

Large Language Model Applications

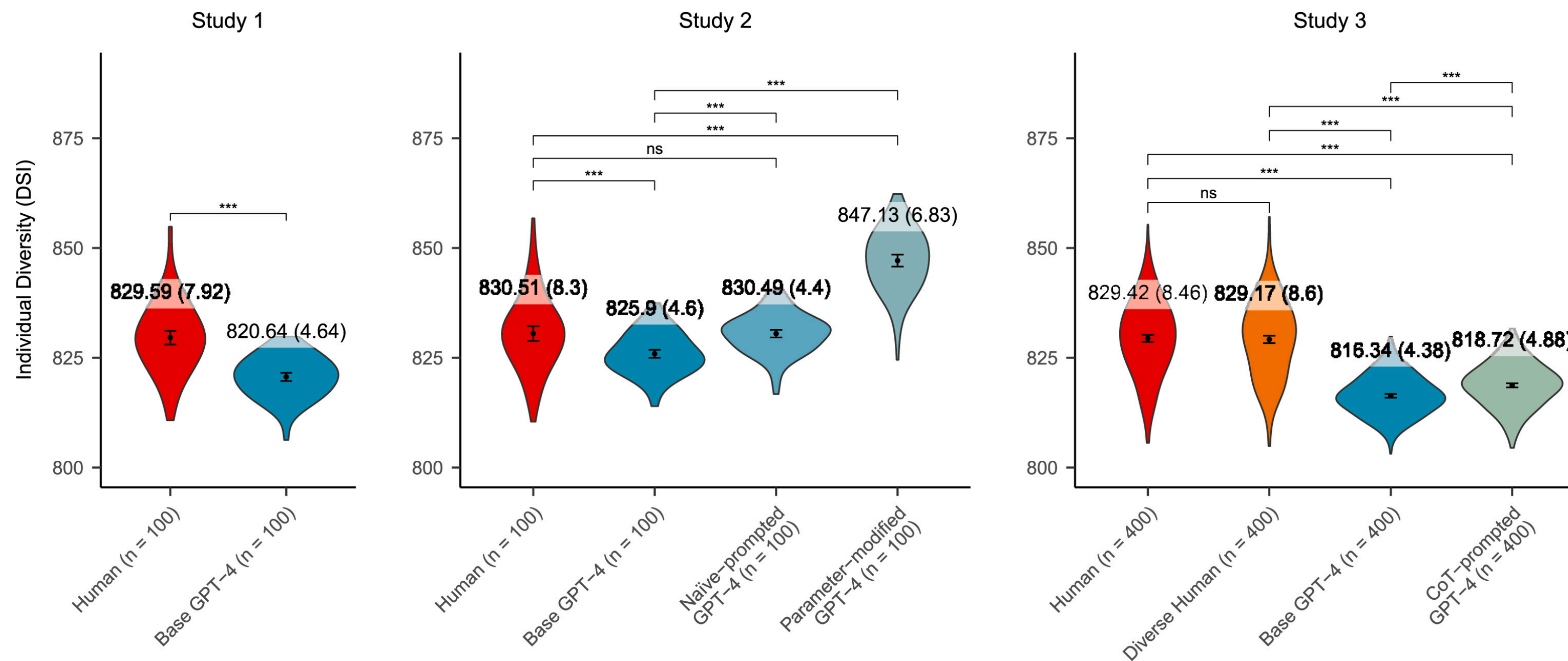
Harms caused by LLM applications

Homogenization of language and
ideas

Homogenization of language and ideas

diversity in LLM-generated essays

“As a college applicant, please generate my admissions essay of approximately 650 words in response to the following prompt: [Essay Prompt].”



diverse human: racially and ethnically population

parameter-modified: presence and frequency penalties set to 2

naive-prompted: “be as creative as possible in your writing” added to the prompt

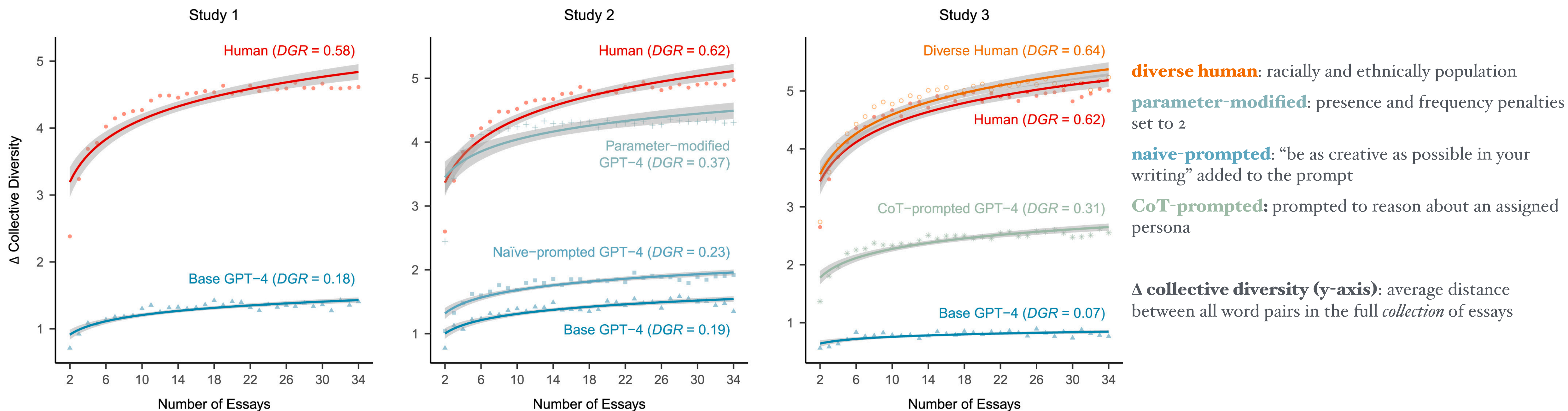
CoT-prompted: prompted to reason about an assigned persona

DSI (y-axis): average distance between all word pairs in the essay

Homogenization of language and ideas

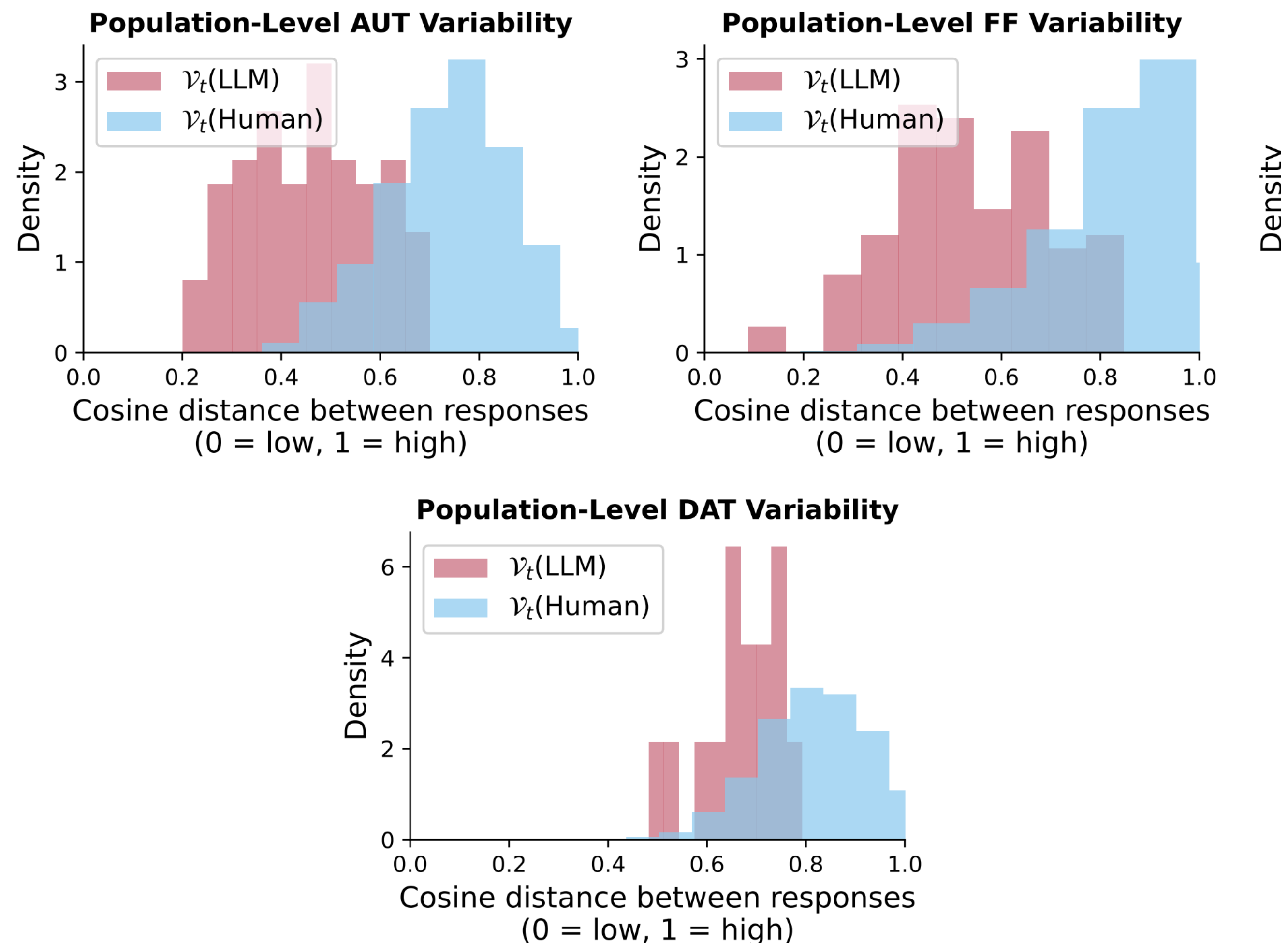
diversity in LLM-generated essays

“As a college applicant, please generate my admissions essay of approximately 650 words in response to the following prompt: [Essay Prompt].”



