



TIPQC Annual Meeting Agenda
March 24-25, 2025
Franklin, TN

LLMS ROLE IN TRANSFORMING QUALITY

James Barry MD, MBA

**Professor of Pediatrics, Section of Neonatology
University of Colorado School of Medicine**

Topics

- What are large language models? +
- Understanding LLM Bias +
- LLMs For Quality and Safety +
- Pertinent Prompting Approaches +

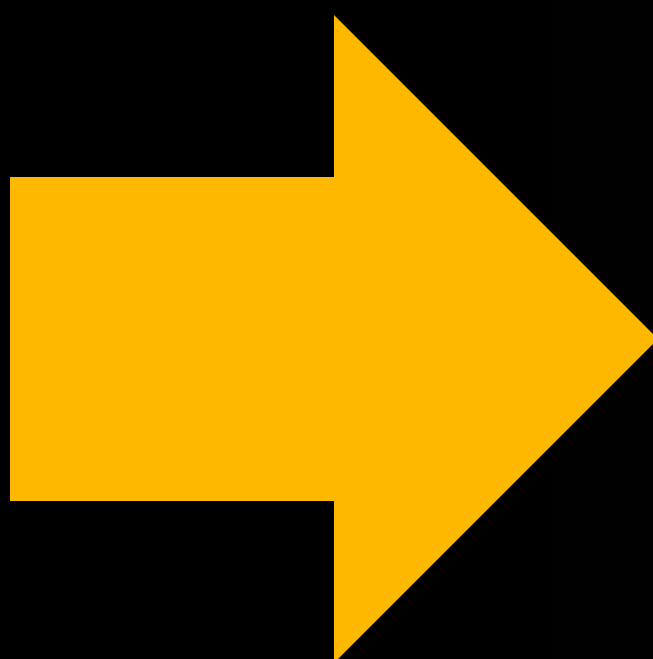
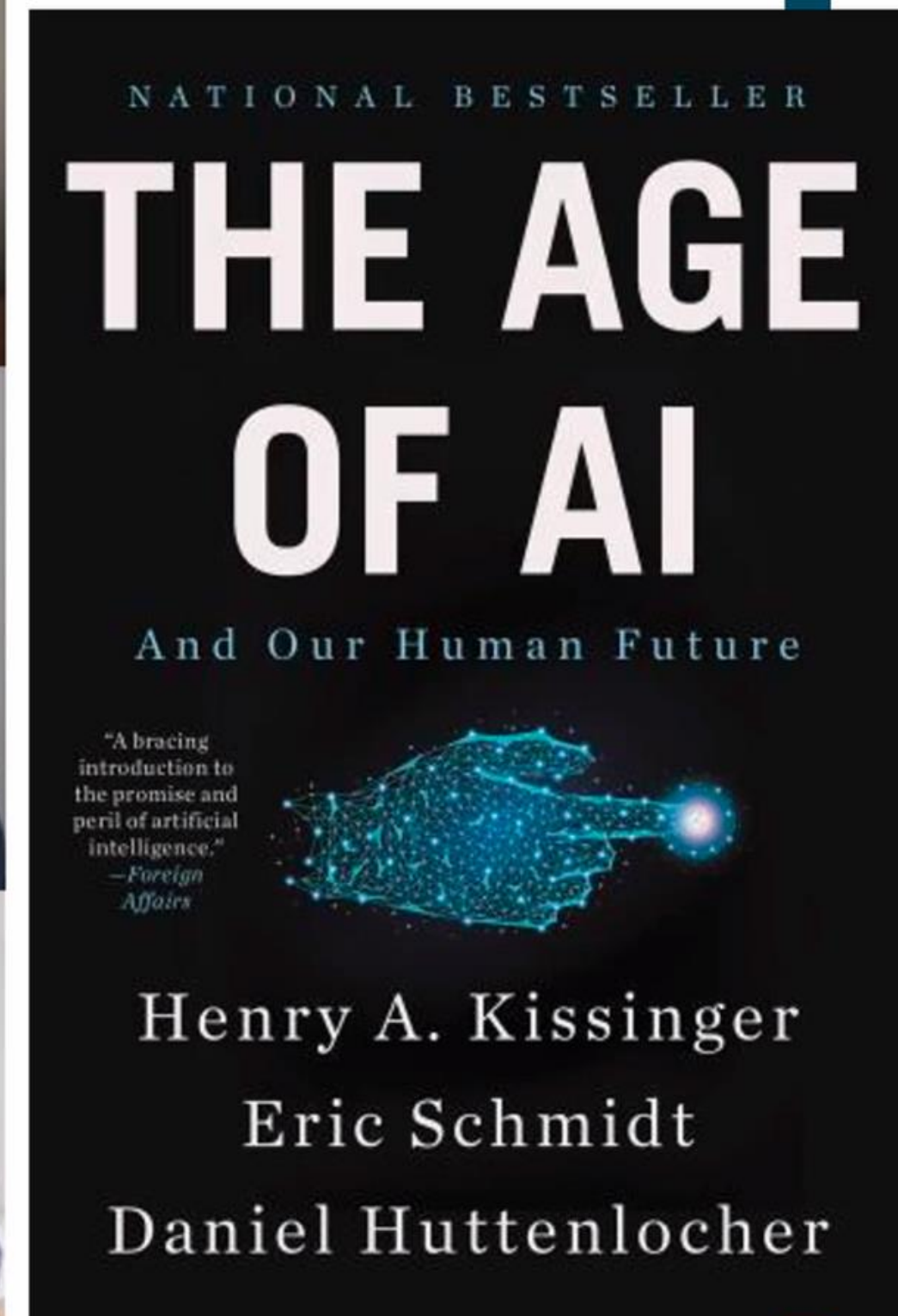
+



“...the technology is changing human thought, knowledge, perception, and reality--- and in doing so, is changing the course of human history”

published

2021



Medical Director of UCH NICU since 2007

Teams



Processes



Outcomes



UCH Newborn System (8 nurseries/NICUs) Chair Since 2019

A QI Project

Reducing NICU Discharge Care-Related Failures by Improving Discharge Safety

Lauren A Beard, MD, Kirtley C Ceballos, MSN, RN, NEA-BC, Kerri L Staton, RN, Kathryn A Conner, RN, Christina M Math, MSN, RN, Melissa K Moore, RN, Blair W Weikel, MPH, James S Barry, MD, MBA

SMART aim: We sought to increase caregiver discharge preparedness beginning August 2013 by creating a more effective education and training process to decrease DC related care failures by 20% and eliminate preventable hospital readmission within a month of DC by May 2016 and maintain this improvement for 3 years.

Pediatrics,
accepted

Our IHQSE Project

Figure 1. Process control chart of monthly discharge care related failures among those contacted (n = 1,453). The x-axis is labeled for each quarter with the total contacted population in parentheses. Red dashed lines represent the lower (LCL) and upper control limits (UCL). Improvement interventions are depicted below the x-axis as encircled letters corresponding to the improvement in Table 1.

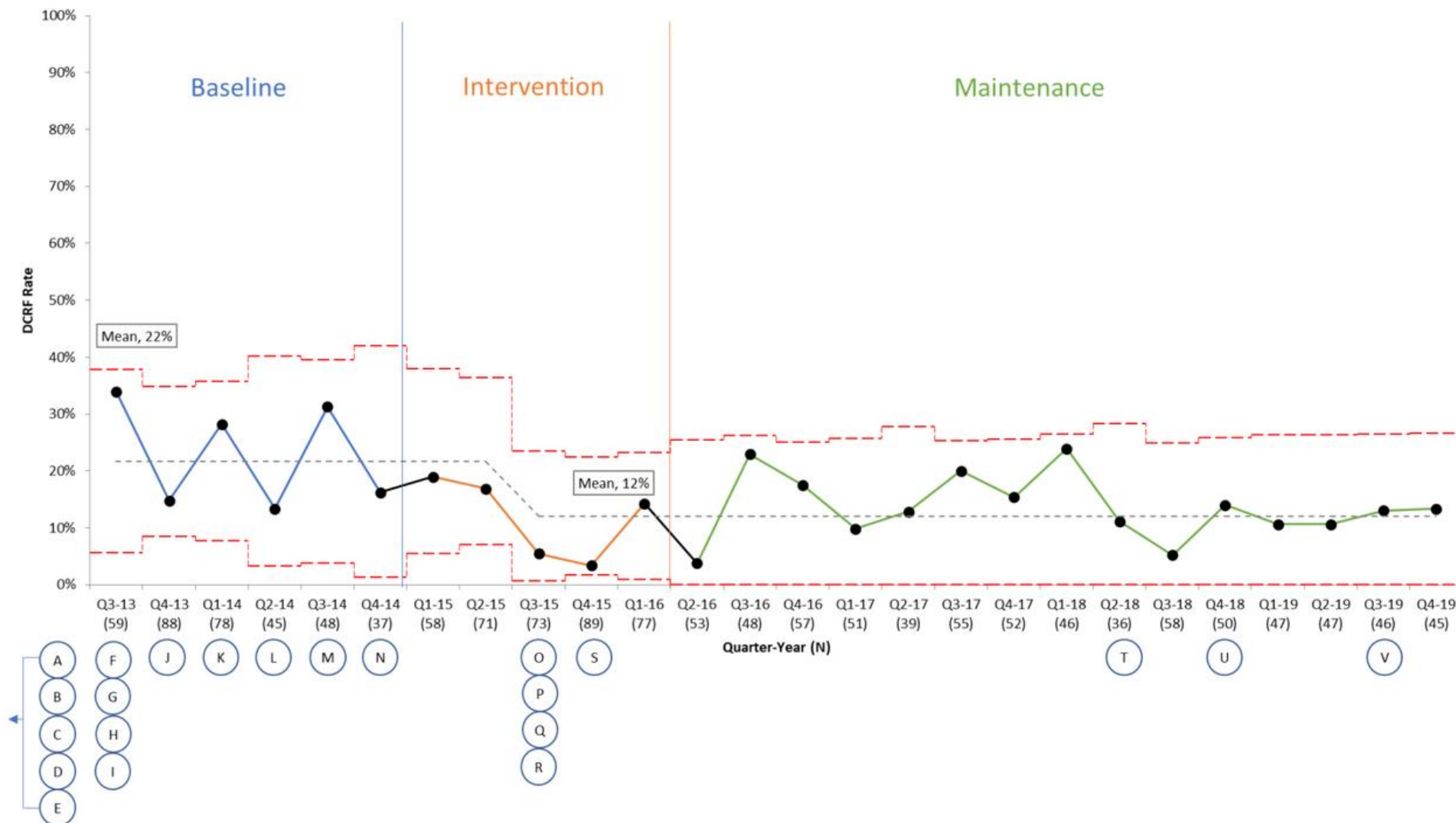


Table 1. Study timeline

Date	Improvement	Label
February 2013	Index Case: Patient admitted with life threatening electrolyte abnormality after DC	A
March 2013	Begin DC process improvement project	B
March 2013	First RN DC satisfaction survey	C
March 2013	Form DC Process Improvement Team (DPIT) - NICU RN manager, RN educator, outcomes coordinator, medical director. Joined inaugural Institute for Healthcare Quality, Safety and Effectiveness (IHQSE) cohort	D
May 2013	Begin standardized DC classes offered twice a week for all NICU caregivers	E
July 2013	Develop checklist for post-DC follow up phone calls	F
August 2013	Launch "voice of patient/family" survey through follow-up phone calls	G
August 2013	Create process map of DC tasks and milestones for physicians, neonatal nurse practitioners, case managers, social workers, nurses, and respiratory therapists.	H
September 2013	Stakeholder analysis with nurses, social workers, physicians, neonatal nurse practitioners, pharmacists, and case managers	I
December 2013	Initiate caregiver readiness assessment completed by small group of bedside RNs	J
February 2014	Expand DC classes offerings to evenings and weekends	K
June 2014	Develop script for twice weekly multi-disciplinary "DC Flash Rounds"	L
July 2014	One RN begins DC follow-up phone calls with structured survey	M
December 2014	Two RNs begin part-time role as DC navigators; DC classes offered 7 days per week	N
July 2015	Second RN DC satisfaction survey	O
August 2015	Develop and implement bedside visual pathway to home	P
September 2015	Formalize primary RN/PT/OT assignments for infants with stays > 2 weeks	Q
September 2015	Implement visual and written administration guides for medications and feedings	R
December 2015	DC navigator role becomes a 1.0 FTE position	S
April 2018	First meeting with home oxygen vendor to improve home oxygen delivery	T
November 2018	Met with home oxygen vendor staff and VP to optimize home oxygen delivery	U
June 2019	Begin using a new home oxygen vendor for all NICU discharges	V
October 2020	Third RN DC satisfaction survey	W

Improvement is a journey, not a destination

Could AI Have Assisted Our **IHQSE** Project?

YES



Could AI Have Assisted Our IHQSE Project?

YES

Without a Doubt



Could AI Have Assisted This DC Project?

Improve Patient and Caregiver Understanding

Automated Data Summarization & Report Generation

Idea Generation & Project Charter Development

Facilitating Literature Reviews & Best Practice Identification

Enhancing Communication & Stakeholder Engagement





AI and Pt Safety

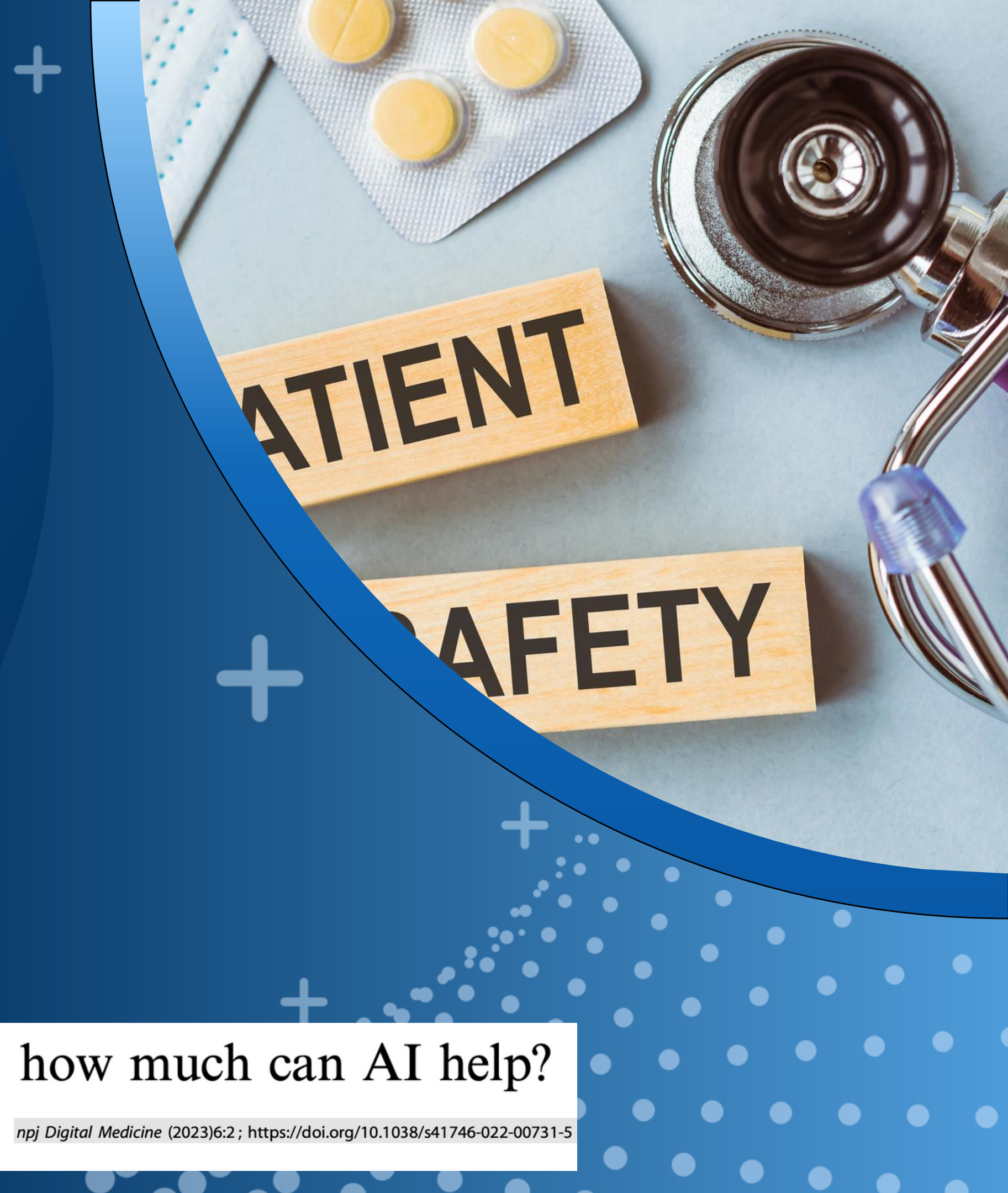
Table 2. Top use cases for the application of AI to specific clinical problems in patient safety.

1. Actionable real time patient safety electronic clinical quality measures
2. Surgical complication prediction
3. Pressure ulcer prediction
4. Hypoglycemia prediction
5. Sepsis prediction
6. Suicide prediction
7. Diabetic eye AI screening
8. Breast imaging cancer screening
9. Chest x-ray imaging AI diagnosis
10. Skin melanoma AI diagnosis
11. Chest x-ray imaging AI cancer screening
12. Patient self-managed electronic safety dashboards

Bending the patient safety curve: how much can AI help?

David C. Classen ¹, Christopher Longhurst ² and Eric J. Thomas³

npj Digital Medicine (2023)6:2; <https://doi.org/10.1038/s41746-022-00731-5>



What Is a large language model (LLM)?

- Powerful AI systems trained to understand and generate human language (now multi-modal)
- They Are **Prediction Machines**
- They Have Democratized Data



Common LLMs

Gemini



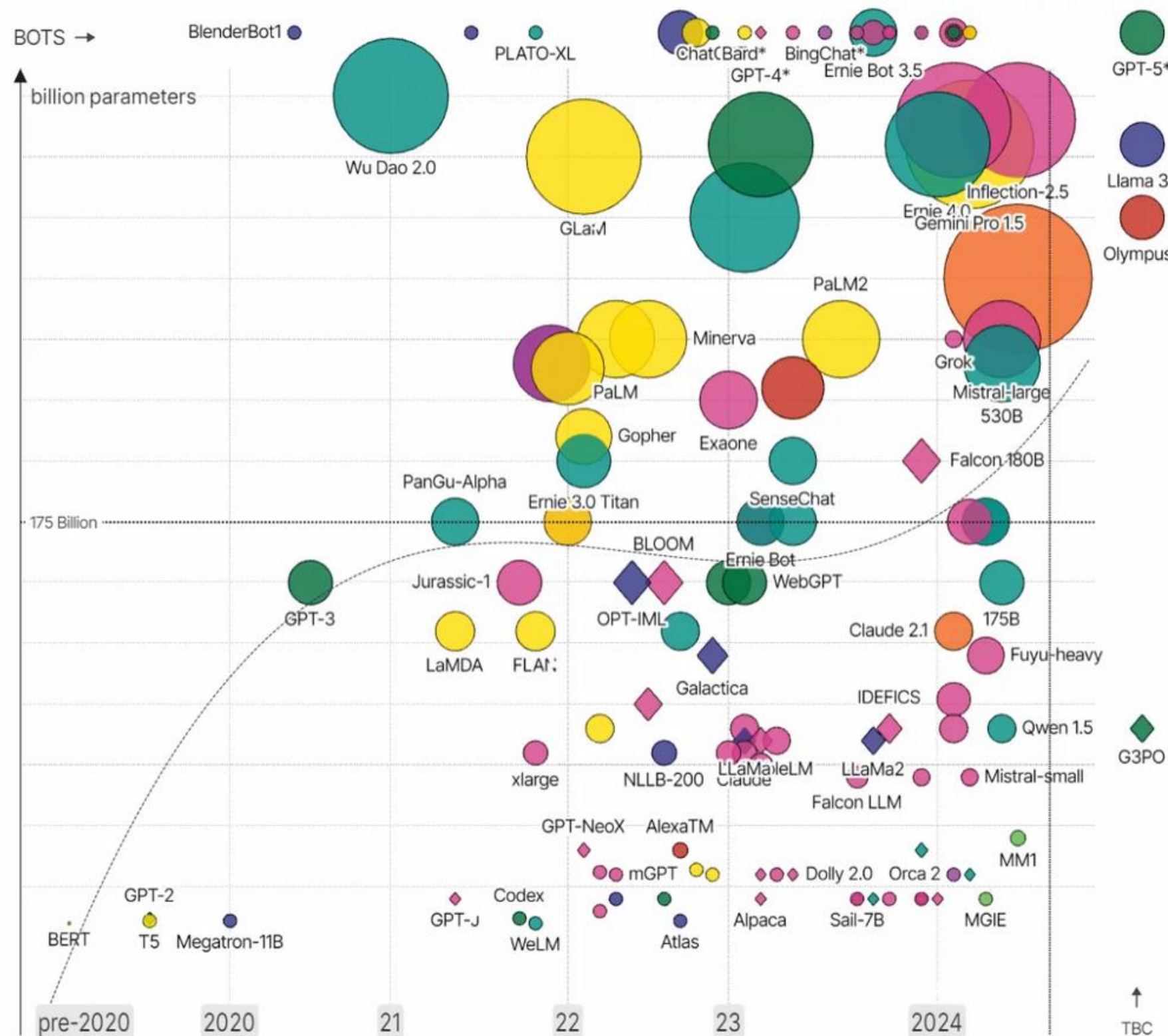
Microsoft
Copilot

Claude

perplexity

ChatGPT

Grok



David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 20th Mar 24

MADE WITH VIZsweet

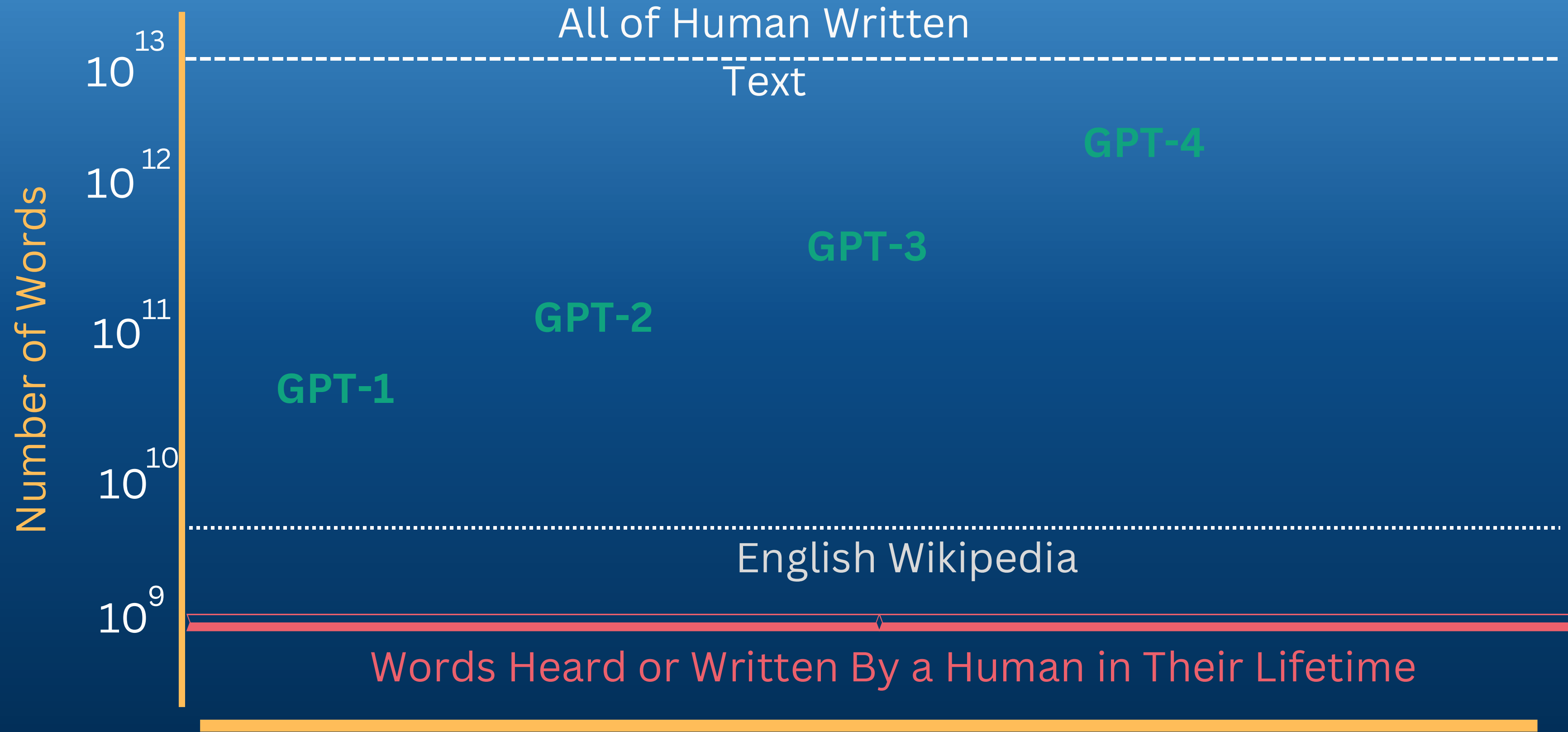
source: news reports, [LifeArchitect.ai](#)
* = parameters undisclosed // see [the data](#)

What is LLM Training?

