Natural Language Processing with Deep Learning

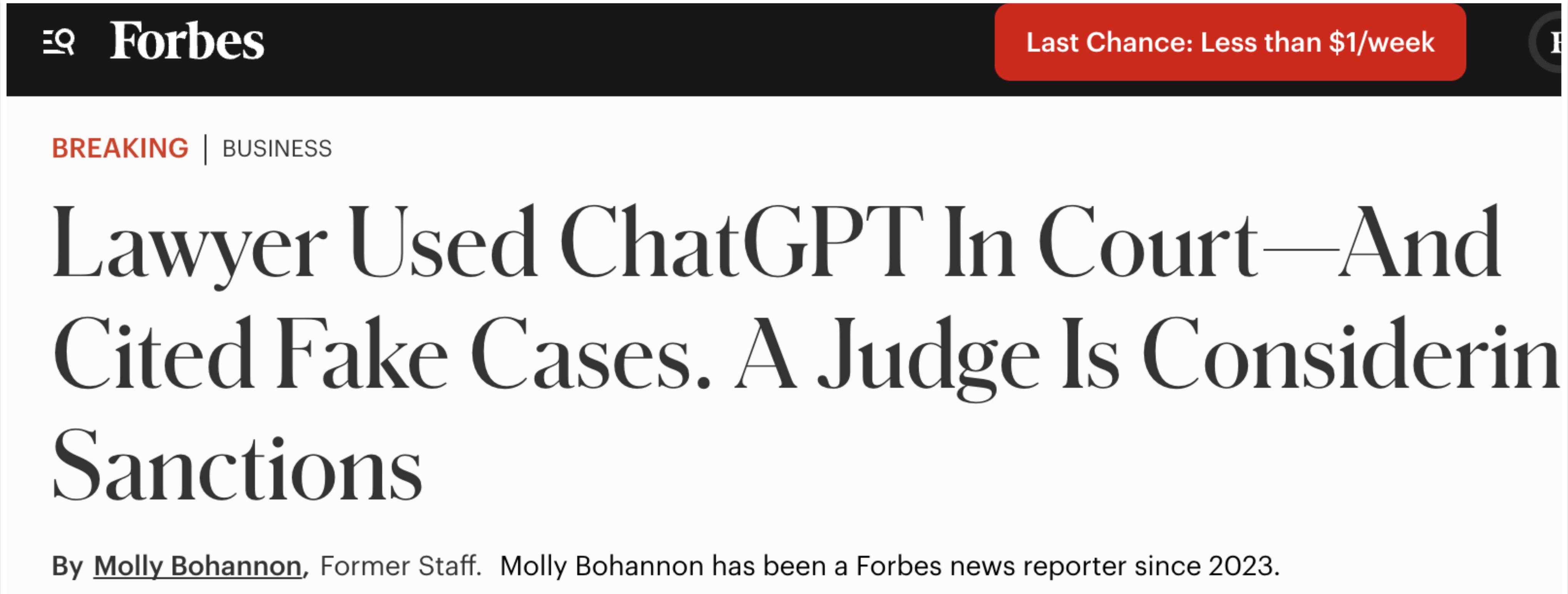
CS224N/Ling284

Yejin Choi

Lecture 16: AI's impact on humanity

Lecture Plan

- Why language models hallucinate
- The paradox of AI-assisted creativity
- AI’s impact on workforce
- The challenges of value alignment