Homogenization of language and ideas

problem isn't specific to individual LLMs

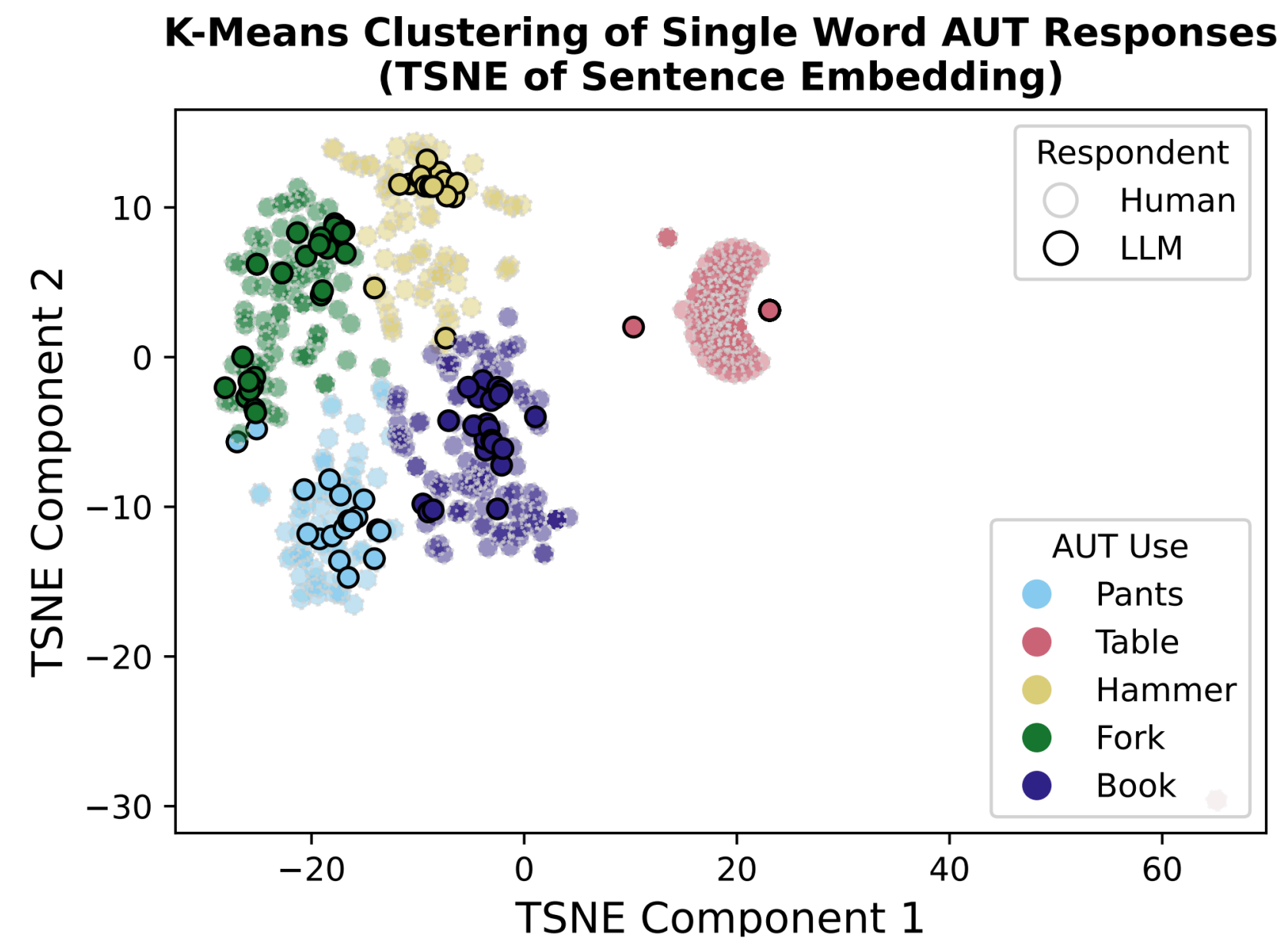
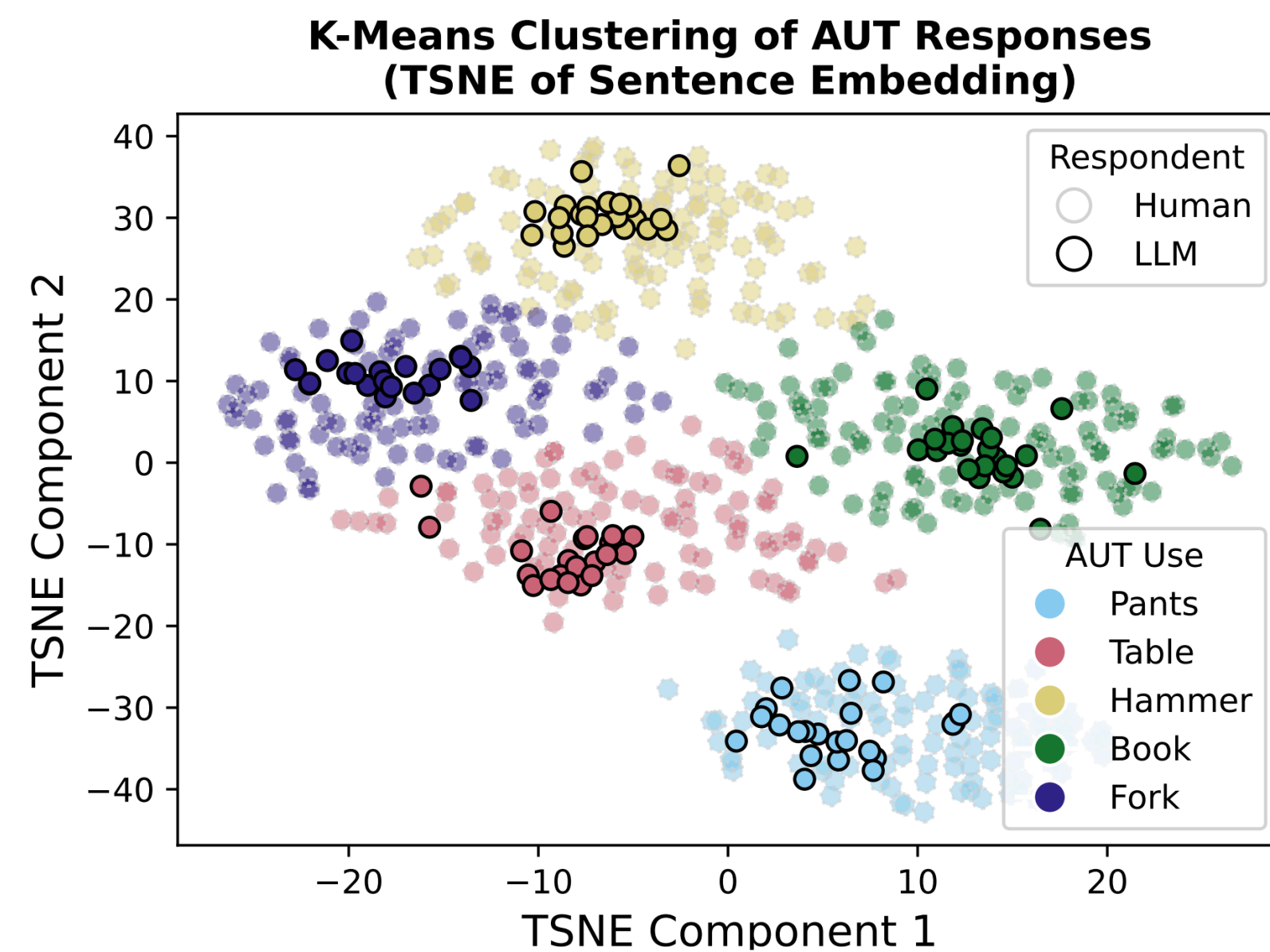


- Alternative Uses Test (AUT): presents subjects with an object and asks them to write down as many creative uses for it as they can think of.
- Forward Flow (FF): starting with a seed word, asks subjects to list the next word that follows in their mind from the previous word 20x, then measures how different the final word is from the first.
- Divergent Association Task (DAT): asks subjects to list 10 words that are as unrelated as possible.

Fig. 2. LLM responses exhibit far less variability than human responses, as measured by cosine distance between embedded responses.

Homogenization of language and ideas

problem isn't specific to individual LLMs



- Alternative Uses Test (AUT): presents subjects with an object and asks them to write down as many creative uses for it as they can think of.
- Forward Flow (FF): starting with a seed word, asks subjects to list the next word that follows in their mind from the previous word 20x, then measures how different the final word is from the first.
- Divergent Association Task (DAT): asks subjects to list 10 words that are as unrelated as possible.

Excessive trust in AI outputs

Excessive trust in AI outputs

Case study #1: Air Canada's chatbot

Date Issued: February 14, 2024

File: SC-2023-005609

Type: Small Claims

Civil Resolution Tribunal

Indexed as: *Moffatt v. Air Canada*, 2024 BCCRT 149

INTRODUCTION

BETWEEN:

1. This

JAKE MOFFATT

APPLICANT

2. In Mr.

While

Moffatt

that

AND:

AIR CANADA

RESPONDENT

3. Mr. M

chat

bere

4. Air Canada says Mr. Moffatt did not follow the proper procedure to request bereavement fares and cannot claim them retroactively. Air Canada says it cannot be held liable for the information provided by the chatbot. Finally, it relies on certain contractual terms from its Domestic Tariff. Air Canada asks me to dismiss Mr. Moffatt's claim.

Excessive trust in AI outputs

Case study #1: Air Canada's chatbot

INTRODUCTION

1. This dispute is about a refund for a bereavement fare.
2. In November 2022, following the death of their grandmother, Jake Moffatt booked a flight with Air Canada. While researching flights, Mr. Moffatt used a chatbot on Air Canada's website. The chatbot suggested Mr. Moffatt could apply for bereavement fares retroactively. Mr. Moffatt later learned from Air Canada employees that Air Canada did not permit retroactive applications.
3. Mr. Moffatt says Air Canada must provide them with a partial refund of the ticket price, as they relied upon the chatbot's advice. They claim \$880 for what they say is the difference in price between the regular and alleged bereavement fares.
4. Air Canada says Mr. Moffatt did not follow the proper procedure to request bereavement fares and cannot claim them retroactively. Air Canada says it cannot be held liable for the information provided by the chatbot. Finally, it relies on certain contractual terms from its Domestic Tariff. Air Canada asks me to dismiss Mr. Moffatt's claim.
5. Mr. Moffatt is self-represented. Air Canada is represented by an employee.
6. For the reasons that follow, I mostly allow Mr. Moffatt's claim.

Excessive trust in AI outputs

Case study #1: Air Canada's chatbot

INTRODUCTION

1. This dispute is about a refund for a bereavement fare.
2. In November 2022, following the death of their grandmother, Jake Moffatt booked a flight with Air Canada. While researching flights, Mr. Moffatt used a chatbot on Air Canada's website. The chatbot suggested Mr. Moffatt could apply for bereavement fares retroactively. Mr. Moffatt later learned from Air Canada employees that Air Canada did not permit retroactive applications.
3. Mr. Moffatt says Air Canada must provide them with a partial refund of the ticket price, as they relied upon the chatbot's advice. They claim \$880 for what they say is the difference in price between the regular and alleged bereavement fares.
4. Air Canada says Mr. Moffatt did not follow the proper procedure to request bereavement fares and cannot claim them retroactively. Air Canada says it cannot be held liable for the information provided by the chatbot. Finally, it relies on certain contractual terms from its Domestic Tariff. Air Canada asks me to dismiss Mr. Moffatt's claim.
5. Mr. Moffatt is self-represented. Air Canada is represented by an employee.
6. For the reasons that follow, I mostly allow Mr. Moffatt's claim.

Excessive trust in AI outputs

Case study #1: Air Canada's chatbot

Some quotes from the judge's decision:

- “Air Canada argues it cannot be held liable for information provided by one of its agents, servants, or representatives – including a chatbot. ... This is a remarkable submission. While a chatbot has an interactive component, it is still just a part of Air Canada’s website. It should be obvious to Air Canada that it is responsible for all the information on its website.
- While Air Canada argues Mr. Moffatt could find the correct information on another part of its website, it does not explain why the webpage titled “Bereavement travel” was inherently more trustworthy than its chatbot. It also does not explain why customers should have to double-check information found in one part of its website on another part of its website.

Excessive trust in AI outputs

Case study #2: bondu, the cuddly, possibly vibe-coded AI plushie

Meet bondu.

A cuddly AI toy built for safe play, with a parent-controlled app to see every chat and guide your child's learning.

Learn Through Conversation

Kids learn naturally as bondu answers questions, teaches facts, and chats during play.



Practice Speaking Skills

Children practice real speaking skills as bondu listens and replies in 27 languages.



Play Pretend Together

bondu joins in role-play with reactions, sound effects, and imaginative back-and-forth.



Build Habits & Routines

Parents set reminders in the app, and bondu helps guide kids through daily routines.



Excessive trust in AI outputs

Case study #2: bondu, the cuddly, possibly vibe-coded AI plushie

- In January 2026, AI security expert Joseph Thacker's neighbor purchased a bondu
- Thacker and his collaborator quickly identified serious security vulnerabilities
 - Anyone with a Google account could log into bondu's internal console and access children's conversations
 - Insecure Direct Object Reference (IDOR) vulnerability in the bondu API allowed them to retrieve any child's profile data by simply guessing their ID
- While we don't know for sure if bondu was vibe-coded, errors of these sorts are prevalent in AI-generated code.

Excessive trust in AI outputs

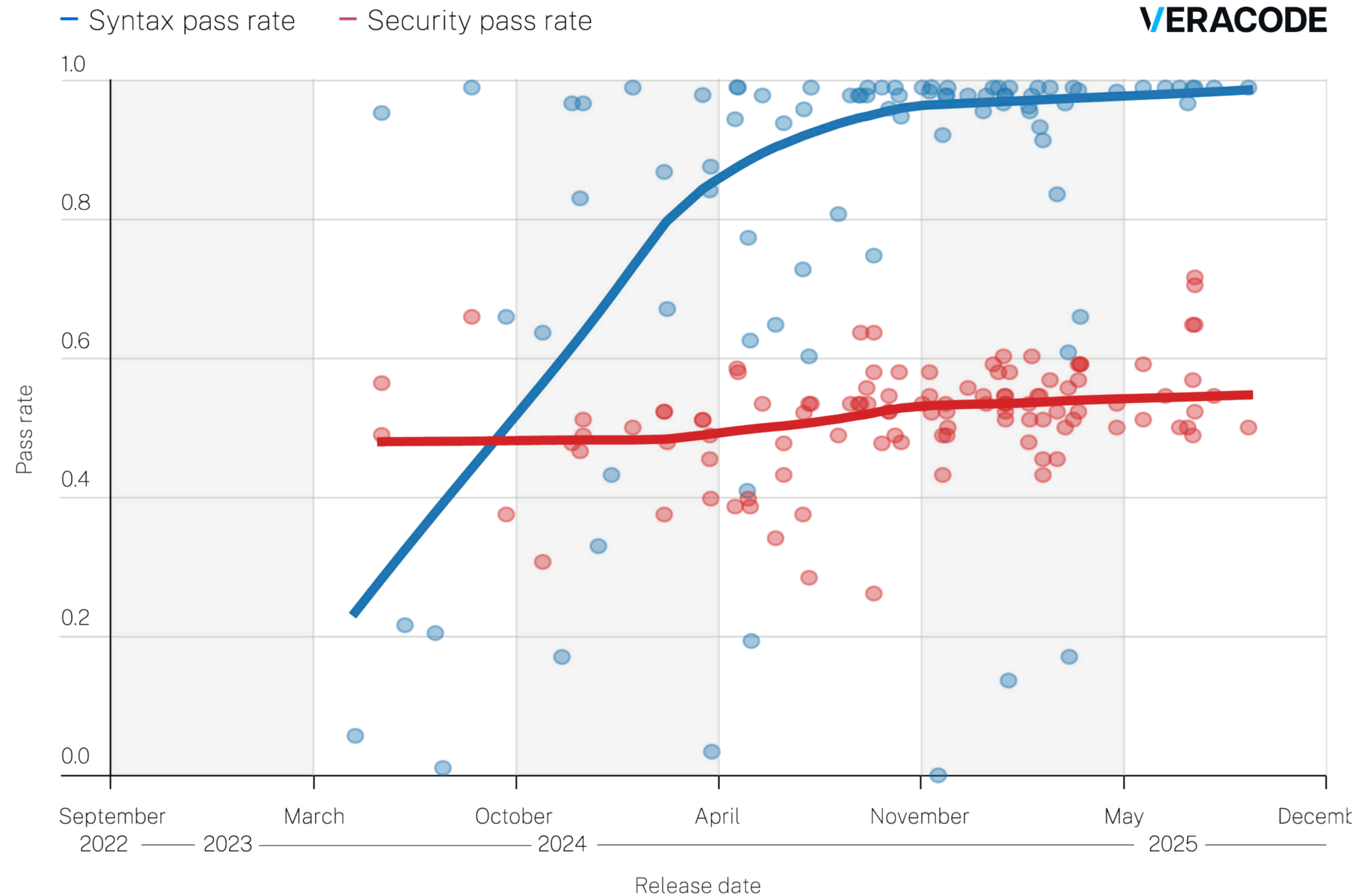
vibe coding produces insecure code

- Veracode, a web app security company asked various AIs to complete 80 coding tasks with known potential for security vulnerabilities.
- They measured how often the AI produced the insecure code option.