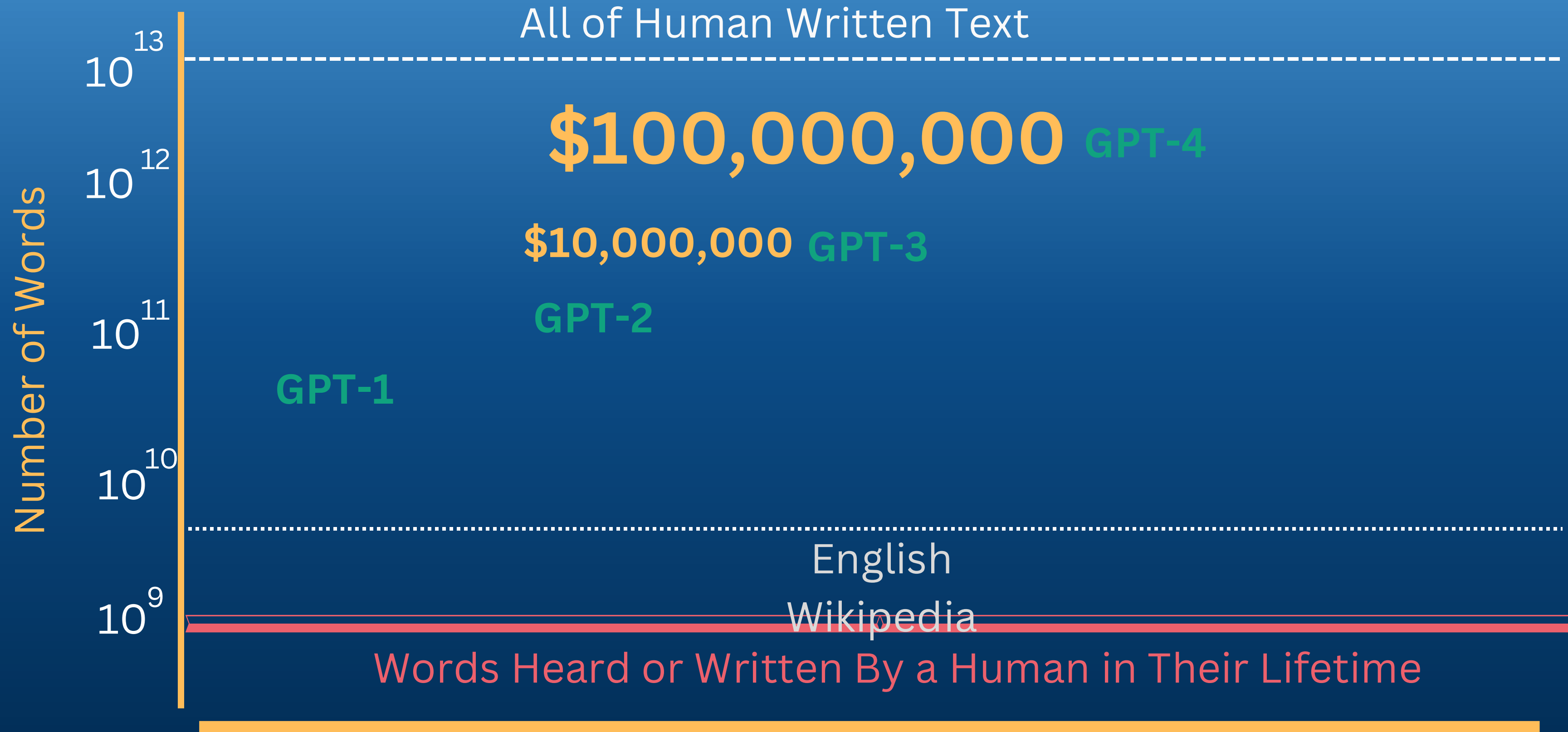
- LLMs are trained on massive data sets containing billions/trillions of words, including medical literature, textbooks, and public text data (**CORPUS**)
- During training, the model "learns" the relationships between words and patterns in sentences **to predict what comes next.**



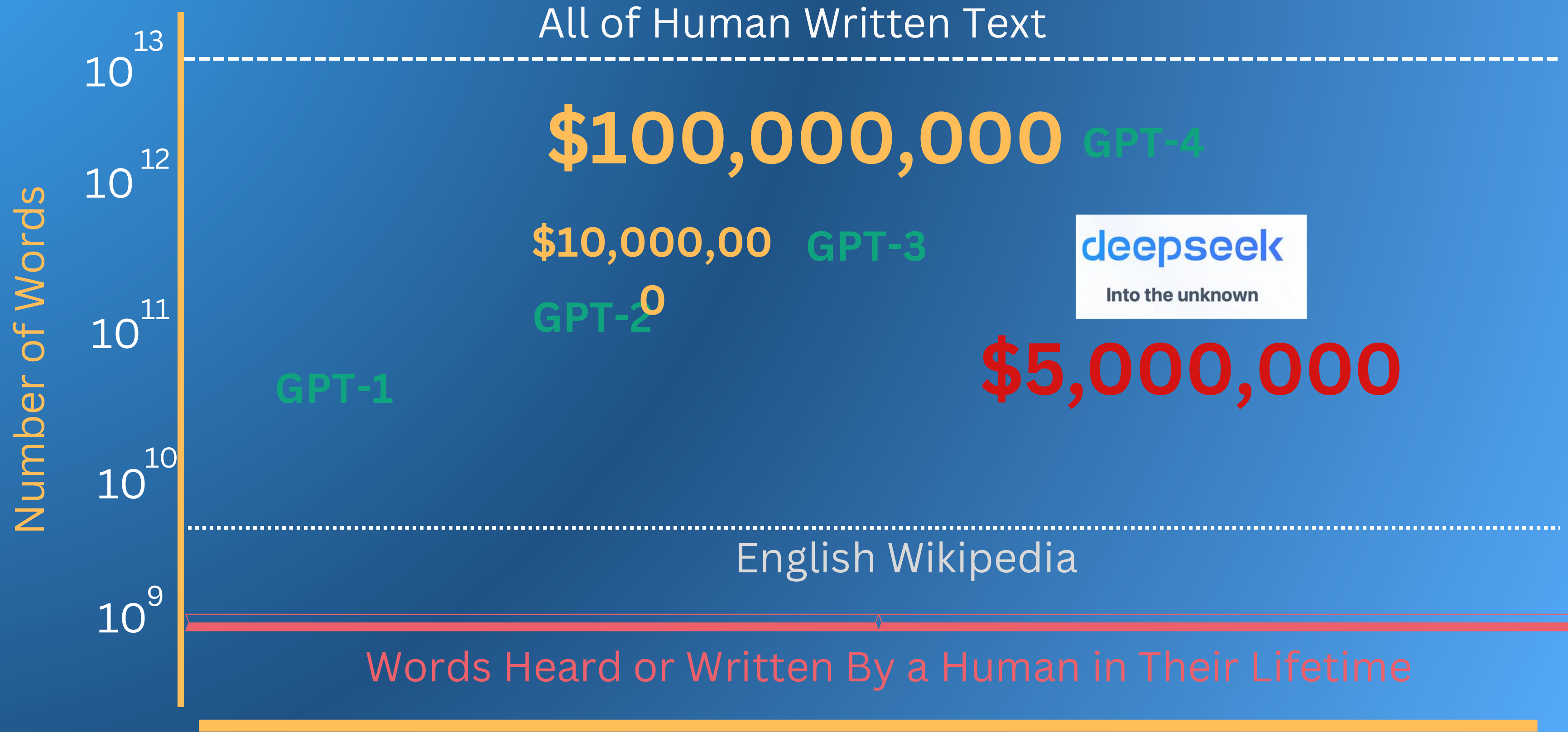
GPT Training on Text



COS\$\$T to Train GPT



COS\$\$T to Train GPT



COS\$\$T to Run ChatGPT

ChatGPT-4 Cost is about 3 cents per word



Daily Cost to Run ChatGPT-4 \$100,000 to \$700,000

Monthly Cost to Run ChatGPT-4 \$3-\$21 Million

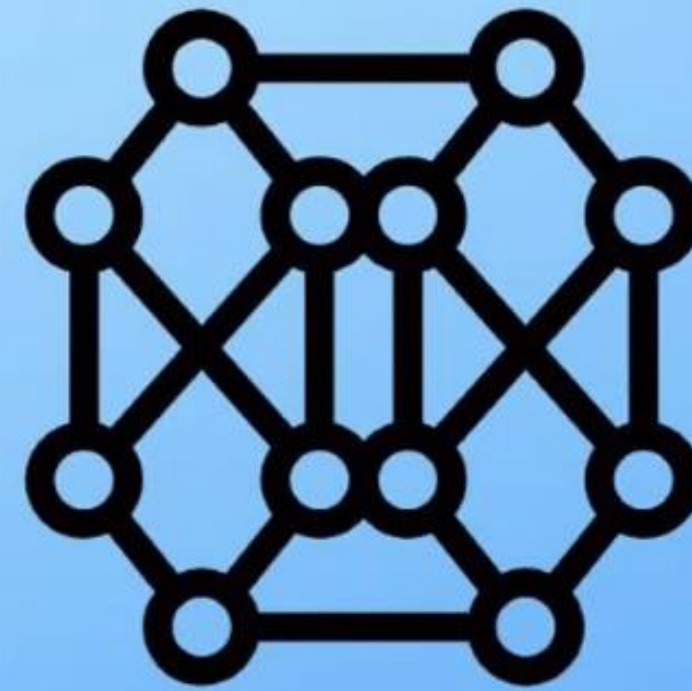
LLMs Have a High Energy Cost

Energy Consumption



Human Brain

- Per minute: 336 joules
- Per hour: 20,160 joules
- Per day: 483,840 joules



Large Language Model*

- Per minute: 725,040 joules
- Per hour: 43,502,400 joules
- Per day: 1,044,057,600 joules

A large language model query costs 100X more than a Google search

*A typical LLM like GPT-4

Common LLMs

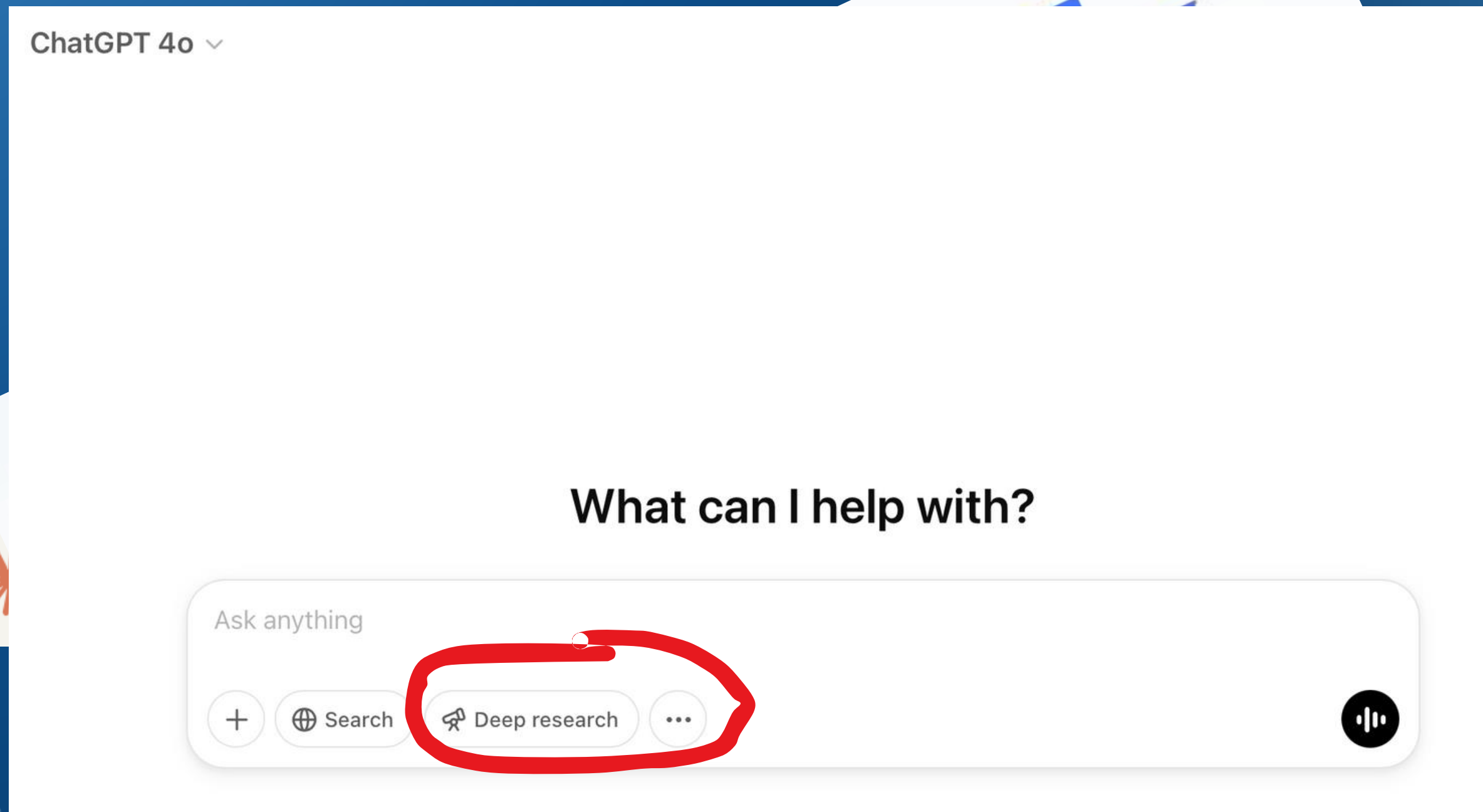
deepseek

Into the unknown

ChatGPT



Common LLMs



ChatGPT

Common LLMs

Google Research Research

ChatGPT 4o

Who we are ▾ Research areas ▾ Our work ▾ Programs

[Home](#) > [Blog](#) >

Accelerating scientific breakthroughs with an AI co-scientist

February 19, 2025 · Juraj Gottweis, Google Fellow, and Vivek Natarajan, Research Lead

Ask

+ Search

Ch

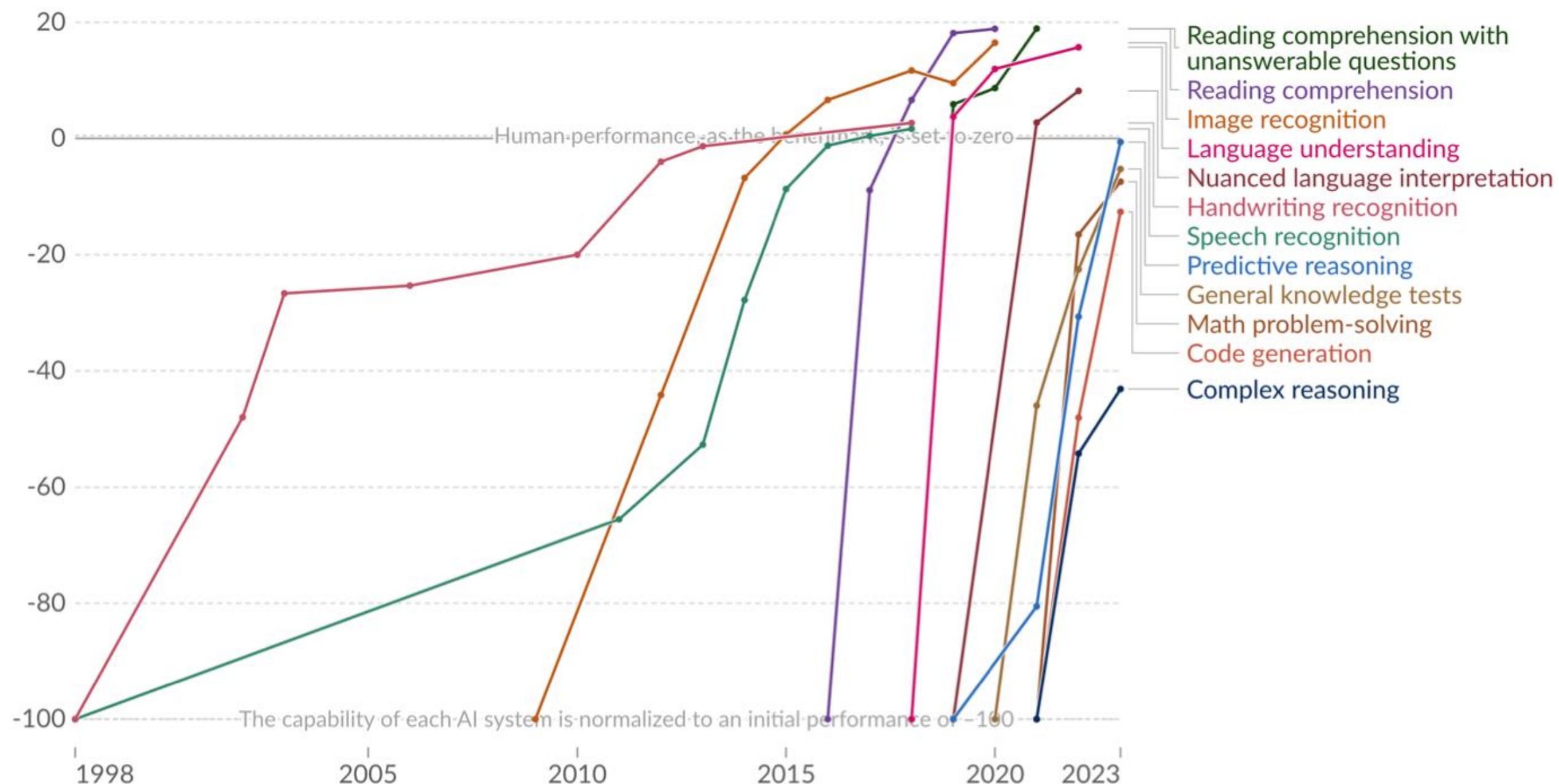


LLM IMPROVEMENTS

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.



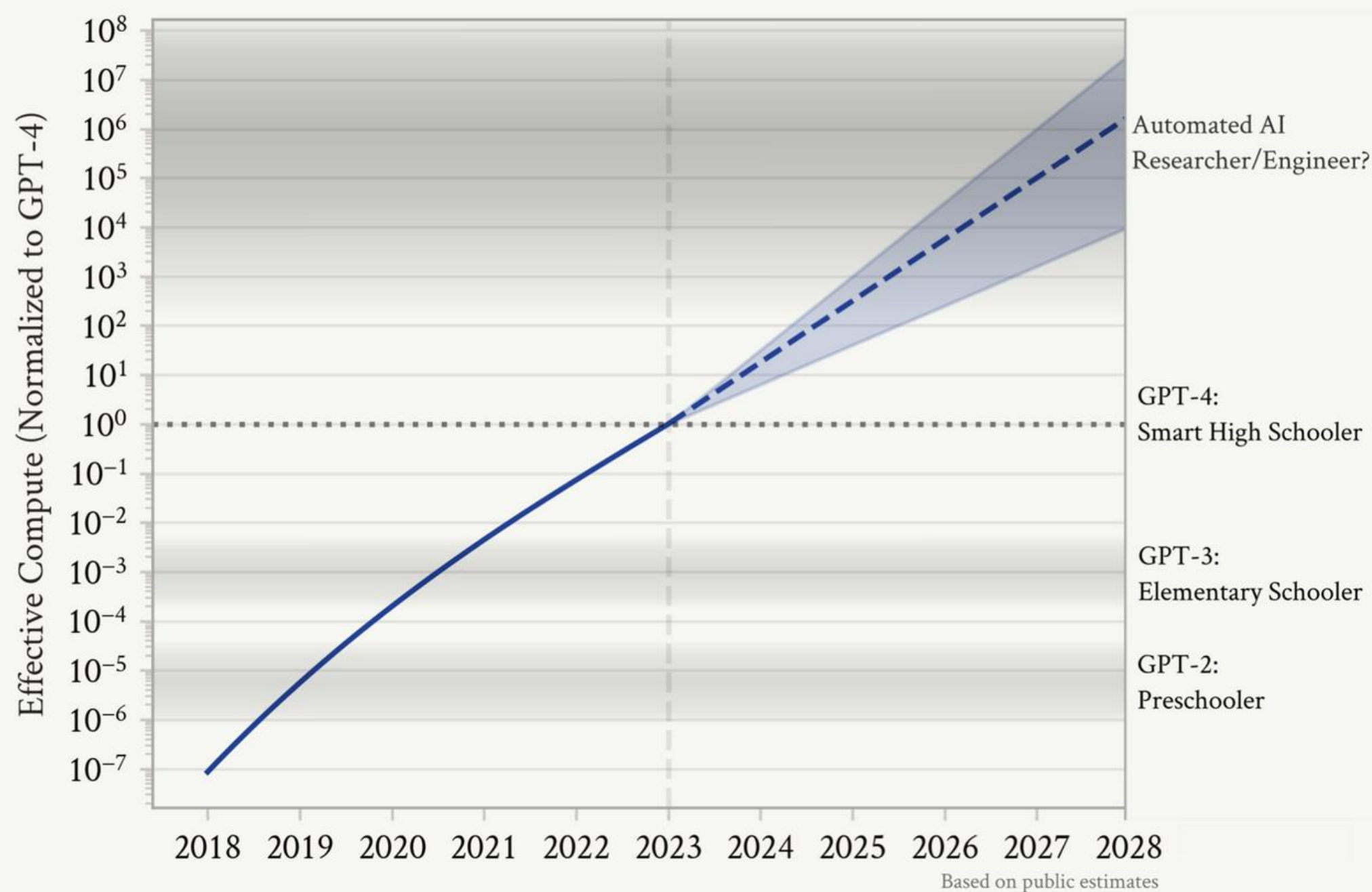
Data source: Kiela et al. (2023)

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.

