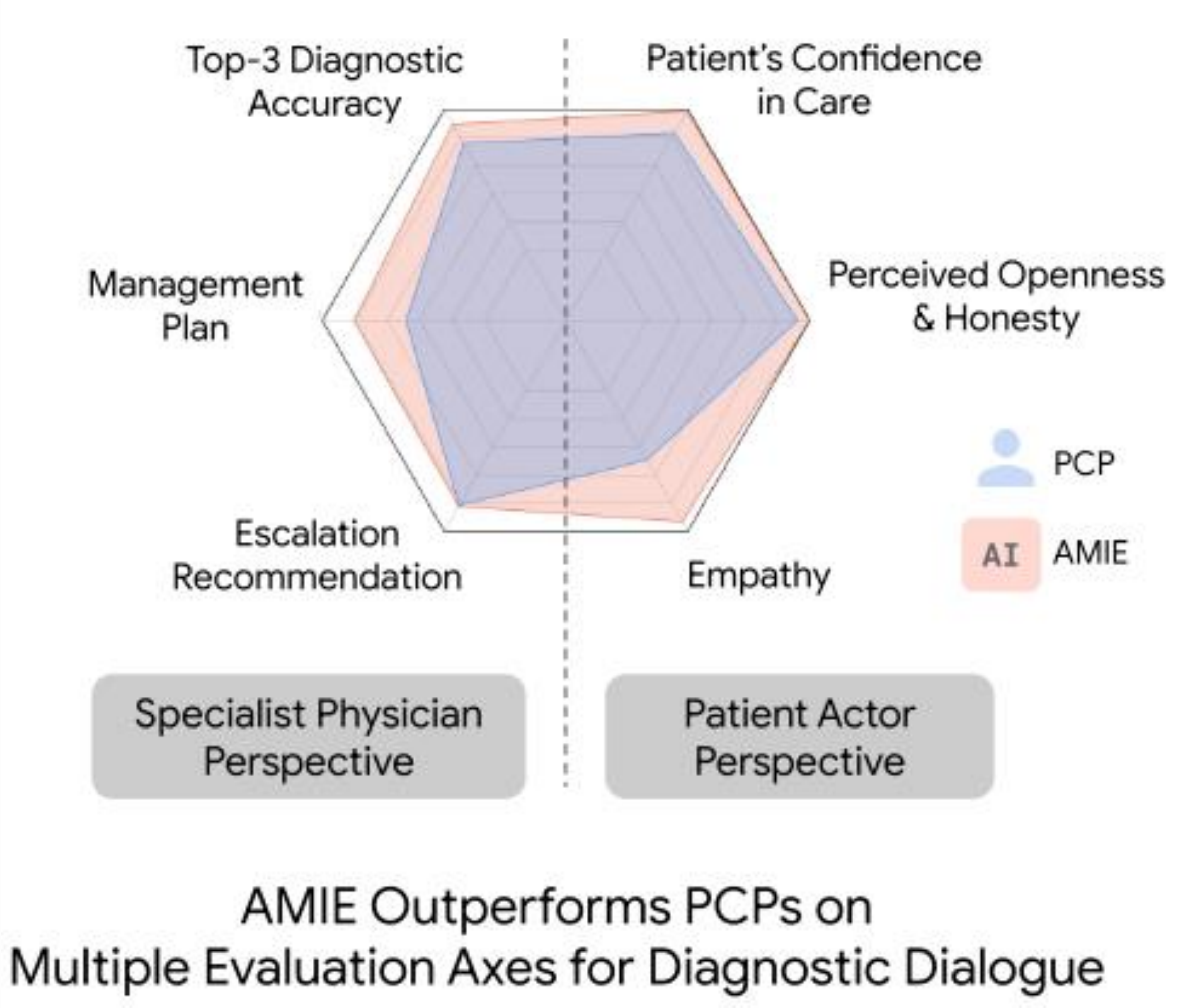
Towards Conversational Diagnostic AI

Tao Tu[*,1], Anil Palepu[*,1], Mike Schaekermann[*,1],
Khaled Saab[1], Jan Freyberg[1], Ryutaro Tanno[2], Amy Wang[1], Brenna Li[1], Mohamed Amin[1],
Nenad Tomasev[2], Shekoofeh Azizi[2], Karan Singhal[1], Yong Cheng[2], Le Hou[1], Albert Webson[2],
Kavita Kulkarni[1], S. Sara Mahdavi[2], Christopher Semturs[1],
Juraj Gottweis[1], Joelle Barral[2], Katherine Chou[1], Greg S. Corrado[1], Yossi Matias[1],
Alan Karthikesalingam[†,1] and Vivek Natarajan[†,1]

[1]Google Research, [2]Google DeepMind

WARNING

AMIE



01 LLMs can already outperform PCPs according to patient actors and specialists (160)

AMIE

02 Diagnostic accuracy has already been surpassed

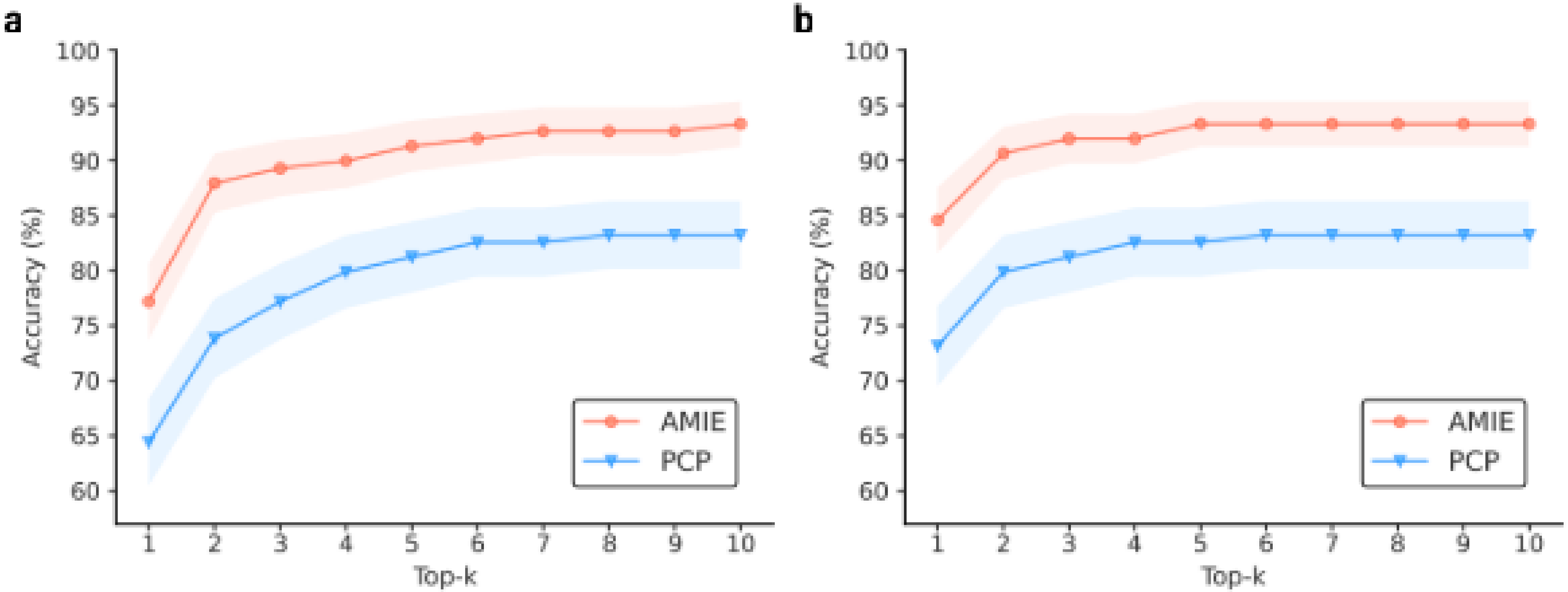
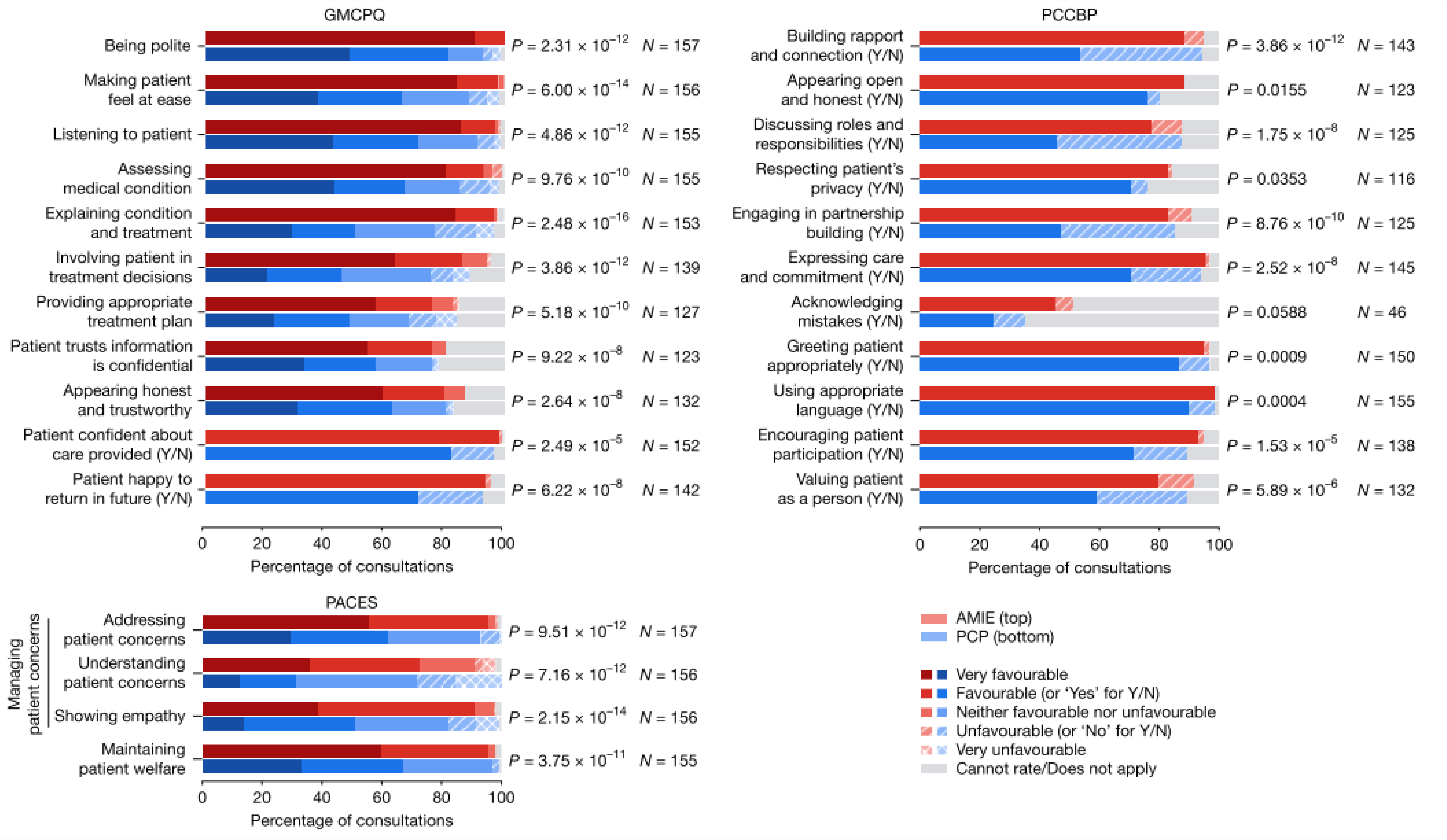


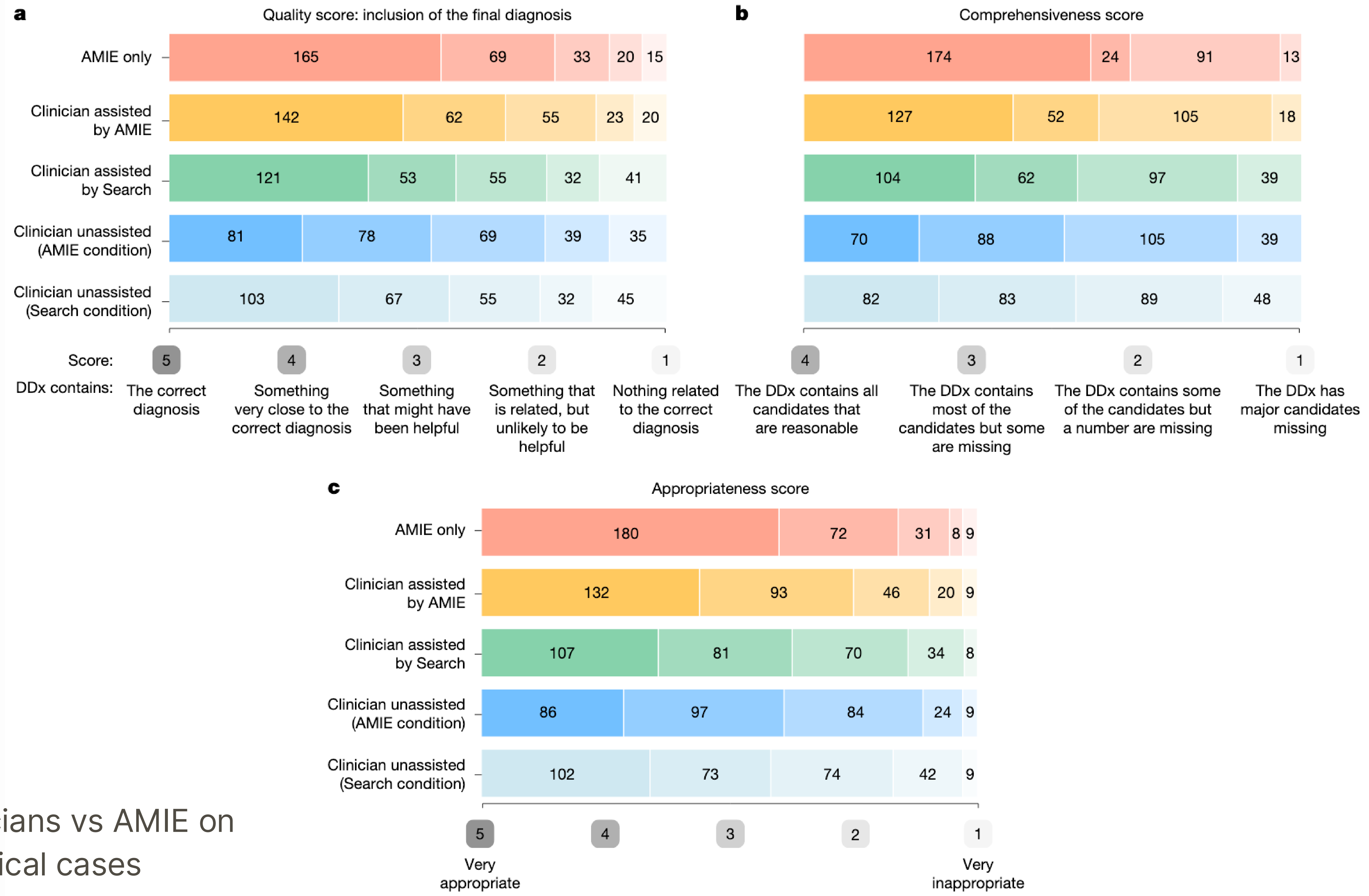
Figure 3 | Specialist-rated top-k diagnostic accuracy. AMIE and PCPs top-k DDx accuracy are compared across 149 scenarios with respect to the ground truth diagnosis (**a**) and all diagnoses in the accepted differential (**b**). Bootstrapping (n=10,000) confirms all top-k differences between AMIE and PCP DDx accuracy are significant with $p < 0.05$ after FDR correction.