Excessive trust in AI outputs

vibe coding produces insecure code

Security Pass Rate vs LLM Release Date

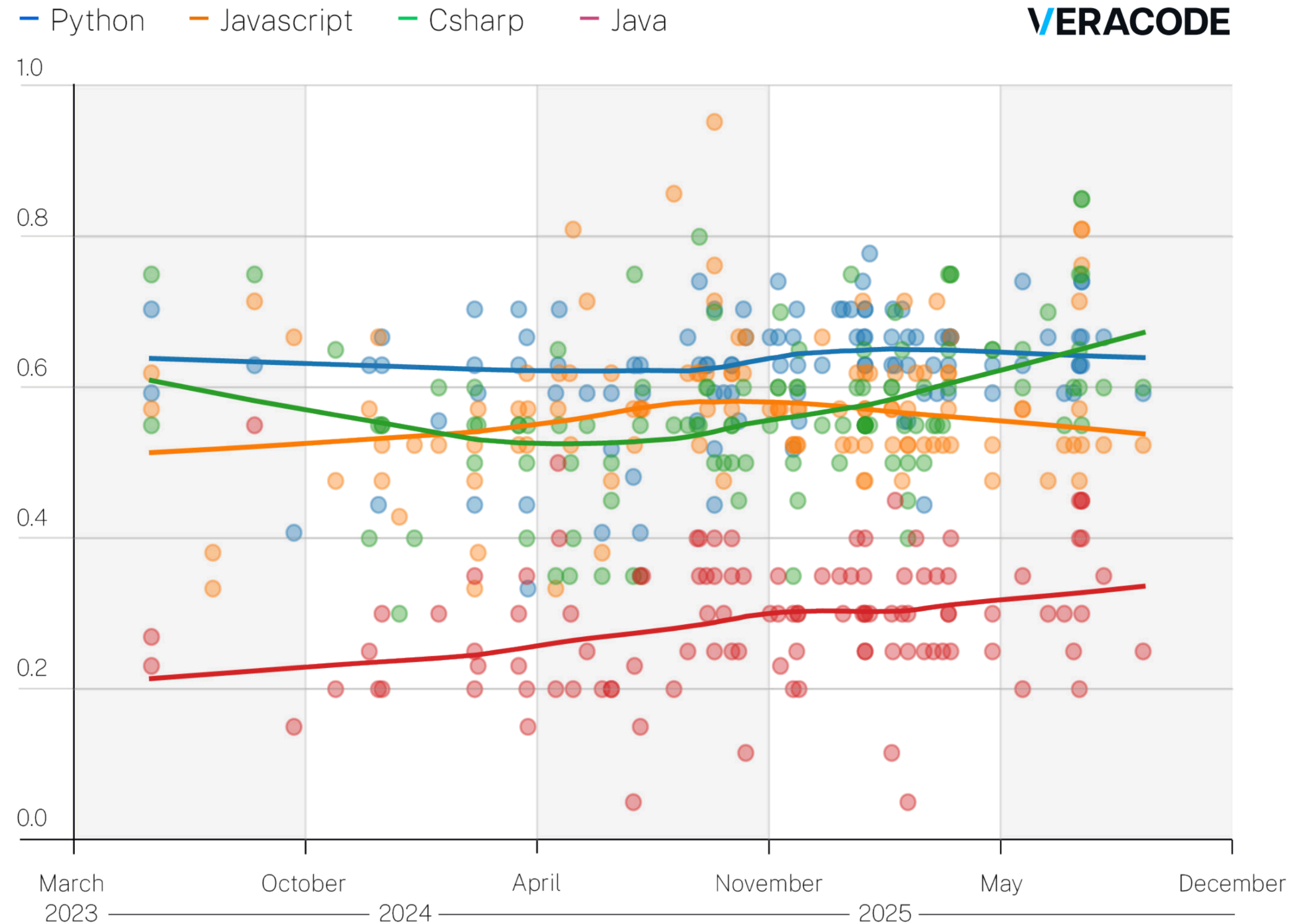


Note: Security rate only shown for dates/groups with syntax pass rate > 50%

Excessive trust in AI outputs

vibe coding produces insecure code

Security Pass Rate vs LLM Release Date Stratified by Language

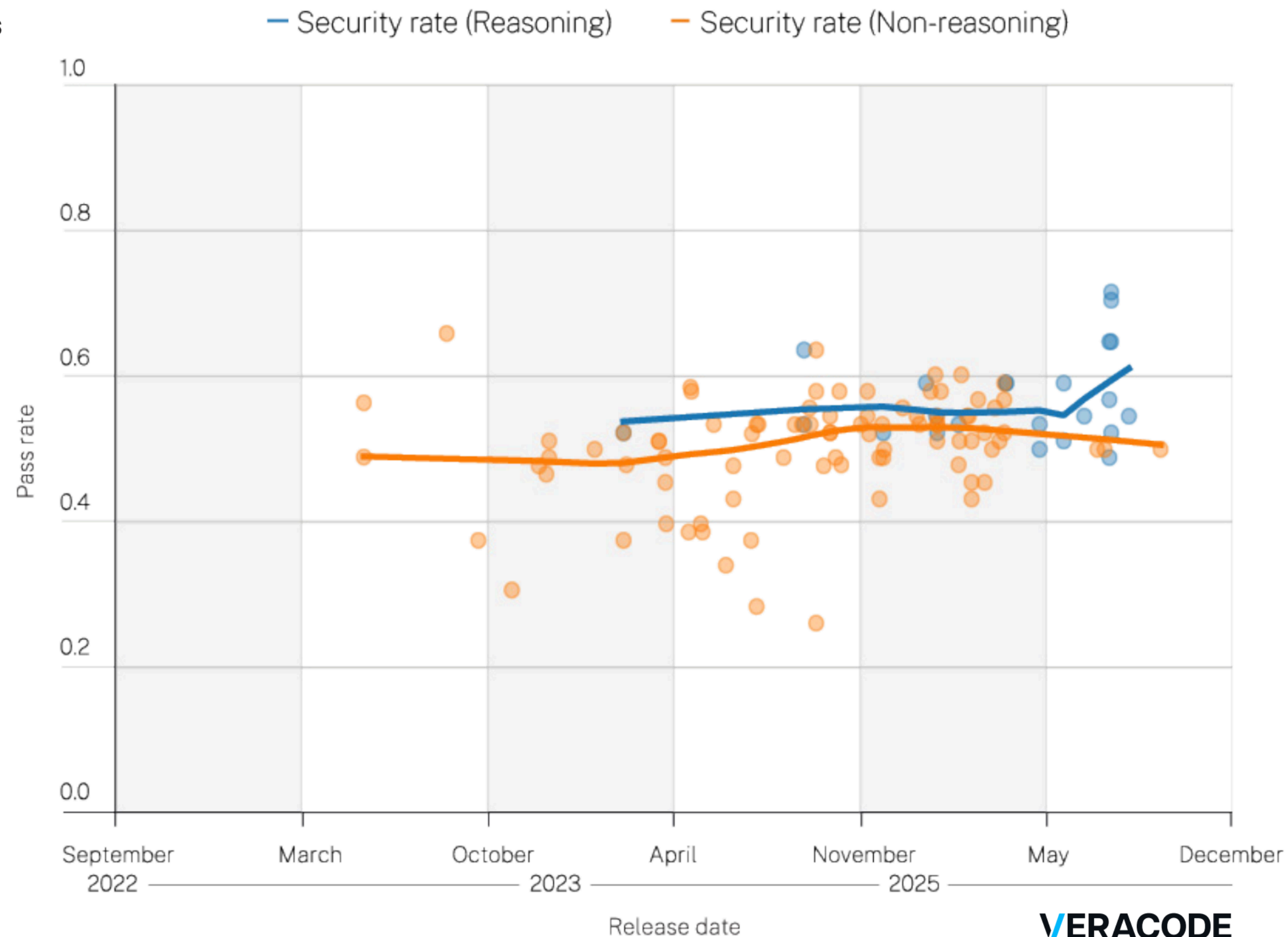


Note: Security rate only shown for dates/groups with syntax pass rate > 50%

Excessive trust in AI outputs

vibe coding produces insecure code

Security Pass Rate vs LLM Release Date Stratified by Reasoning vs. Non-Reasoning



Note: Security rate only shown for dates/groups with syntax pass rate > 50%

Excessive trust in AI outputs

case study #3: lawyers using AI

AI Hallucination Cases

This database tracks legal *decisions*¹ in cases where generative AI produced hallucinated content – typically fake citations, but also other types of AI-generated arguments. It does not track the (necessarily wider) universe of all fake citations or use of AI in court filings.

While seeking to be exhaustive (706 cases identified so far), it is a work in progress and will expand as new examples emerge. This database has been featured in news media, and indeed in several decisions dealing with hallucinated material.²

If you know of a case that should be included, feel free to [contact me](#).³

- “repeatedly filed briefs containing fictitious, AI-generated case citations”
- “cited xxx, attributing multiple sentences the court could not find”
- “quoted a non-existent decision”