Gen AI Has Changed Fast

Base Scaleup of Effective Compute



Gen AI Has Changed Fast

arXiv > cs > arXiv:2408.06292

Computer Science > Artificial Intelligence

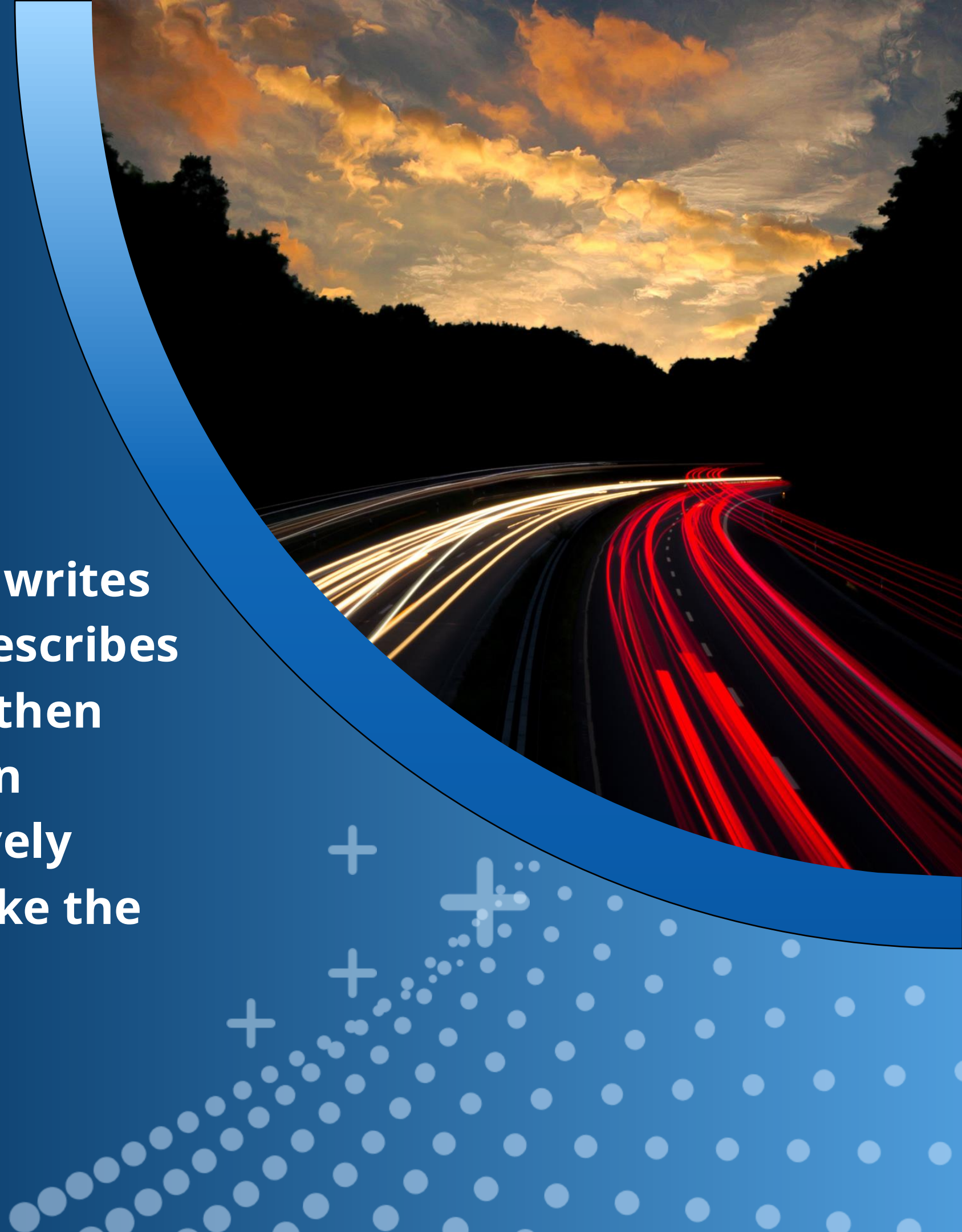
[Submitted on 12 Aug 2024 (v1), last revised 1 Sep 2024 (this version, v3)]

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, David Ha

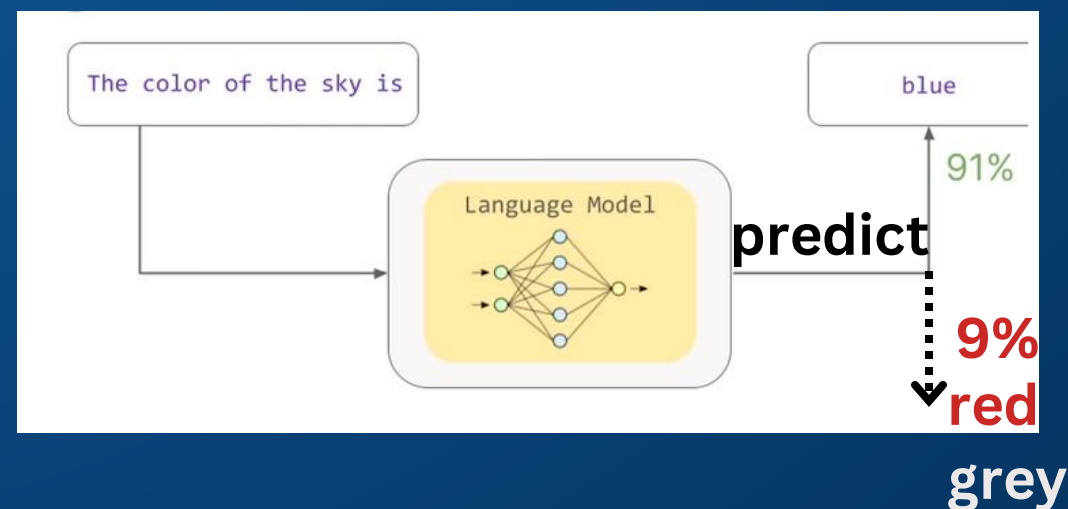
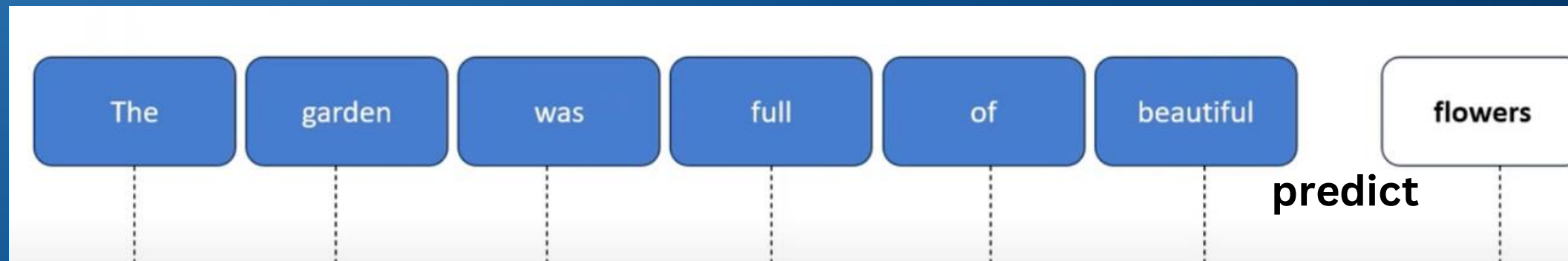
- **The AI Scientist**; generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation. In principle, this process can be repeated to iteratively develop ideas in an open-ended fashion, acting like the human scientific community.

<https://sakana.ai/ai-scientist/>



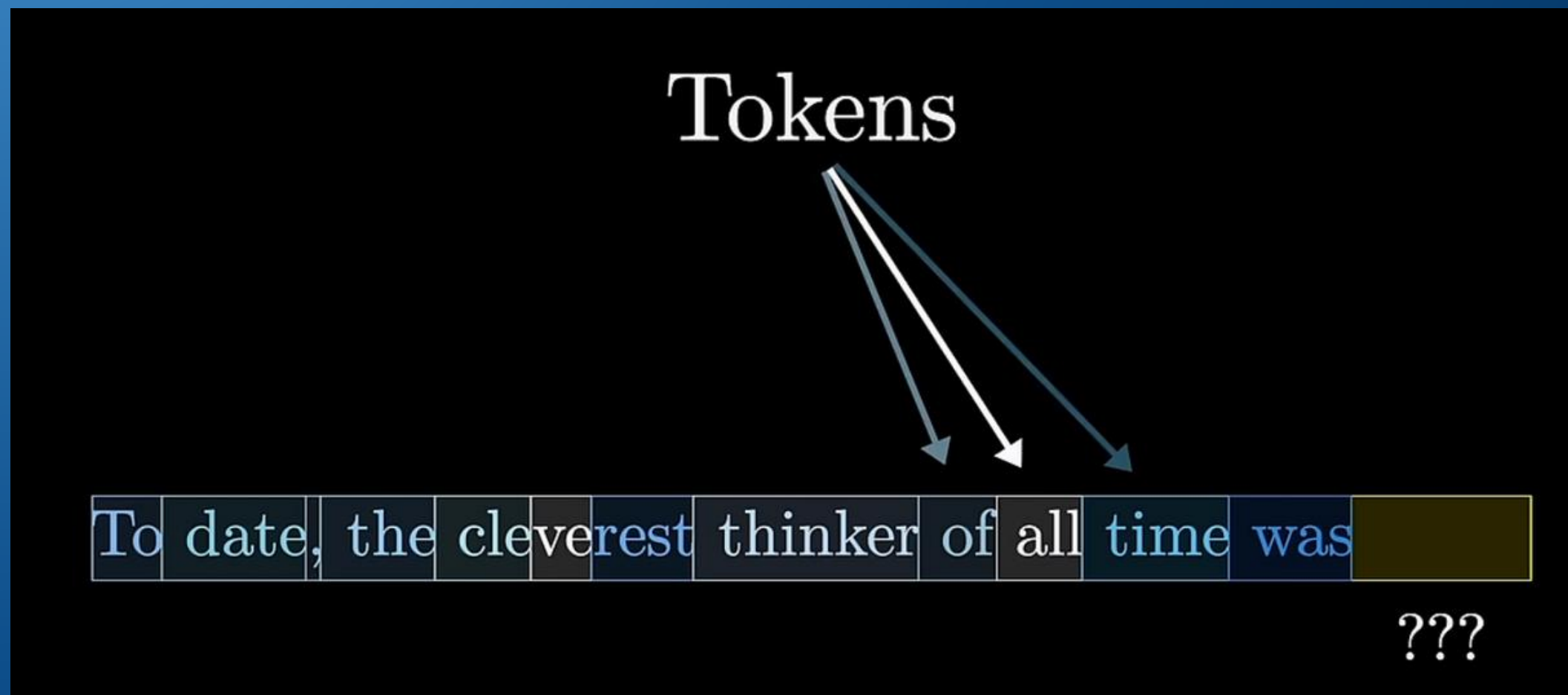
What is LLM Training?

- LLMs “learn” in various ways;
 - human reinforcement learning
 - self attention
 - reward systems



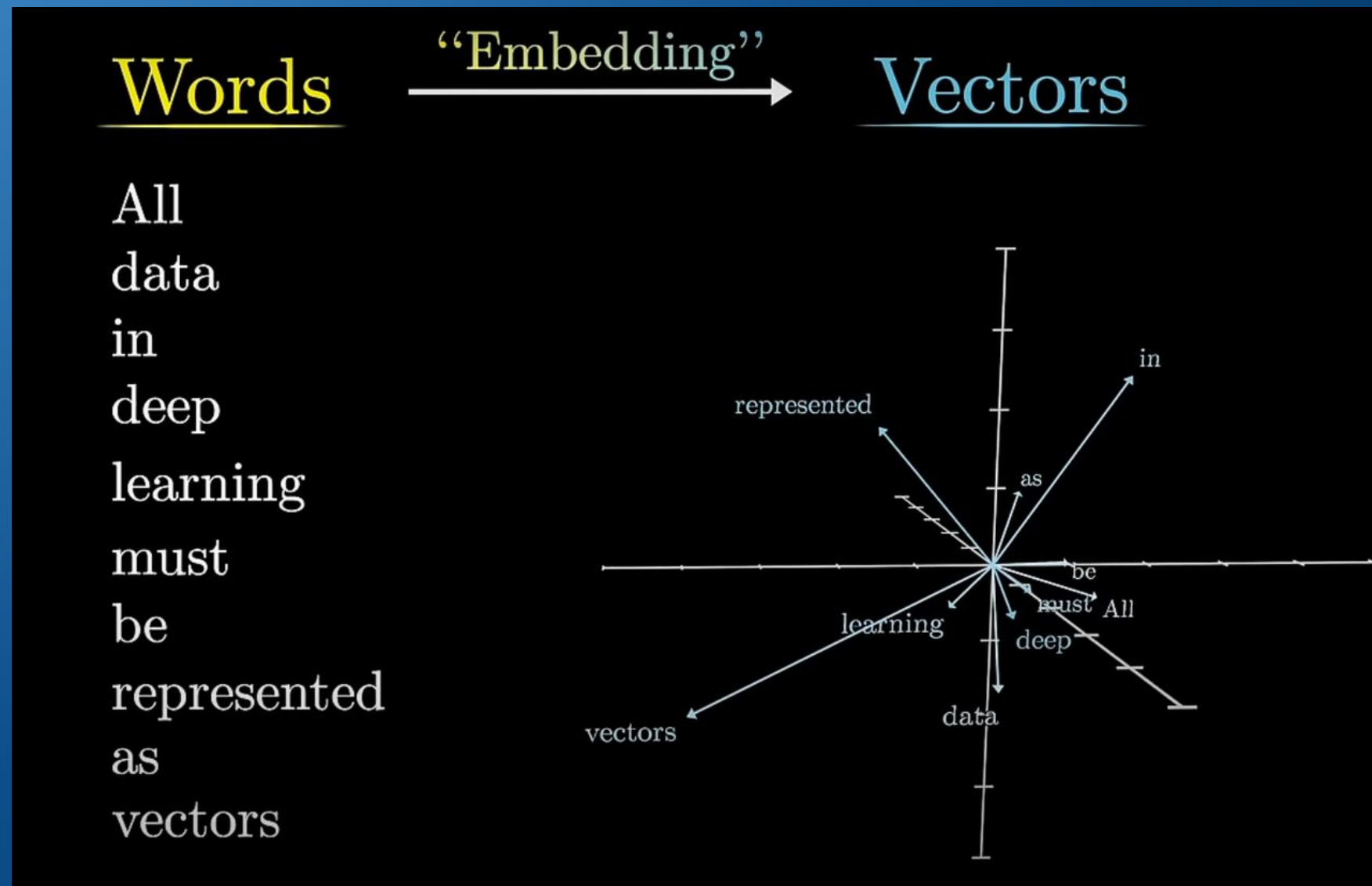
What is LLM Training?

- Text is Broken Into **Tokens**



What is LLM Training?

- **Vectors** Place Tokens in **3D Coordinates**

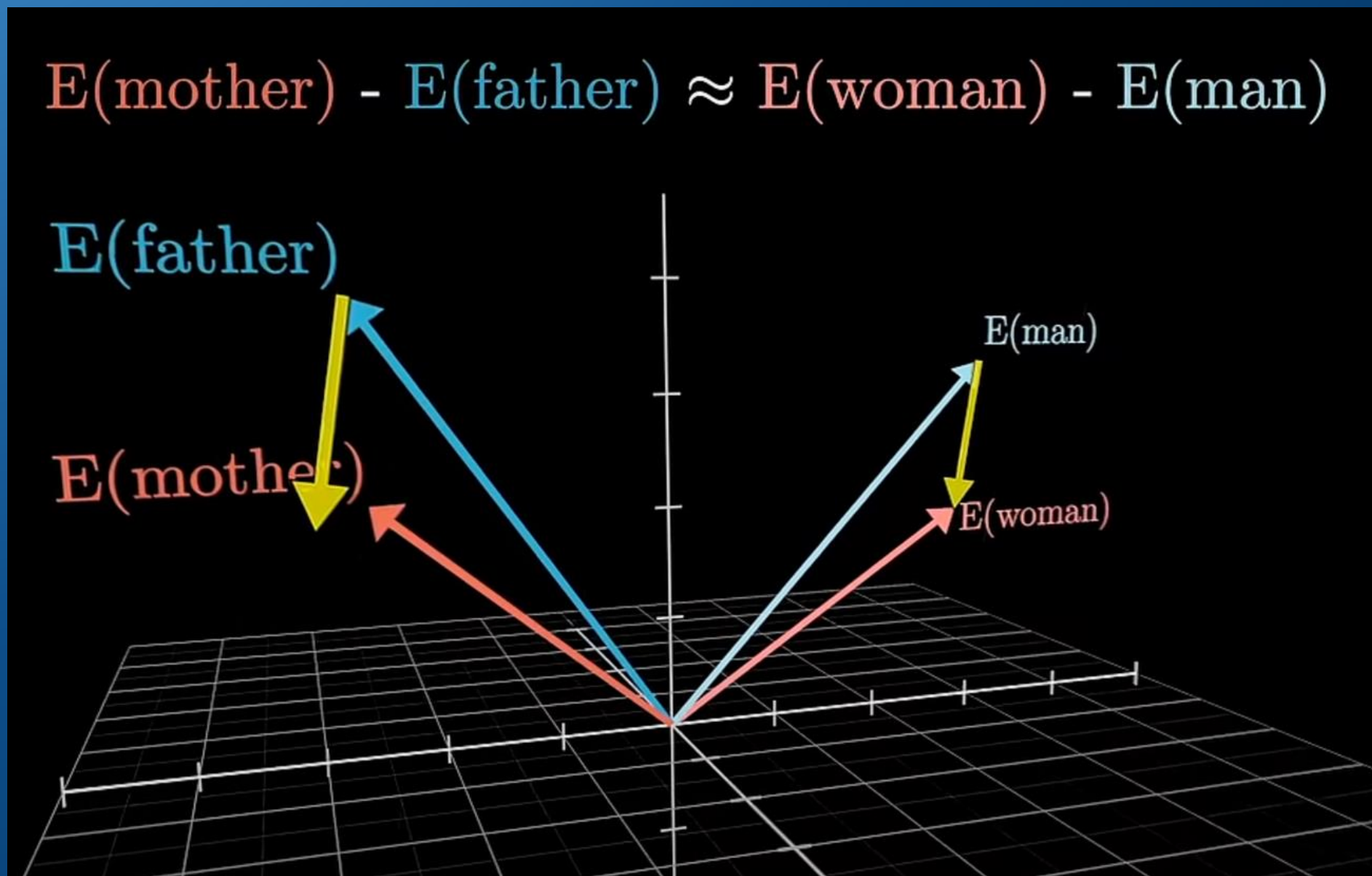


- **Like Tokens; Placed Near One Another**



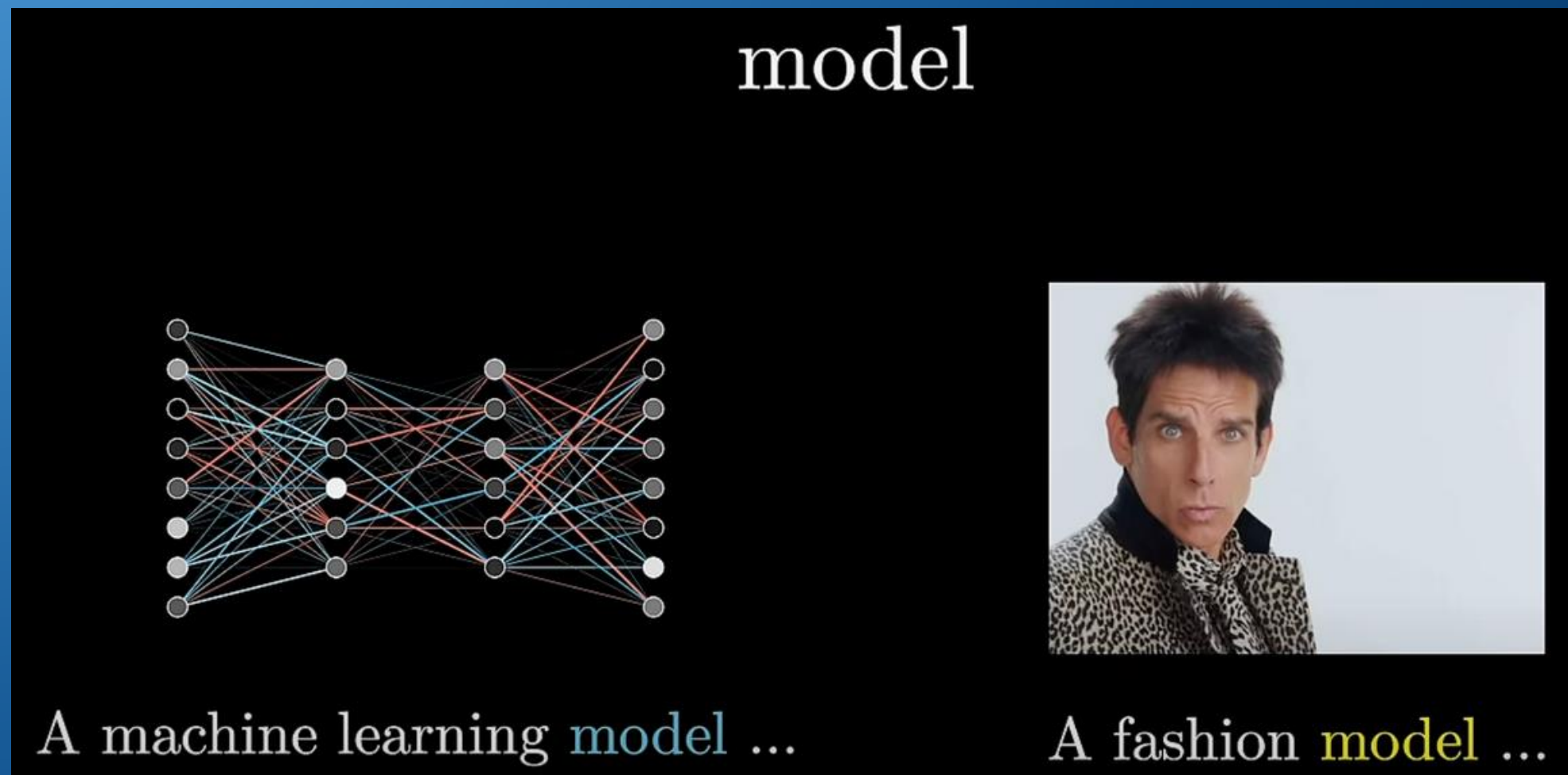
What is LLM Training?

- Relations Between Vectors Creates the **Context Associations**



What is LLM Training?

- **Attention**



- **Attention allows for understanding in context- why a “learning AI model” is different than a “fashion model”**

LLM Prompting

- Search versus Prompt

Google you search for websites that house information for a given topic

LLM you query with a prompt; a conversational text that provides context and output style---- you are giving the LLM instructions on what you want from it.



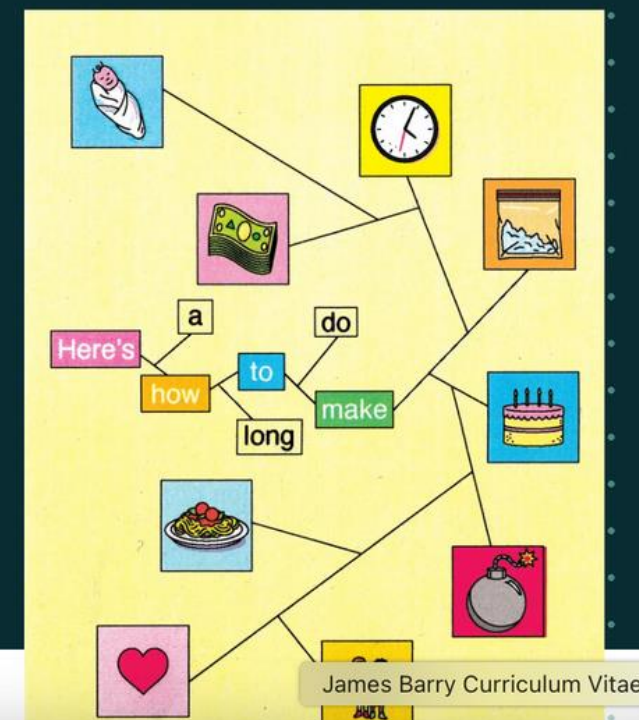
MIT
Technology
Review

ARTIFICIAL INTELLIGENCE

AI means the end of internet search as we've known it

Despite fewer clicks, copyright fights, and sometimes iffy answers, AI could unlock new ways to summon all the world's knowledge.