AMIE

03 Majority of axes outperformed

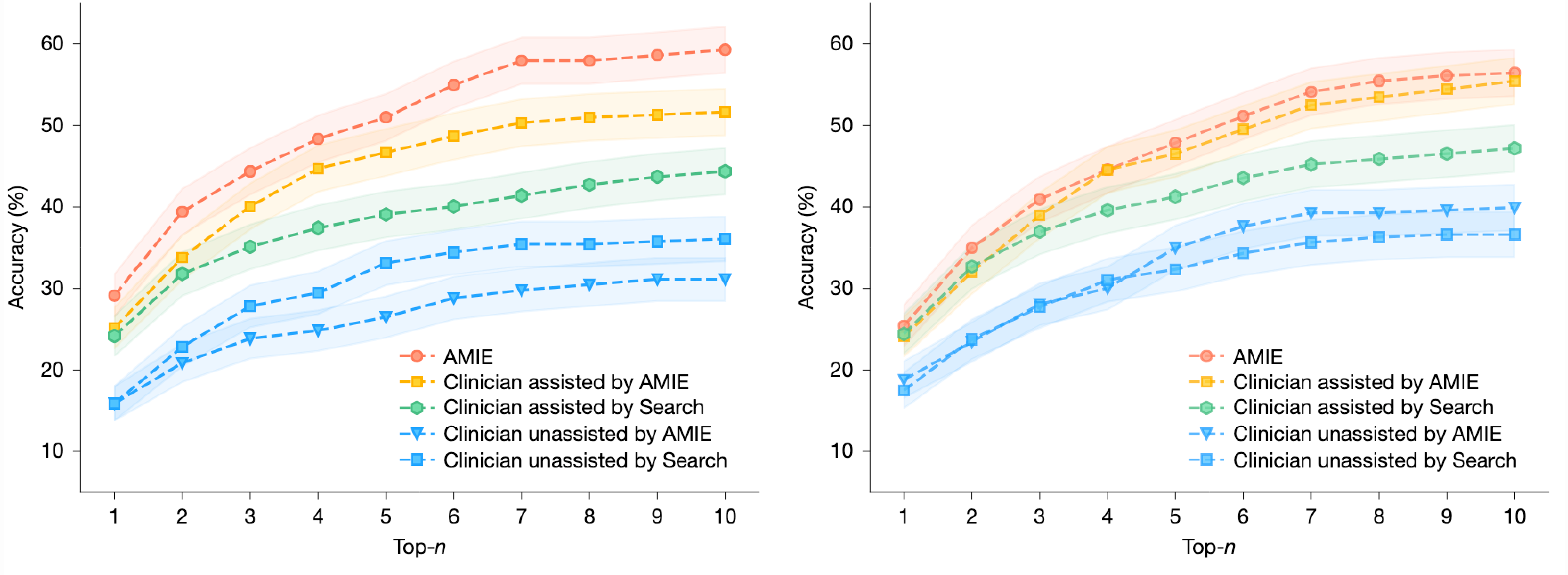


AMIE



01 Twenty Clinicians vs AMIE on 302 real medical cases

AMIE



Accuracy (%)
Top-n
AMIE
Clinician assisted by AMIE
Clinician assisted by Search
Clinician unassisted by AMIE
Clinician unassisted by Search

O1-PREVIEW

Figure 1. Performance of Differential Diagnosis Generators and LLMs on *NEJM* Clinicopathologic Case Conferences (CPCs) 2012-2024

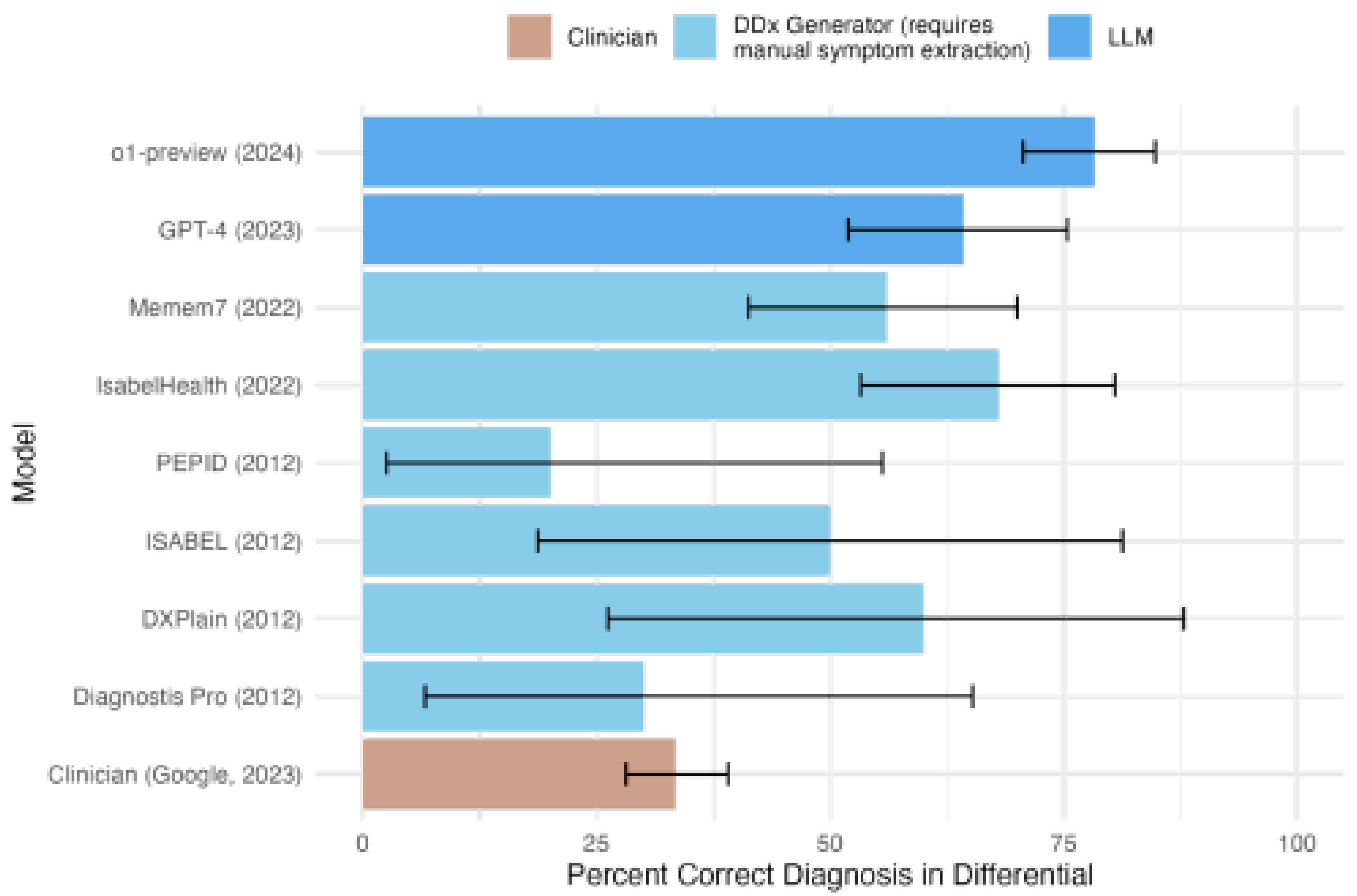
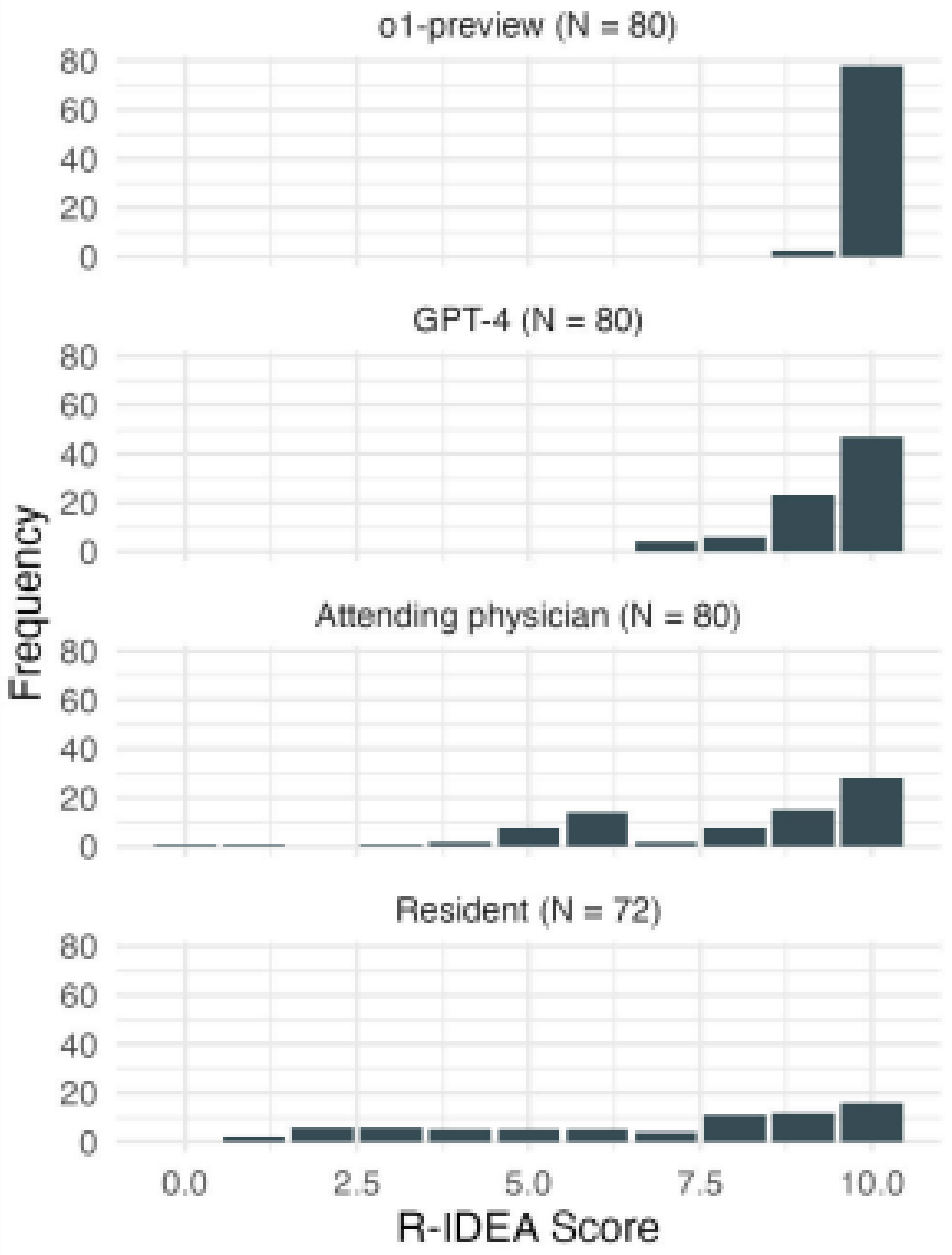
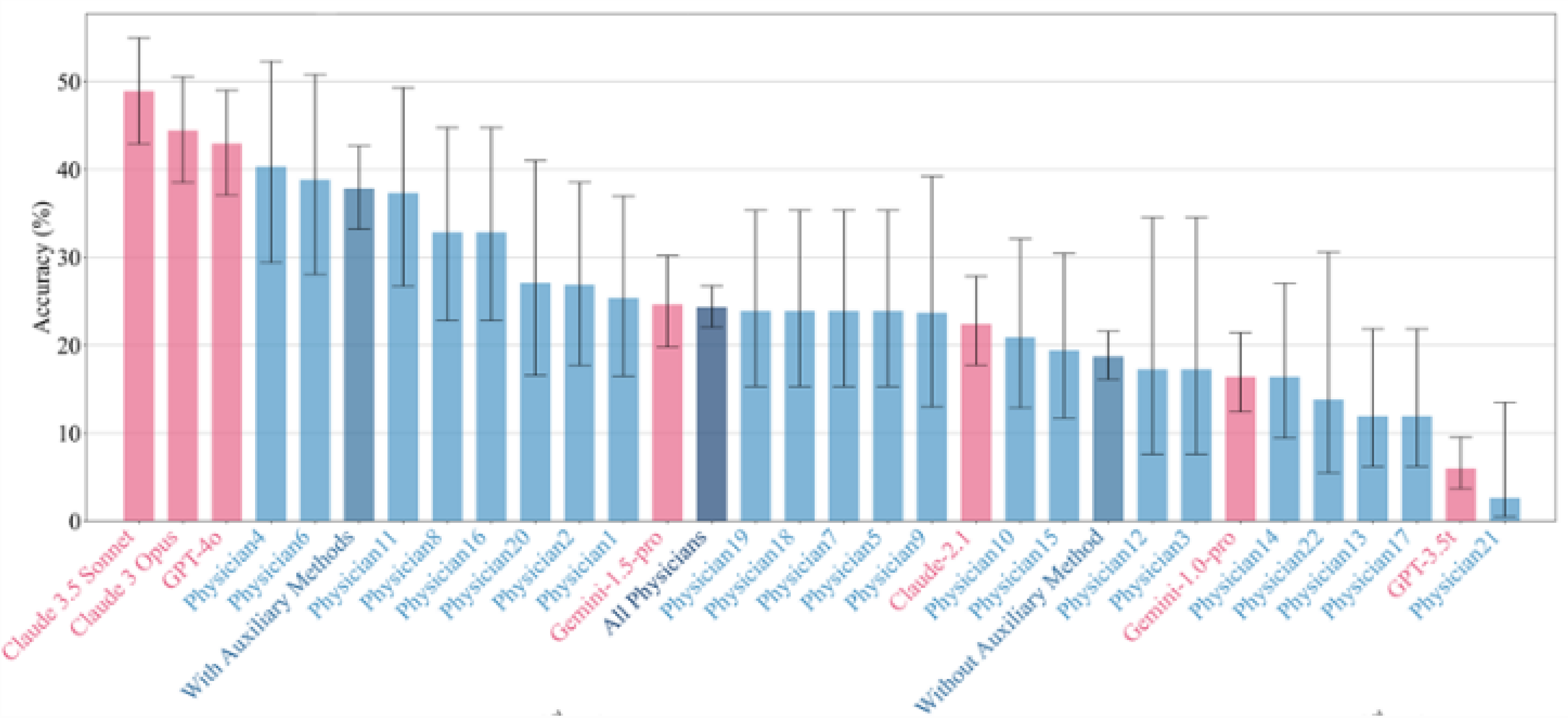


Figure 1: Barplot showing the accuracy of including the correct diagnosis in the differential for differential diagnosis (DDx) generators and LLMs on the *NEJM* CPCs, sorted by year. Data for other LLMs or DDx generators was obtained from the literature.[36 23 8] The 95% confidence intervals were computed using a one-sample binomial test.