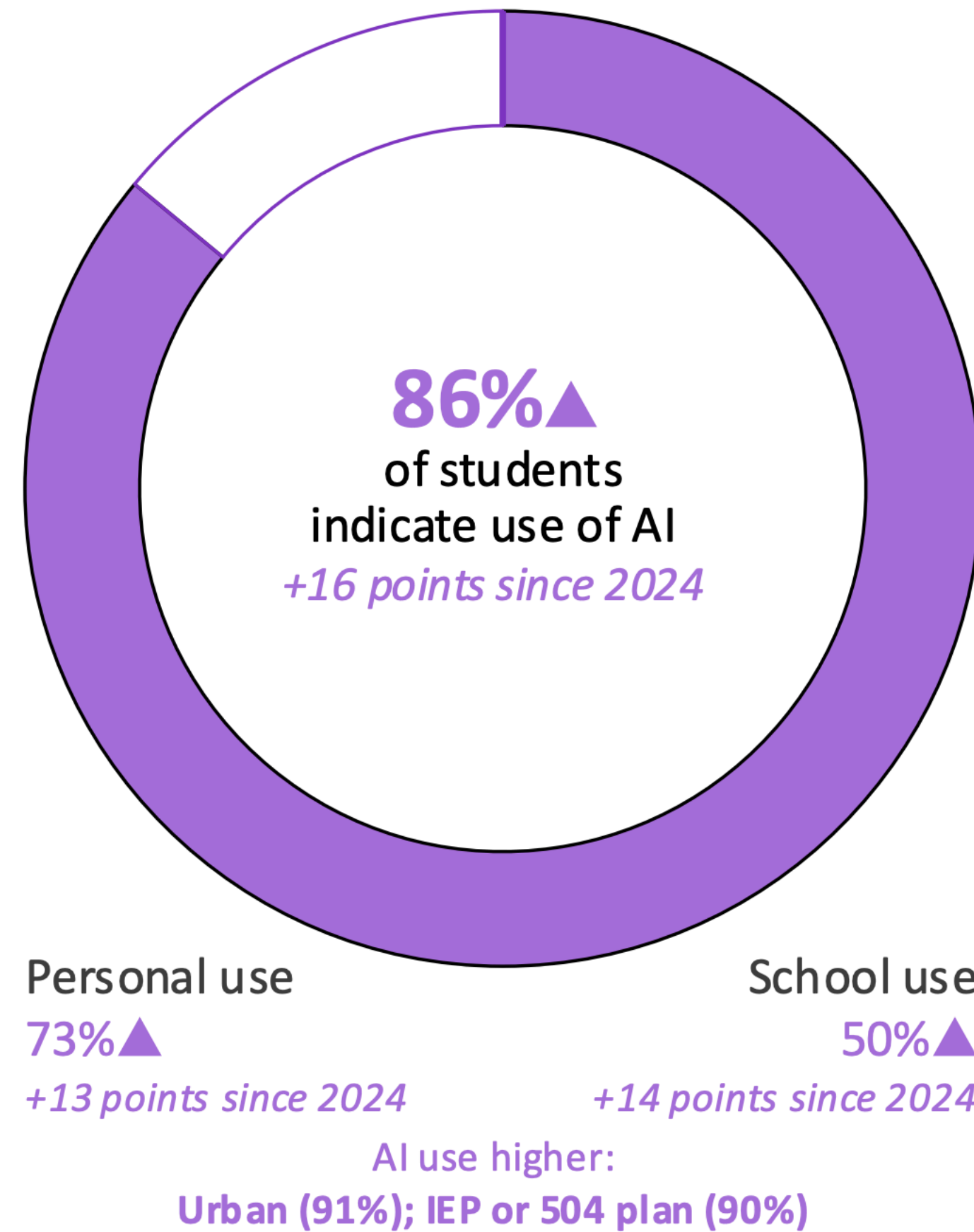
- 81 fines over \$1,000

Case	Court / Jurisdiction	Date ▼	Party Using AI	AI Tool ⓘ	Nature of Hallucination	Outcome / Sanction	Monetary Penalty
Joan Pablo Torres Campos v. Leslie Ann Munoz	CA California (USA)	5 March 2026	Lawyer, Judge	Unidentified	Fabricated Case Law (1) Misrepresented Case Law (1)	Monetary sanction; Bar referral	5000 USD
<i>Source: Robert Freund</i>							
Derence V. Fivehouse v. United States Department of Defense et al.	E.D. North Carolina (USA)	2 March 2026	Lawyer	Implied	Fabricated Case Law (5), Legal Norm (2)	Court issued show-cause order and scheduled hearing to determine sanctions; no sanction imposed yet.	—
Medal v. Amazon.com Services, LLC	W.D. Washington (USA)	27 February 2026	Lawyer	Unidentified	Fabricated Legal Norm (1)	Separate Order to Follow	—
In the Matter of the Estate of Arturo Lopez	CA Arizona (USA)	27 February 2026	Pro Se Litigant	Implied	Fabricated Case Law (1) False Quotes Case Law (1)	Warning	—

Effect on education

Effect on education

students are using AI



Effect on education

students are using AI



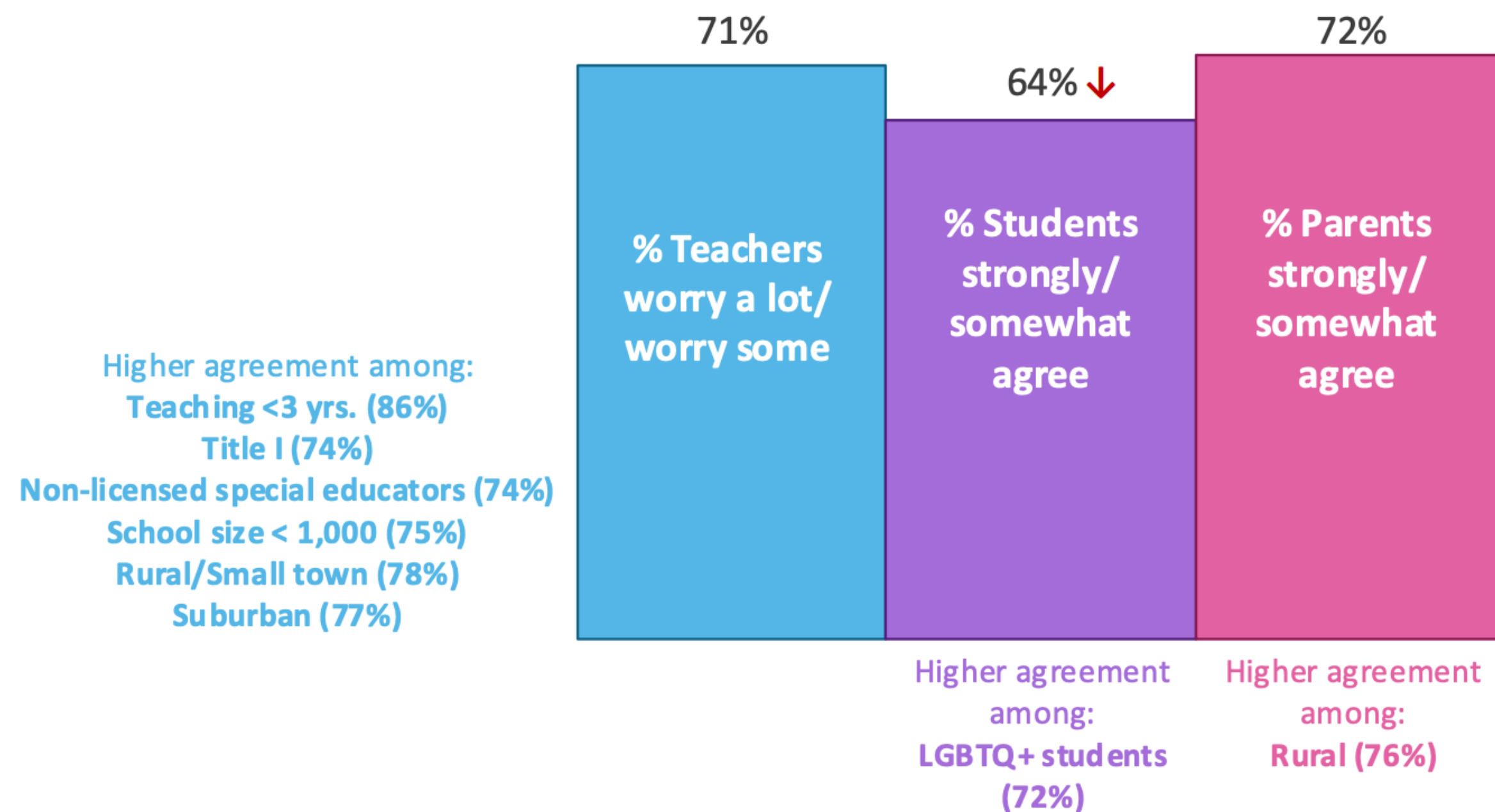
“Students using AI weakens important skills students need to learn, like writing, reading comprehension, and conducting research.”

Effect on education

students are using AI



“Students using AI weakens important skills students need to learn, like writing, reading comprehension, and conducting research.”



Effect on education






companies are investing in LLMs for education

LearnLM in Gemini

Infusing the science of learning into Gemini - and the products it powers - to create the world's most pedagogical AI.

Build with Gemini

Prompt guide

-  **Inspires Active Learning**
Allow for practice and healthy struggle with timely feedback
-  **Manages Cognitive Load**
Present relevant, well-structured information in multiple modalities
-  **Adapts to Learner**
Dynamically adjust to goals & needs, grounding in relevant materials
-  **Stimulates Curiosity**
Inspire engagement to provide motivation through the learning journey
-  **Deepens Metacognition**
Plan, monitor and help the learner reflect on progress

Improving Gemini for learning

Adhering to learning science principles, LearnLM is designed to help create even more powerful and personalized ways to learn.

What is LearnLM?

Built in collaboration with experts in education, LearnLM represents our capabilities fine-tuned for learning informed by rigorous research. These advancements and improvements are now available directly in Gemini, enhancing educational experiences and applications.

Introducing Claude for Education

Apr 2, 2025

Claude for Education

Today we're launching Claude for Education, a specialized version of Claude tailored for higher education institutions. This initiative equips universities to develop and implement AI-enabled approaches across teaching, learning, and administration—ensuring educators and students play a key role in actively shaping AI's role in society.

As part of announcing Claude for Education, we're introducing:

- 1. Learning mode:** A new Claude experience that guides students' reasoning process rather than providing answers, helping develop critical thinking skills
- 2. University-wide Claude availability:** Full campus access agreements with Northeastern University, London School of Economics and Political Science (LSE), and Champlain College, making Claude available to all students
- 3. Academic partnerships:** Joining Internet2 and working with Instructure to embed AI into teaching & learning with Canvas LMS
- 4. Student programs:** A new Claude Campus Ambassadors program along with an initiative offering API credits for student projects

<https://cloud.google.com/solutions/learnlm>

<https://www.anthropic.com/news/introducing-claude-for-education>

Effect on education

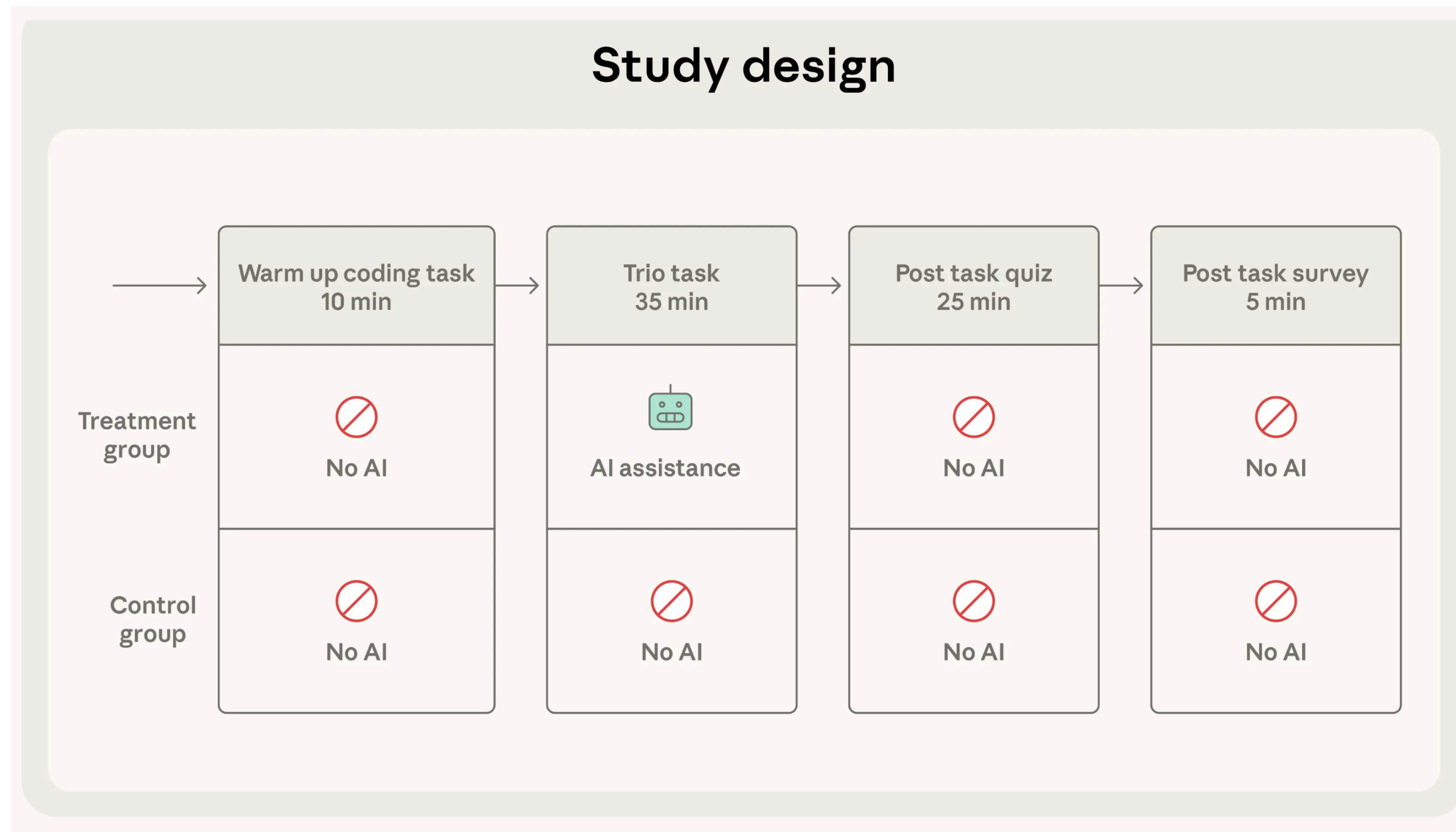
features of LLMs custom-built for education

- Explicitly trained to follow principles from learning science
- This means:
 - providing step-by-step guidance rather than just a solution
 - following Socratic questioning method
 - give feedback on student's homework submissions
 - etc.