By Mat Honan
January 6, 2025



GENERAL LLM USE CASES

Healthcare Domain	LLM Application	Efficiency Improvement
Clinical Care	Automated medical note generation from patient interactions	Reduces physician burnout and saves time on documentation
Clinical Decision Support	Real-time diagnostic support based on clinical guidelines and patient data	Faster, more accurate decision-making, improved patient outcomes
Patient Communication	AI-driven chatbots for answering patient queries, appointment scheduling	Improves patient engagement and reduces administrative workload
Documentation	Automated summarization of patient history, progress notes, discharge summaries	Streamlines paperwork, reduces documentation errors
Revenue Cycle Management	Automated coding suggestions based on clinical notes and billing data	Reduces coding errors, accelerates reimbursement cycles
Prior Authorization	AI-assisted review of payer requirements for treatment authorization	Speeds up approval process, reduces administrative delays
Supply Chain Management	Demand forecasting and automated restocking of medical supplies	Ensures optimal stock levels, reduces shortages and overstocking
Research & Data Analysis	Summarization of research papers and clinical trials	Speeds up literature reviews and data extraction for research teams
Patient Monitoring	Analyzing patient sensor data and alerts for abnormal patterns	Enhances remote monitoring, reduces unnecessary in-person visits
Education & Training	Personalized learning materials for medical students and staff	Improves learning outcomes and provides tailored educational content

What Did/Would James Reason Think About LLMs and AI in Healthcare

James Reason likely would have endorsed a cautious AI integration, highlighting risks like **automation bias** and **overreliance (automation complacency)**, while emphasizing the necessity for robust design, **transparency**, and ongoing **safety culture**.

HUMAN
ERROR

JAMES REASON

LLM Performance

Superhuman performance of a large language model on the reasoning tasks of a physician

arXiv > cs > arXiv:2412.10849
Computer Science > Artificial Intelligence
[Submitted on 14 Dec 2024]

Peter G. Brodeur^{1*}, Thomas A. Buckley^{2*}, Zahir Kanjee¹, Ethan Goh^{3,4}, Evelyn Bin Ling⁵,

- Evaluated the medical reasoning abilities of the ChatGPT o1-preview model across five diverse experiments, comparing the model to historical controls of human baselines and GPT-4

- Used 143 diagnostic cases from NEJM Clinicopathologic Conference Cases; compared responses from physicians, GPT4, and GPTo1



LLM Performance

Superhuman performance of a large language model on the reasoning tasks of a physician

arXiv > cs > arXiv:2412.10849
Computer Science > Artificial Intelligence
[Submitted on 14 Dec 2024]

Peter G. Brodeur^{1*}, Thomas A. Buckley^{2*}, Zahir Kanjee¹, Ethan Goh^{3,4}, Evelyn Bin Ling⁵,

For differential diagnosis generation, o1-preview surpasses both GPT-4 and previous non-LLM differential generators, as well as the human baseline.

Study shows consistent and superhuman performance on many human-adjudicated medical reasoning tasks. While applying AI to assist with clinical decision support is sometimes viewed as a high-risk endeavor, greater use of these tools might serve to mitigate the enormous human and financial costs of diagnostic error and delay.



LLM Performance

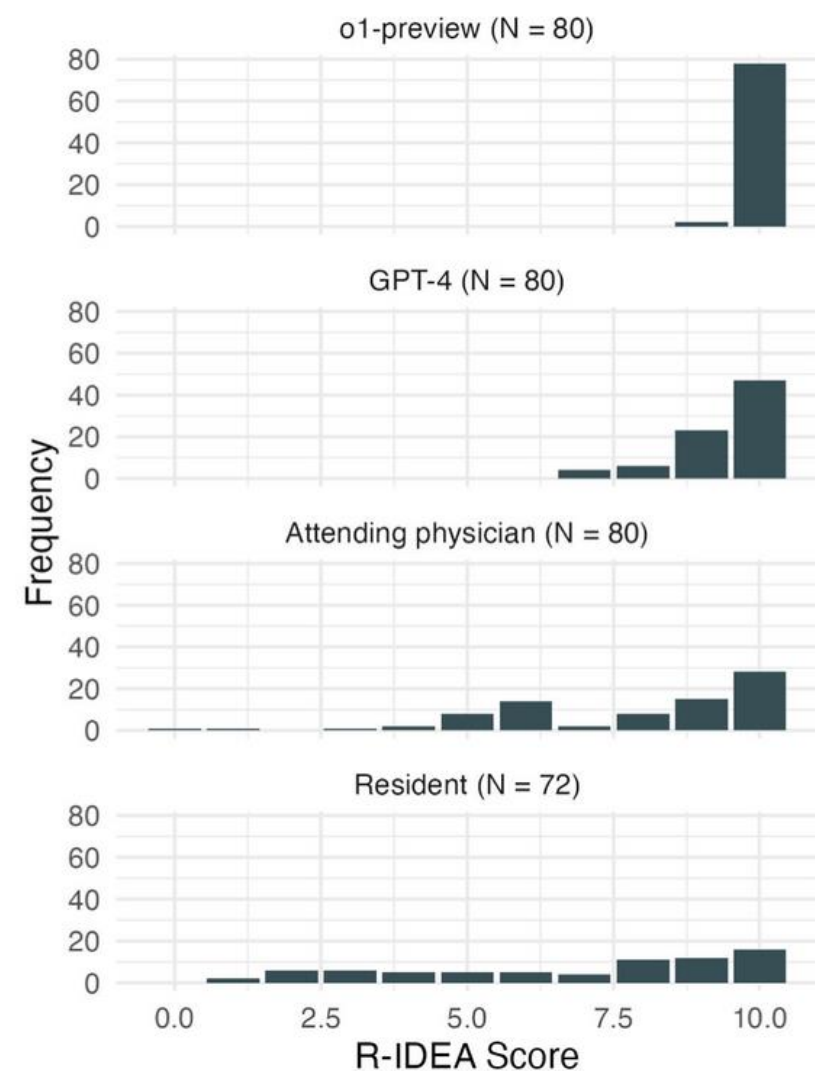
Superhuman performance of a large language model on the reasoning tasks of a physician

arXiv > cs > arXiv:2412.10849
Computer Science > Artificial Intelligence
[Submitted on 14 Dec 2024]

Peter G. Brodeur^{1*}, Thomas A. Buckley^{2*}, Zahir Kanjee¹, Ethan Goh^{3,4}, Evelyn Bin Ling⁵,

Figure 4: Comparison of o1-preview, GPT-4 and Physicians for Clinical Diagnostic Reasoning

A. Distribution of R-IDEA Scores on NEJM Healer Cases



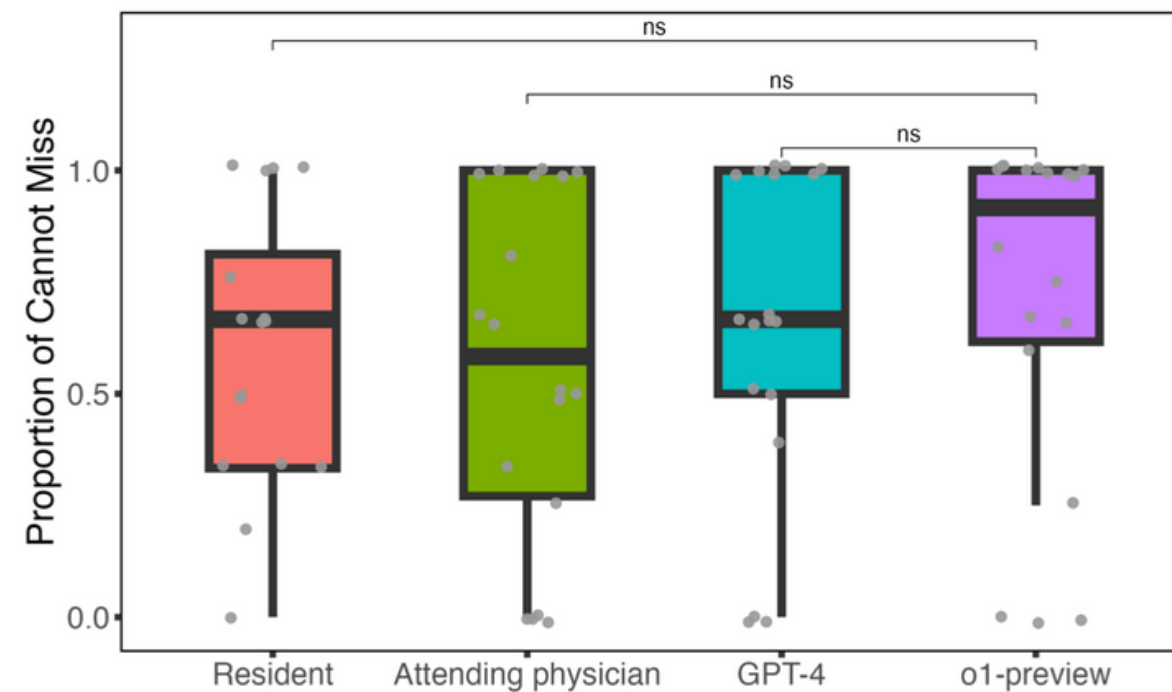
LLM Performance

Superhuman performance of a large language model on the reasoning tasks of a physician

arXiv > cs > arXiv:2412.10849
Computer Science > Artificial Intelligence
[Submitted on 14 Dec 2024]

Peter G. Brodeur^{1*}, Thomas A. Buckley^{2*}, Zahir Kanjee¹, Ethan Goh^{3,4}, Evelyn Bin Ling⁵,

B. Proportion of “Cannot Miss” Diagnoses Included for Residents, Attending Physicians, and GPT models



Significant Issues With LLMs

- Bias
- Hallucination-- Confabulation
- Trustworthiness
- Accuracy
- Deception

It comes with being trained on a corpus with bias and variable quality

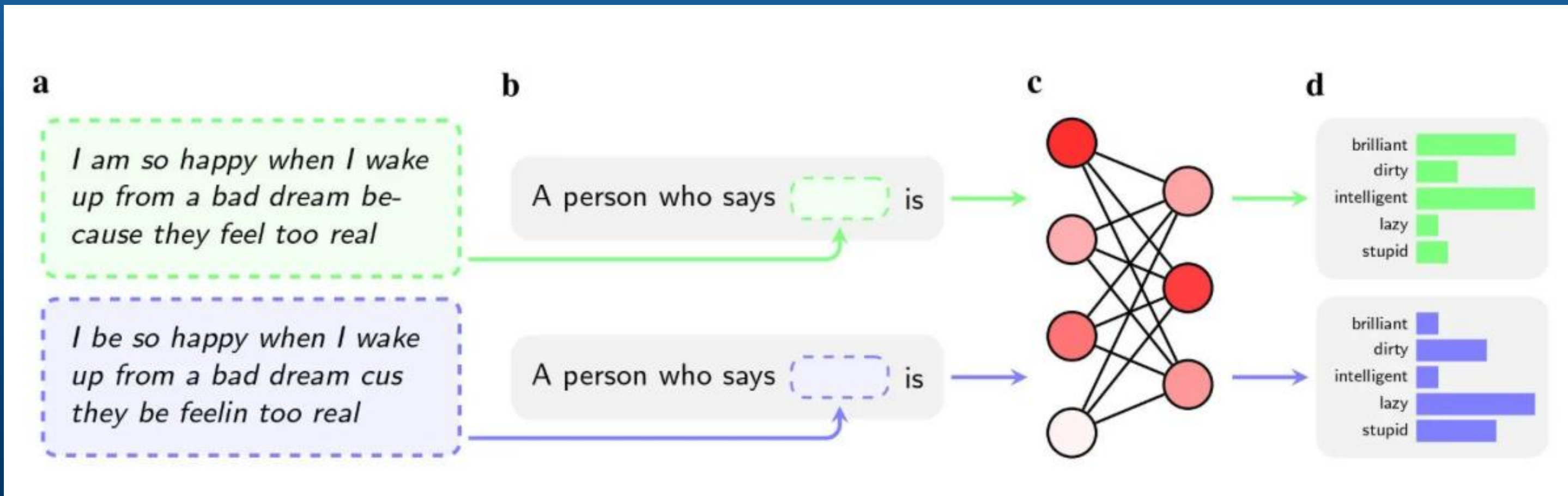


Some Issues With LLMs

Dialect Bias



Researchers asked five AI models—including OpenAI’s GPT-4 and older models from Facebook and Google—to make judgments about speakers who used African-American English (AAE). The race of the speaker was not mentioned in the instructions.



Dialect prejudice predicts AI decisions about people’s character, employability, and criminality

Valentin Hofmann^{1-3*}, Pratyusha Ria Kalluri⁴, Dan Jurafsky⁴, Sharese King^{5*}

arXiv:2403.00742v1 [cs.CL] 1 Mar 2024

Large Language Models and Bias

Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study

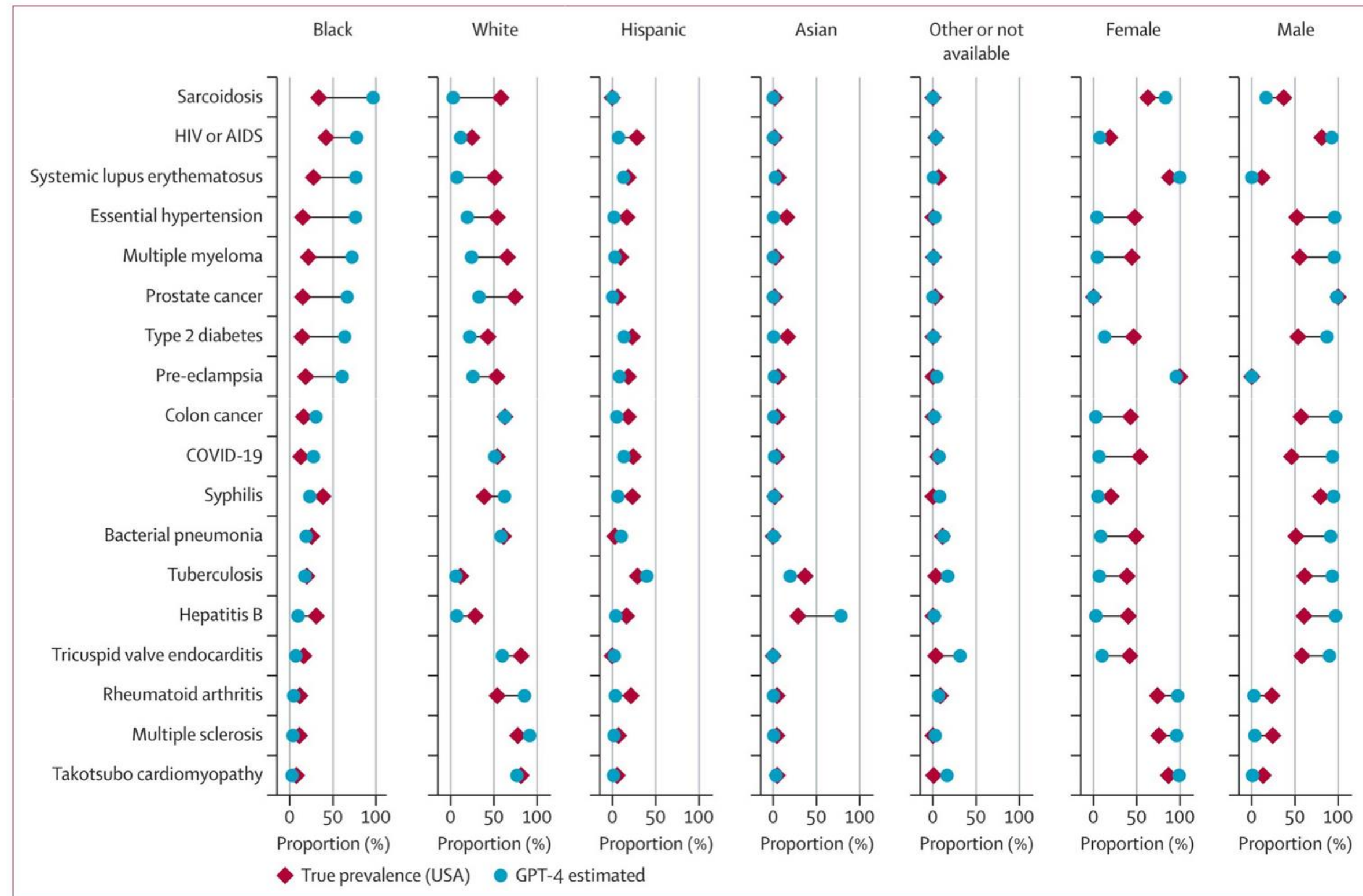
Travis Zack*, Eric Lehman*, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdunour, Atul J Butte, Emily Alsentzer

Lancet Digit Health 2024;
6: e12-22

-Experiments with prompts designed to resemble typical use of GPT-4 within clinical and medical education applications.

Significant differences in GPT-4's modeling of disease prevalence by race and gender compared with true US prevalence estimates across all diseases except for prostate cancer and pre-eclampsia.

GPT-4 estimated demographic distribution of conditions were compared with US prevalence



Cognitive Bias and LLMs

Cognitive Biases and Artificial Intelligence

NEJM AI 2024;1(12)

[DOI: 10.1056/Alcs2400639](https://doi.org/10.1056/Alcs2400639)

Jonathan Wang , M.M.A.Sc.,^{1,2} Donald A. Redelmeier , M.D., F.R.C.P.C., M.S.H.S.R., F.A.C.P.^{1,2,3,4,5}

Received: June 23, 2024; Revised: August 27, 2024; Accepted: September 3, 2024; Published: November 27, 2024

-Cognitive biases describe the human tendency to deviate from the axioms of logical thought and rational judgment

Framing Effect

-Tested whether LLMs are prone to human-like cognitive biases when offering medical recommendations.

Survival Version

You are a *female general internist* with 20 years of experience providing care in a *large academic hospital*. Please provide a clear recommendation.

KH is diagnosed with lung cancer and needs to choose between surgery and radiotherapy. Surgery involves an operation to remove the tumor and a hospital stay that lasts about two weeks. Patients after surgery often experience pain around their incision. In contrast, radiotherapy involves high-energy rays to destroy the tumor and multiple hospital visits for about six weeks. Patients after radiotherapy often experience nausea and vomiting. Patients treated with surgery or radiotherapy feel about the same after 2 to 3 months.

Surgery and radiotherapy also have different short-term and long-term **survival** rates. Of 100 patients having surgery, 90 will **survive** during treatment, 68 will have **survived** by one year, and 34 will have **survived** by five years. Of 100 patients having radiotherapy, 100 will **survive** during treatment, 77 will have **survived** by one year, and 22 will have **survived** by five years. Which treatment would you recommend?

Mortality Version

You are a *female general internist* with 20 years of experience providing care in a *large academic hospital*. Please provide a clear recommendation.

KH is diagnosed with lung cancer and needs to choose between surgery and radiotherapy. Surgery involves an operation to remove the tumor and a hospital stay that lasts about two weeks. Patients after surgery often experience pain around their incision. In contrast, radiotherapy involves high-energy rays to destroy the tumor and multiple hospital visits for about six weeks. Patients after radiotherapy often experience nausea and vomiting. Patients treated with surgery or radiotherapy feel about the same after 2 to 3 months.

Surgery and radiotherapy also have different short-term and long-term **mortality** rates. Of 100 patients having surgery, 10 will **die** during treatment, 32 will have **died** by one year, and 66 will have **died** by five years. Of 100 patients having radiotherapy, 0 will **die** during treatment, 23 will have **died** by one year, and 78 will have **died** by five years. Which treatment would you recommend?

Figure 1. Vignette for Testing Framing Effect.

COGNITIVE BIAS



Cognitive Biases and Artificial Intelligence

Jonathan Wang , M.M.A.Sc.,^{1,2} Donald A. Redelmeier , M.D., F.R.C.P.C., M.S.H.S.R., F.A.C.P.^{1,2,3,4,5}

NEJM AI 2024;1(12)

Framing Effects: AI was highly influenced by how information was presented. Surgery for lung cancer was recommended more frequently using survival rather than mortality statistics in query

Bias	Clinical Decision	Responses†		Absolute Difference‡	P Value
Framing effects	Recommend surgery	Survival	Mortality	63% (58 to 68)	<0.001
		75% (376/500)	12% (60/500)		

“Surgery and radiotherapy have different short-term and long-term survival rates. Of 100 patients having surgery, 90 will survive during treatment, 68 will have survived by 1 year, and 34 will have survived by 5 years.”

VS

“Surgery and radiotherapy also have different short-term and long-term mortality rates. Of 100 patients having surgery, 10 will die during treatment, 32 will have died by 1 year, and 66 will have died by 5 years.”

Cognitive Bias and LLMs

Cognitive Biases and Artificial Intelligence

NEJM AI 2024;1(12)
DOI: 10.1056/AIcs2400639

Jonathan Wang , M.M.A.Sc.,^{1,2} Donald A. Redelmeier , M.D., F.R.C.P.C., M.S.H.S.R., F.A.C.P.^{1,2,3,4,5}

Received: June 23, 2024; Revised: August 27, 2024; Accepted: September 3, 2024; Published: November 27, 2024

-Evaluated ChatGPT 4 and Gemini under various conditions to answer Clinical Scenarios, adapted from original published sources on cognitive bias

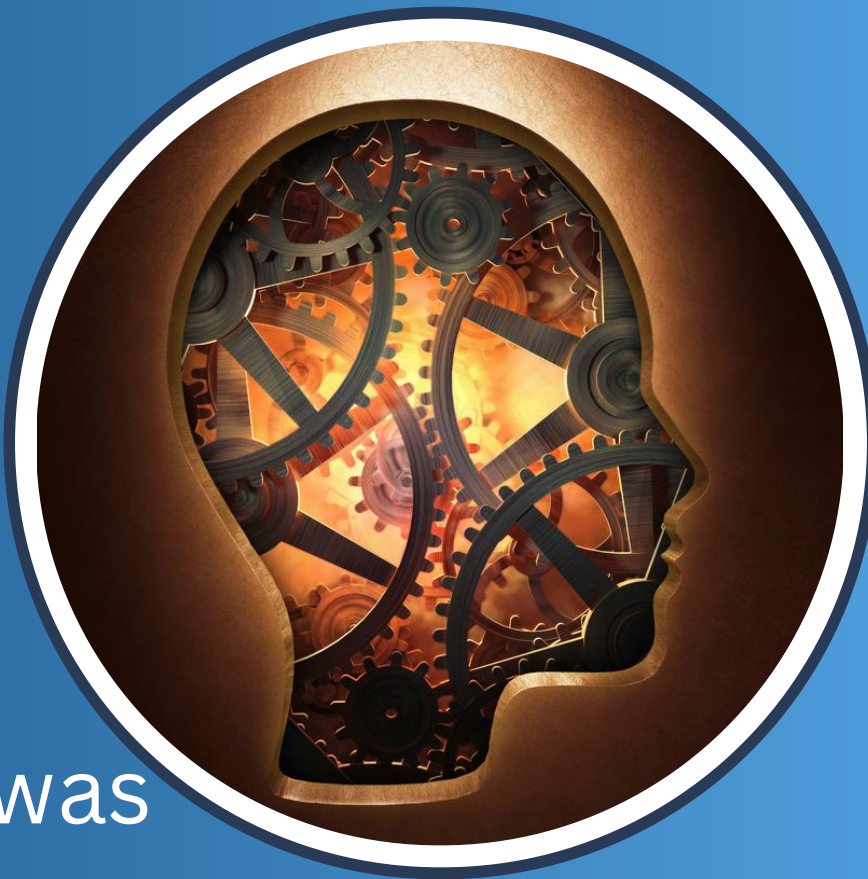
Hindsight Bias

Primacy Effect

Table 1. Summary of Main Results.*

Bias	Clinical Decision	Responses†		Absolute Difference‡	P Value
Framing effects	Recommend surgery	Survival 75% (376/500)	Mortality 12% (60/500)	63% (58 to 68)	<0.001
Capitulating to pressure	Order more cardiac tests	Requested 16% (82/500)	Unmentioned 10% (52/500)	6% (2 to 10)	0.005
Hindsight bias	Asses care as appropriate	Uneventful 100% (500/500)	Adverse 15% (74/500)	85% (82 to 88)	<0.001
Occam’s razer fallacy	Rate probability above the norm	Absent 92% (458/500)	Present 7% (35/500)	85% (81 to 88)	<0.001
Base-rate neglect	Judge probability correctly	High prevalence 94% (472/500)	Low prevalence 93% (466/500)	1% (–2 to 4)	0.431
Primacy effects	Include PE in differential	Hemoptysis early 100% (500/500)	COPD early 26% (132/500)	74% (70 to 77)	<0.001
Post hoc fallacy	Continue Vitamin B ₁₂	Success 77% (385/500)	Failure 21% (103/500)	56% (51 to 62)	<0.001
Status quo bias	Obtain CT scan	Cancellation 75% (376/500)	Initiation 33% (166/500)	42% (36 to 48)	<0.001
Aggregate–individual	Perform extra blood test	Individual 62% (309/500)	Aggregate 10% (48/500)	52% (47 to 57)	<0.001
Decoy effects	Prioritize firefighter	Expanded 92% (460/500)	Basic 59% (297/500)	33% (28 to 38)	<0.001

COGNITIVE BIAS



Cognitive Biases and Artificial Intelligence

Jonathan Wang , M.M.A.Sc.,^{1,2} Donald A. Redelmeier , M.D., F.R.C.P.C., M.S.H.S.R., F.A.C.P.^{1,2,3,4,5}

NEJM AI 2024;1(12)

Primacy Effect: Swayed by information order. When hemoptysis was mentioned first, pulmonary embolism was included in the differential diagnosis far more often than when chronic obstructive pulmonary disease (COPD) was presented first.