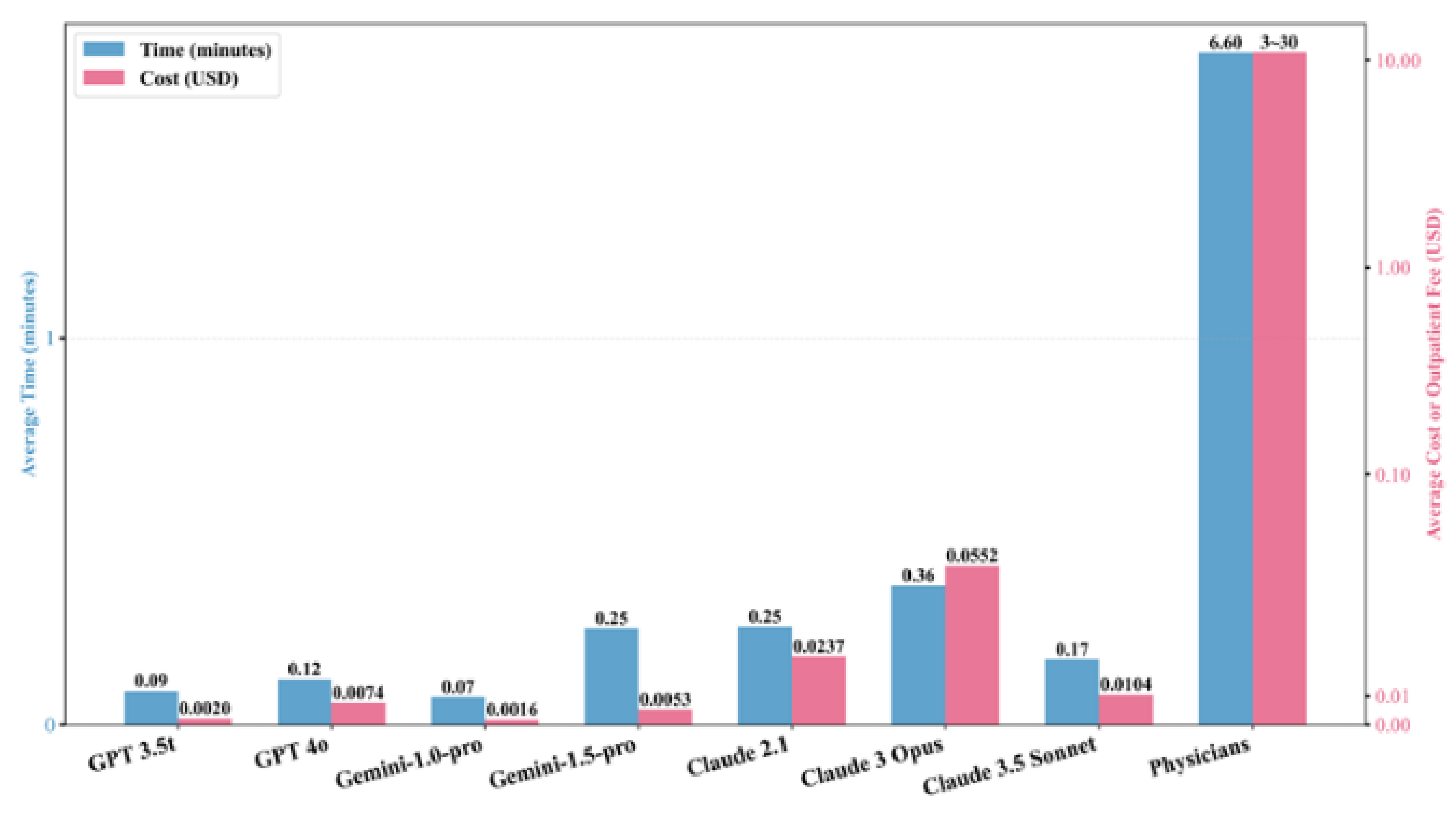


FRONTIER MODELS



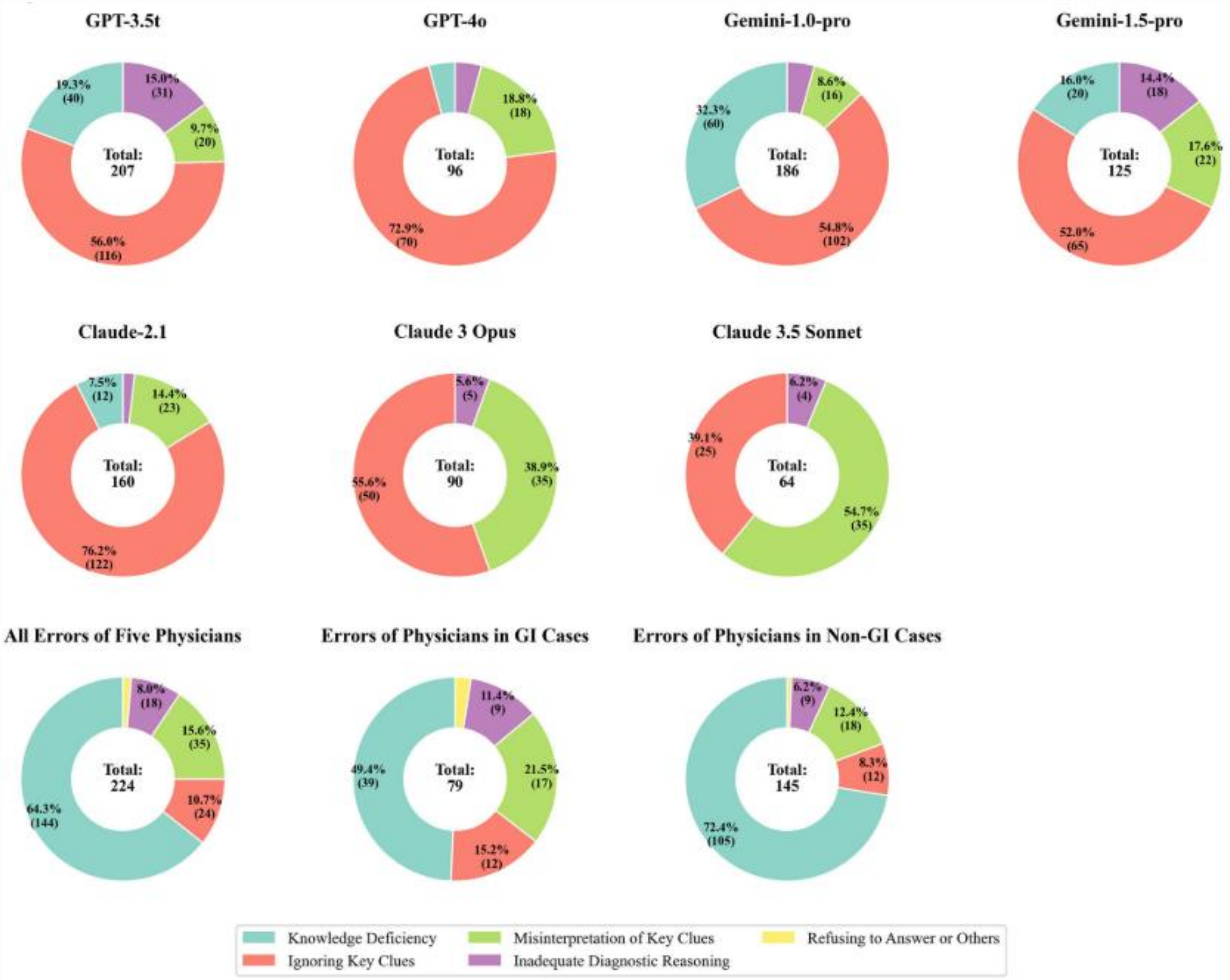
01 LLMs outperform specialists in their own specialty*

FRONTIER MODELS

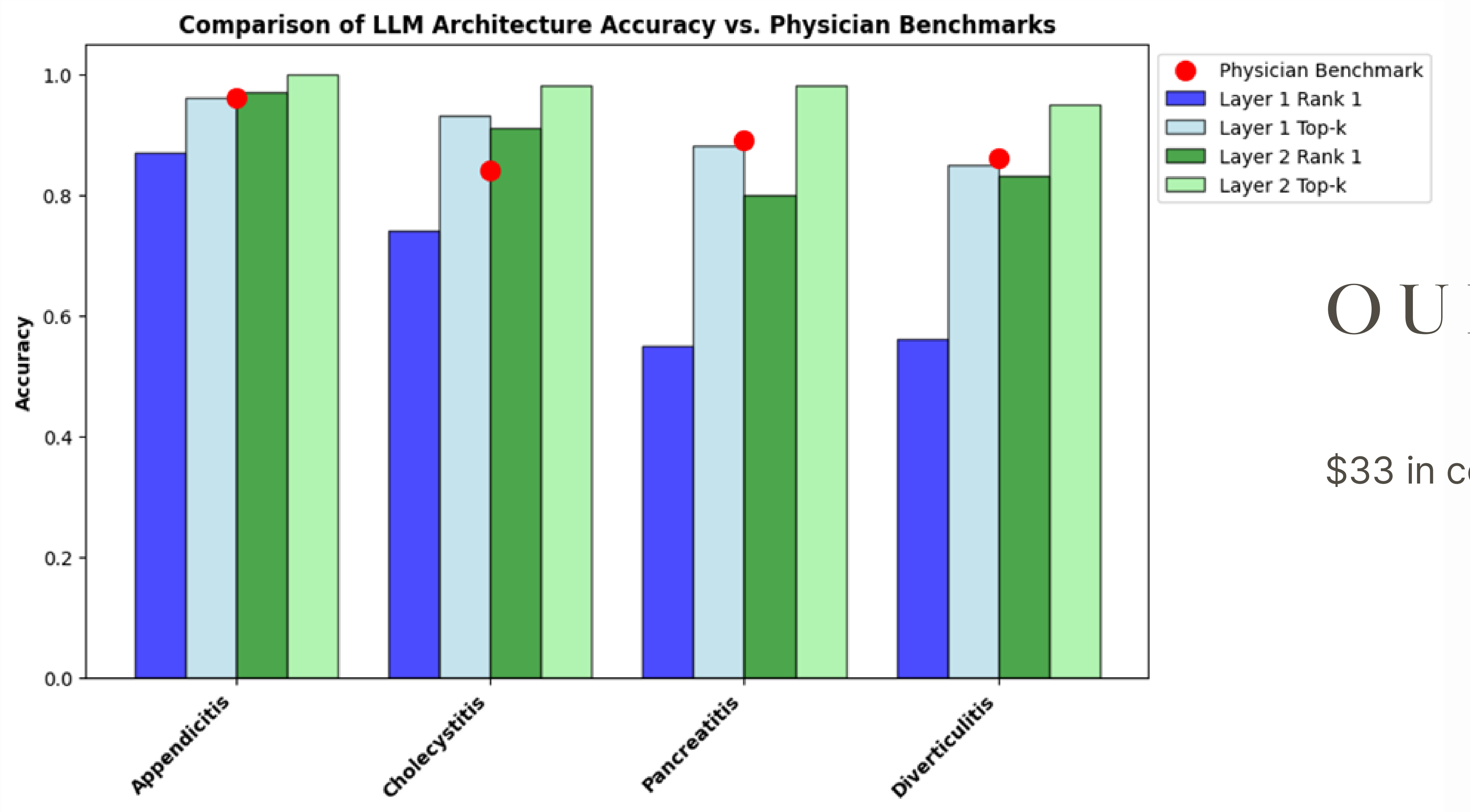


02 They do it faster and cheaper

FRONTIER MODELS



03 They fail from modifiable factors

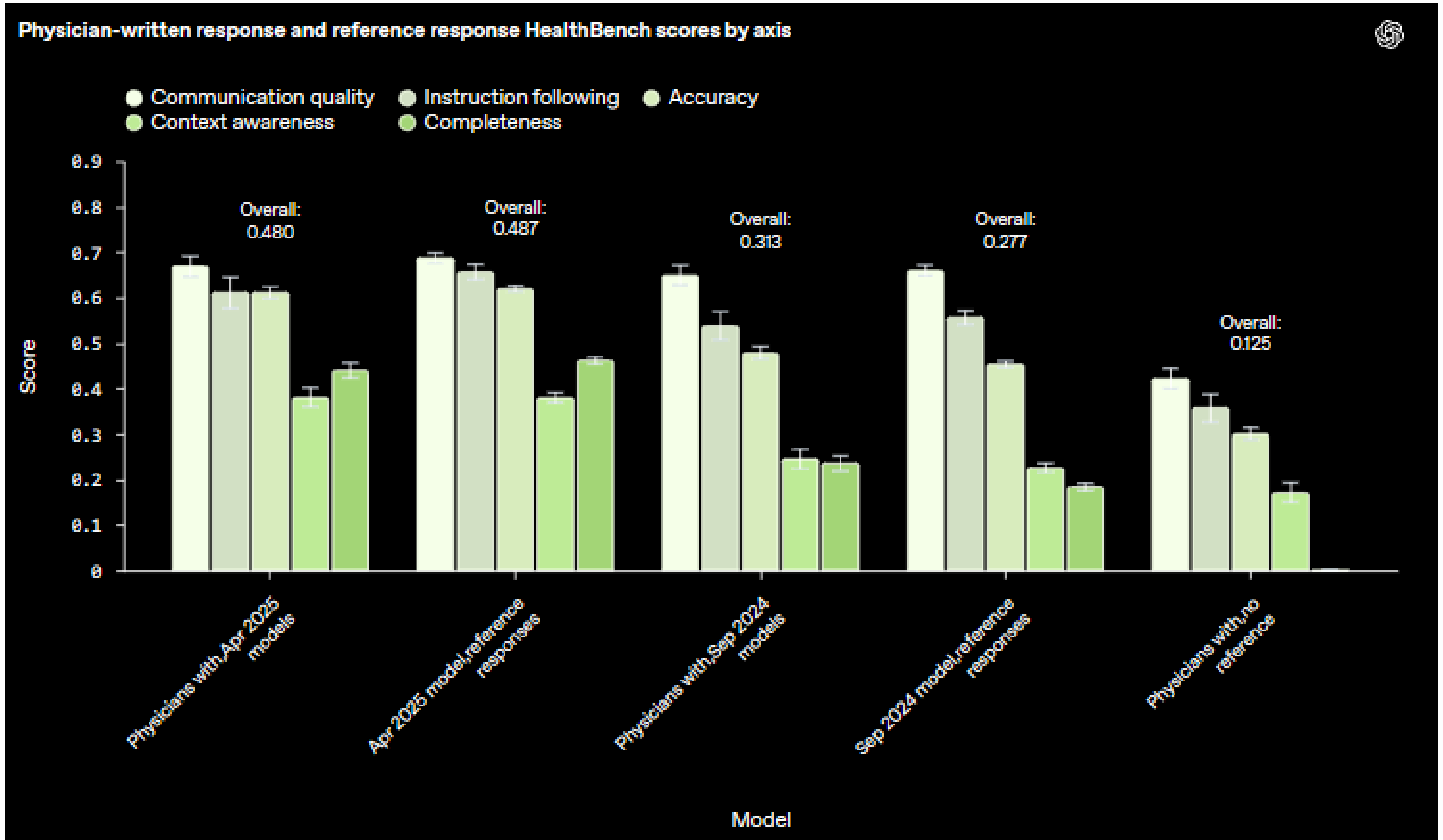


OUR DATA

$33 in compute cost, 1 hour in compute time. 2400 cases.