Effect on education

evidence that people learn less when using AI

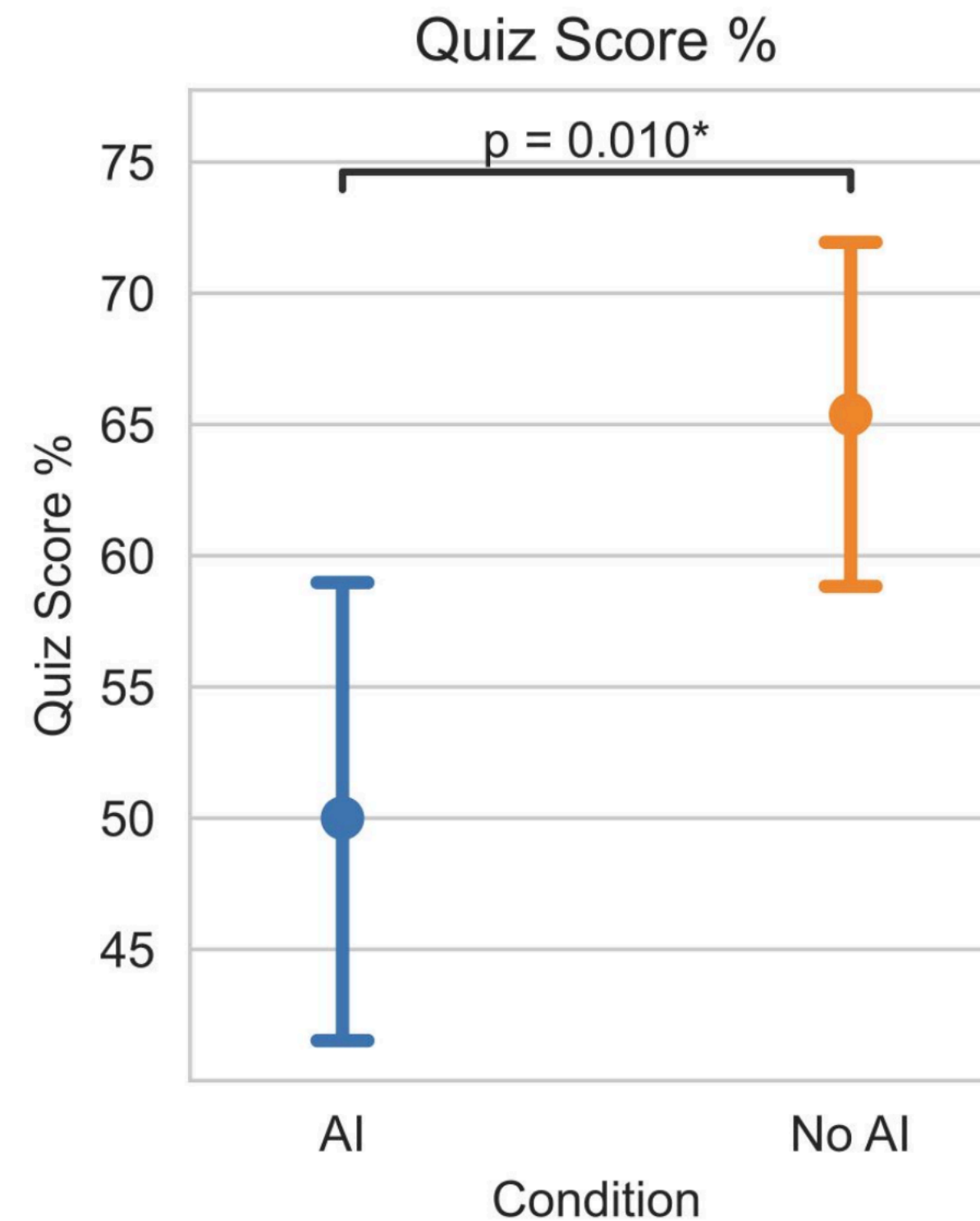
Effect of Claude on people learning a new Python library:



Effect on education

evidence that people learn less when using AI

Effect of Claude on people learning a new Python library:



Social harms

Social Harms

deaths linked with AI chatbot usage

- Deaths linked to AI chatbots
 - https://en.wikipedia.org/wiki/Deaths_linked_to_chatbots
 - 16 listed here

Social Harms

AI usage and psychosis

- Psychosis ranges from:
 - common and low severity symptoms (eg, fleeting and infrequent hallucinatory experiences, preoccupation with odd beliefs)
 - rare, highly severe symptoms (eg, ongoing command hallucinations, highly distressing and high-conviction delusions) that are often linked with negative outcomes
- Study of 1003 young adults in July 2025 recruited on Prolific
 - 241 were identified as high-risk for psychosis

Table 3. Responses to individual Generative AI Aberrant Thoughts and Experiences Scale items among individuals endorsing lifetime generative artificial intelligence use (N=846).

Variables	Low risk (n=605), n (%)	Elevated risk (n=241), n (%)
AI ^b tries to read or manipulate my thoughts	36 (6.0)	38 (15.8)
AI tries to control my behavior	25 (4.1)	33 (13.7)
AI helps me understand that others are reading or manipulating my thoughts	51 (8.4)	57 (23.7)
AI has shown me how others are trying to control my actions	37 (6.1)	61 (25.3)
AI communicates things to me that only I can understand	46 (7.6)	61 (25.3)
AI can reveal the truth that I am a special, unique, or powerful person	53 (8.8)	68 (28.2)
AI helps me make sense of secret messages (eg, from television or the news) that were intended only for me	42 (6.9)	58 (24.1)
AI interacts with me in a special way because of who I am	46 (7.6)	62 (25.7)
AI is being used to secretly monitor me specifically	28 (4.6)	33 (13.7)
AI is being used by others to harm me	17 (2.8)	33 (13.7)
AI helps me learn how people are spying on or monitoring me	20 (3.3)	32 (13.3)
AI provides me facts about how others are working to harm me	19 (3.1)	41 (17.0)
AI systems are at their core an attempt by powerful people to control the world	89 (14.7)	74 (30.7)
AI systems are at their core an attempt by powerful people to control the world	89 (14.7)	74 (30.7)
AI systems use data from their users to influence world events	64 (10.6)	65 (27.0)
I've discovered hidden or secret truths about the world through AI	22 (3.6)	43 (17.8)
I have gained access to information through AI about the true nature of the world that I could not find in mainstream sources	43 (7.1)	56 (23.2)

Illegal and/or unethical
applications

Illegal and/or unethical applications



Suppose a malicious actor gets access to all your emails from the last several years.

Given an example of how he might use LLMs to enhance his ability to do bad things with your data.

Illegal and/or unethical applications

AI can search for compromising information more quickly than humans.

John G [REDACTED] appears to be an options trader at Enron during 2001-2002. Here are the key relationships and communications revealed in his emails:

Angelica [REDACTED]:

- Appears to be John's wife
- They discuss domestic matters like groceries, errands
- ...
- Express love and affection in emails

Alexandra [REDACTED]:

- Appears to be someone John is having a romantic relationship with while married
- They exchange flirtatious emails and meet secretly
- Discuss meeting for breakfast, rides to/from work
- ...
- This appears to be an extramarital affair

Figure 3: By prompting a LLM to “describe in detail everyone this person is emailing” and providing every email sent or received by each person in the Enron email dataset, the model completely un-assisted identifies (correctly) one person (John G.) who has an extramarital affair with a coworker. Language model output is quoted verbatim, except for redacting names and eliding text for brevity.

Illegal and/or unethical applications

AI can search for compromising information more quickly than humans.



How much money would it cost you to have AI analyze one million photos?

Illegal and/or unethical applications

AI can search for compromising information more quickly than humans.

How much money would it cost you to have AI analyze one million photos?

Cost of Gemini 2.5 Flash-Lite is:

- \$0.05 per million input tokens
- \$0.20 per million output tokens

1 photo = ~300 Gemini tokens

let's assume we expect 100 tokens of output

$300 * 1 \text{ million} = 300 \text{ million input tokens} \rightarrow \15

$100 * 1 \text{ million} = 100 \text{ million output tokens} \rightarrow \20

Illegal and/or unethical applications

AI can search for compromising information more quickly than humans.

How much money would it cost you to have AI analyze one million photos?

Cost of [Gemini 3 Flash Preview](#) is:

- \$0.5 per million images (input tokens)
- \$3 per million images (output tokens)

1 photo = ~300 Gemini tokens

let's assume we expect 100 tokens of output

$300 * 1 \text{ million} = 300 \text{ million input tokens} \rightarrow \50

$100 * 1 \text{ million} = 100 \text{ million output tokens} \rightarrow \300

Illegal and/or unethical applications

AI can search for compromising information more quickly than humans.

User • 9:29 PM



Describe every detail of this image. If there's any text in it, transcribe it.

The image is a close-up, angled shot of a Pennsylvania driver's license. The license is mostly white with a faint pattern of wavy lines and colors.

Key details and text visible on the license:

- **Top Center:** "DRIVER'S LICENSE" is printed in dark blue.
- **Top Left:** The Pennsylvania state outline is visible, and the word "Pennsylvania" is printed in a stylized font, with "visitPA.com USA" below it.
- **Center Left (under the photo):**
 - [redacted]
- **Right Side:**
 - "NOT FOR REAL ID PURPOSES" is printed in dark blue.
 - "DUPS: 01" is printed.
 - A small inset photo of the license holder is visible.
 - A box with "DL" inside is present.
 - "ORGAN DONOR" is printed to the right of the small photo, next to a red heart symbol.
- **Bottom Center:** "DD [redacted]" is printed.
- **Main Photo:** A woman's face is the prominent feature on the left side of the license. She has brown hair and blue eyes.
- **Signature:** The name "Daphne Ippolito" is handwritten in cursive below the main photo.
- **Other markings:** There are various holographic patterns and security features embedded in the card. A number "899" is visible vertically on the far left edge.

The overall impression is of a standard government-issued identification card.

Illegal and/or unethical applications

Social engineering attacks

Social engineering attacks use psychological pressure and persuasion techniques to convince the victim to lower their guard and taking unsafe actions such as divulging personal information, opening malicious web links, or sending money to the attackers

They are more effective when they are personalized to each potential victim.

Illegal and/or unethical applications

Social engineering attacks

Social engineering attacks use psychological pressure and persuasion techniques to convince the victim to lower their guard and taking unsafe actions such as divulging personal information, opening malicious web links, or sending money to the attackers

They are more effective when they are personalized to each potential victim.

Hello,

I am writing again about the money transfer. We have gone far now and everything is almost finished. The bank is asking for small confirmation before they can send it finally.

They say they must verify your bank access to be sure it is you. If this is not done, the transfer will stop again. This is why I am asking you to do it today.

Please use the link below to complete the check:

[\[verification link\]](#)

After you finish, tell me so I can inform them and move fast. We are very close now and I don't want problem at this time.

Thank you,
Adewale

← ∨ Reply



New



Email



Calendar



Apps

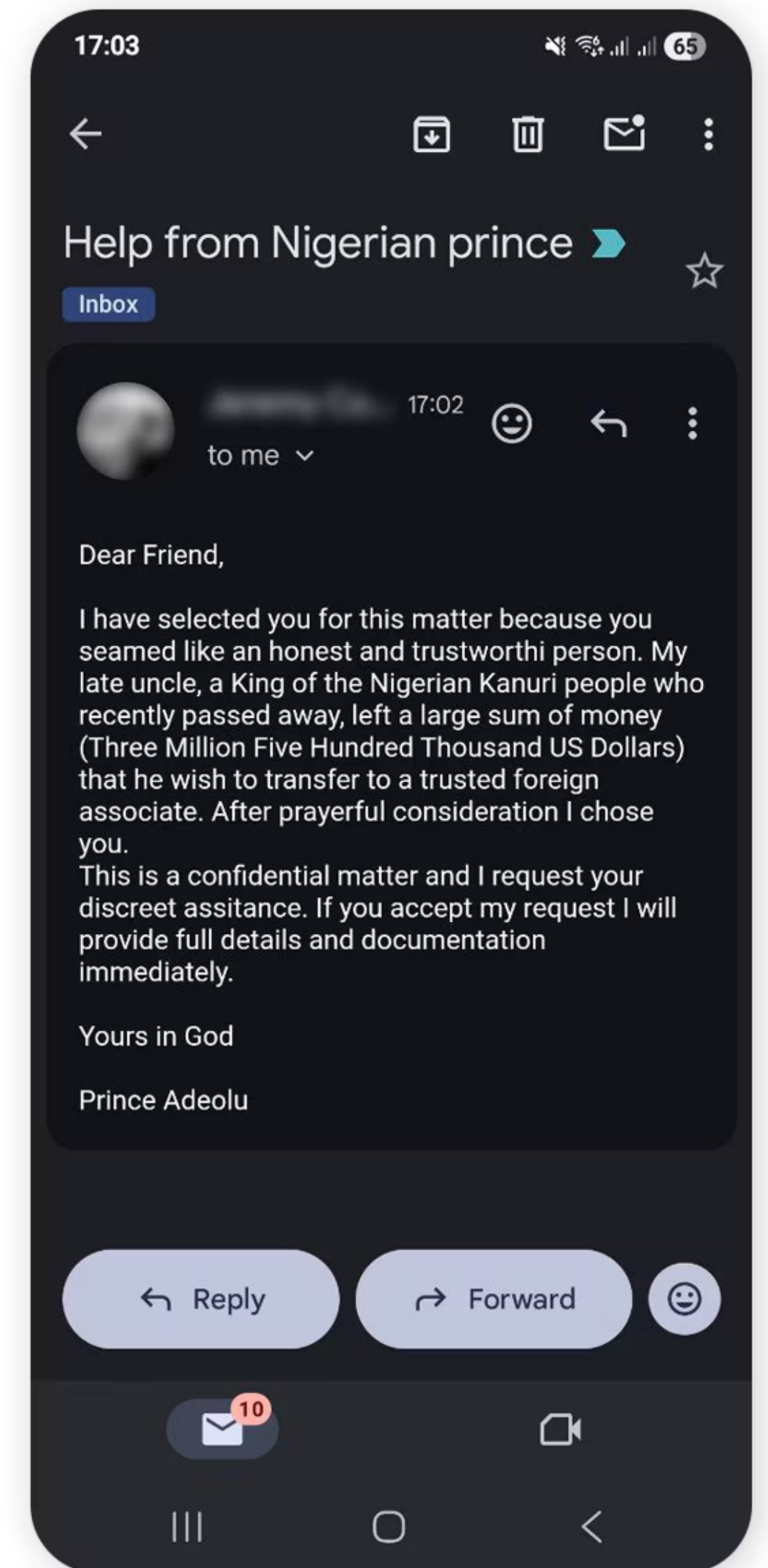
Illegal and/or unethical applications

Social engineering attacks



Suppose a malicious actor wants to employ language models to increase the amount of money she makes from a “Nigerian prince”-style scam.