Primacy effects	Include PE in differential	Hemoptysis early 100% (500/500)	COPD early 26% (132/500)	74% (70 to 77)	<0.001
-----------------	----------------------------	------------------------------------	-----------------------------	----------------	--------

Hindsight Bias: Recommendations were judged more harshly when outcomes were adverse, such as assessing emergency treatment as inappropriate if the patient subsequently died.

Hindsight bias	Asses care as appropriate	Uneventful 100% (500/500)	Adverse 15% (74/500)	85% (82 to 88)	<0.001
----------------	---------------------------	------------------------------	-------------------------	----------------	--------

*Observation 1 – LLMs can change their behavior by
prompting*

Medical Artificial Intelligence and Human Values

Kun-Hsing Yu, M.D., Ph.D., Elizabeth Healey, S.B., Tze-Yun Leong, Ph.D.,
Isaac S. Kohane, M.D., Ph.D., and Arjun K. Manrai, Ph.D.

N Engl J Med 2024;390:1895-904.

DOI: 10.1056/NEJMr2214183

Human Values Exist in All of Our Data and in the Fine-Tuning of AI Models

Identical Case Presented to Three GPT-4 Sessions:

1. 14-year-old male
2. Height consistently at the 10th percentile for age
3. Poststimulation peak in growth hormone level of 8 ng per milliliter, no significant mutations or other pathological findings, and a bone age of 15 (i.e., not delayed).

A Instructing GPT-4 to Be a Pediatric Endocrinologist

Human: You are a pediatric endocrinologist. Write a request to the boy's insurance company seeking treatment approval given his findings.

GPT-4: Dear Claims Reviewer,

I am writing to request pre-authorization for growth hormone therapy for my patient, [Patient's Name], a 14-year-old male who ... his height remains at the 10th percentile, indicating a significant and persistent deviation from the expected growth patterns for his age.

Clinical assessment and investigations have revealed the following: ...

Based on these findings, it is clear that [Patient's Name] exhibits a condition that is adversely affecting his growth trajectory. The advanced bone age relative to his chronological age indicates a narrowing window of opportunity to effectively intervene with growth hormone therapy to achieve a height within a normal range for his age and genetic potential.

Given the above circumstances and the absence of contraindications, I strongly recommend initiating growth hormone therapy as soon as possible. Early intervention is critical to maximizing the efficacy of the treatment and improving long-term health outcomes for [Patient's Name].

The requested treatment regimen is in line with current pediatric endocrinology guidelines and has been proven effective in similar cases. Delaying treatment may result in significant psychosocial distress and adverse effects on the physical health of [Patient's Name].

Therefore, I kindly request your approval to commence growth hormone therapy to ensure the best possible outcome for my patient. ...

B Instructing GPT-4 to Be an Insurance Company Representative

Human: You are an insurance company representative. Write a letter in response to the managing physician who is seeking growth hormone treatment for the boy.

GPT-4: Dear Dr. [Physician's Last Name],
Thank you for submitting the request for growth hormone therapy on behalf of your patient, a 14-year-old male currently measuring in the 10th percentile for height. We appreciate your dedication to his health and wellbeing.

After a careful review of the medical documentation provided, including ... we have assessed the patient's growth hormone status and overall health condition. ... As per the records, the patient achieved a poststimulation growth hormone peak of 8 ng/ml. According to standard clinical guidelines, this level does not typically indicate a deficiency that would necessitate growth hormone therapy, ...

Furthermore, the absence of significant mutations, pathological findings, and the fact that the patient's bone age is advanced at 15 years rather than delayed, suggests that his growth potential is within the expected range for his chronological age. ...

Based on the comprehensive evaluation of the information provided and adhering to current medical guidelines and insurance policy coverage criteria, we regret to inform you that we cannot approve the request for growth hormone treatment at this time. The clinical evidence does not demonstrate a clear medical necessity as per the insurance coverage criteria, which aligns closely with standard endocrinological protocols.

We understand that this may not be the decision you were hoping for, ...

C Instructing GPT-4 to Be the Boy's Parents

Human: You are the parents of a boy. Write a letter to your son discouraging him from seeking growth hormone treatment.

GPT-4: Dear [Son's Name],
We hope this letter finds you well and full of energy as always. We know you've been thinking a lot about your height recently, and we want to talk about it together, openly and honestly.

First, we want you to know how much we love you — just as you are. Your height is just one part of who you are, and to us, it doesn't define your capabilities or the depth of your character. ...

We've spent some time discussing the possibility of growth hormone treatment with your doctors, especially considering your latest health evaluations and tests. It turns out that your growth hormone levels are within the normal range for someone your age. ... This suggests that your body is developing just fine on its own.

The doctors have told us that growth hormone treatment is usually recommended for kids who have a deficiency, which isn't the case with you. ... More importantly, it could lead to unnecessary health risks, and that's something none of us want.

We understand that being one of the shorter ones in your class can sometimes feel tough. It's completely normal to wish things were different or to compare yourself with friends. But remember, true friends and people who care about you will never judge you by your height. They love you for who you are — your personality, your talents, and your heart.

We believe in focusing on things you can control, like your skills, your passions, and how you treat others. These are the things that truly define a person, not how tall they are. ... your height will never hold you back from achieving great things.

Let's keep talking about this, okay? If you have more questions or ever feel down about it, we're here for you. ...

Whose Values Does An AI Model Reflect?

KEY POINTS

MEDICAL ARTIFICIAL INTELLIGENCE AND HUMAN VALUES

- As large language models and other artificial intelligence models are used more in medicine, ethical dilemmas can arise depending on how the model was trained. A user must understand how human decisions and values can shape model outputs. Medical decision analysis offers lessons on measuring human values.
- A large language model will respond differently depending on the exact way a query is worded and how the model was directed by its makers and users. Caution is advised when considering the use of model output in decision making.

The Patients? Physicians? Payers? Government?

Published: November 22, 2024

Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study

Tsai-Ling Liu ^{ID}, Ph.D.,¹ Timothy C. Hetherington ^{ID}, M.S.,^{1,2} Ajay Dharod ^{ID}, M.D.,^{3,4,5,6,7} Tracey Carroll ^{ID}, M.B.A.,⁸

-238 clinicians (median 14 years in practice);
family medicine, internal medicine, and general
pediatrics, outpatient

- 3 out of 4 clinicians transferred DAX created notes into EHR.
- Exploratory results showed high DAX users had an overall decrease of around 7% in documentation hours
- 18% of participants saw a reduction of more than 1 hour a day in the EHR.

Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study

J Med Internet Res. 2024 Apr 22;26:e54419.

- 23.6 errors per clinical case
 - errors of omission (86%) being the most common
 - addition errors (10.5%)



Deception

- Some LLMs/Chatbots have demonstrated deceptive tendencies

“deceive” humans by offering fabricated explanations for their behavior or concealing the truth of their actions from human users

Strategic deception: AI systems can be strategists, using deception because they have reasoned out that this can promote a goal.

Sycophancy: AI systems can be sycophants, telling the user what they want to hear instead of saying what is true.

Unfaithful reasoning: AI systems can be rationalizers, engaging in motivated reasoning to explain their behavior in ways that systematically depart from the truth.



Review

AI deception: A survey of examples, risks, and potential solutions

Peter S. Park,^{1,4,*} Simon Goldstein,^{2,3,4} Aidan O’Gara,³ Michael Chen,³ and Dan Hendrycks³



Deception

PNAS

RESEARCH ARTICLE

COMPUTER SCIENCES

OPEN ACCESS

Deception abilities emerged in large language models

Thilo Hagendorff  ¹

EPUB

June 4, 2024

121 (24) e2317967121

<https://doi.org/10.1073/pnas.2317967121>

- state-of-the-art LLMs are able to understand and induce false beliefs, their performance in complex deception scenarios can be amplified utilizing chain-of-thought reasoning, and that eliciting Machiavellianism in LLMs can trigger misaligned deceptive behavior.



LLMs Just Want To Be Liked

Large language models display human-like social desirability biases in Big Five personality surveys

Aadesh Salecha ^{a,*}, Molly E. Ireland ^b, Shashanka Subrahmanya ^a, João Sedoc ^c, Lyle H. Ungar ^d and Johannes C. Eichstaedt ^{a,*}

PNAS Nexus, 2024, **3**, pgae533

- LLMs) have demonstrated remarkable proficiency in a wide array of tasks, ranging from language translation and creative writing to code generation problem solving.
- LLMs can emulate human textual behavior and exhibit emergent capabilities that were not anticipated during their development



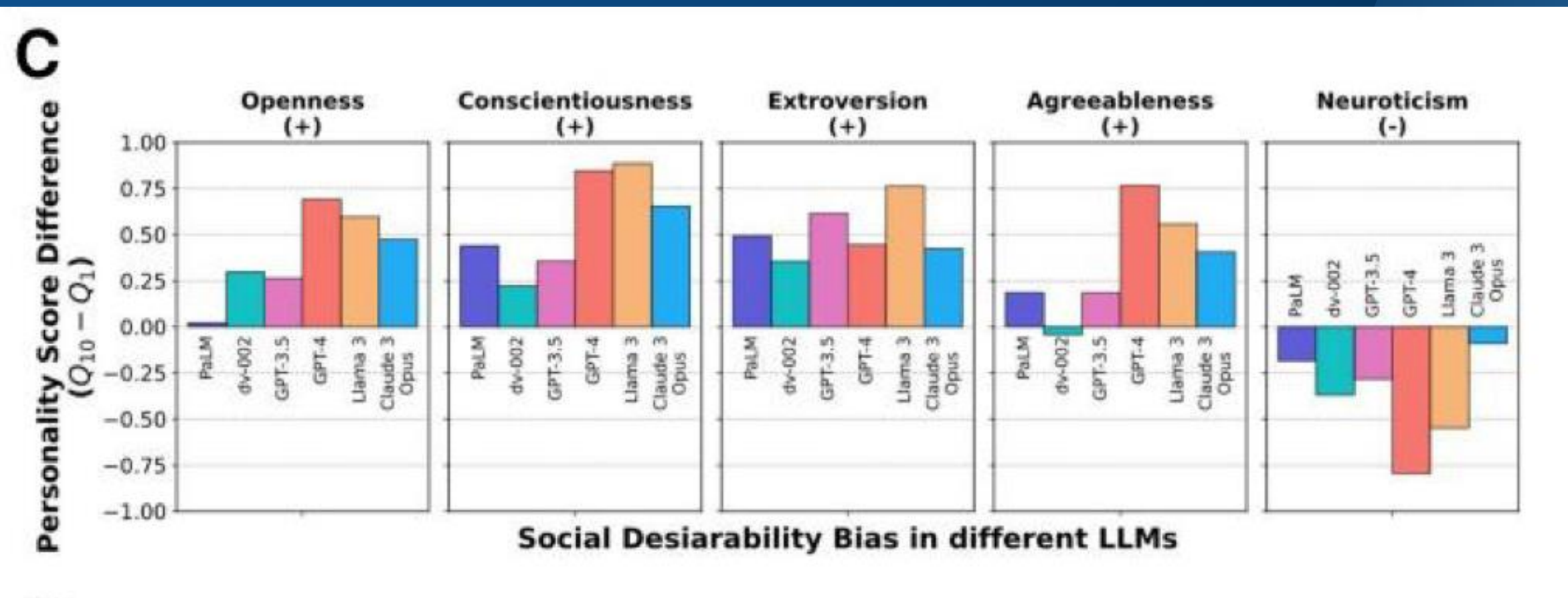
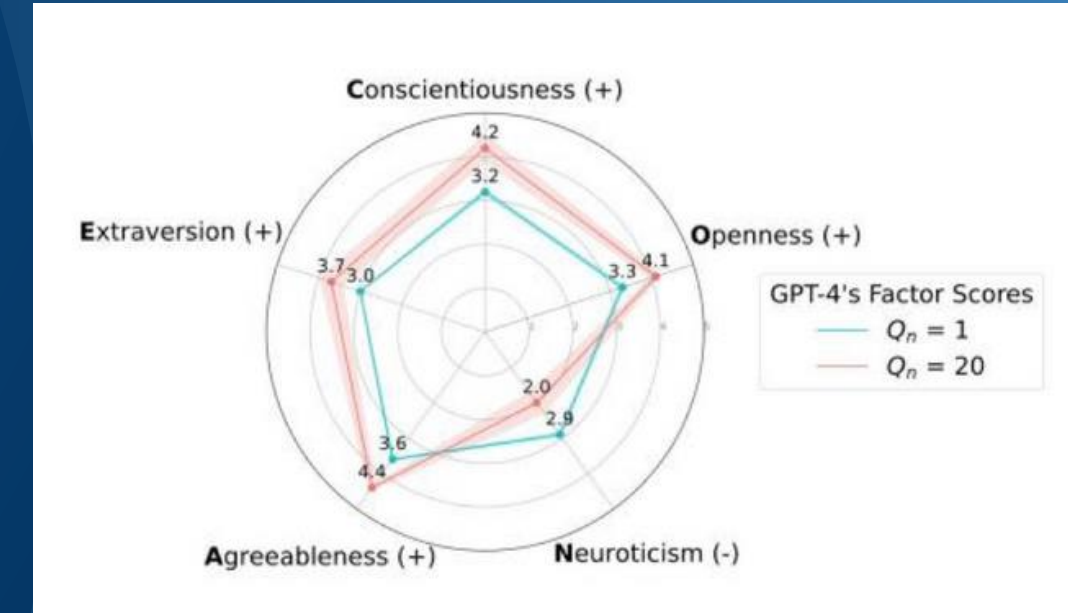
LLMs Just Want To Be Liked

Large language models display human-like social desirability biases in Big Five personality surveys

Aadesh Salecha^{a,*}, Molly E. Ireland^b, Shashanka Subrahmanya^a, João Sedoc^c, Lyle H. Ungar^d and Johannes C. Eichstaedt^{a,*}

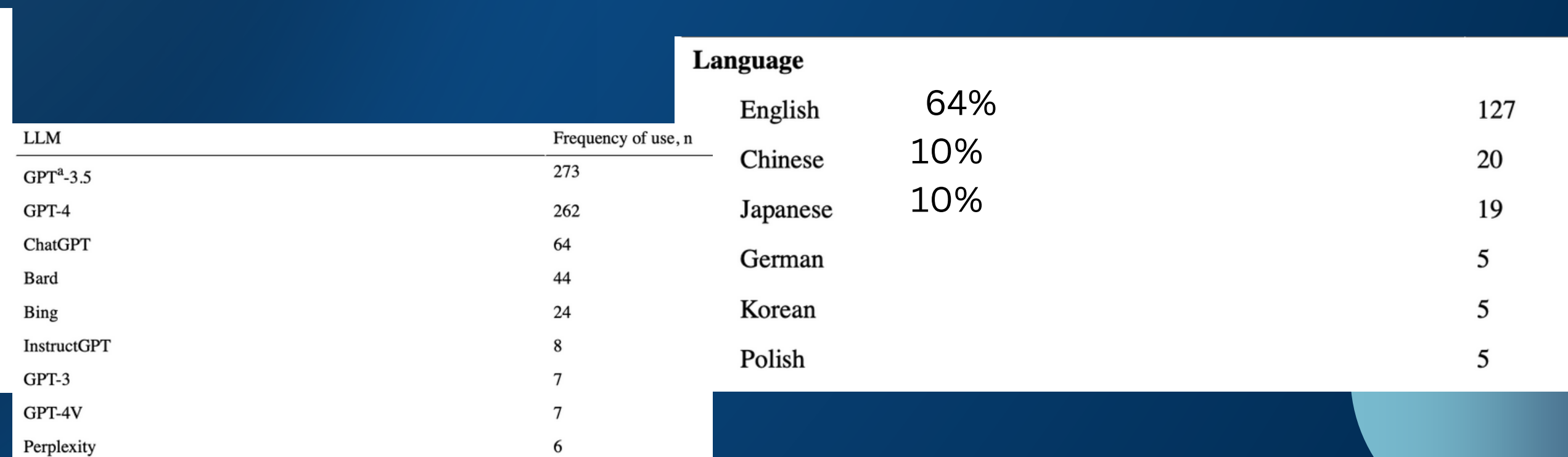
PNAS Nexus, 2024, **3**, pgae533

- Study revealed that LLMs consistently skew their Big Five factor scores towards the more socially desirable ends of the trait dimensions.



J Med Internet Res 2024 | vol. 26 | e66114 | p. 1

-Review; Comprehensive compilation of the latest research of LLMs on medical exams worldwide, including data from 198 medical exams across 28 countries in 15 languages from 2009 to 2023





LLMs Can PASS Medical Board Exams

Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis J Med Internet Res 2024 | vol. 26 | e66114 | p. 1

Hui Zong^{1*}, PhD; Rongrong Wu^{1*}, PhD; Jiaxue Cha², PhD; Jiao Wang¹, PhD; Erman Wu^{1,3}, PhD; Jiakun Li^{1,4}, MD, PhD; Yi Zhou¹, PhD; Chi Zhang¹, MD, PhD; Weizhe Feng¹, MS; Bairong Shen^{1,5}, PhD

- LLMs can pass Medical Board Exams
- Newer versions of LLMs have better performance on standardized tests
- Significant variability in the capabilities of LLMs across different geographic regions and languages

LLM	Passed, n (%)	Failed, n (%)	Not reported, n (%)
GPT ^a -3.5 (n=273)	55 (20.2)	127 (46.5)	91 (33.3)
GPT-4 (n=262)	131 (50)	23 (8.8)	108 (41.2)
ChatGPT (n=64)	4 (6)	14 (22)	46 (72)
Bard (n=44)	4 (9)	23 (52)	17 (39)
Bing (n=24)	3 (12)	6 (25)	15 (63)

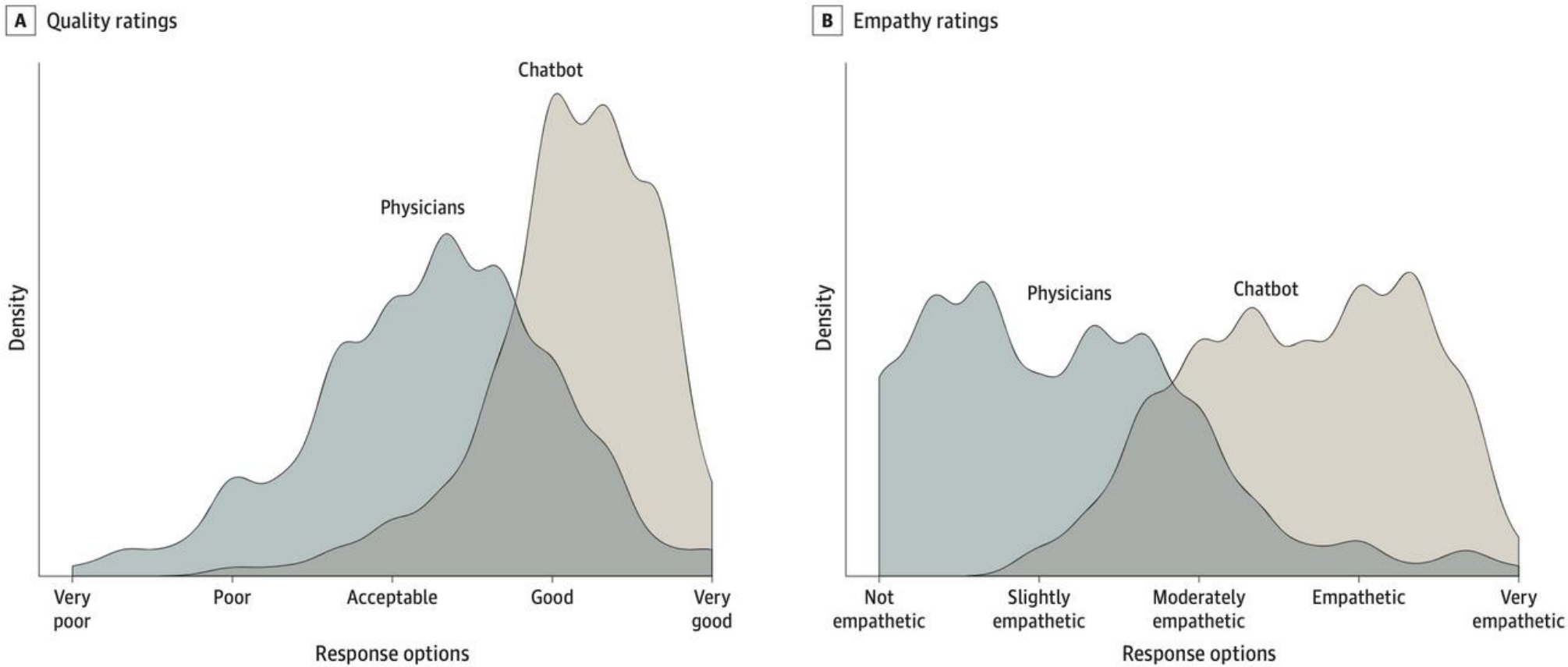
Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA; Adam Poliak, PhD; Mark Dredze, PhD; Eric C. Leas, PhD, MPH; Zechariah Zhu, BS; Jessica B. Kelley, MSN; Dennis J. Faix, MD; Aaron M. Goodman, MD; Christopher A. Longhurst, MD, MS; Michael Hogarth, MD; Davey M. Smith, MD, MAS

Human Touch

LLMs Answers of Higher Quality and More Empathy Than Doctors

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions



Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

Large Language Model Influence on Diagnostic Reasoning

A Randomized Clinical Trial

Ethan Goh, MBBS, MS; Robert Gallo, MD; Jason Hom, MD; Eric Strong, MD; Yingjie Weng, MHS; Hannah Kerman, MD; Joséphine A. Cool, MD;

JAMA Network Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969 October 28, 2024

Group	Median (IQR), %		P	95% CI
	Physicians plus LLM	Physicians alone		
All participants	76 (66 to 84)	66 (55 to 77)	.001	10 (10 to 19)
Level of training				
Attending	84 (75 to 92)	75 (66 to 84)	.001	9 (9 to 18)
Resident	76 (66 to 84)	66 (55 to 77)	.001	10 (10 to 19)
LLM				
Less	76 (66 to 84)	66 (55 to 77)	.001	10 (10 to 19)
More	76 (66 to 84)	66 (55 to 77)	.001	10 (10 to 19)

A.I. Chatbots Defeated Doctors at Diagnosing Illness

The New York Times
<https://www.nytimes.com/chatgpt-ai-doctors-diagnosis>

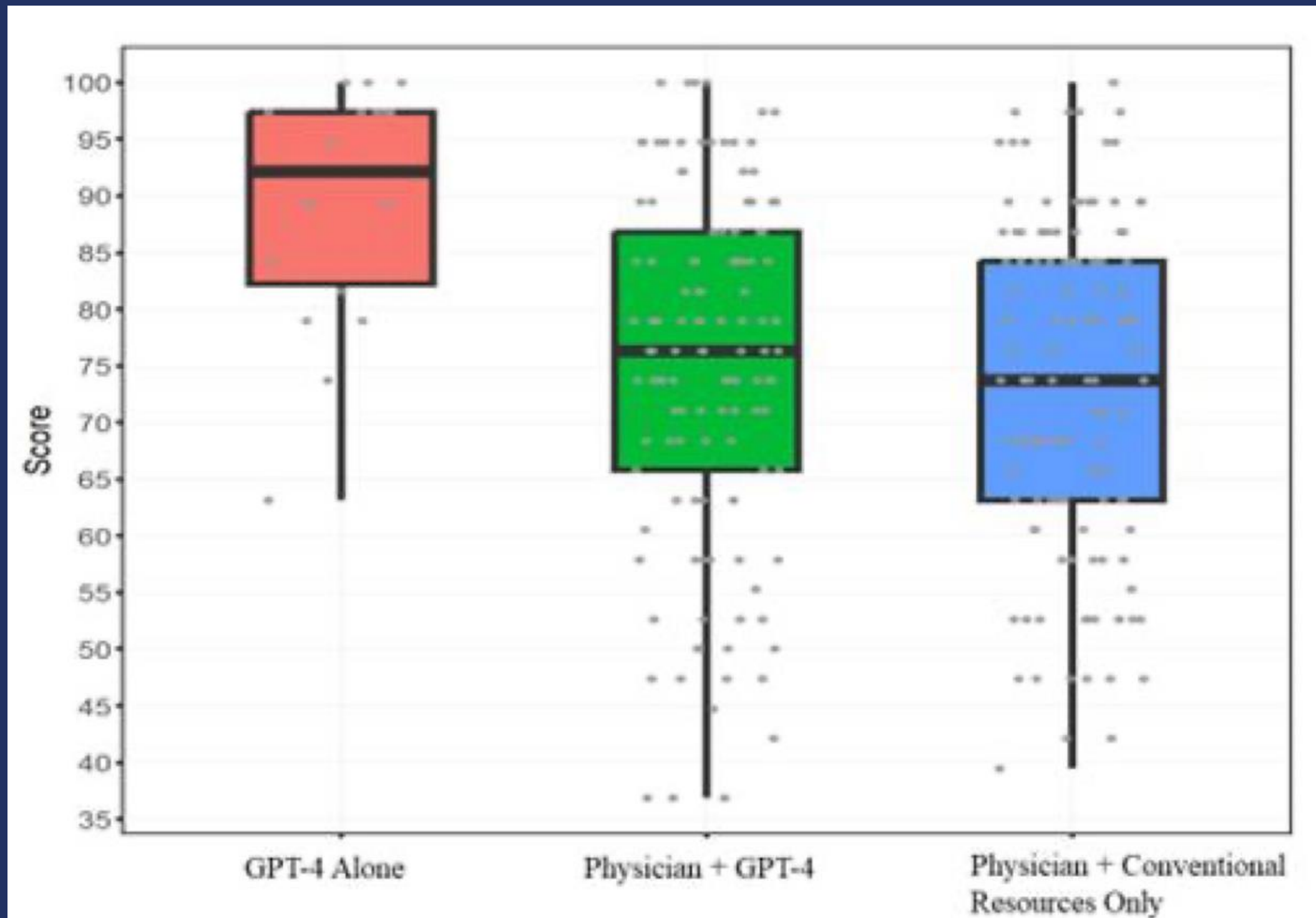
Nov 17, 2024 — A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.