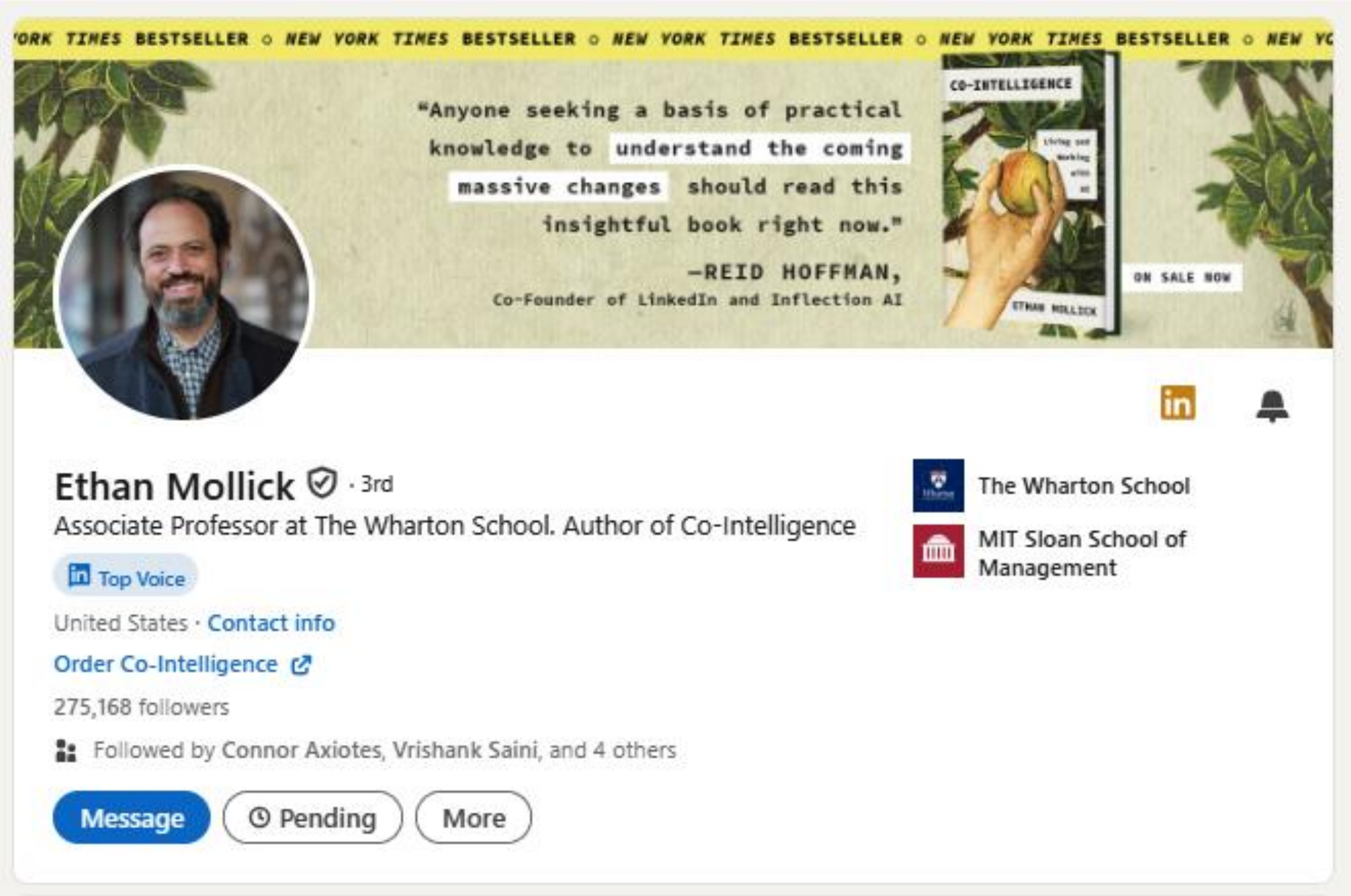
HEALTHBENCH

01 Physicians no longer improve quality of responses

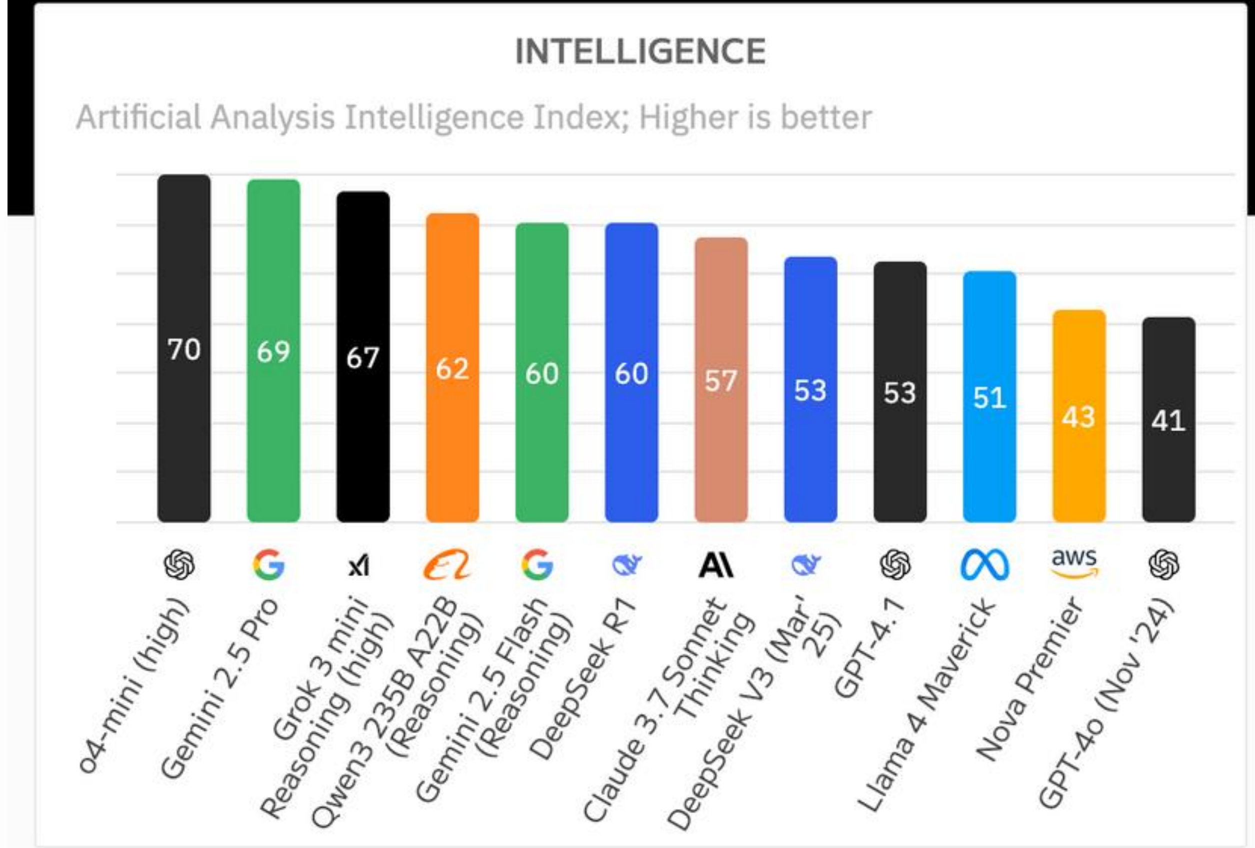


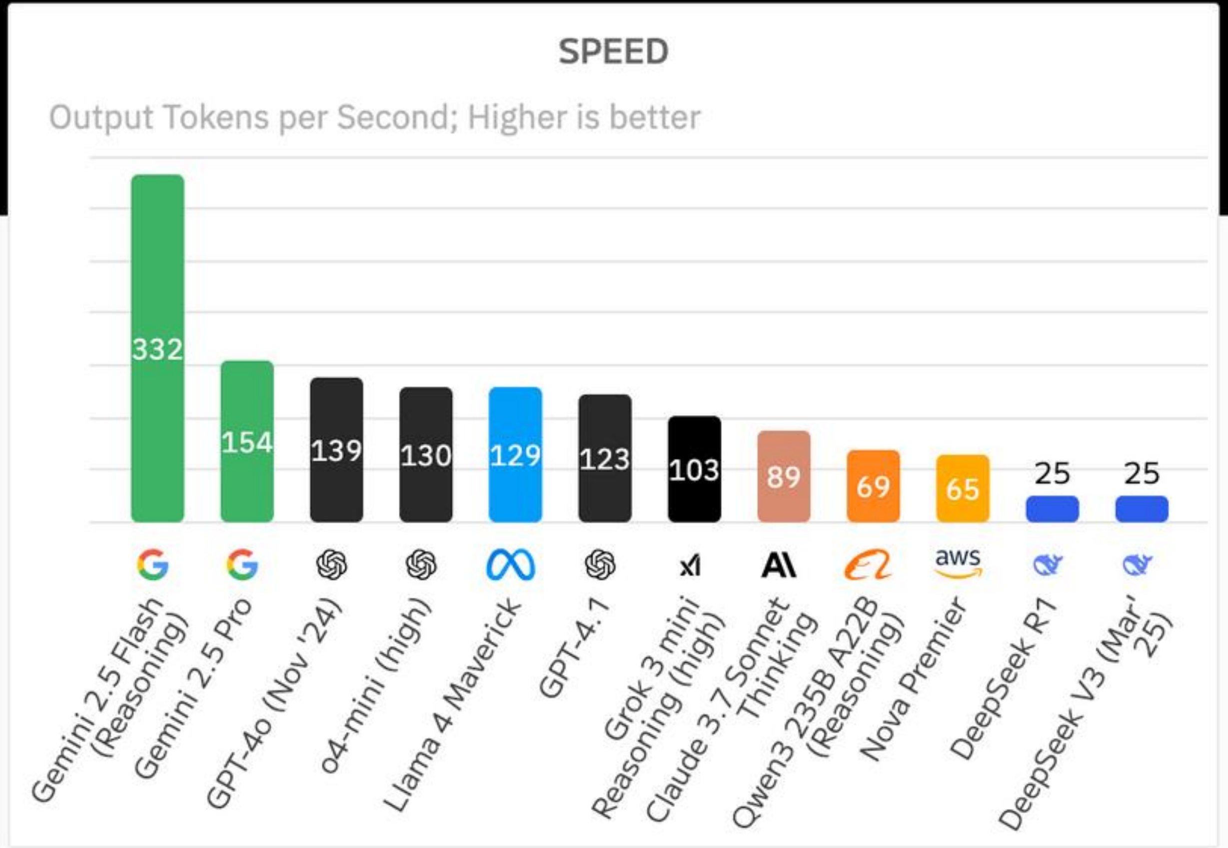
We performed an additional experiment to measure whether human physicians could further improve the quality of responses from our April 2025 models – comparing reference responses from o3 and GPT-4.1 with expert responses written by physicians with access to those references. We found that on these examples, physicians' responses no longer improved over the responses from the newer models.