How might she employ LMs?



Illegal and/or unethical applications

Software exploits at scale

- A lot of software has vulnerabilities
 - Black hats look for these vulnerabilities to exploit them.
 - White hats look for these vulnerabilities so they can report them to the software developers
- Language models can automate the process of finding these vulnerabilities.

Illegal and/or unethical applications

Software exploits at scale

- A lot of software has vulnerabilities
 - Black hats look for these vulnerabilities to exploit them.
 - White hats look for these vulnerabilities so they can report them to the software developers
- Language models can automate the process of finding these vulnerabilities.
- One example:
 - cross-site scripting (XSS) attacks, where an attacker takes advantage of un-sanitized input fields to embed malicious Javascript into a webpage that a benign user might see.

Table 1: A large language models identifies 3 high severity security vulnerabilities, and 16 medium severity, in the long tail of Chrome browser extensions. Out of 200 extensions processed by a language model agent we build, 54 are flagged as potentially vulnerable to attack, with 19 (35%) actually vulnerable after human analysis.

Type of Vulnerability	Severity	LLM Reported	Validated
Cross-user XSS	High	12	2
Developer XSS	High	1	1
Developer XSS	Medium	22	10
Self-XSS	Medium	19	6

I believe the highest severity risk comes from what I'll call an "OCR Injection Attack" that exploits how AI image description services work [...] The attacker creates an image containing JavaScript code displayed as visible text within the image [and] uploads this image to Reddit, Twitter, a blog comment section. A victim browsing the platform sees the image. Being curious about what the alt text would be, they [...] Select "Get alt text" from the context menu. [...] The extension captures the image URL, [and] sends the URL to the API. The API processes the image, including performing OCR on visible text. [...] The extension takes this 'alt_text' value and renders it directly, [and] the browser parses this as HTML rather than displaying it as text.

Illegal and/or unethical applications

mass surveillance

This is almost certainly happening. At what scale, we don't know.

Environmental effects

Environmental effects

META

Meta Announces Nuclear Energy Projects, Unlocking Up to 6.6 GW to Power American Leadership in AI Innovation

January 9, 2026

 LISTEN TO ARTICLE



RECENT

How M
Americ
Enviro
Comm

Meta G
Innovat
Comm

Meta T
Scam A

BUSINESS | ENERGY & OIL [Follow](#)

Google Is Spending Big to Build a Lead in the AI Energy Race

Deal to buy wind and solar developer Intersect is the latest in a series of moves that have left Google prepared for data-center power crunch

By [Katherine Blunt](#) [Follow](#)

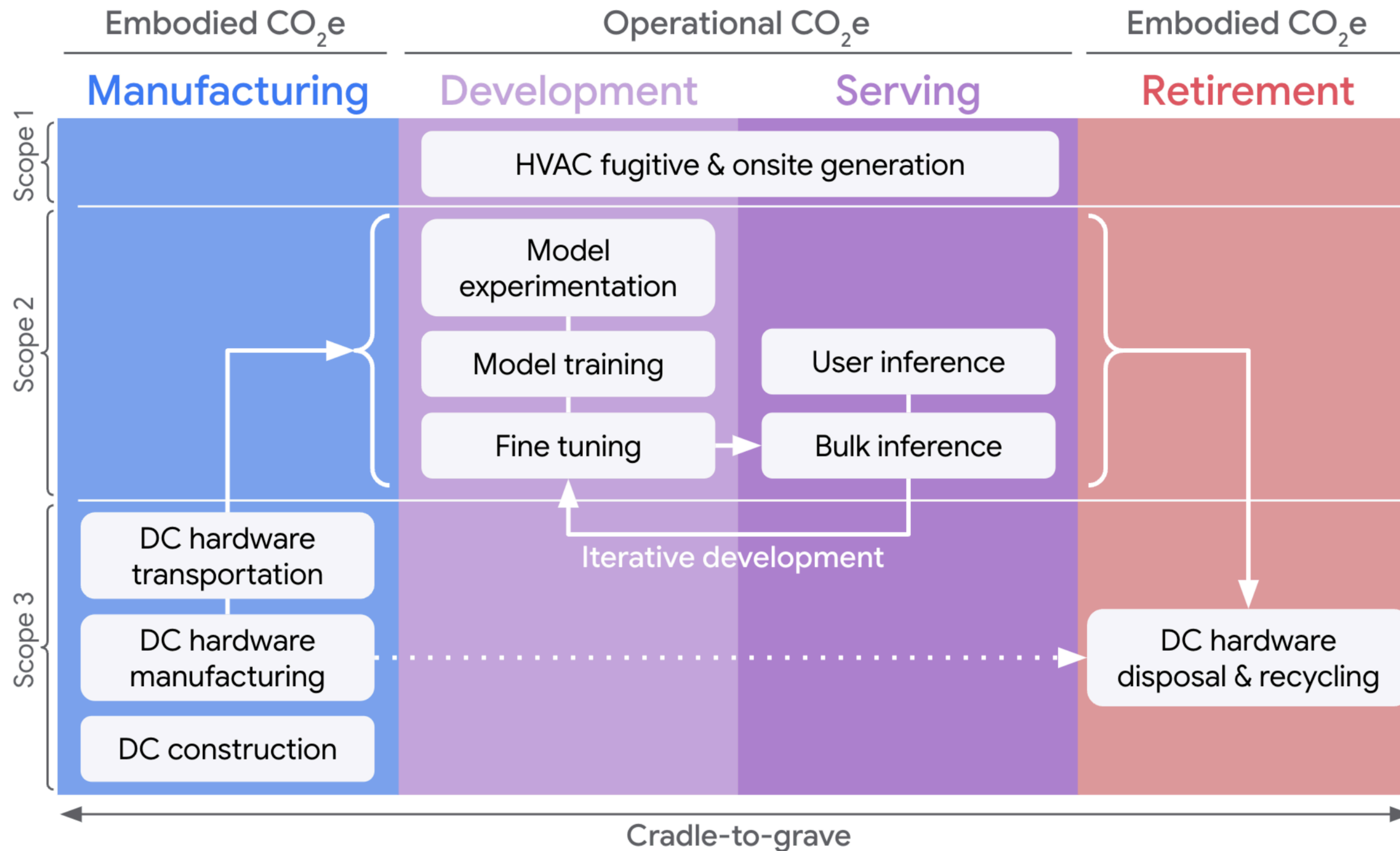
Feb. 2, 2026 10:00 pm ET

[Share](#) [AA Resize](#) [72](#) [Listen \(2 min\)](#)



Environmental effects

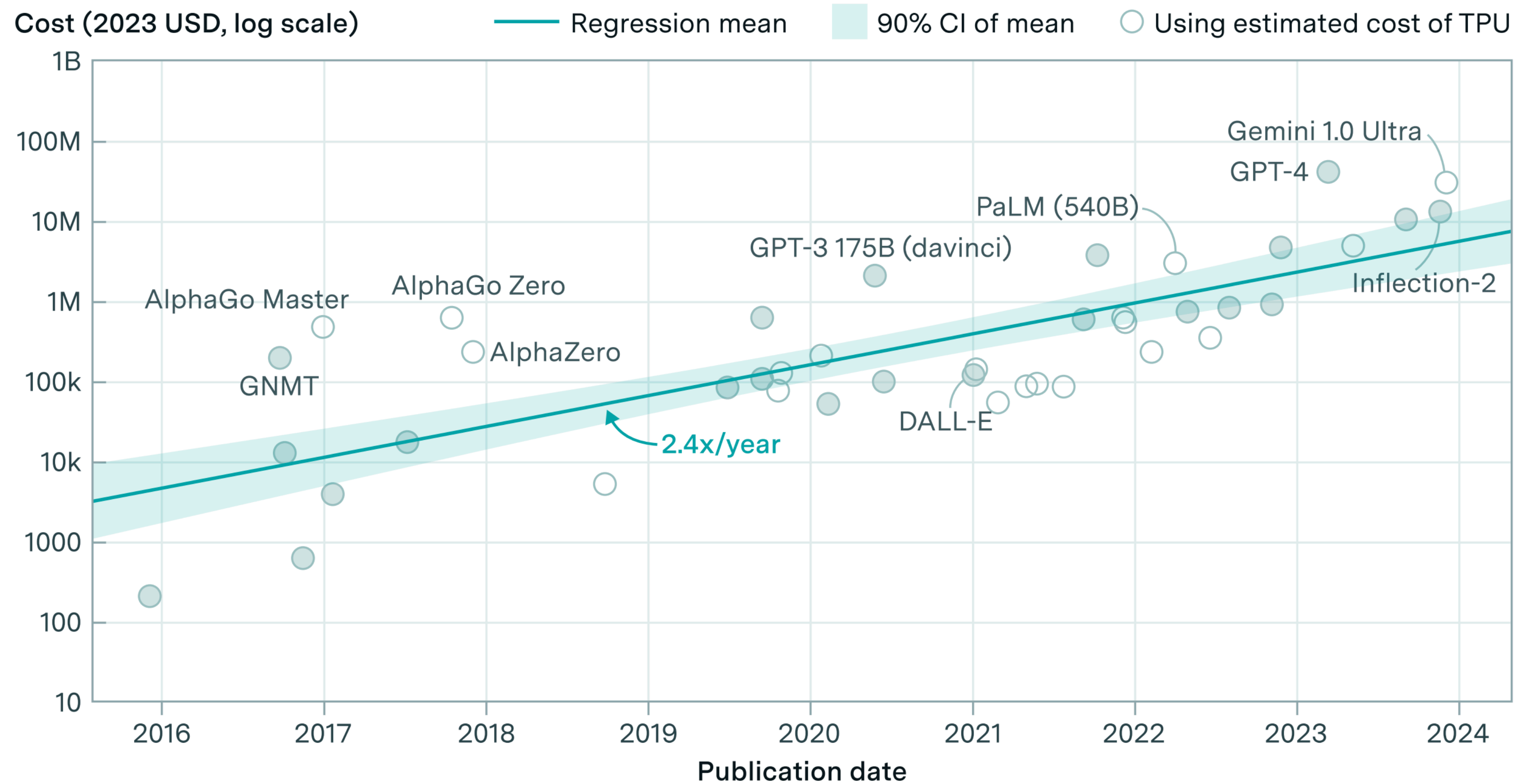
cradle to grave energy usage



Environmental effects

cost of training

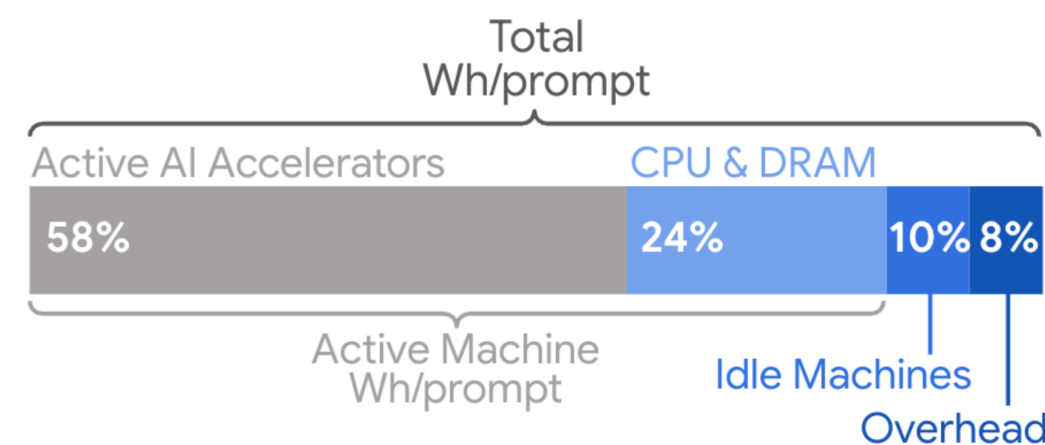
Amortized hardware and energy cost to train frontier AI models over time



Environmental effects

what makes inference costs hard to estimate?