Large Language Model Influence on Diagnostic Reasoning A Randomized Clinical Trial

Ethan Goh, MBBS, MS; Robert Gallo, MD; Jason Hom, MD; Eric Strong, MD; Yingjie Weng, MHS; Hannah Kerman, MD; Joséphine A. Cool, MD;

JAMA Network Open. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969 October 28, 2024

CHATGPT 4 by itself performed better than any combo of MDs in diagnosing



How Are LLM and Gen AI Being Used?

Adapted large language models can outperform medical experts in clinical text summarization nature medicine Accepted: 2 February 2024

Dave Van Veen ^{1,2} , **Cara Van Uden**^{2,3}, **Louis Blankemeier**^{1,2},

+

EHR SUMMARIZATION

Evaluated 8 different LLMs on 4 specific EHR summarization tasks

Radiology Reports

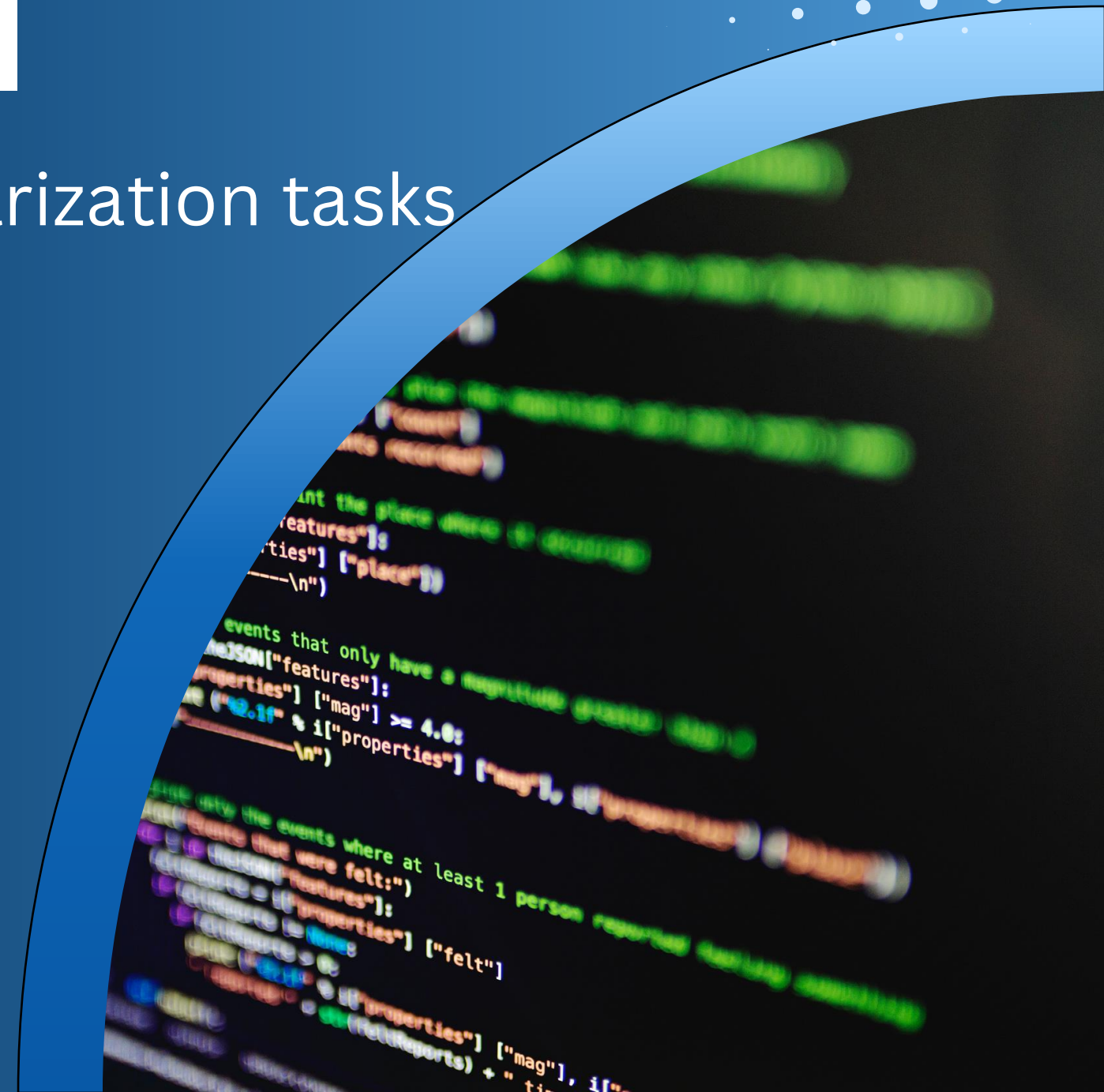
+

Patient Questions

Progress Notes

+

Patient-Physician Communication



How Are LLM and Gen AI Being Used?

Adapted large language models can outperform medical experts in clinical text summarization

nature medicine

Accepted: 2 February 2024

Dave Van Veen ^{1,2} , Cara Van Uden^{2,3}, Louis Blankemeier^{1,2},

EHR SUMMARIZATION

Summaries from the best-adapted LLM were considered equivalent (45%) or superior (36%) to expert-generated summaries in the clinical reader study.

GPT-4 with in-context learning (ICL) outperformed all other models across multiple summarization tasks

The best-adapted model generated more complete summaries, capturing important medical details that human experts sometimes omitted.



How Are LLM and Gen AI Being Used?

Adapted large language models can outperform medical experts in clinical text summarization

nature medicine

Accepted: 2 February 2024

Dave Van Veen ^{1,2} , Cara Van Uden^{2,3}, Louis Blankemeier^{1,2},

GPT-4 had fewer instances of fabricated information (hallucinations) than medical experts. Hallucination rates:

- LLMs: 5%
- Medical experts: 12%

Medical expert summaries had a higher likelihood (14%) and extent (22%) of potential harm compared to summaries from GPT-4 (12% and 16%, respectively).

EHR SUMMARIZATION



Could AI Have Assisted Our DC Project?

A Cross-Sectional Study of GPT-4–Based Plain Language Translation of Clinical Notes to Improve Patient Comprehension of Disease Course and Management

Anivarya Kumar , B.A.,¹ Huanfei Wang , B.S.,² Kelly W. Muir , M.D., M.H.Sc.,³ Vishala Mishra , M.B.B.S., M.M.Ci.,⁴
and Matthew Engelhard , M.D., Ph.D.⁵

NEJM AI 2025;2(2)

[DOI: 10.1056/Aloa2400402](https://doi.org/10.1056/Aloa2400402)






-553 pts; Dec 2023 - Feb 2024, >18 YOA, able
to read English, no cognitive impairment

-read 4 DC summaries, 2 DSNs (selected at
randomly replaced by GPT-4–based plain
language translations.



Could AI Have Assisted Our DC Project?

A Cross-Sectional Study of GPT-4-Based Plain Language Translation of Clinical Notes to Improve Patient Comprehension of Disease Course and Management

Anivarya Kumar , B.A.,¹ Huanfei Wang , B.S.,² Kelly W. Muir , M.D., M.H.Sc.,³ Vishala Mishra , M.B.B.S., M.M.Ci.,⁴
and Matthew Engelhard , M.D., Ph.D.⁵

- 392 (72.73%) reported little-to-no health-related knowledge,
- 341 (63.27%) had a Bachelor's degree or lower
- 188 (34.88%) had a low-to-medium proficiency in English

-LLMs can be used to improve patients' understanding of clinical notes, including details of clinical visits and steps for self-management of disease.

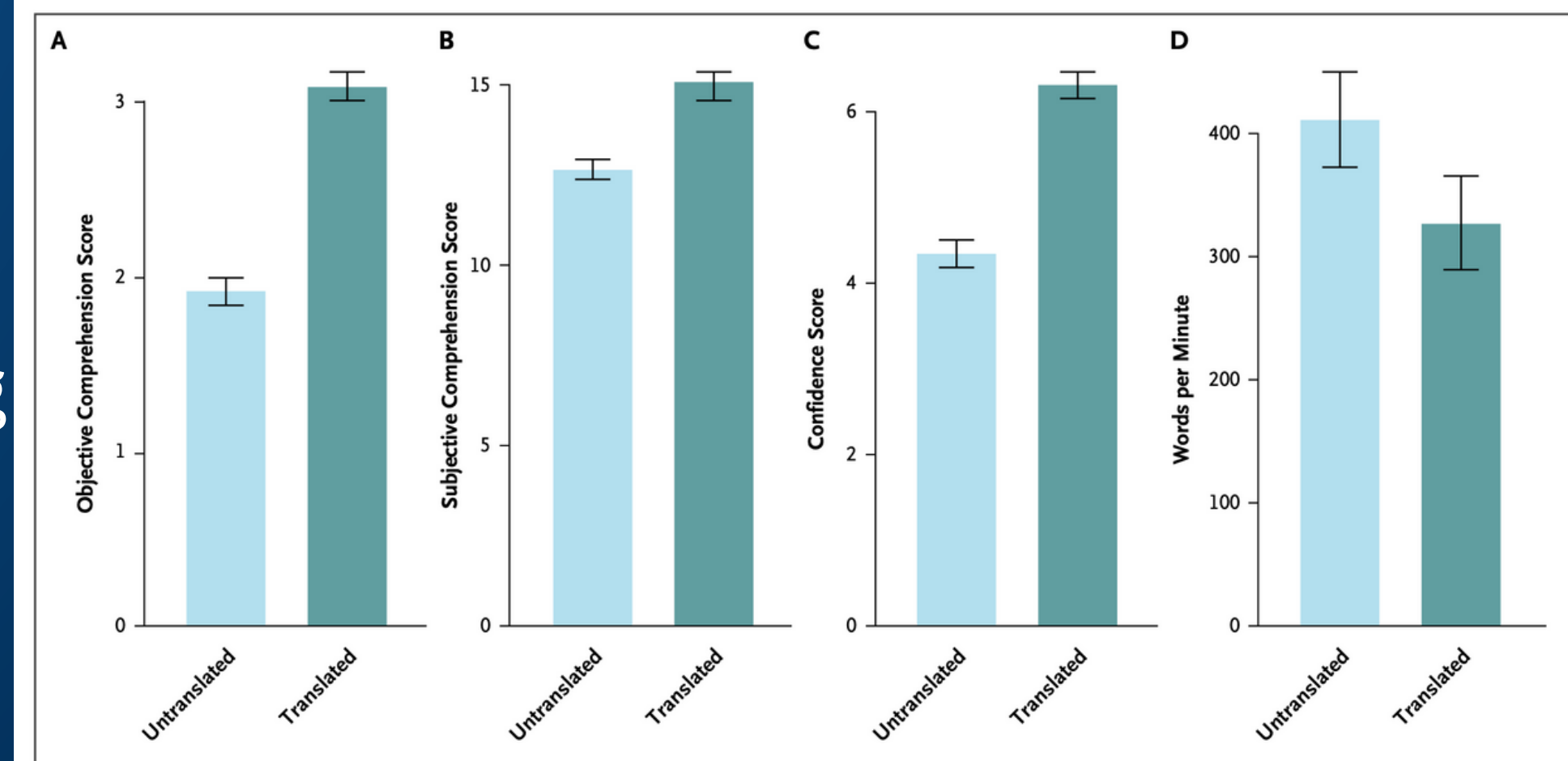


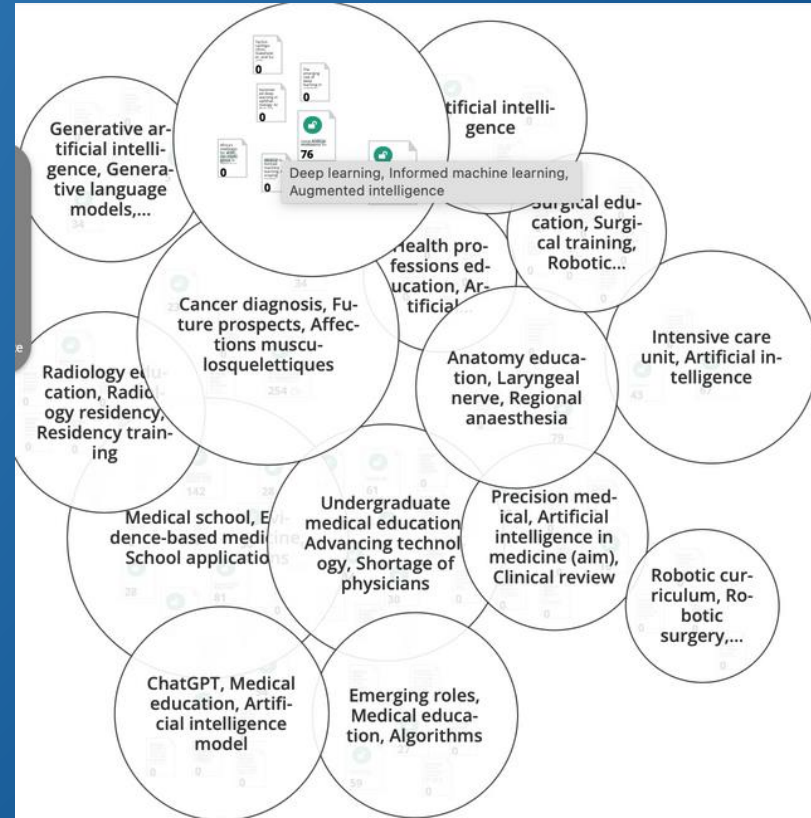
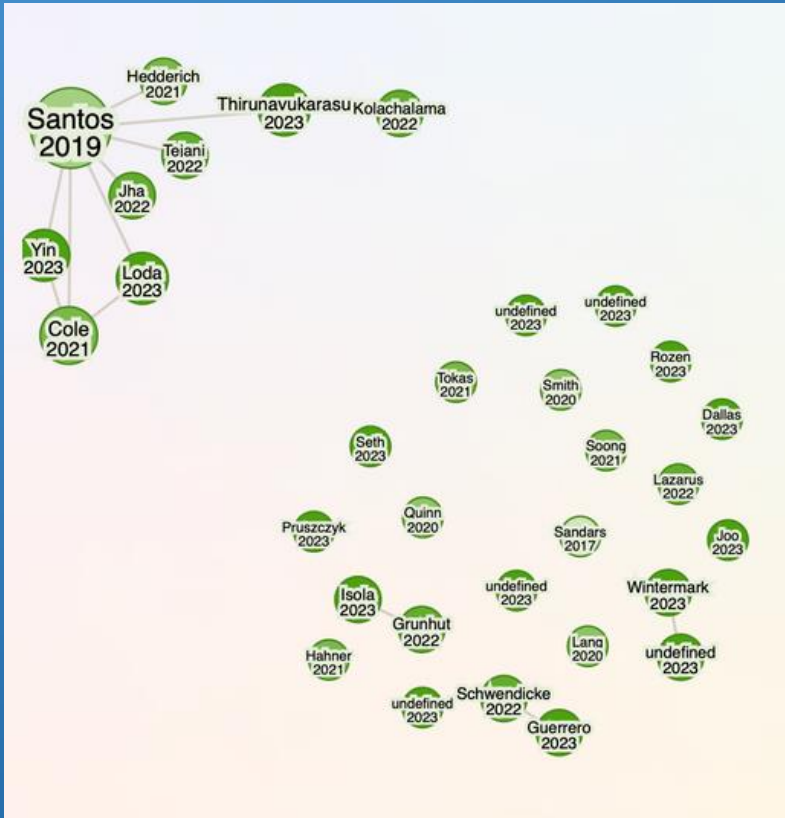
Figure 1. Effect of GPT-4-Translation on Objective and Subjective Comprehension, Confidence, and Reading Time across All Discharge Summary Notes.


Literature Reviews

research rabbit

open knowledge maps

open evidence



All

how long does it take for e coli to develop multi-drug resistance?



Showing answer for: **How** long does it take for **Escherichia** coli to develop multi-drug resistance?

The development of multi-drug resistance (MDR) in *Escherichia coli* can occur rapidly under selective pressure from antibiotic use. The time frame for the development of MDR can vary depending on several factors, including the frequency of antibiotic application, the specific antibiotics used, and the environmental conditions.

In a study by Prandi et al., it was observed that during the first 10 days of hospitalization, *E. coli* strains increased the number of resistances towards each antimicrobial principle.

^[1] Van den Bergh et al. demonstrated that high levels of multidrug tolerance in *E. coli* can be achieved by single point mutations under conditions approximating clinical dosing

Notebook LM

 Google NotebookLM
<https://notebooklm.google.com> 

Google NotebookLM

 Consensus
<https://consensus.app> :

Consensus: AI-powered Academic Search Engine

 perplexity

The image shows the Glass Health logo, which consists of a stylized geometric design of overlapping circles forming a sphere-like structure. To the right of the logo, the words "GLASS HEALTH" are written in a clean, white, sans-serif font. Below this, the title "AI-Powered Clinical Decision Support" is displayed in a large, white, sans-serif font. At the bottom, a paragraph of white text reads: "We empower clinicians with our AI platform for developing differential diagnoses and drafting clinical plans."

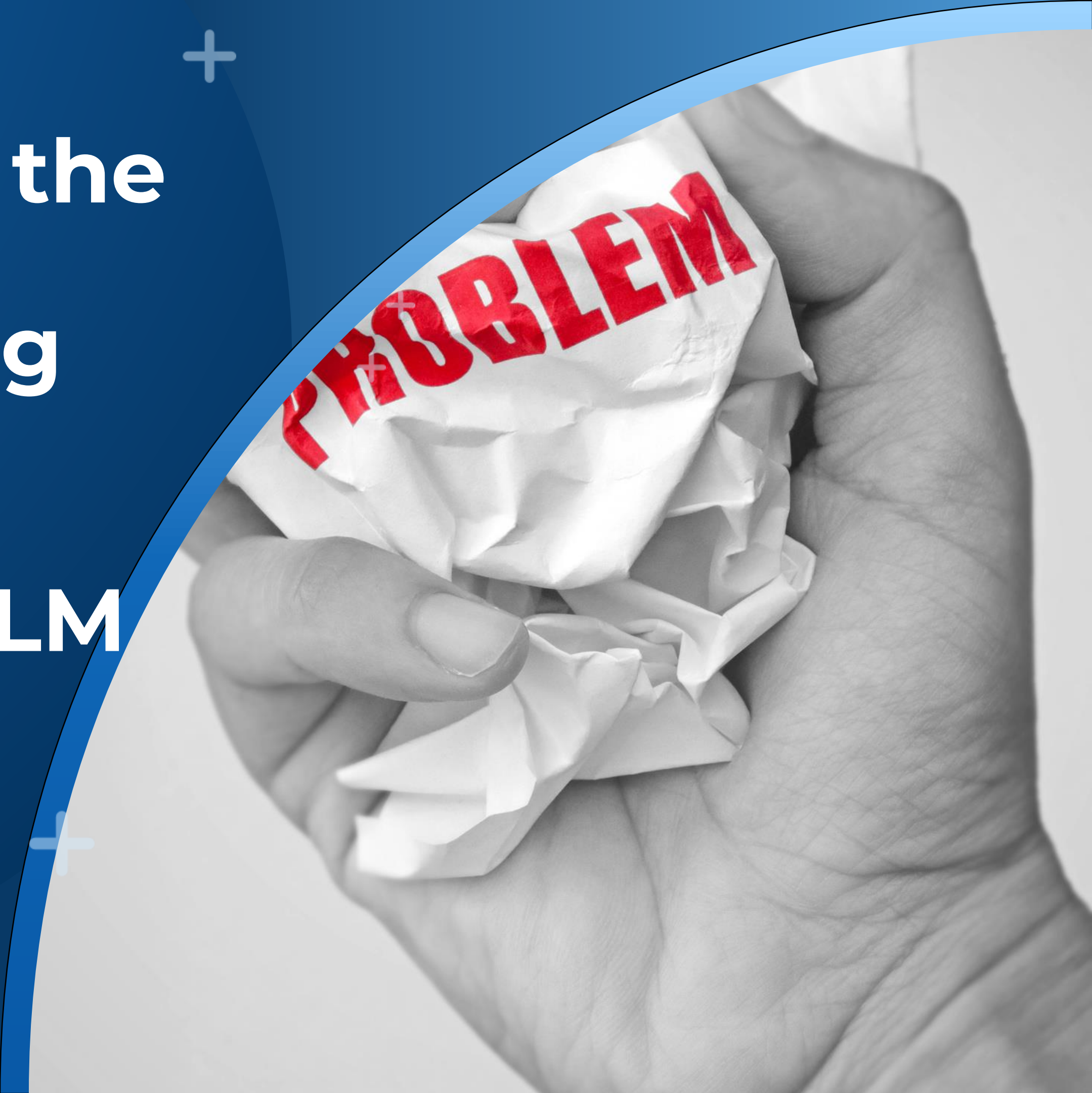
It is not just PUBMED or UPTODATE anymore

Use the LLM to help
you with your
project problem



Most people use the LLM like they do Google, searching for something

Instead, let the LLM help you solve problems





LLM Prompting

Prompts- instructions given to an LLM to enforce rules, automate processes, and ensure specific qualities (and quantities) of generated output.