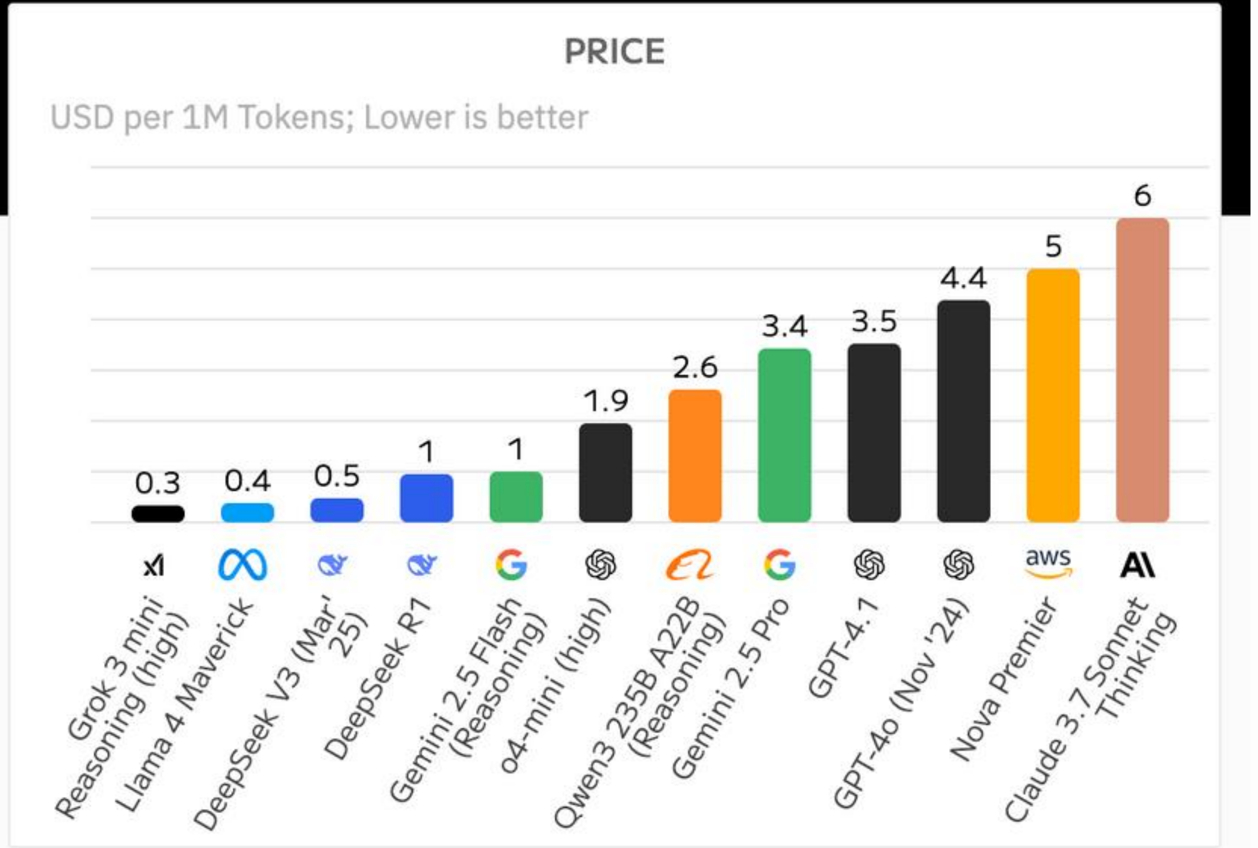
TABLE OF CONTENTS

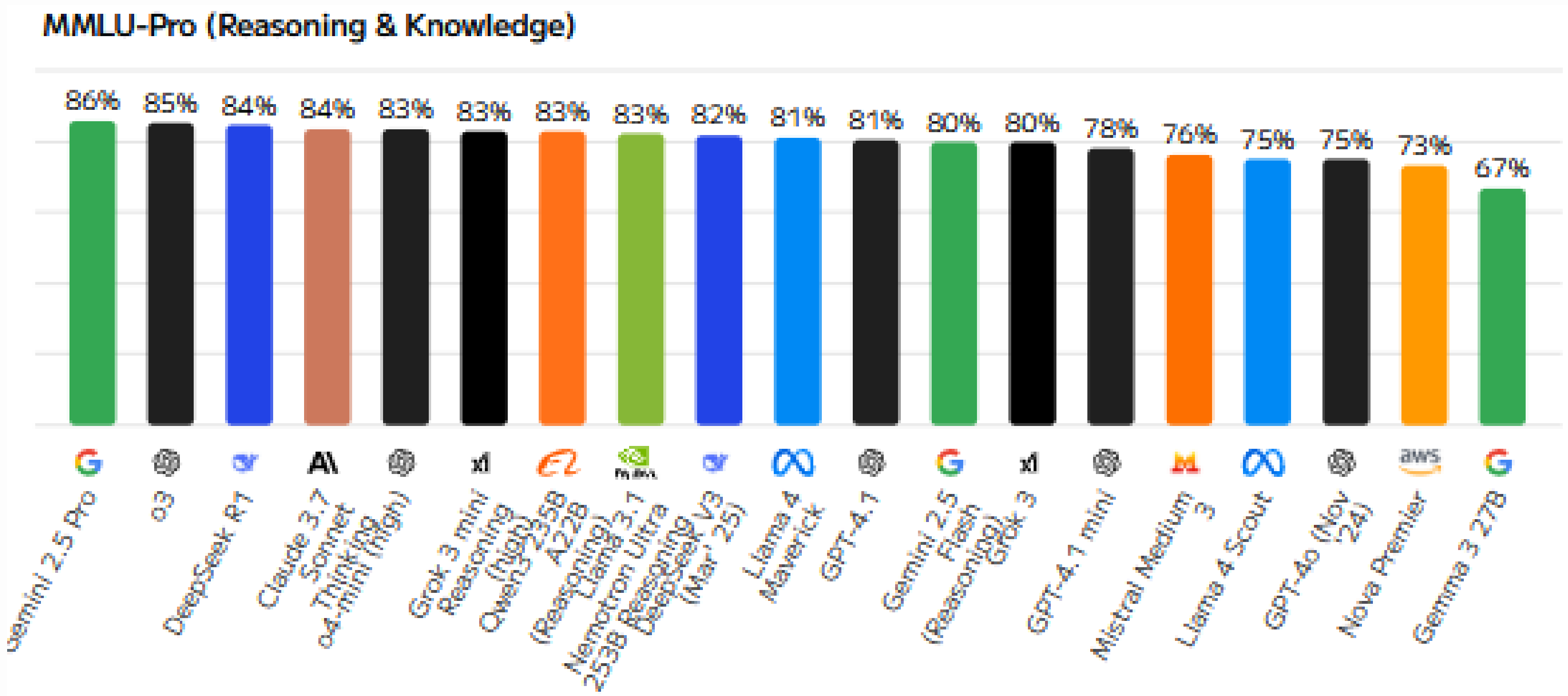

NEW YORK TIMES BESTSELLER
"Anyone seeking a basis of practical knowledge to understand the coming massive changes should read this insightful book right now."
—REID HOFFMAN,
Co-Founder of LinkedIn and Inflection AI
CO-INTELLIGENCE
ETHAN MOLLICK
ON SALE NOW
Ethan Mollick · 3rd
Associate Professor at The Wharton School. Author of Co-Intelligence
Top Voice
United States · Contact info
Order Co-Intelligence
275,168 followers
Followed by Connor Axiotes, Vrishank Saini, and 4 others
Message
Pending
More
The Wharton School
MIT Sloan School of Management

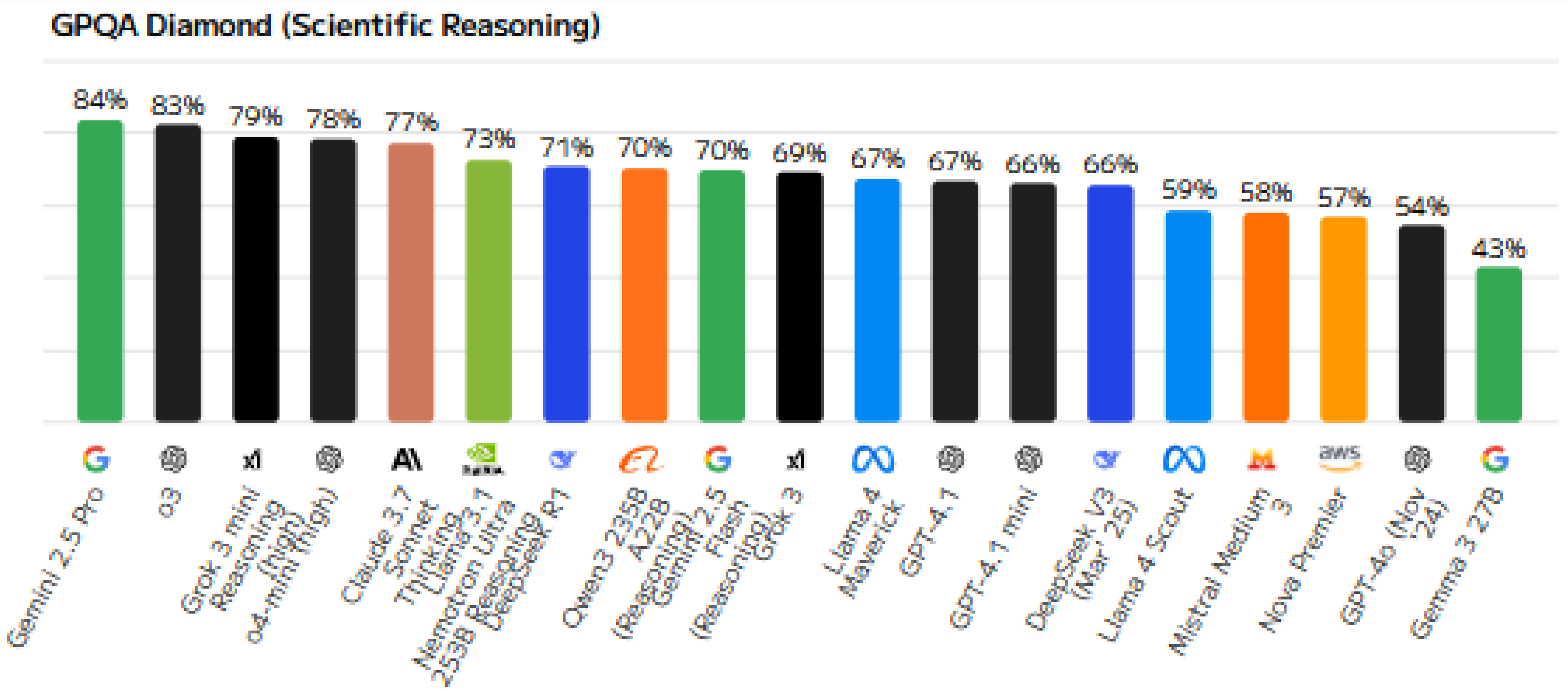
Highlights

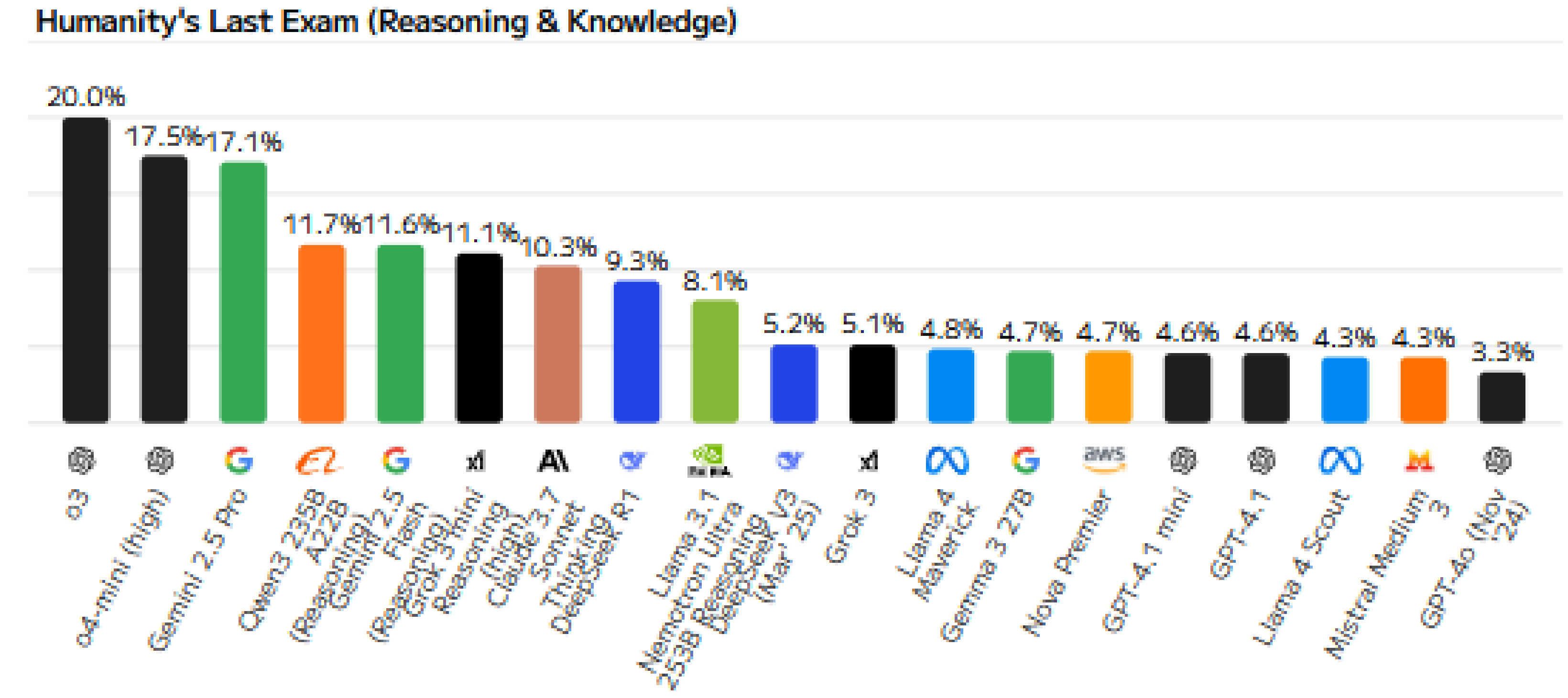

INTELLIGENCE
Artificial Analysis Intelligence Index; Higher is better
70 69 67 62 60 60 57 53 53 51 43 41
o4-mini (high)
Gemini 2.5 Pro
Grok 3 mini Reasoning (high)
Qwen3 235B A22B (Reasoning)
Gemini 2.5 Flash (Reasoning)
DeepSeek R1
Claude 3.7 Sonnet Thinking
DeepSeek V3 (Mar' 25)
GPT-4.1
Llama 4 Maverick
Nova Premier
GPT-4o (Nov '24)