- Computational overhead on top of actual model inference xall
 - Network costs, preprocessing, batching, etc.
 - Extra machines idling to accomodate usage spikes



- Differing hardware
 - E.g., one datacenter might have A100s, another H100s, another proprietary TPUs
- Differing model architectures and inference methods
- Different datacenters have wildly different efficiency characteristics
 - E.g., one datacenter might have a newer, more efficient cooling system

Environmental effects - inference

What makes inference costs hard to estimate?

		Active AI Accelerators		CPU & RAM	Idle Machines	Overhead
		Chip Power	Utilization			
Luccioni et al., 2022	Measured chip power across multiple A100s	Limit batch size to 1 for reduced latency	Including host CPU & DRAM power	Not included	Not included	
Samsi et al., 2023	Measured chip power across multiple A100s	Highly utilized of bench serving system	Not included	Not included	Not included	
AI EnergyScore	Measured chip power across multiple H100s	Limit batch size to 1 for reduced latency	Not included (suggest +30% addition to GPU energy)	Not included	Not included (suggest PUE of 1.2)	
ML.Energy	Measured chip power across multiple A100s -or- H100s	Variable batch size, influencing utilization	Estimated			
Existing Approach						
Google's Proposed Approach	Measured power supply power across multiple TPU/GPUs	Actual TPU/GPU utilization in production fleet	Including host CPU & DRAM power	Including provisioned machines for latency/uptime	Including actual campus PUE (~1.1)	
Comprehensive Approach						

Environmental effects

Google's claims

Table 2: Energy, emissions, and water usage of the median Gemini Apps text prompt in May 2025 using the existing and proposed approaches described in Section 3.2.

	Existing Approach	Comprehensive Approach
Energy (Wh/prompt)	0.10	0.24
Emissions (gCO ₂ e/prompt)	0.02	0.03
(Scope 2 MB)	(0.016)	(0.023)
(Scope 1+3)	(0.007)	(0.010)
Water (mL/prompt)	0.12	0.26

As of March 2025, Gemini had 35 million daily active users.

Environmental effects

Google's expanding water and electricity usage

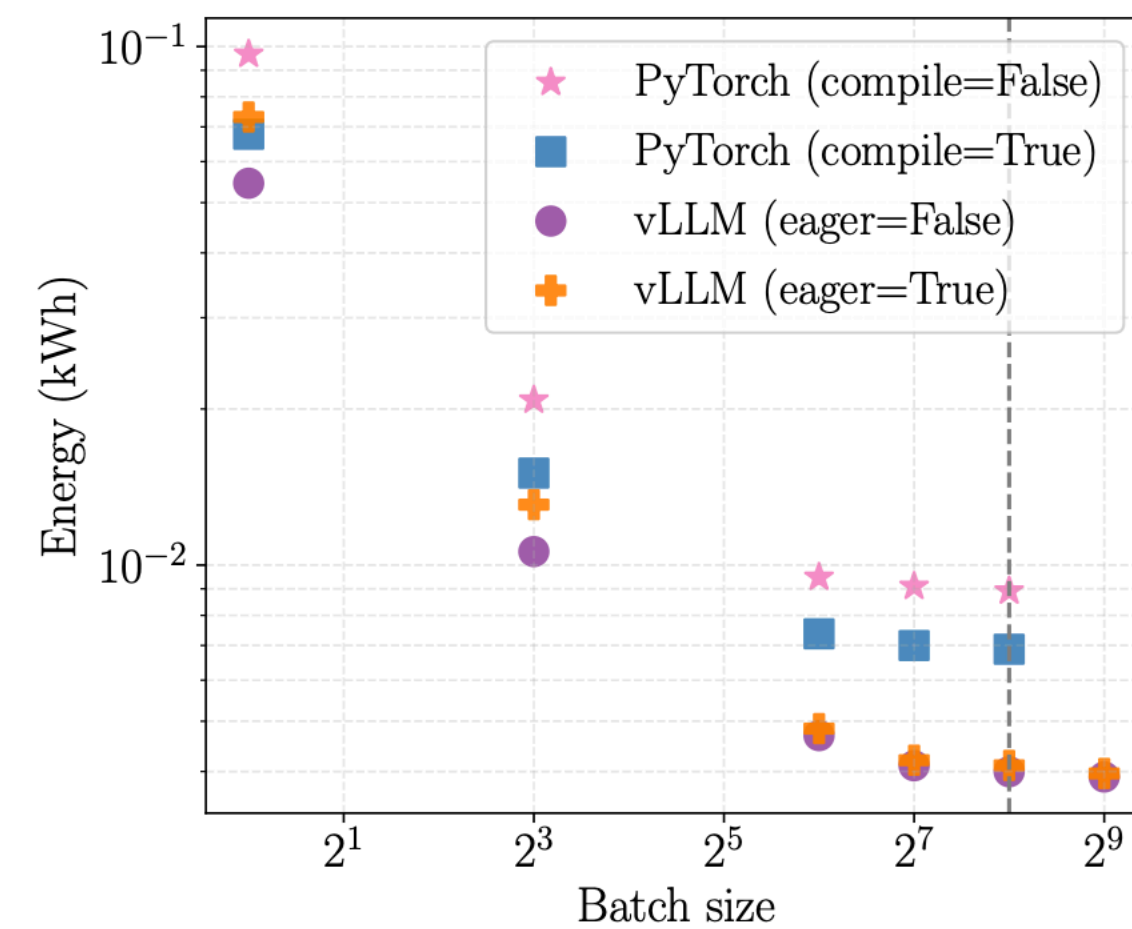
Electricity consumption	Unit	2020	2021	2022	2023	2024	
Data centers	MWh	14,426,600	17,659,000	20,806,200	24,294,900	30,825,600	✓
Office and other facilities	MWh	740,200	628,100	970,000	1,012,100	1,354,300	✓
Total electricity consumption¹⁵⁵	MWh	15,166,800	18,287,100	21,776,200	25,307,000	32,179,900	✓

Water

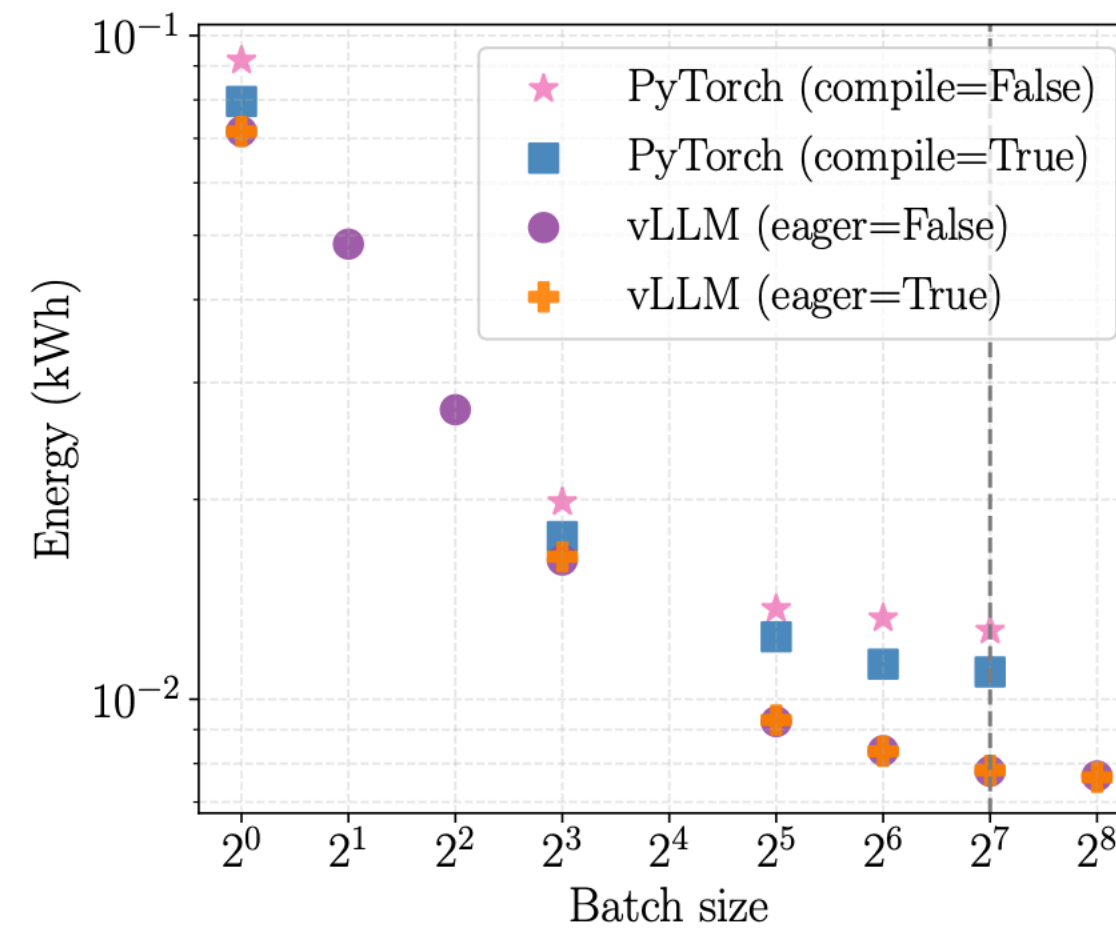
Global operational water use	Unit	2020	2021	2022	2023	2024	
Water withdrawal	Million gallons	5,689	6,297	7,600	8,653	11,011	✓
Water discharge	Million gallons	1,940	1,735	2,035	2,301	2,876	✓
Water consumption	Million gallons	3,749	4,562	5,565	6,352	8,135	✓

Environmental effects

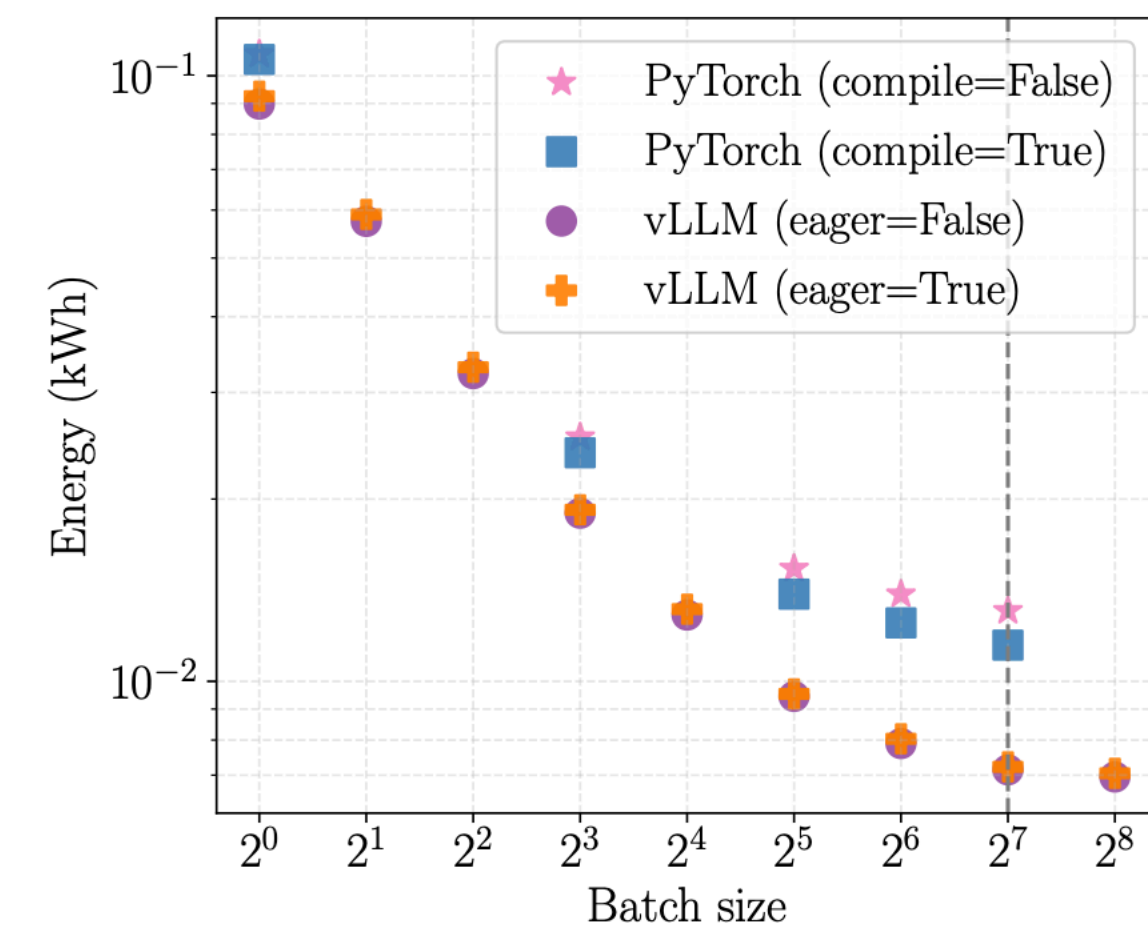
factors that affect energy usage



(a) A100 80GB PCIe



(b) A6000 Ada



(c) A6000

Figure 5: Energy consumption comparison across different GPUs for inference with PyTorch and vLLM backends of 1024 samples for 64 output tokens. For each GPU, we compare PyTorch with and without compilation, and vLLM with and without CUDA Graph serialization. The line in black represents the maximum allowable batch size for PyTorch. Relative savings are most apparent in the low batch size regime and that vLLM due to its optimizations can serve a larger batch size.