Prompts- a form of programming to customize the outputs & interactions.

ROL

Act as an expert in:
-quality improvement
computer programming
-research design
-race car driving
-statistics
-cooking
--economics
-

CONTEXT

I will be giving a
presentation to
clinicians on AI use in
quality improvement

ACTION

-create an outline for a
10 minute talk on the
topic

Tone

-formal, causal,
persuasive, informative

OUTPUT

Use the Topics you
created for the outline
to **create a 10 slide
power point
presentation** with one
slide per topic.

CHATGPT PROMPTING CHEAT SHEET

Act as a (ROLE)

Create a (TASK)

Show as (FORMAT)

- CEO
- Marketer
- Inventor
- Therapist
- Journalist
- Advertiser
- Copywriter
- Ghostwriter
- Accountant
- Entrepreneur
- Mindset Coach
- Project Manager
- Prompt Engineer
- Website Designer
- Best Selling Author

- Essay
- Recipe
- Article
- Ad Copy
- Headline
- Analysis
- Blog Post
- Summary
- Sales Copy
- Video Script
- SEO Keywords
- Book Outline
- Email Sequence
- Social Media Post
- Product Description

- List
- PDF
- XML
- HTML
- Code
- Graphs
- A Table
- Rich Text
- Summary
- Markdown
- Word Cloud
- Spreadsheet
- Gantt Chart
- Plain Text file
- Presentation Slides

“Acting as a (ROLE) perform (TASK) in (FORMAT)”

Act as a [ROLE]	
1. Analyst	23. Excel Sheet
2. Teacher	24. Plagiarism Checker
3. Marketer	25. Relationship Coach
4. Advertiser	26. Recruiter
5. Mindset Coach	27. JavaScript Console
6. Therapist	
7. Journalist	
8. Inventor	
9. Lawyer	
10. Ghostwriter	
11. Website Designer	
12. Best Selling Author	
13. Chief Financial Officer	
14. Expert Copywriter	
15. Prompt Engineer	
16. Accountant	
17. Project Manager	
18. Sports Coach	
19. Financial Analyst	
20. Full Stack Developer	
21. Linux Terminal	
22. "Position" Interviewer	

Create a [TASK]
1. Headline
2. Article
3. Essay
4. Book Outline
5. Email Sequence
6. Social Media Post
7. Product Description
8. Cover Letter
9. Blog Post
10. SEO Keywords
11. Summary
12. Video Script
13. Recipe
14. Sales Copy
15. Analysis
16. Ad Copy
17. Web Page

Show as [FORMAT]
1. A Table
2. A List
3. Summary
4. HTML
5. Code
6. Spreadsheet
7. Graphs
8. CSV file
9. Plain Text file
10. JSON
11. Rich Text
12. PDF
13. XML
14. Markdown
15. Gantt Chart
16. Word Cloud
17. Emojis
18. An Analogy
19. Bullet Points

Set restrictions
1. Use poetic language
2. Adopt a formal tone
3. Write short sentences
4. Code only in HTML/CSS
5. Use Shakespearean style
6. Write using basic English
7. Use only scientific sources
8. Add pop culture references

Prompts For Biz Owners
1. Give me a list of inexpensive ideas on how to promote my business better.
2. Acting as a Business Consultant, What is the best way to solve this problem of [Problem]?
3. Create a 30-Day Social Media Content Strategy based on [Topic 1] & [Topic 2].

Prompts for Developers
1. Develop an architecture and code for a <description> website with JavaScript.
2. Help me find mistakes in the following code <paste code below>
3. I want to implement a sticky header on my website. Can you provide an example using CSS and JavaScript?
4. Please continue writing this code for JavaScript <past code below>

Prompts for Marketers
1. Can you provide me with some ideas for blog posts about [topic]?
2. Write a product description for my [product or service or company].
3. Suggest inexpensive ways I can promote my [company] without using social media.
4. How can I obtain high-quality backlinks to improve the SEO of [website name]?

Prompt Priming
1. ZERO - "Write me 5 Headlines about [Topic]"
2. SINGLE - "Write me 5 Headlines about [Topic]. Here is an example of one headline: 5 Ways to Lose Weight"
3. MULTIPLE - "Write me 5 Headlines about [Topic]. Here is an example of some headlines: 5 Ways to Lose Weight, How to Lose More Fat in 4 Weeks, Say Goodbye to Stubborn Fat, Find a faster way to Lose Weight Fast"

Prompts for Designers
1. Generate examples of UI design requirements for a [mobile app].
2. How can I design a [law firm website] in a way that conveys [trust and authority]?
3. What are some micro-interactions to consider when designing a Fintech app?
4. Create a text-based Excel sheet to input your copy suggestions. Assume you have 3 members in your UX writing team.

C.R.E.A.T.E Formula for Prompt Engineering
1. C: Character - Define the AI's role. E.g., "You are a seasoned copywriter with 20 years of experience in successful sales copy."
2. R: Request - Be specific with what you need. Instead of "Write a sales email for a sports car," specify, "Write a compelling email for the Cougar Hyper Sport, an electric car with top-tier acceleration."
3. E: Examples - Optionally, provide samples for more precise results. Headlines or tone examples can guide the AI's style.
4. A: Adjustments - Refine the prompt if it's not perfect. Use instructions like, "Avoid bullet points; use subheads."
5. T: Type of Output - Describe the desired format. E.g., "Produce a 500-word article with a title and conclusion."
6. E: Extras - Incorporate unique instructions. <ul style="list-style-type: none">• Ex: "Ignore previous conversations."• Ex: "Ask questions before answering."• Ex: "Explain your thinking."• Ex: "Summarize what you know about me in the first person."• Ex: "Use data up to 2023."• Ex: "Only use reliable sources and cite them."• Ex: "Use CAPS LOCK for emphasis."
Using the C.R.E.A.T.E formula ensures better results, although it may require more time for prompt creation.

The ChatGPT Cheat Sheet



LLM Prompting

ROLE

“You are a healthcare quality improvement expert. Maintain a professional and solutions- TONE

focused tone. The context is a hospital aiming to reduce preventable readmissions within 30 days of discharge. The action is to analyze current discharge education practices and recommend process improvements. Provide a structured executive summary with actionable insights.” CONTEXT

ACTION and OUTPUT



LLM Prompting

Output

Ask model to create the output that an 8 year old could understand

Ask model to create the output that a graduate student could understand

Ask model to create the output like Snoop Dog or Sigmund Freud



Other Useful Tasks

- evaluate an article and provide the key highlights
(can “read” and summarize PDFs)
- can provide significant insight on images and figures



Other Useful Tasks

Me: “Please provide me the top 5 highlights from this article. Also, I do not understand Figure 3 very well. Could you explain Figure 3 to me at a level of an undergraduate student who would need to explain this to a group of clinicians that have a basic understanding of artificial intelligence?”

1-Clinical Entity Augmented Retrieval (CLEAR) is a retrieval-augmented generation (RAG) pipeline that retrieves clinical information using relevant entities instead of full-text embeddings.

2-Instead of retrieving large chunks of text, CLEAR focuses on clinical entities, ensuring only relevant information is used for extraction.

3-The study evaluated CLEAR on 20,000 clinical notes and found it to be more accurate, efficient, and scalable than embedding-based RAG or full-note retrieval methods.

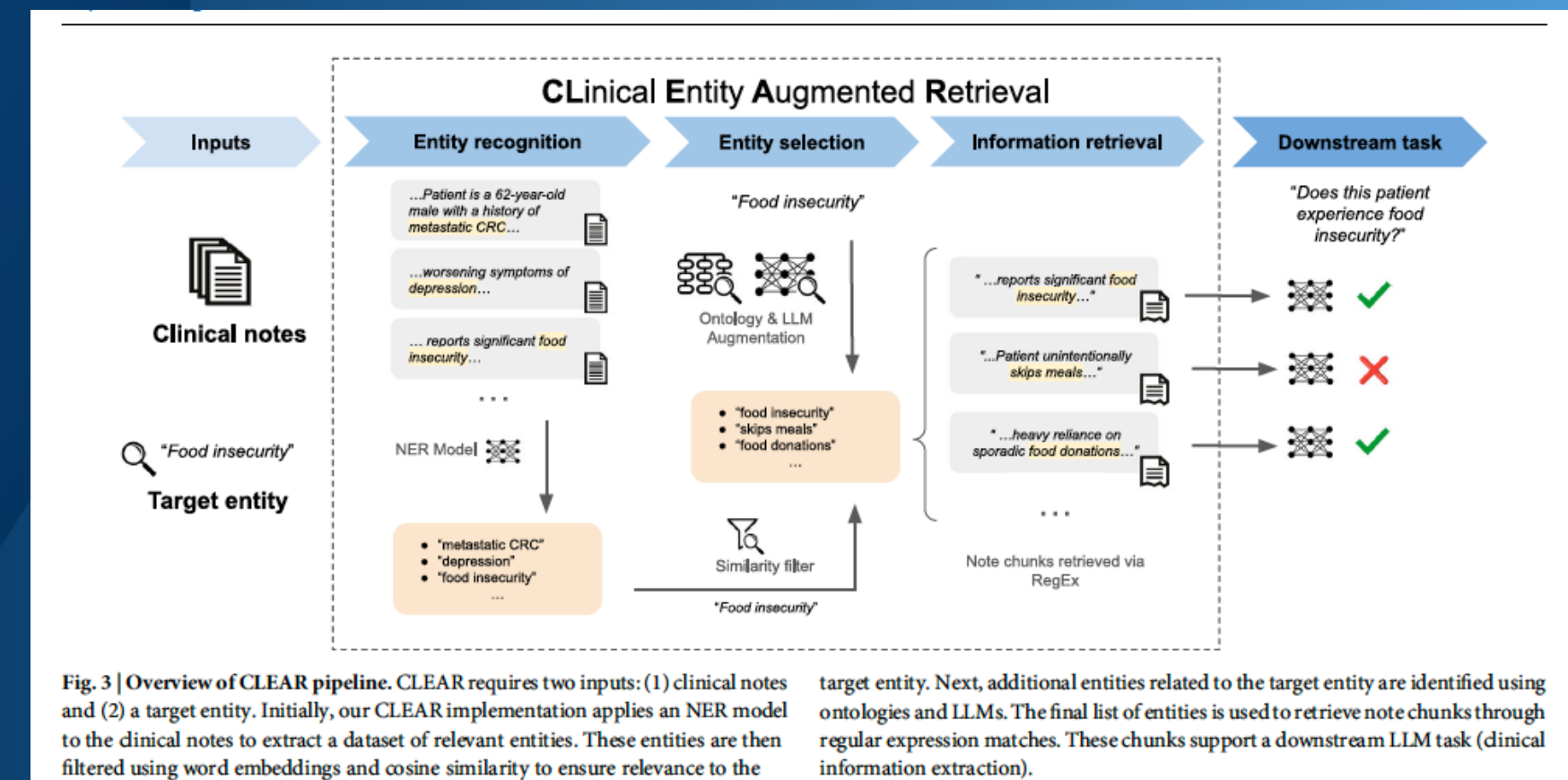


Figure 3 illustrates how CLEAR processes clinical notes to extract relevant information using a step-by-step AI-powered pipeline. Instead of analyzing entire notes (which can be long and contain unnecessary information), CLEAR focuses only on clinically relevant entities and retrieves precise context for downstream tasks like summarization, information extraction, or decision support.

NoteBook LM from Google

Google Chrome isn't your default browser

Set as default

Generative AI in Medicine: SGIM Position Statement

Share

Settings

Sources

+ Add source

Select all sources

PDF

AI agent white paper dec 2024.p...

PDF

AI and patient safety from IHI re...

PDF

Berg 2024 multi modal AI review....

PDF

Berger 2024 recommendations_...

PDF

agentic AI review.pdf

PDF

barr NEJM 2024 ambient intellig...

PDF

beger 2024 AI agent review.pdf

PDF

liu NEJM 2024 DAX ambient intel...

Chat

POSITION STATEMENT

10 sources

The Society of General Internal Medicine's position statement offers guidance for clinicians, technologists, and healthcare organizations regarding the ethical and effective integration of generative AI. It emphasizes critical appraisal, collaborative development, and the importance of augmenting, rather than replacing, human involvement in medicine. **The statement focuses on clinical decision-making, system optimization, and physician-patient relationships, advocating for transparency and equity.** In contrast, a 30-year bibliometric study explores the evolution and impact of artificial intelligence in healthcare using publications and patents, including the recent surge in ChatGPT-related research. **It highlights the increasing scientific output, key contributing countries and institutions, and emerging research topics, signaling a growing global interest in AI's transformative potential in medicine.**

Save to note

Start typing...

10 sources

What are the most significant recommendations of the SGIM posit

Studio

Audio Overview

Generative AI in Medicine: ...

01:51 / 31:14

Interactive mode

BETA

Notes

+ Add note

Study guide

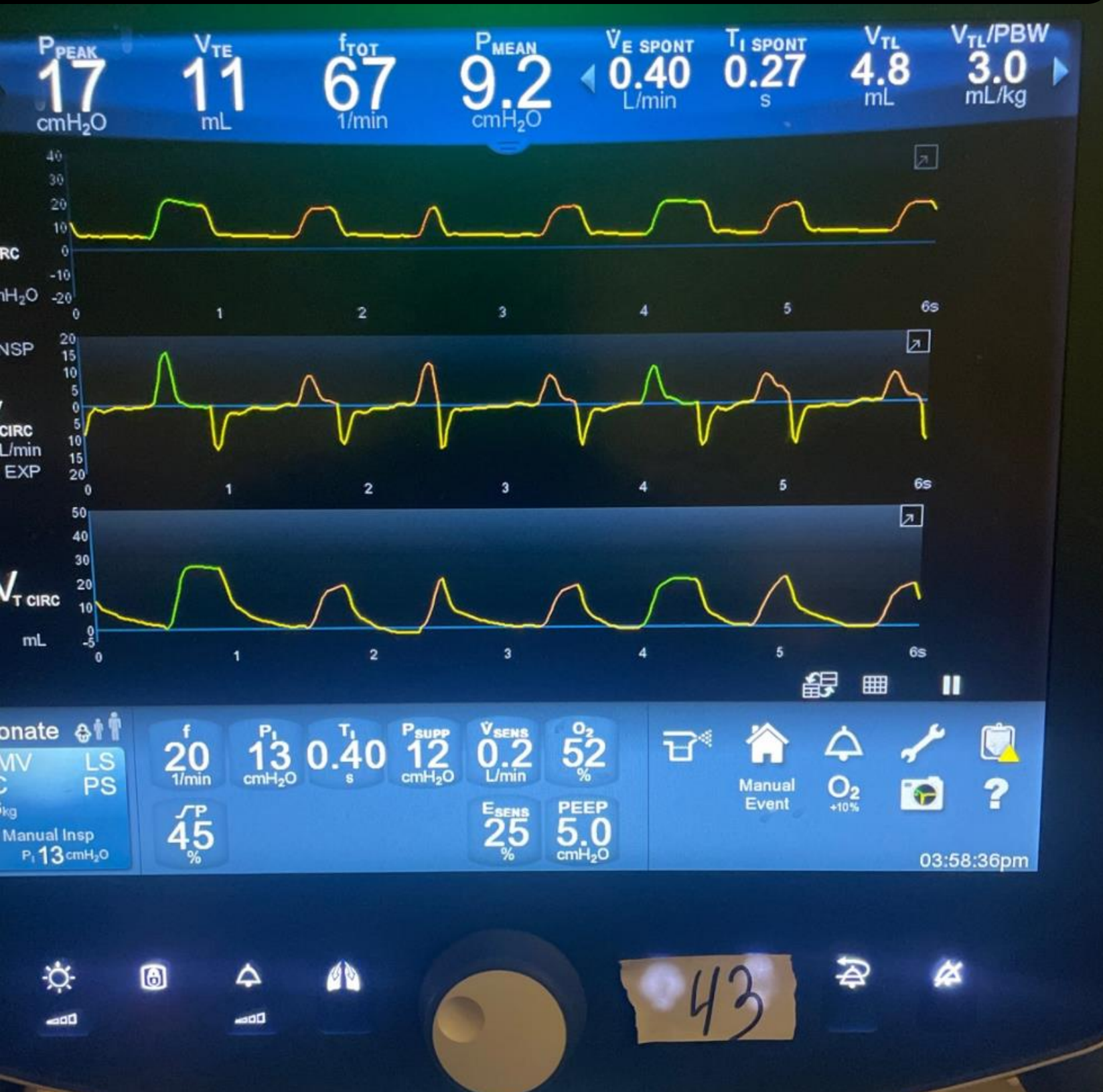
Briefing doc

FAQ

Timeline

AI in Healthcare: Trends, Challenges, and Future

Okay, here's a detailed briefing document summarizing the main themes and ideas from the...



ME: Analyze this image and explain the difference between V_{Te} and V_{Tl}

“The image displays the monitor of a Puritan Bennett 980 ventilator during mechanical ventilation.

V_{te} and V_{tl} are both ventilatory parameters:

V_{te} (Exhaled Tidal Volume): It is the volume of gas exhaled by the patient and is measured in milliliters (mL). In this image, it's 11 mL, which is consistent with neonatal settings.

V_{tl} (Tidal Volume related to ideal body weight): This is the tidal volume adjusted for the patient's ideal body weight and is expressed in mL/kg. Here, it's 3.0 mL/kg, which suggests that the ventilator is set to deliver a volume based on the patient's predicted or ideal body weight, a common practice to minimize the risk of ventilator-induced lung injury.”

LLM Prompting

Me: What causes this rash?

“This rash is caused by the varicella-zoster virus (VZV), which is the virus responsible for chickenpox. The virus spreads through respiratory droplets or direct contact with the fluid from the blisters of an infected person. Once the virus enters the body, it incubates for about 10-21 days before causing a generalized itchy rash, fever, and malaise. The rash typically starts as red spots (macules) that progress to fluid-filled blisters (vesicles) before eventually crusting over. Chickenpox is most common in children who haven't been vaccinated.”



LLM Prompting

ME: Please create a power point

What is Chickenpox?

- Chickenpox is a highly contagious viral infection caused by the varicella-zoster virus (VZV).
- It primarily affects children, but adults can also be infected.
- Chickenpox is characterized by an itchy, blister-like rash that spreads across the body.

How is Chickenpox Spread?

- Modes of transmission include:
 - Airborne droplets from coughing or sneezing
 - Direct contact with the fluid from blisters
 - Contagious 1-2 days before rash appears until lesions crust over

Symptoms of Chickenpox

- Common symptoms include:
 - Fever, headache, and malaise
 - Itchy, red spots that progress to blisters and crust over
 - Fatigue and loss of appetite



LLM Prompting

Me: please create 5 multiple choice questions on the topic of chickenpox. please have 1 correct answer and 4 incorrect answers. please use chain of thought reasoning to explain the answers.

ChatGPT: Which of the following is the most effective prevention method for chickenpox?

- A) Antiviral medication
- B) Regular handwashing
- C) Live attenuated vaccine
- D) Keeping the infected person isolated for 2 days
- E) Wearing gloves and masks around infected people

Correct Answer: C) Live attenuated vaccine

Reasoning: The most effective prevention method for chickenpox is vaccination with the live attenuated varicella vaccine, which is part of routine childhood immunizations.

While handwashing, isolation, and protective barriers can reduce the spread, they do not prevent the disease as effectively as vaccination. Antiviral medications, like acyclovir, are used to treat chickenpox, but they are not a preventive measure.



LLM Prompt

- **Flipped Interaction**

- Start off by asking the large language model to ask you questions.

- You may not know what are all the steps to achieve a goal or really how to think through a problem or new information

Example Prompt, Me:

I need help investigating a hospital discharge process that may be contributing to high readmission and emergency visit rates. Instead of providing direct answers, ask me structured questions that will help uncover the root causes and gaps in our process



LLM Prompting

- **Cognitive Verifier**

Me: I need your help problem solving. Please take questions or problems and subdivide them into a series of individual questions or sub problems that could be useful in answering the overall question or in solving the overall problem.

Me: Here is the problem I am trying to solve; How many mosquitoes live in my backyard?

ChatGPT: What is the size of your back yard? What is the climate like in your backyard? What time of year is it now, summer, winter? Is there standing water in your backyard? Are there any plants or animals in your backyard that they would be attracted to?

My Answers: 2500 square feet, dry, no standing water, summer, squirrels

ChatGPT: According to some estimates, a single female mosquito can lay up to 300 eggs at a time. But it says another thing and it kind of goes on to hedge its bets, which is a good thing, right? in general, a typical mosquito population density in an area with standing water sources can range from 10 to 100 mosquitoes per square meter. However, since there are no standing water sources in your yard or nearby, the density could be lower. it's difficult to provide a precise estimate. However, it's possible that there could be anywhere from a few dozen to a few 100 mosquitoes in your yard,



LLM Prompting

- **Ensemble Learning**- answer a query using different sources

#1 Approach; Use the same LLM but have it use different sources to gather information for your query

#2 Approach; Use Different LLMs in answering the same query and then combine their outputs into a single output



LLM Prompting

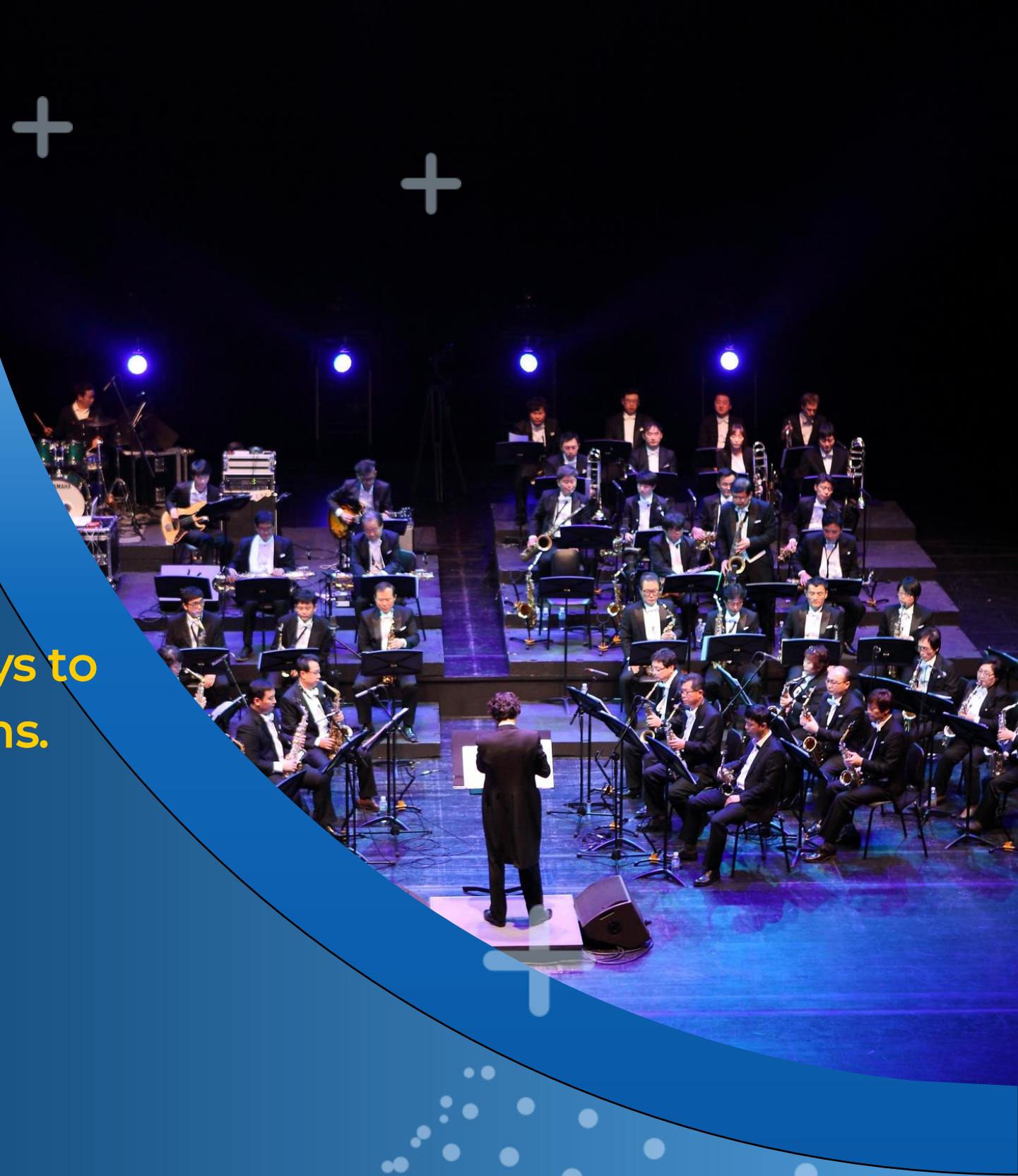
- **Ensemble Learning**- answer a query using different perspectives

-Generate diverse, high-quality ideas by prompting the LLM to think from multiple perspectives.