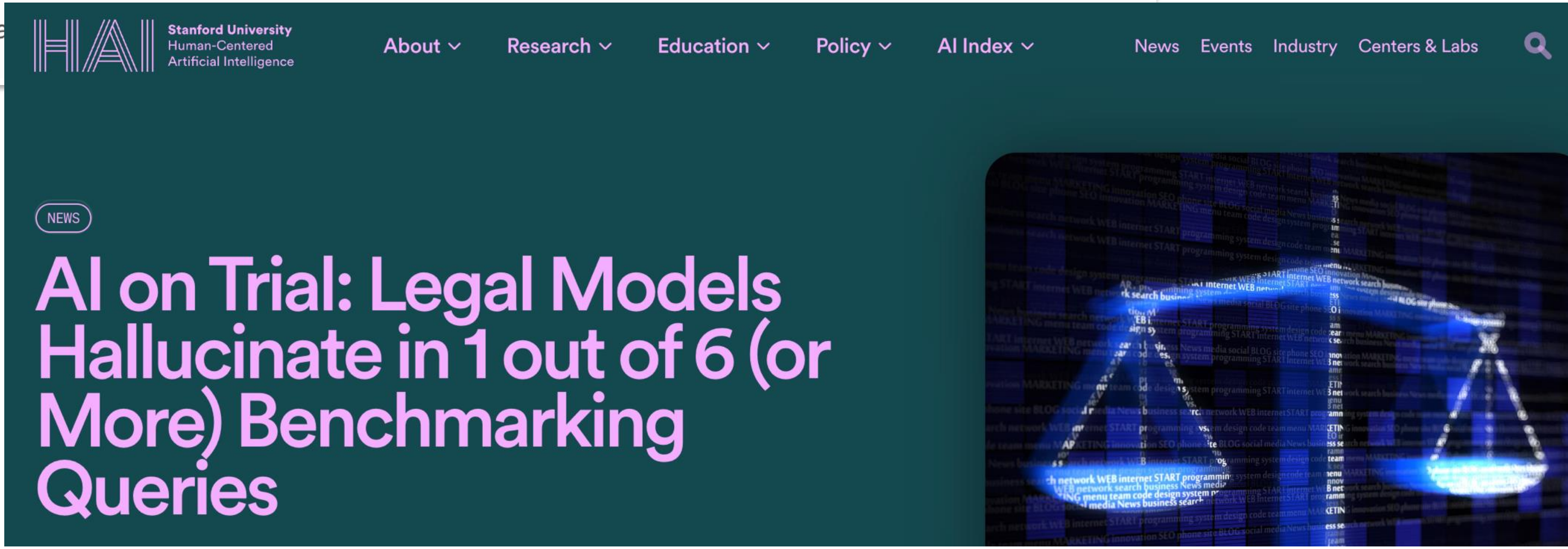


BREAKING | BUSINESS

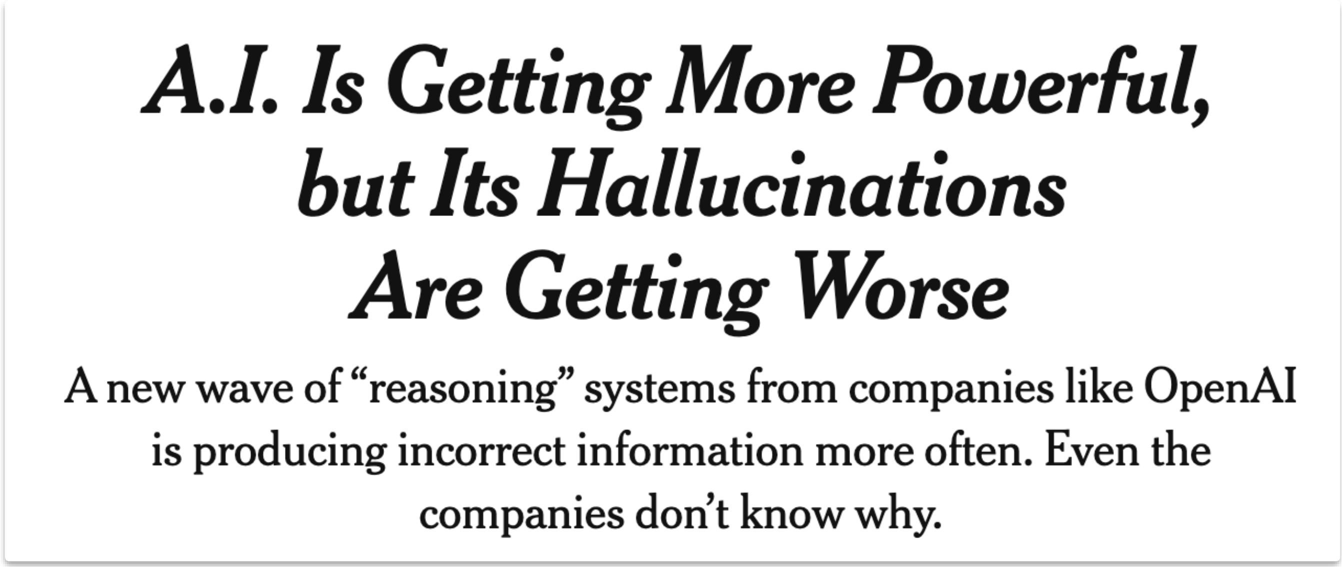
Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considerin Sanctions

By Molly Bohannon, Former Staff. Molly Bohannon has been a Forbes news reporter since 2023.