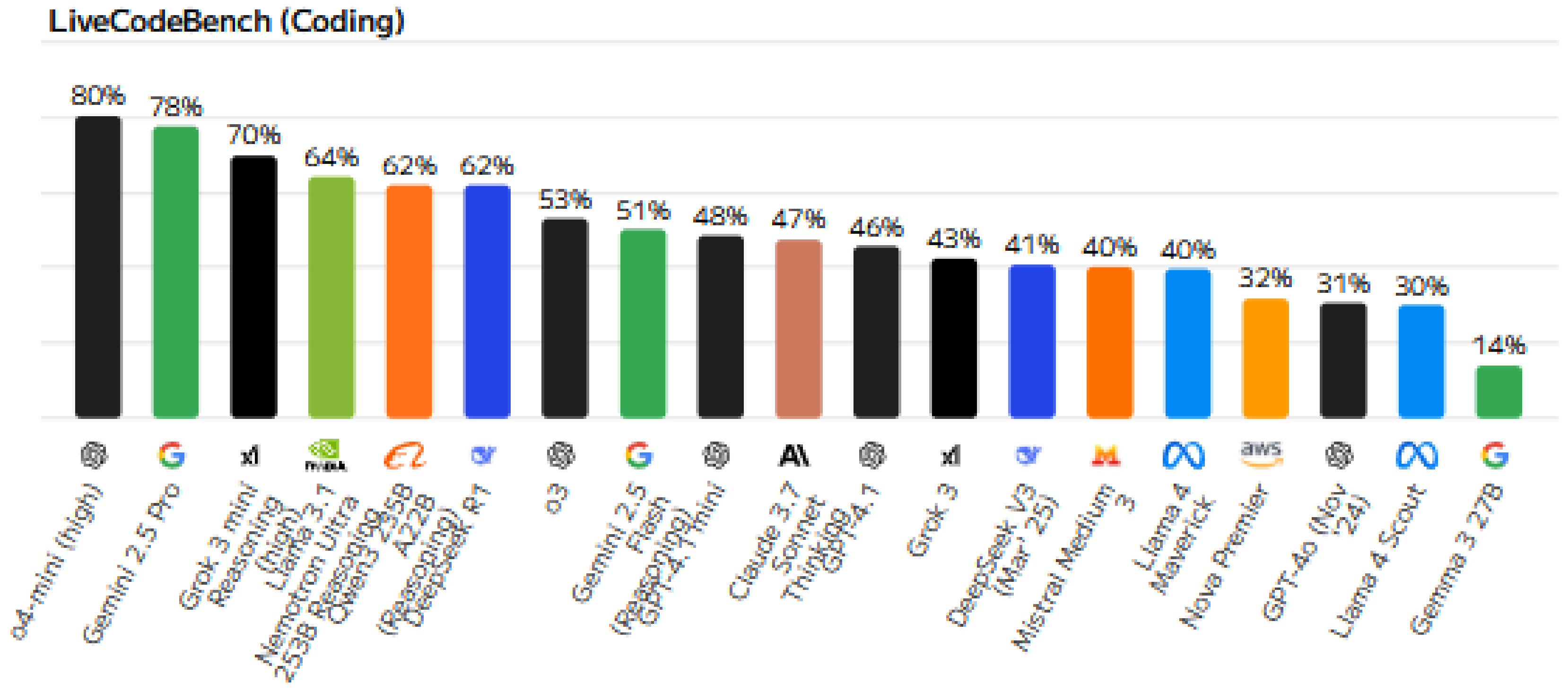

SPEED
Output Tokens per Second; Higher is better
332 154 139 130 129 123 103 89 69 65 25 25
Gemini 2.5 Flash (Reasoning)
Gemini 2.5 Pro
GPT-4o (Nov '24)
o4-mini (high)
Llama 4 Maverick
GPT-4.1
Grok 3 mini Reasoning (high)
Claude 3.7 Sonnet Thinking
Qwen3 235B A22B (Reasoning)
Nova Premier
DeepSeek R1
DeepSeek V3 (Mar' 25)


PRICE
USD per 1M Tokens; Lower is better
0.3 0.4 0.5 1 1 1.9 2.6 3.4 3.5 4.4 5 6
Grok 3 mini Reasoning (high)
Llama 4 Maverick
DeepSeek V3 (Mar' 25)
DeepSeek R1
Gemini 2.5 Flash (Reasoning)
o4-mini (high)
Qwen3 235B A22B (Reasoning)
Gemini 2.5 Pro
GPT-4.1
GPT-4o (Nov '24)
Nova Premier
Claude 3.7 Sonnet Thinking


MMLU-Pro (Reasoning & Knowledge)
86%
85%
84%
84%
83%
83%
83%
83%
82%
81%
81%
80%
80%
78%
76%
75%
75%
73%
67%
emini 2.5 Pro
o3
DeepSeek R1
Claude 3.7 Sonnet Thinking
o4-mini (high)
Grok 3 mini Reasoning (high)
Qwen3 235B A22B (Reasoning)
Llama 3.1 Nemotron Ultra 253B Reasoning
DeepSeek V3 (Mar' 25)
Llama 4 Maverick
GPT-4.1
Gemini 2.5 Flash (Reasoning)
Grok 3
GPT-4.1 mini
Mistral Medium 3
Llama 4 Scout
GPT-4o (Nov '24)
Nova Premier
Gemma 3 27B


GPQA Diamond (Scientific Reasoning)
84%
83%
79%
78%
77%
73%
71%
70%
70%
69%
67%
67%
66%
66%
59%
58%
57%
54%
43%
Gemini 2.5 Pro
o3
Grok 3 mini Reasoning (high)
o4-mini (high)
Claude 3.7 Sonnet Thinking
Llama 3.1 Nemotron Ultra 253B Reasoning
DeepSeek R1
Qwen3 235B A22B (Reasoning)
Gemini 2.5 Flash (Reasoning)
Grok 3
Llama 4 Maverick
GPT-4.1
GPT-4.1 mini
DeepSeek V3 (Mar' 25)
Llama 4 Scout
Mistral Medium 3
Nova Premier
GPT-4o (Nov '24)
Gemma 3 27B


Humanity's Last Exam (Reasoning & Knowledge)
20.0%
17.5%
17.1%
11.7%
11.6%
11.1%
10.3%
9.3%
8.1%
5.2%
5.1%
4.8%
4.7%
4.7%
4.6%
4.6%
4.3%
4.3%
3.3%
o3
o4-mini (high)
Gemini 2.5 Pro
Qwen3 235B A22B (Reasoning)
Gemini 2.5 Flash (Reasoning)
Grok 3 mini Reasoning (high)
Claude 3.7 Sonnet Thinking
DeepSeek R1
Llama 3.1 Nemotron Ultra 253B Reasoning
DeepSeek V3 (Mar' 25)
Grok 3
Llama 4 Maverick
Gemma 3 27B
Nova Premier
GPT-4.1 mini
GPT-4.1
Llama 4 Scout
Mistral Medium 3
GPT-4o (Nov '24)


LiveCodeBench (Coding)
80%
78%
70%
64%
62%
62%
53%
51%
48%
47%
46%
43%
41%
40%
40%
32%
31%
30%
14%
o4-mini (high)
Gemini 2.5 Pro
Grok 3 mini Reasoning (high)
Llama 3.1 Nemotron Ultra 253B Reasoning
Qwen3 235B A22B (Reasoning)
DeepSeek R1
o3
Gemini 2.5 Flash (Reasoning)
GPT-4.1 mini
Claude 3.7 Sonnet Thinking
GPT-4.1
Grok 3
DeepSeek V3 (Mar' 25)
Mistral Medium 3
Llama 4 Maverick
Nova Premier
GPT-4o (Nov '24)
Llama 4 Scout
Gemma 3 27B

OpenEvidence®

Ask a medical question...

OpenEvidence is HIPAA compliant.

If HIPAA applies to you, review and accept our Business Associate Agreement to begin including PHI in your conversations.

Accept

The leading medical information platform

what are the subtypes of PDAs and their frequency

Expanded question: What are the subtypes of *patent ductus arteriosus* and their frequency?

Patent ductus arteriosus (PDA) is morphologically classified into subtypes based on angiographic appearance, most commonly using the Krichenko classification, which includes Types A–E, with an additional Type F described for preterm infants. The subtypes are:

• **Type A (conical/funnel-shaped):** Most common, characterized by a well-defined aortic ampulla and constriction at the pulmonary end.

• **Type B (window):** Short and wide, with minimal constriction.

• **Type C (tubular):** Uniform diameter without constriction.

• **Type D (complex):** Multiple constrictions.

• **Type E (elongated):** Long ductus with constriction remote from the pulmonary artery.

• **Type F (fetal type):** Described in preterm infants, characterized by a long and large ductus, not fitting Types A–E.[1]

In a cohort of 100 children undergoing transcatheter closure, the distribution was approximately: Type A (29%), Type C (7%), Type E (32%), and Type F (26%).[1] Type F is particularly prevalent in premature infants and is associated with larger and longer PDAs compared to other types.

Additionally, echocardiographic and clinical classifications are used to describe **hemodynamic significance** (e.g., trivial/silent, small/audible, moderate, or large/hsPDA), but these are not morphological subtypes.[2]

The Krichenko classification remains the standard for morphological subtyping, with the addition of Type F for preterm infants now recognized in the literature.[1][3]

References

1. Morphologic Characterization of the Patent Ductus Arteriosus in the Premature Infant and the Choice of Transcatheter Occlusion Device.
   Philip R, Waller BR, Agrawal V, et al.
   Catheterization and Cardiovascular Interventions : Official Journal of the Society for Cardiac Angiography & Interventions. 2016;87(2):310-7. doi:10.1002/ccd.26287.
2. ACC/AHA/ASE/HRS/ISACHD/SCAI/SCCT/SCMR/SOPE 2020 Appropriate Use Criteria for Multimodality Imaging During the Follow-Up Care of Patients With Congenital Heart Disease: A Report of the American College of Cardiology Solution Set Oversight Committee and Appropriat...
   Sachdeva R, Valente AM, Armstrong AK, et al.
   Journal of the American College of Cardiology. 2020;75(6):657-703. doi:10.1016/j.jacc.2019.10.002.
   ☆ Leading Journal
3. Value of the Echocardiographic Suprasternal View for Diagnosis of Patent Ductus Arteriosus Subtypes.
   Zhang XY, Cao TS, Yuan LJ, Liu J, Duan YY.
   Journal of Ultrasound in Medicine : Official Journal of the American Institute of Ultrasound in Medicine. 2012;31(9):1421-7. doi:10.7863/jum.2012.31.9.1421.

Follow-Up Questions

- List the most common complications associated with each subtype of patent ductus arteriosus?
- Specify which patient populations are at highest risk for each subtype?
- Describe the typical pharmacological treatments and their side effect profiles for each subtype?

what are the subtypes of PDAs and their frequency?

Patent ductus arteriosus (PDA) can be classified into several **morphologic subtypes** based on angiographic appearance, which may influence treatment approach (e.g., catheter-based closure vs surgery). The most widely used classification is the **Krichenko classification**, which identifies **five types (A–E)** of PDA based on angiography:

Krichenko Classification of PDA (Types A–E)

Type	Description	Frequency
A	**Conical**: Most common. Narrow at the pulmonary artery end, wide at the aortic end (classic funnel shape).	**~70%**
B	**Window**: Short and wide, with no constriction at the pulmonary artery end.	**~2–3%**
C	**Tubular**: Long and cylindrical, no narrowing; equal diameter throughout.	**~10–12%**
D	**Complex**: Multiple constrictions (irregular shape).	**~5–7%**
E	**Elongated**: Long ductus with a constriction remote from the pulmonary artery end.	**~5–8%**

Additional Subtypes or Descriptions

Some studies and classifications go beyond Krichenko and describe PDAs in neonates and preterm infants by **hemodynamic significance** or **size**:

- **Hemodynamically significant PDA (hsPDA)** – causes symptoms or affects organ perfusion.

Additional Subtypes or Descriptions

Some studies and classifications go beyond Krichenko and describe PDAs in neonates and preterm infants by **hemodynamic significance** or **size**:

- **Hemodynamically significant PDA (hsPDA)** – causes symptoms or affects organ perfusion.
- **Silent PDA** – small, with no audible murmur or symptoms, often found incidentally.
- **Moderate to large PDA** – visible shunt on imaging and more likely to require intervention.

Summary of Frequencies (Approximate):

- **Type A (Conical) – 70%** (most frequent, especially in term infants and children)
- **Type B (Window-like) – 2–3%**
- **Type C (Tubular) – 10–12%**
- **Type D (Complex) – 5–7%**
- **Type E (Elongated) – 5–8%**

These numbers are based on observational angiographic series, such as:

- **Krichenko A et al., Circulation, 1989**: PMID: 2910058
- **Baruteau AE et al., Arch Cardiovasc Dis. 2014**: PMID: 25443873

Let me know if you'd like a figure illustrating the PDA types or info on device selection based on subtype.

what are the subtypes of PDAs and their frequency

Perplexity Sources · 8 Tasks

Annals of medicine
Pancreatic Cancer Subtypes: A Roadmap for...

Nature medicine
Subtypes of Pancreatic Ductal Adenocarcinoma...

structuralheart.abbott
Types of Congenital Heart Defects | ASD, VSD, PDA

en.wikipedia
Personal digital assistant - Wikipedia

Subtypes of PDA and Their Frequency

Patent Ductus Arteriosus (PDA) is a congenital heart defect with several recognized subtypes based on angiographic appearance, commonly referred to as the Krichenko classification. The main subtypes are:

- **Type A (Conical):** Features a prominent aortic ampulla with a constricted pulmonary end.
- **Type B (Window):** Characterized by a large width and very short length.
- **Type C (Tubular):** Long and without any evidence of constriction.

There are two additional types in the Krichenko classification (Types D and E), but the search results specifically mention Types A, B, and C [5].

Frequency

- **General Prevalence:** Patent ductus arteriosus occurs in approximately 1 in 2000 live births and accounts for 5% to 10% of all congenital heart disease cases [8].
- **Subtypes Frequency:** The search results do not provide precise frequency data for each angiographic subtype (A, B, C, etc.). However, Type A (conical) is generally considered the most common in clinical practice, while Types B and C are less frequent [5].

Summary Table

Do pediatric deceased donor kidneys hypertrophy when transplanted

Answer · Sources · 8 · Tasks

pmc.ncbi.nlm.nih
Successful en bloc transplantation of pediatri...

Pediatric clinics of Nort...
Long-term Outcomes of Kidney Transplantation in...

Translational Pediatrics
Successful single kidney transplantation from...

Sage Journals
Excellent Clinical Long-Term Outcomes of Kidney...

Yes, pediatric deceased donor kidneys do undergo hypertrophy after transplantation. Clinical studies demonstrate that these kidneys exhibit compensatory growth to adapt to the recipient's metabolic demands, particularly when transplanted into adult recipients.

Key Findings on Hypertrophy

- **Size Increase:** Transplanted pediatric kidneys show a median length increase of **20–37 mm** within the first 12–18 months post-transplant [3] [4] [7]. This growth is attributed to nephron hypertrophy in response to the adult recipient's higher metabolic needs [3] [4].
- **Functional Adaptation:** Alongside structural growth, glomerular filtration rate (eGFR) improves steadily, reaching **~100 mL/min/1.73 m²** by 2 years post-transplant in single-kidney transplants from donors ≤10 kg [3] [4].
- **Long-Term Adaptation:** Hypertrophy in pediatric donor kidneys persists for **up to 6 years**, unlike adult donor kidneys, where growth plateaus within months [4] [7].