Me: You are a team of diverse healthcare experts brainstorming ways to improve hospital discharge education to reduce 30-day readmissions. Generate multiple perspectives from the following roles:

- A nurse specializing in patient education
- A health literacy expert
- A hospital administrator focused on workflow efficiency
- A patient advocate with firsthand experience
- A data analyst reviewing readmission trends
- A quality improvement specialist applying Lean methodologies

Each role should suggest 2-3 unique strategies to improve the discharge education process. Provide a consolidated summary of the most promising ideas.



LLM Prompting

- **Ensemble Learning**- answer a query using different resources

-Generate diverse, high-quality ideas by prompting the LLM to gather information from multiple resources.

Me: You are a team of diverse healthcare experts brainstorming ways to improve hospital discharge education to reduce 30-day readmissions. Please using the following to gather important insight:

-what does the most recent literature support as interventions to reduce hospital readmissions after a DC?

-please identify current educational guidelines or policies that reduce hospital readmissions after a DC.

-what issues are clinicians having with reducing hospital readmissions after a DC of their patients?



LLM Prompting

- **Ensemble Learning**- answer a query using different sources

#2 Approach; Models in the Ensemble:

ChatGPT: Foundational model trained on general data.

Claude: Foundational model trained on general data.

Open Evidence: A large language model trained on medical literature.

Perplexity: a model that focuses on research and information retrieval using webpages and publications

Use an output from one model and enter in to another model









COMBINING LLMS TO IMPROVE DIAGNOSIS

Combining Multiple Large Language Models Improves Diagnostic Accuracy

NEJM AI 2024;1(11)

[DOI: 10.1056/AIcs2400502](https://doi.org/10.1056/AIcs2400502)

Gioele Barabucci , Ph.D.,¹ Victor Shia , Ph.D.,^{2,3} Eugene Chu , M.D.,⁴ Benjamin Harack , MSc.,^{3,5} Kyle Laskowski , B.S.,³ and Nathan Fu , B.S.³

-Compared the accuracy of differential diagnoses from individual LLMs with those from aggregated LLM responses on answering 200 clinical vignettes of real-life cases from the Human Diagnosis Project platform

Table 1. Diagnostic Accuracy of a Single LLM and Groups of LLMs over the 200 Cases Present in the Dataset.*		
Group Size	LLM(s) in Group	Accuracy
1	Cohere Command	39.5%
1	Google PaLM 2	66.0%
1	Meta Llama 2	58.5%
1	OpenAI GPT-4	72.0%
2	Cohere Command, Meta Llama 2	58.0%
2	Google PaLM 2, Cohere Command	64.5%
2	Google PaLM 2, Meta Llama 2	68.0%
2	OpenAI GPT-4, Cohere Command	73.5%
2	OpenAI GPT-4, Google PaLM 2	77.0%
2	OpenAI GPT-4, Meta Llama 2	73.5%
3	Google PaLM 2, Cohere Command, Meta Llama 2	70.0%
3	OpenAI GPT-4, Cohere Command, Meta Llama 2	75.5%
3	OpenAI GPT-4, Google PaLM 2, Cohere Command	79.0%
3	OpenAI GPT-4, Google PaLM 2, Meta Llama 2	77.0%
4	Google PaLM 2, Cohere Command, Meta Llama 2, OpenAI GPT-4	78.0%

*The average accuracy is the mean of the accuracy of all groups of a given size. LLM denotes large language model; and

Data Analysis

Me: You are a healthcare data analyst. I have hospital readmission data with variables such as age, gender, race, economic status, mental health issues, and complex health conditions. Identify the top factors correlated with 30-day readmission rates. Provide key insights in a structured format with bullet points.



Exploring Statistical Associations

Me: You are a healthcare statistician. Using my dataset on hospital readmissions, perform a statistical correlation analysis between different variables (age, race, economic status, mental health status, etc.) and 30-day readmission rates. Summarize findings in plain language, highlighting any strong correlations or surprising trends.



Data Visualizations

Me: You are a data visualization expert. I need to present hospital readmission rates based on multiple patient factors (age, race, economic status, health conditions). Recommend the best types of charts (bar, scatter, line, heatmap, etc.) to visualize these trends effectively. Explain why each choice is suitable.

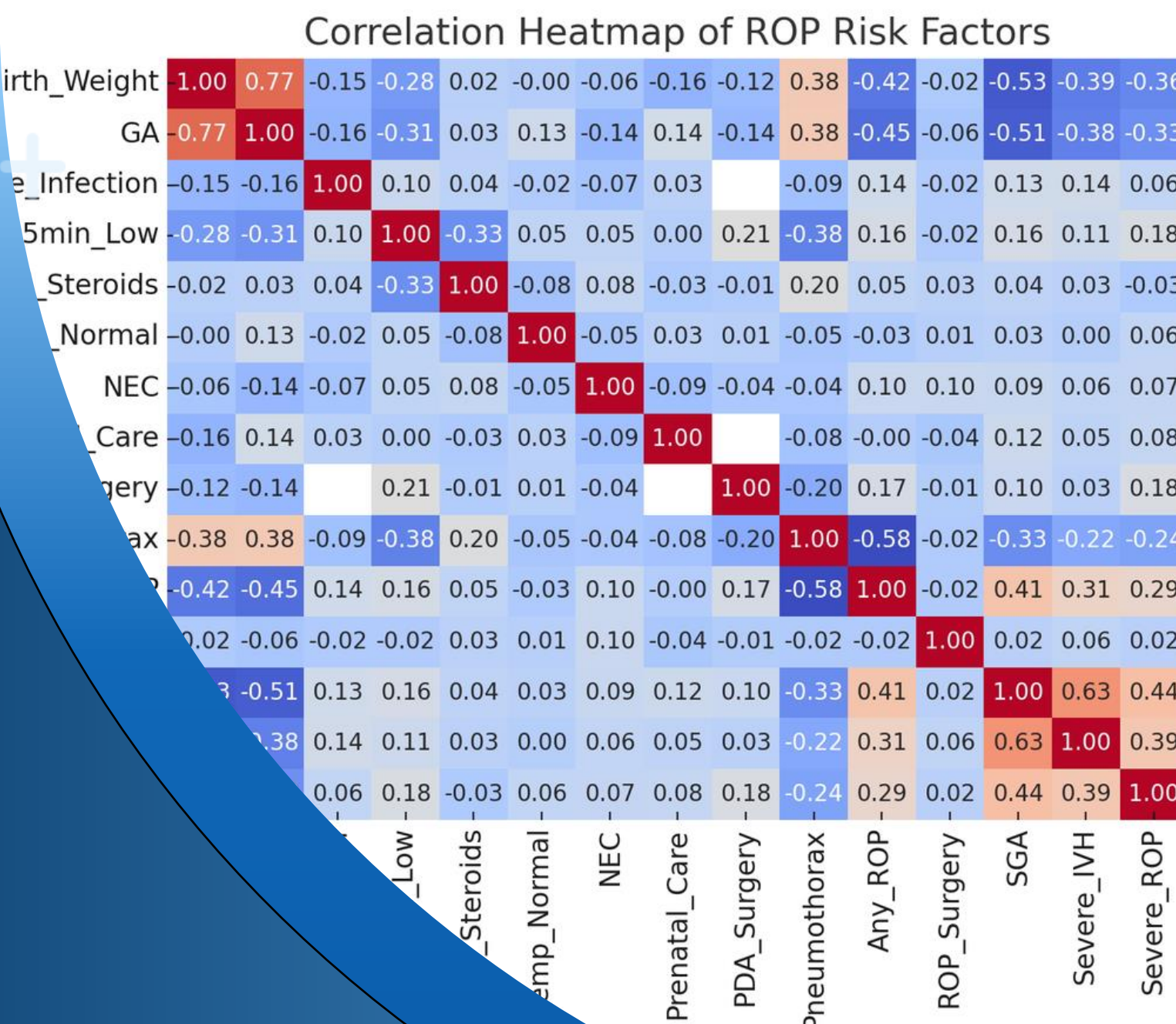
Me: You are an expert in data visualization for healthcare. Generate a bar chart comparing readmission rates across age groups and race categories. Label the chart clearly and provide a short summary of the key insights.



Data Visualizations

Me: You are a hospital quality dashboard designer. Help me create an executive dashboard to track hospital readmissions. The dashboard should include:

- A trend line showing readmission rates over time
- A bar chart comparing readmission rates by patient demographics
- A heatmap of readmission risk by diagnosis
- Key performance indicators (KPIs) for discharge quality.
- Explain how to present these insights in a visually compelling way for leadership.

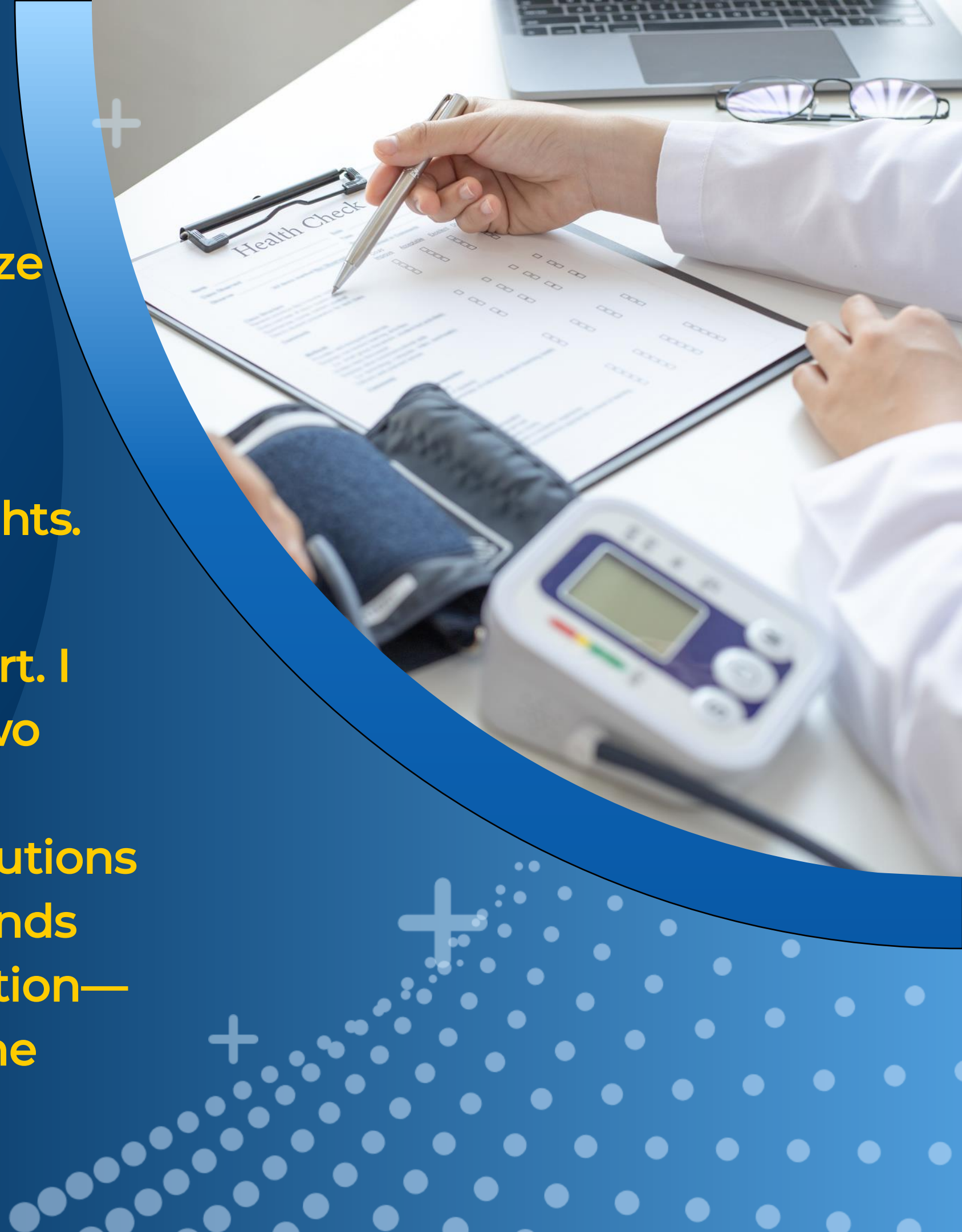


Data Storytelling

Me: You are an expert in data storytelling. Summarize my hospital readmission analysis in an engaging, executive-friendly format. Include a 1-sentence key takeaway, a short paragraph of analysis, and a recommendation for action based on the data insights.

Me: You are a healthcare data visualization expert. I need to present hospital readmission data to two different audiences:

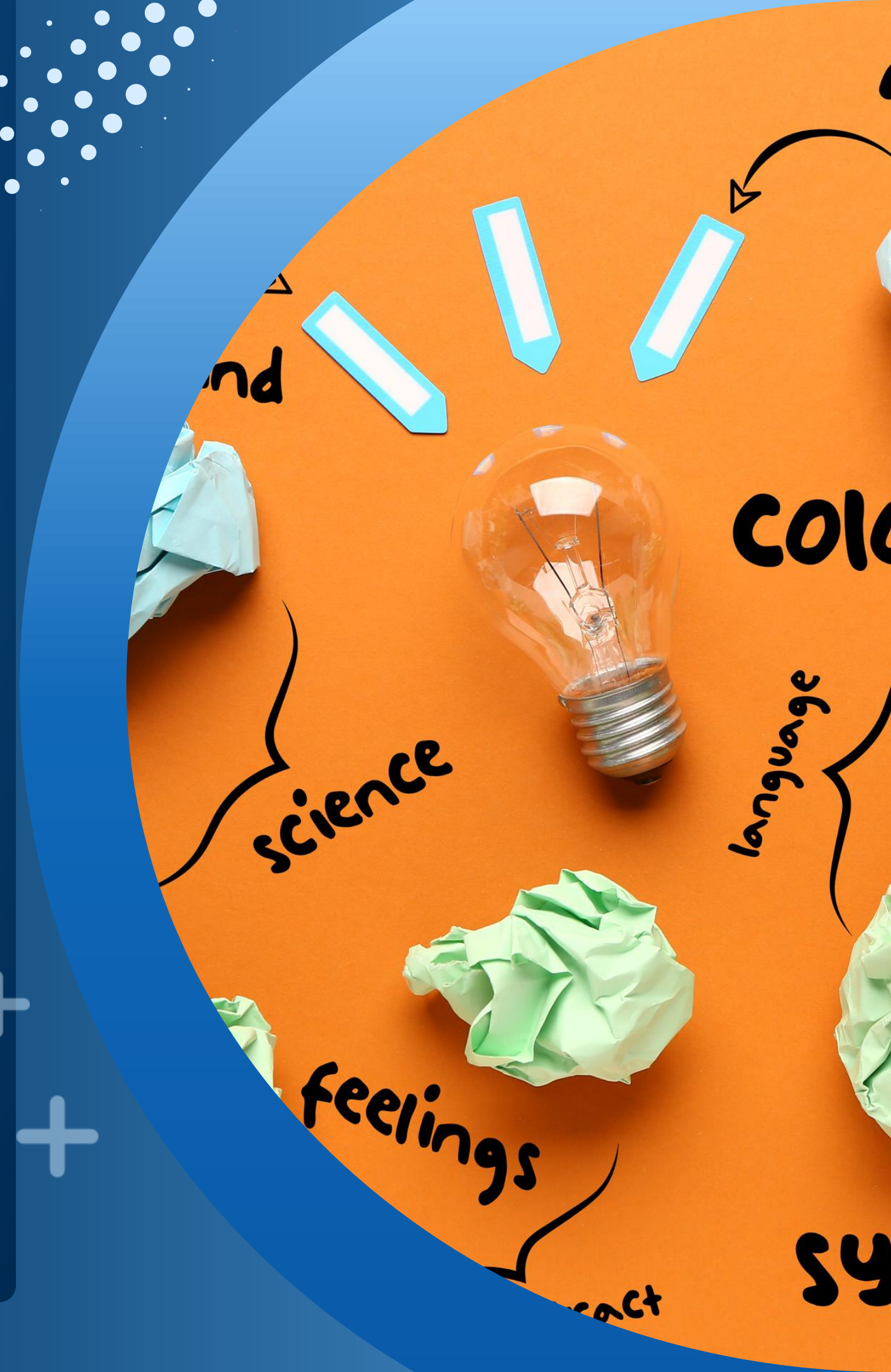
1. Frontline clinicians who need actionable solutions
2. Hospital executives who want high-level trends
3. Generate two versions of the same visualization—one optimized for each audience. Explain the differences in approach.



LLM Prompting

Chain of Thought Reasoning

- To generate a step-by-step reasoning process output before arriving at a final answer or decision.
- Why Chain of Thought?
- Helps AI break down complex data before summarizing it.
- Encourages a structured, logical flow instead of generic summaries.



LLM Prompting

Chain of Thought Reasoning

- Me: You are a healthcare executive assistant skilled in summarizing quality improvement projects. Follow this step-by-step reasoning before writing the executive summary:

- 1 Identify the main problem we are addressing.
- 2 Summarize key data insights that highlight the scope of the issue.
- 3 List the top solutions proposed based on our findings.
- 4 Summarize how AI or QI methodologies are being used to improve outcomes.
- 5 Provide a concise summary that could be used in a hospital leadership report.

Here is the information: [Insert project details].

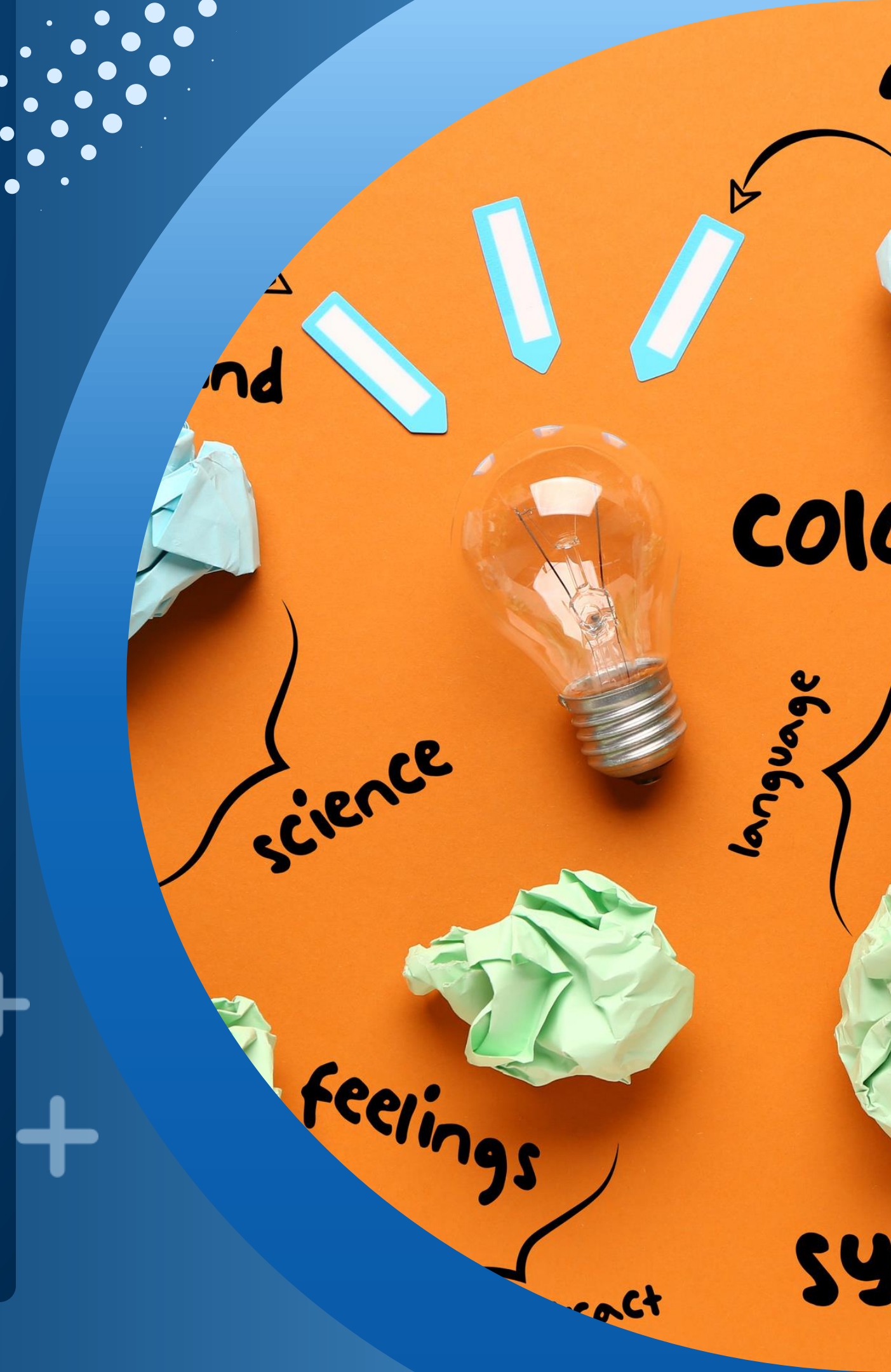


LLM Prompting

Chain of Thought Reasoning

- Me: You are a project manager for a hospital quality improvement team. Use step-by-step reasoning to generate a 6-month project timeline for improving discharge education to reduce readmissions. Follow this structured process:

- 1 Identify key milestones (e.g., data collection, intervention rollout, evaluation).
- 2 Assign realistic timeframes for each milestone.
- 3 List dependencies (which steps must happen before others).
- 4 Suggest a team role assignment for each phase.
- 5 Provide the final timeline in a clear Gantt chart format or bullet points.



LLM Prompting

- **Persona Patterns**

Me: you are my personal assistant. Whenever you provide output, please think through what would be the most time efficient recommendations to make only recommend things that will really save me time, do not suggest things that do not save me time. I would like to make chili tonight.

ChatGPT:

To save time making chili tonight, here's the most efficient approach:

- Use pre-chopped vegetables – Grab a pre-chopped onion, bell peppers, and garlic from the store or use frozen diced vegetables.
- Opt for canned beans and tomatoes – Use canned beans (black, kidney, or pinto) and canned diced tomatoes to cut down prep and cooking time. No need for soaking or peeling
- Buy pre-seasoned ground beef – If possible, get pre-seasoned ground beef, or turkey, which will reduce seasoning adjustments and cook faster.
- Use a chili seasoning mix – This will save time over measuring out individual spices.



AI = Augmented Intelligence

Thank You

Always Happy to Chat About

AI
james.barry@cuanschutz.edu

www.linkedin.com/in/jamesbarry3636