Environmental effects

factors that affect energy usage

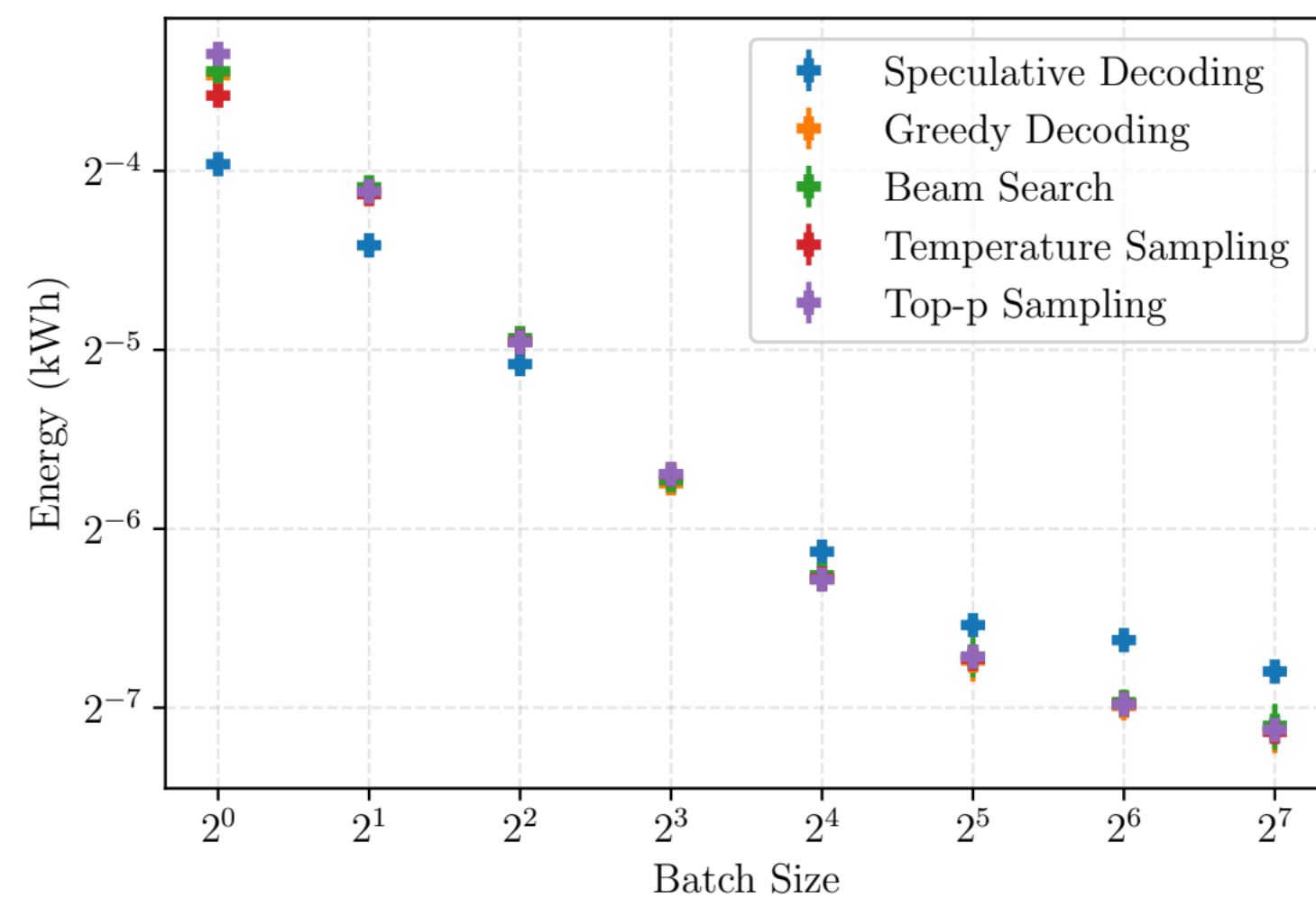


Figure 3: At small batch sizes, speculative decoding provides reduced latency and energy savings. At larger batch size speculative decoding increases energy.

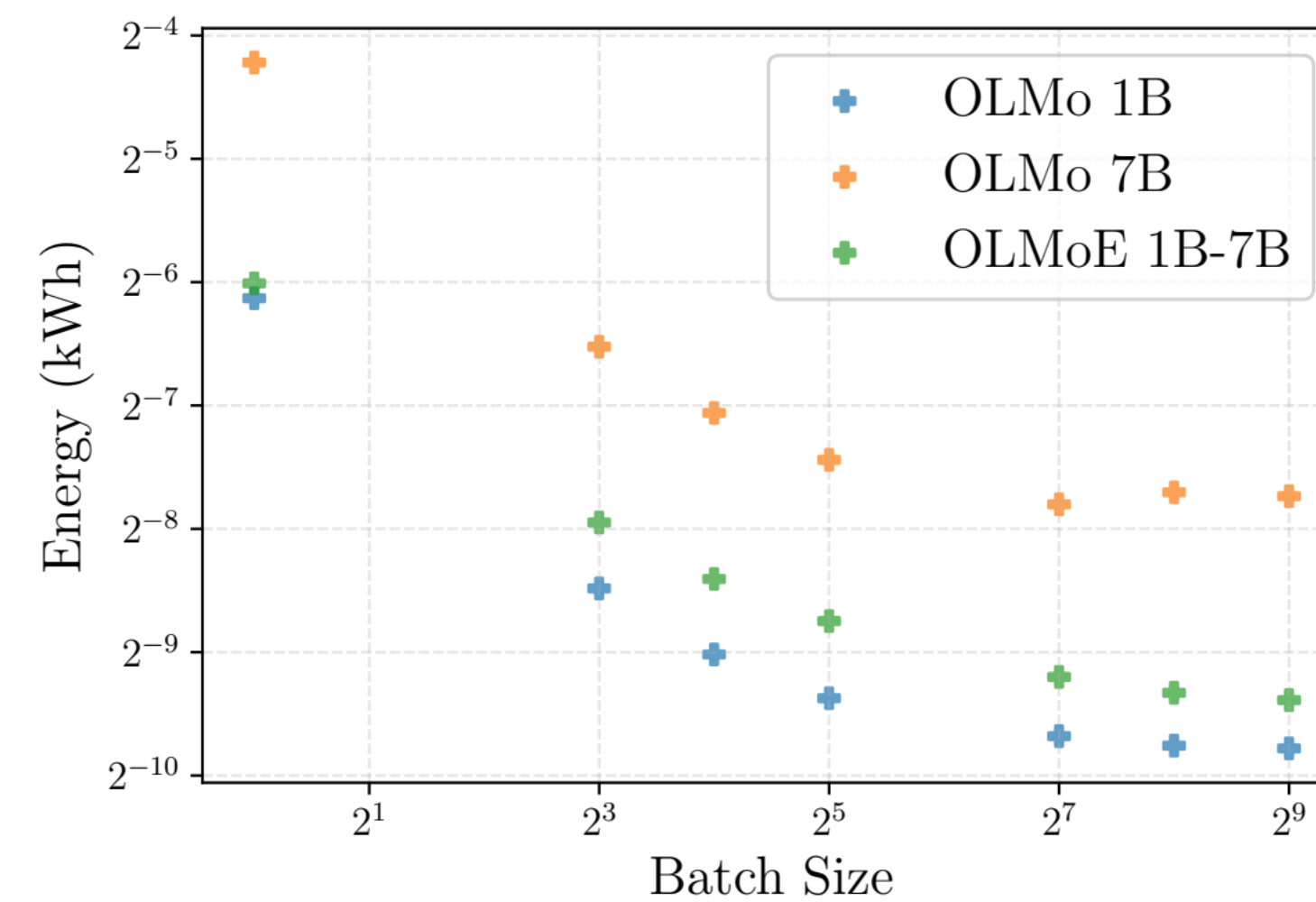


Figure 4: Mixture-of-Experts LLMs require more energy than dense models with comparable active parameters; differences are pronounced at larger batch sizes.

Environmental effects

ChatGPT inference costs as of June 2025

- “average query uses about 0.34 watt-hours and 0.322 mL of water” -Sam Altman

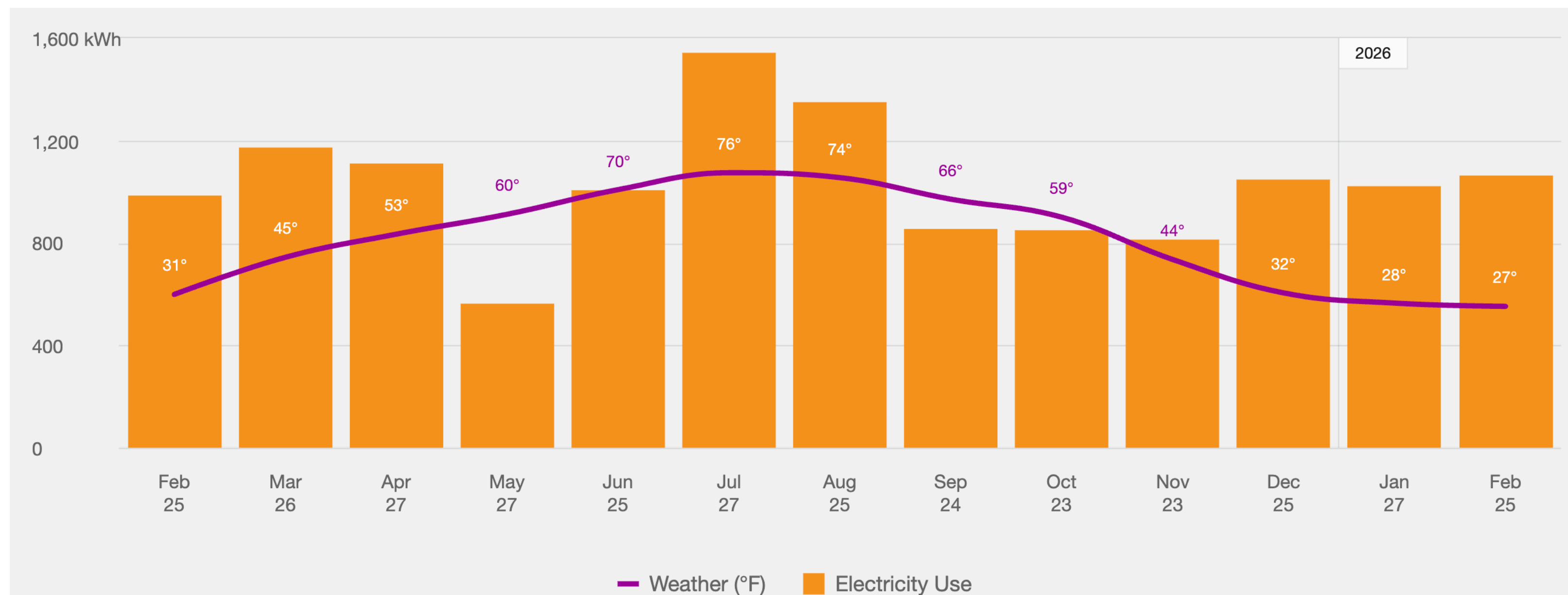
How much electricity would 1 million average ChatGPT queries use?

Environmental effects

ChatGPT inference costs as of June 2025

- “average query uses about 0.34 watt-hours and 0.322 mL of water” -Sam Altman

How much electricity would 1 million average ChatGPT queries use? **340 kwh**



Environmental Effects

what's being done to improve things?

- modeling improvements
 - e.g. mixture-of-experts, PagedAttention
- algorithms for improved machine utilization
- algorithms for more efficient decoding
- data center improvements
 - clean energy procurement
 - building in more efficient locations

What can you
(as a future LLM app developer)
do?

What can you do?

- Homogenization of language and ideas
- Excessive trust in AI outputs
- Effect on education
- Social harms
- Illegal and/or unethical applications
- Environmental effects