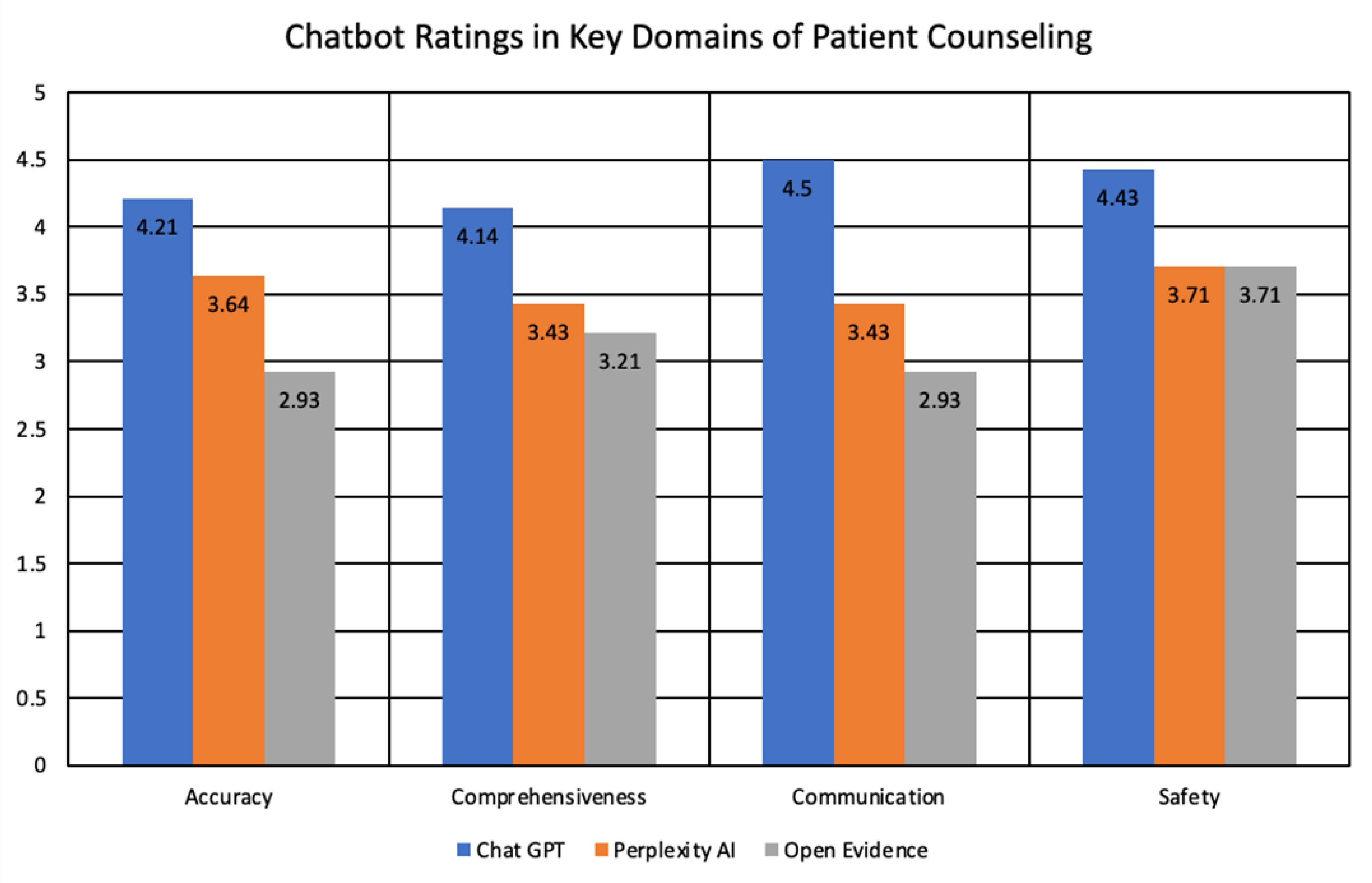
Mechanisms and Clinical Implications

- **Hyperfiltration Mitigation:** Early hypertrophy helps counteract hyperfiltration injury, which can cause proteinuria. Studies show transient proteinuria (40–70% of cases) typically resolves as kidneys adapt [3] [5] [7].
- **Donor-Recipient Matching:** Successful outcomes depend on selecting recipients with BMI <30 kg/m² and weight <100 kg to minimize mismatch-related stress [4] [7].

Evidence from Studies

- A single-center study of 47 transplants from donors ≤10 kg found comparable outcomes between single and en bloc transplants, with no graft loss due to hyperfiltration [3].
- Pediatric kidneys transplanted into adults achieved **higher eGFR** (69.9 vs. 59.7 mL/min/1.73 m² at 1 year) compared to standard adult donor kidneys [4].

This adaptive hypertrophy enables pediatric deceased donor kidneys to function effectively in adult recipients, expanding the donor pool while maintaining graft survival [3] [4] [7].


Chatbot Ratings in Key Domains of Patient Counseling
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
4.21
3.64
2.93
4.14
3.43
3.21
4.5
3.43
2.93
4.43
3.71
3.71
Accuracy
Comprehensiveness
Communication
Safety
Chat GPT
Perplexity AI
Open Evidence

AMIE Improves Diagnostic Accuracy

Sources

+ Add

Discover

Select all sources

s41586-025-08866-7.pdf

s41586-025-08869-4.pdf

s41746-025-01486-5.pdf

s41746-025-01519-z (1).pdf

Chat

Refresh

Discuss what these sources say about Diagnostic Performance Comparison (LLMs vs Physicians), in the larger context of LLMs in Medical Diagnosis and Assessment.

Based on the provided sources, a comprehensive comparison of the diagnostic performance of Large Language Models (LLMs) and physicians reveals nuanced findings across different studies and medical tasks.

LLM Performance in Diagnostic Dialogue (AMIE Study)

One study introduced AMIE (Articulate Medical Intelligence Explorer), an LLM-based AI system optimized for diagnostic dialogue [1]. This system was evaluated against primary care physicians (PCPs) in a randomized, double-blind crossover study involving text-based consultations with validated patient-actors [1]. The study design, while enabling potentially large-scale LLM–patient interactions, utilized a synchronous text chat interface which was unfamiliar to the PCPs for remote consultation and not representative of usual practice or telemedicine [1] [...].

Key findings from this study regarding diagnostic performance include:

- **Superior Diagnostic Accuracy:** AMIE demonstrated greater diagnostic accuracy than the PCPs [1] [...]. Specifically, AMIE exhibited superior differential diagnosis (DDx) accuracy as assessed by various measures, including top-1 and top-3 accuracy of the DDx list [5]. AMIE showed significantly higher top-k accuracy across all values of k compared to PCPs, considering matches with both the ground-truth diagnosis and accepted differentials [2]. The performance difference was statistically significant ($P < 0.05$) [2] [...].
- **Information Interpretation vs. Acquisition:** Results suggested that AMIE was approximately equivalent to PCPs in information acquisition during the simulated consultations, eliciting similar amounts of information from patients [7] [...]. However, AMIE was **better than PCPs at interpreting that information to produce an accurate or complete DDx** [7] [...]. When truncated conversations were used, neither AMIE nor PCPs showed a significant advantage in the speed, efficiency, or diagnostic utility of information acquisition [8] [...].
- **Evaluation by Experts and Patients:** According to specialist physicians, AMIE exhibited superior performance on 30 out of 32 evaluation axes, and was non-inferior on the rest [5] [...]. Patient-actors rated AMIE's consultations significantly better than those with PCPs across 25 of 26 axes, including conversation qualities like empathy and communication skills, despite the text-based format potentially disadvantaging human clinicians accustomed to voice/non-verbal communication [1] [...].
- **Performance Across Specialties and Non-Disease States:** AMIE's performance matched or surpassed PCP performance across most specialties, with the most pronounced improvements in respiratory and internal medicine, although it did not outperform PCPs in obstetrics and gynaecology/urology [15]. AMIE appeared to maintain the trend of better performance on mostly negative or non-disease state scenarios, although results were not statistically significant for this small subset of cases [16]. The majority of scenarios in this evaluation assumed an underlying disease state, which is a limitation as it doesn't reflect the reality of primary care where ruling out disease is common [17].

This study represents a milestone towards conversational diagnostic AI, demonstrating that an LLM optimized for the task can outperform PCPs in simulated diagnostic conversations when evaluated along multiple clinically meaningful axes [1] [...]. However, the limitations, particularly the text-chat interface, mean the findings should be interpreted with caution and are not representative of usual clinical practice [1] [...].

LLM Performance in Challenging Case Diagnosis (AMIE and Other LLMs)

Another study evaluated AMIE's ability to generate a differential diagnosis alone or as an aid to clinicians using 302 challenging, real-

Start typing...

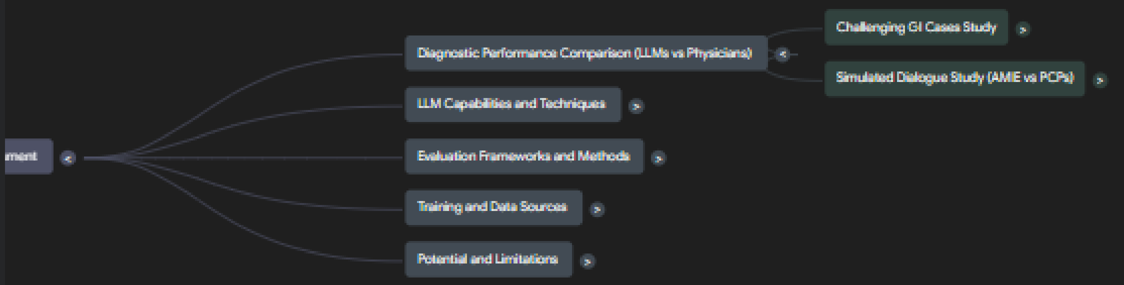
4 sources

What are the key capabilities that make AMIE effective for medical diagnosis?

How was AMIE evaluated to demonstrate its impact on d

LLMs vs. Physicians: Diagnostic Performance and Assessment

Based on 4 sources



Good content

Bad content

Studio > Note

Evaluating AI in Medical Diagnosis and Practice

(Saved responses are view only)

Understanding AI in Medical Diagnosis: A Study Guide

Quiz

1. What was the primary finding regarding AMIE's diagnostic accuracy compared to PCPs across 159 scenarios according to the first source?
2. Based on the first source, how did AMIE's differential diagnosis performance compare to PCPs for non-disease-state scenarios versus positive disease-state scenarios?
3. According to the first source, how did patient-actors rate AMIE compared to PCPs on conversational qualities?
4. What did specialist physicians rate AMIE as significantly better than PCPs on, according to the first source's Fig. 5?
5. According to the second source, how did Claude 3.5 Sonnet's coverage rate of instructive diagnoses compare to experienced gastroenterologists in the study?
6. In the second source's subgroup analysis, how did LLMs and gastroenterologists differ in their performance on GI vs. non-GI cases?
7. According to the second source, what was the most common error type made by the five participating physicians when analyzing their diagnostic errors?
8. What is Retrieval Augmented Generation (RAG) and why is it relevant to the third source's study?
9. According to the third source, which LLM-RAG model achieved the highest accuracy in determining surgical fitness using international guidelines?
10. What evaluation framework was used in the third source to qualitatively assess LLM responses in a medical context?

Answer Key

1. The first source found that AMIE had superior top-k differential diagnosis (DDx) accuracy compared to PCPs across all 159 scenarios, with statistically significant differences for all k values tested.
2. AMIE appeared to perform better on non-disease-state scenarios, maintaining the trend of better performance seen in the larger set of positive disease-state scenarios, although the results for non-disease states were not statistically significant due to the small sample size.
3. Patient-actors rated AMIE significantly more favorably than PCPs on several conversational qualities, such as being polite, making the patient feel at ease, listening, explaining conditions, and involving the patient in decisions.
4. Specialist physicians rated AMIE significantly more favorably than PCPs on numerous conversation and reasoning qualities, including gathering and providing information, DDx appropriateness and comprehensiveness, clinical judgement, showing empathy, and maintaining patient welfare.
5. According to the second source, Claude 3.5 Sonnet's coverage rate of instructive diagnoses (76.1%) significantly surpassed that of all 22 participating experienced gastroenterologists (average 29.5%).
6. The second source's subgroup analysis showed that LLMs performed better on non-GI cases than GI cases, while gastroenterologists performed better on GI cases than non-GI cases.
7. Based on the second source's error analysis, the most common error type made by the five participating physicians was Knowledge Deficiency (64.3%).
8. Retrieval Augmented Generation (RAG) is a technique that integrates specialized knowledge into LLMs, making them more capable in domain-specific applications like medicine, which is relevant to the third source's study on surgical fitness.
9. According to the third source, the GPT-4 LLM-RAG model using international guidelines achieved the highest accuracy (96.4%) in determining surgical fitness, performing significantly better than human-generated responses.
10. The third source used the S.C.O.R.E. evaluation framework to qualitatively assess LLM-RAG model responses based on safety, clinical consensus, objectivity, reproducibility, and explainability.

Essay Questions

1. Compare and contrast the methodologies and key findings of the first two sources regarding the diagnostic capabilities of LLMs compared to human physicians. Discuss the types of cases studied, the metrics used for evaluation, and the overall conclusions reached in each paper.
2. Analyze the different aspects of AI performance evaluated across the three sources (diagnostic accuracy, conversational skills, surgical fitness assessment). How do these different evaluations contribute to a broader understanding of the potential and limitations of LLMs in healthcare?
3. The third source highlights the use of Retrieval Augmented Generation (RAG) to improve LLM performance in a specific medical domain. Discuss the implications of RAG for the future development and application of LLMs in medicine, referencing insights from all three sources about areas where LLMs currently demonstrate strengths and weaknesses.

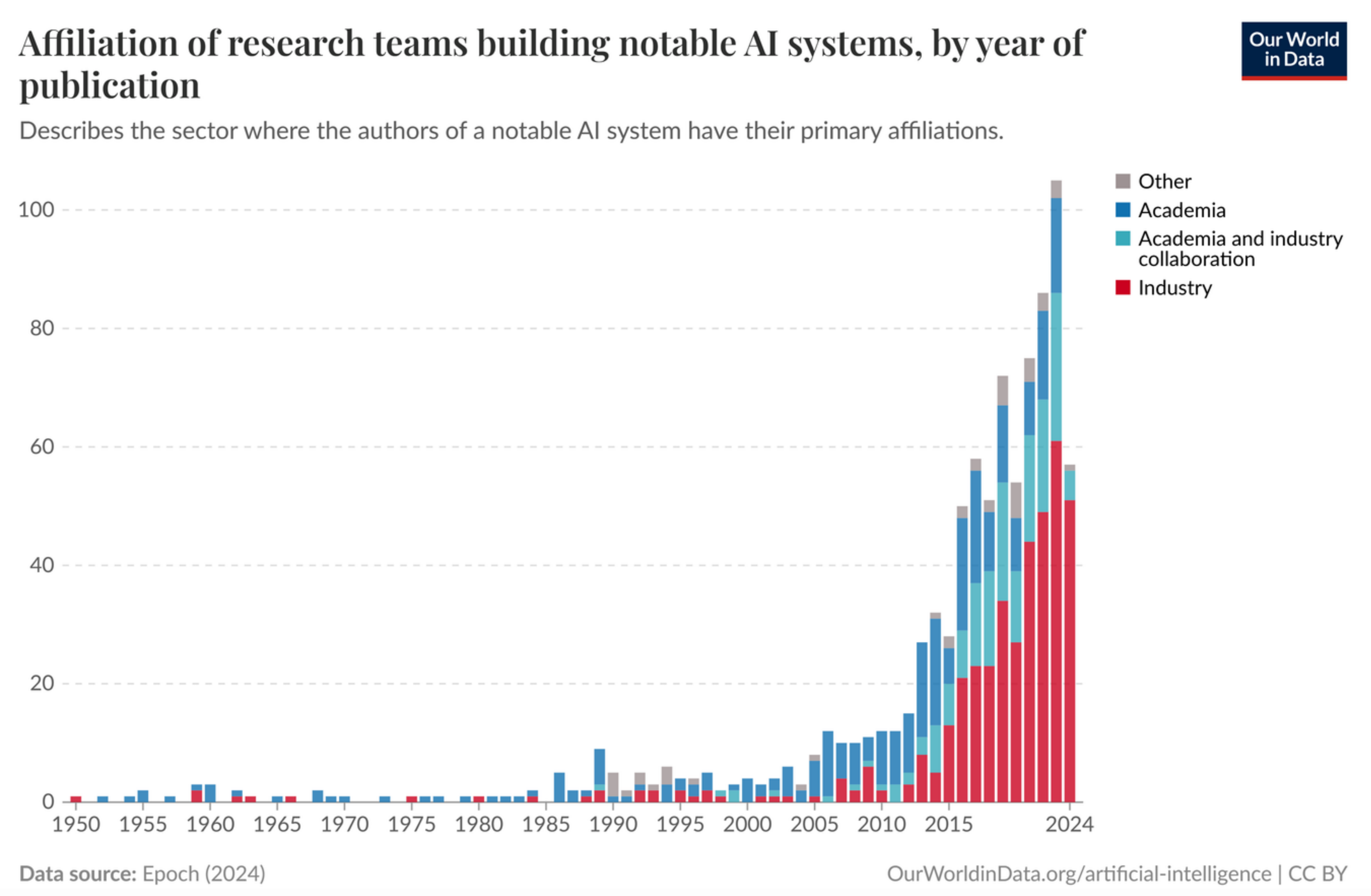
Convert to source

TABLE OF CONTENTS

BE INVOLVED

Now is the best time to become a domain expert.

Use it to educate yourself.



Young Field

The field is less than 5 years old. You can quickly become a domain expert. Most physicians are not paying attention.

Influence

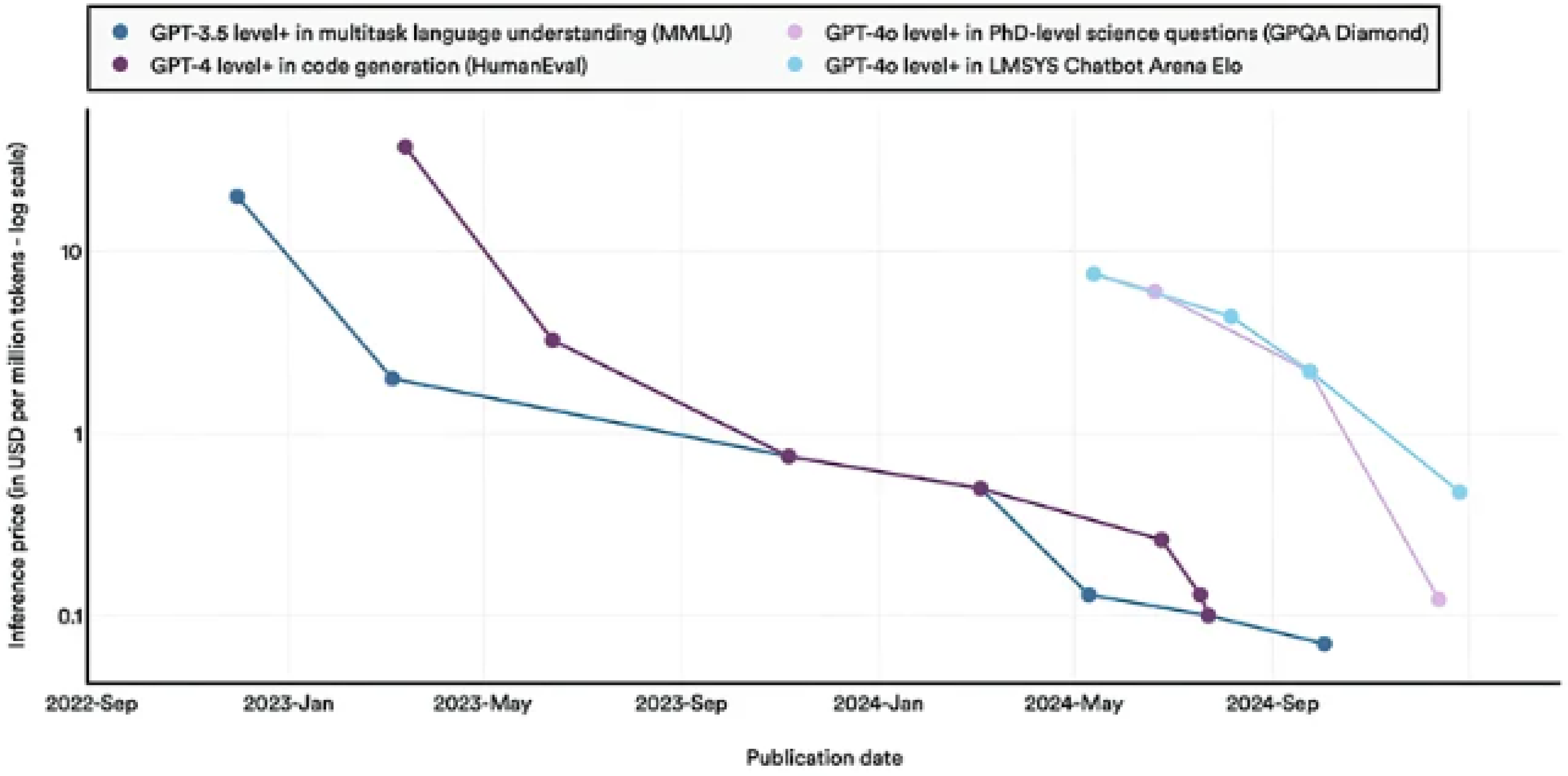
You are in a position to discover and guide the use of these tools.

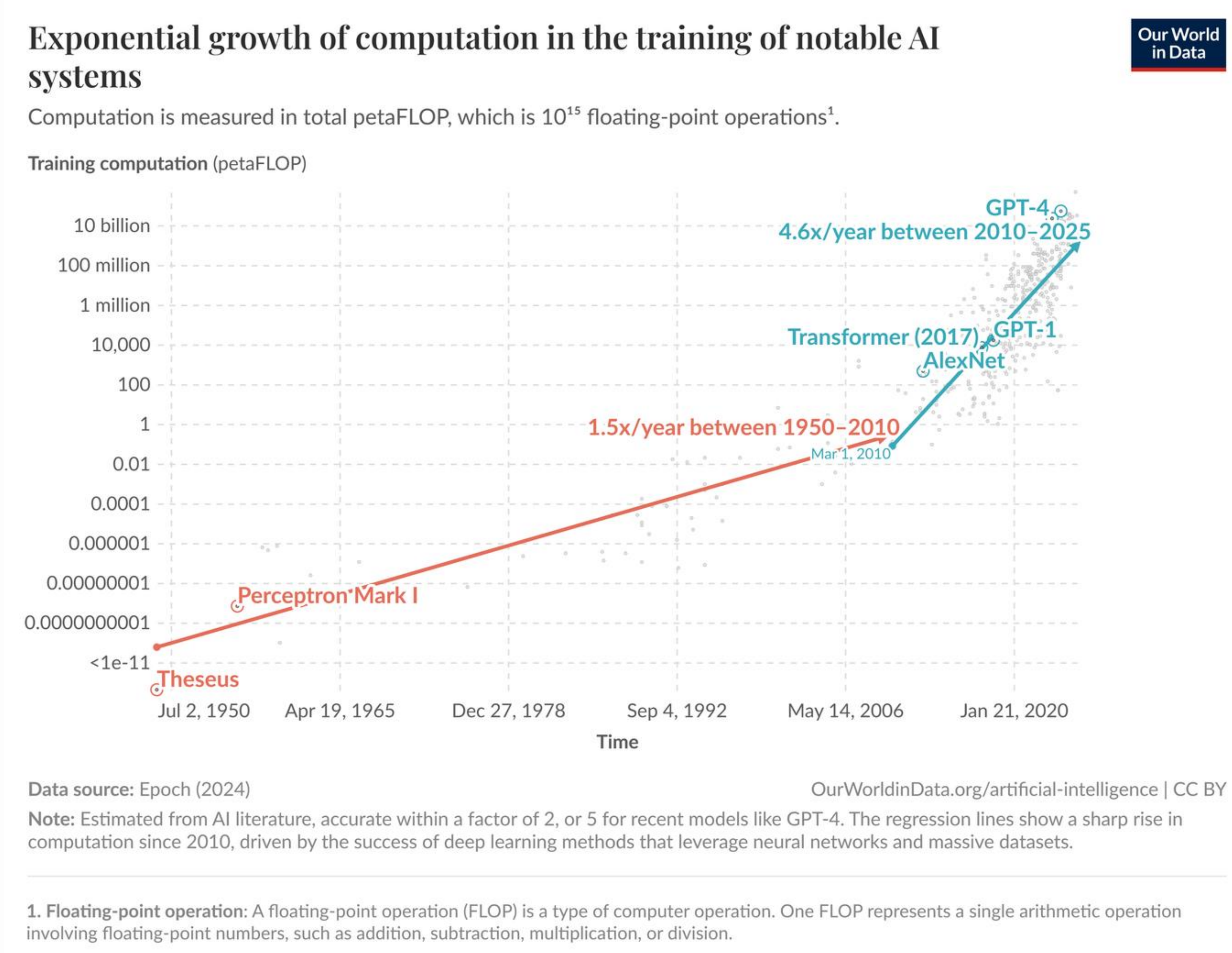
Drive Values

Your involvement will determine how AI is implemented. Embrace Change. You have access to patients - Industry does not.

Inference price across select benchmarks, 2022–24

Source: Epoch AI, 2025; Artificial Analysis, 2025 | Chart: 2025 AI Index report





QUESTIONS TO CHEW ON

What level of accuracy would LLMs have to provide to make checking your work the standard of care?

Do you remove humans from the loop?

Should patients be double checking their doctors work?

If the value of care delivery is quality divided by cost, and AI is cheaper while preserving quality, is this best for patients?

$$\text{Value} = \frac{\text{Quality}}{\text{Cost}}$$

TABLE OF CONTENTS

RESEARCH AID

Living Donor Nephrectomies typically get Lasix and Mannitol prior to cross-clamping the kidney.

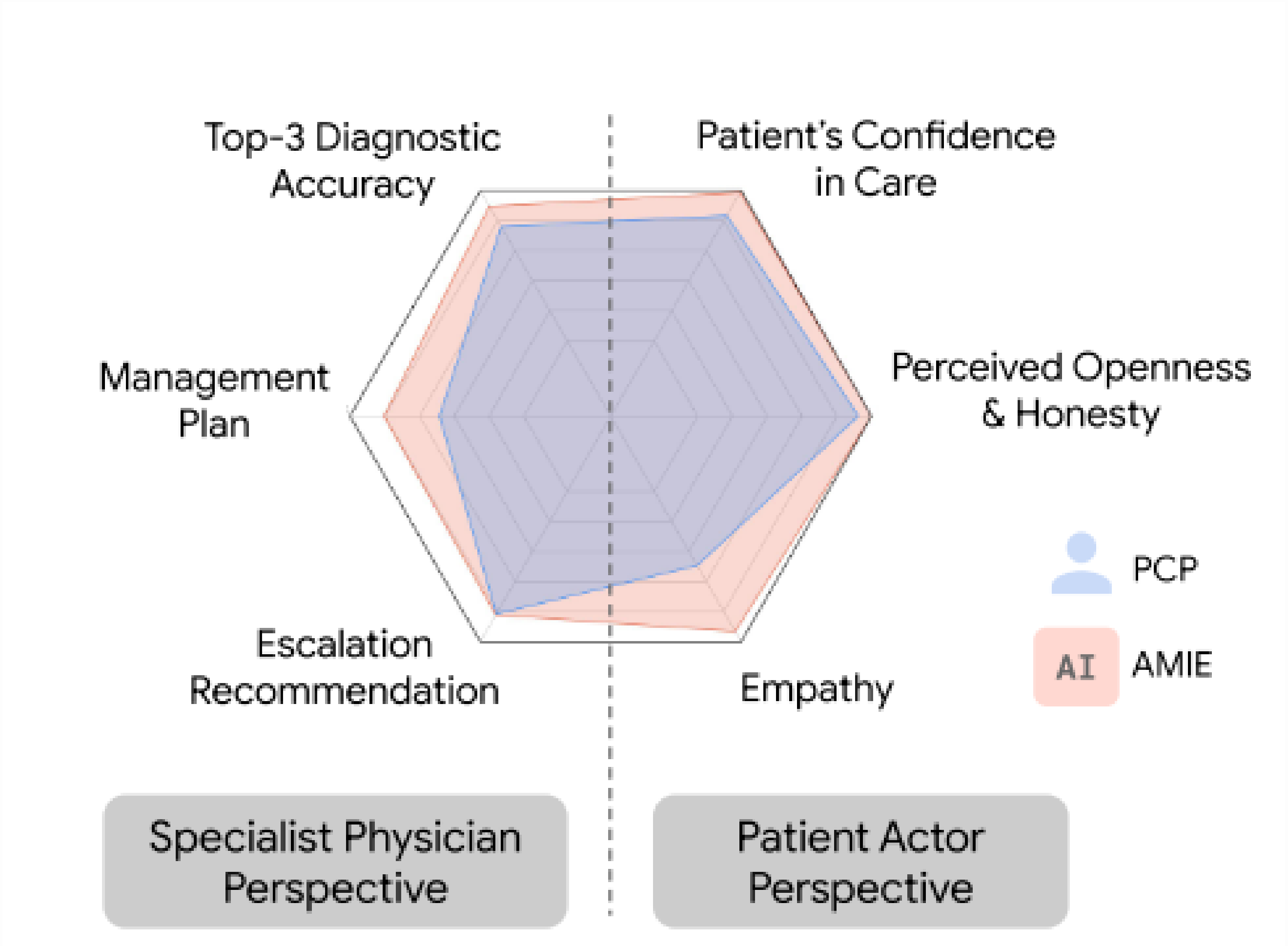
Our program is debating eliminating this practice.

What evidence supports or doesnt support this practice?
What are the likely effects we will see if we abandon this practice?

SPECIALIST CONSULT

A patient has been in the hospital for 3 weeks with recurrent liver abscesses and the primary team, the infectious disease consult service and GI is calling because they cannot figure out why she keeps getting abscesses. Her notable history is that she is 5 years out from a liver transplant and GI did an ERCP and stented multiple strictures but she is still having abscesses despite drainage.

What is going on?
What is one test which would confirm your diagnosis?



MEDICAL AI

A PRIMER