Published Jun 08, 2023, 02:06pm EDT, Upda

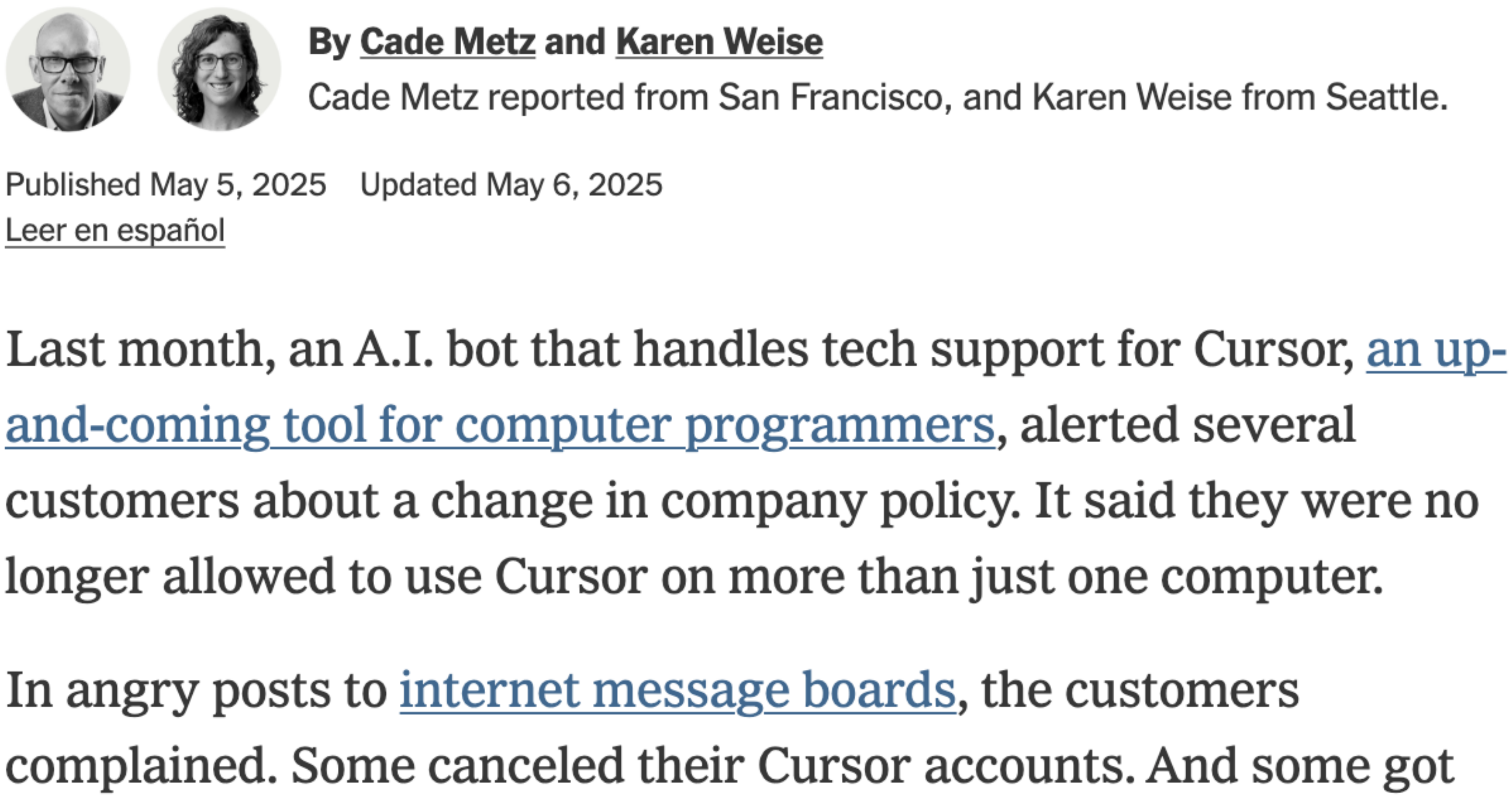


AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries



A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse

A new wave of “reasoning” systems from companies like OpenAI is producing incorrect information more often. Even the companies don’t know why.



By Cade Metz and Karen Weise

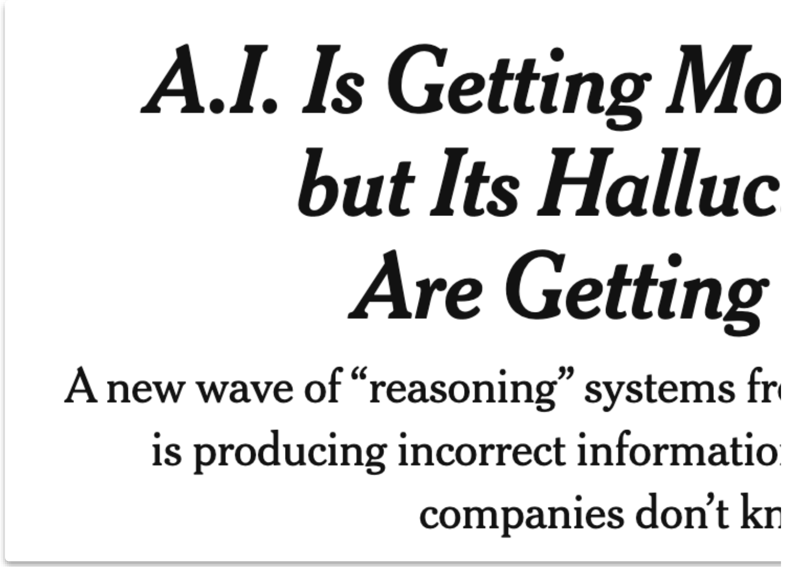
Cade Metz reported from San Francisco, and Karen Weise from Seattle.

Published May 5, 2025 Updated May 6, 2025

Leer en español

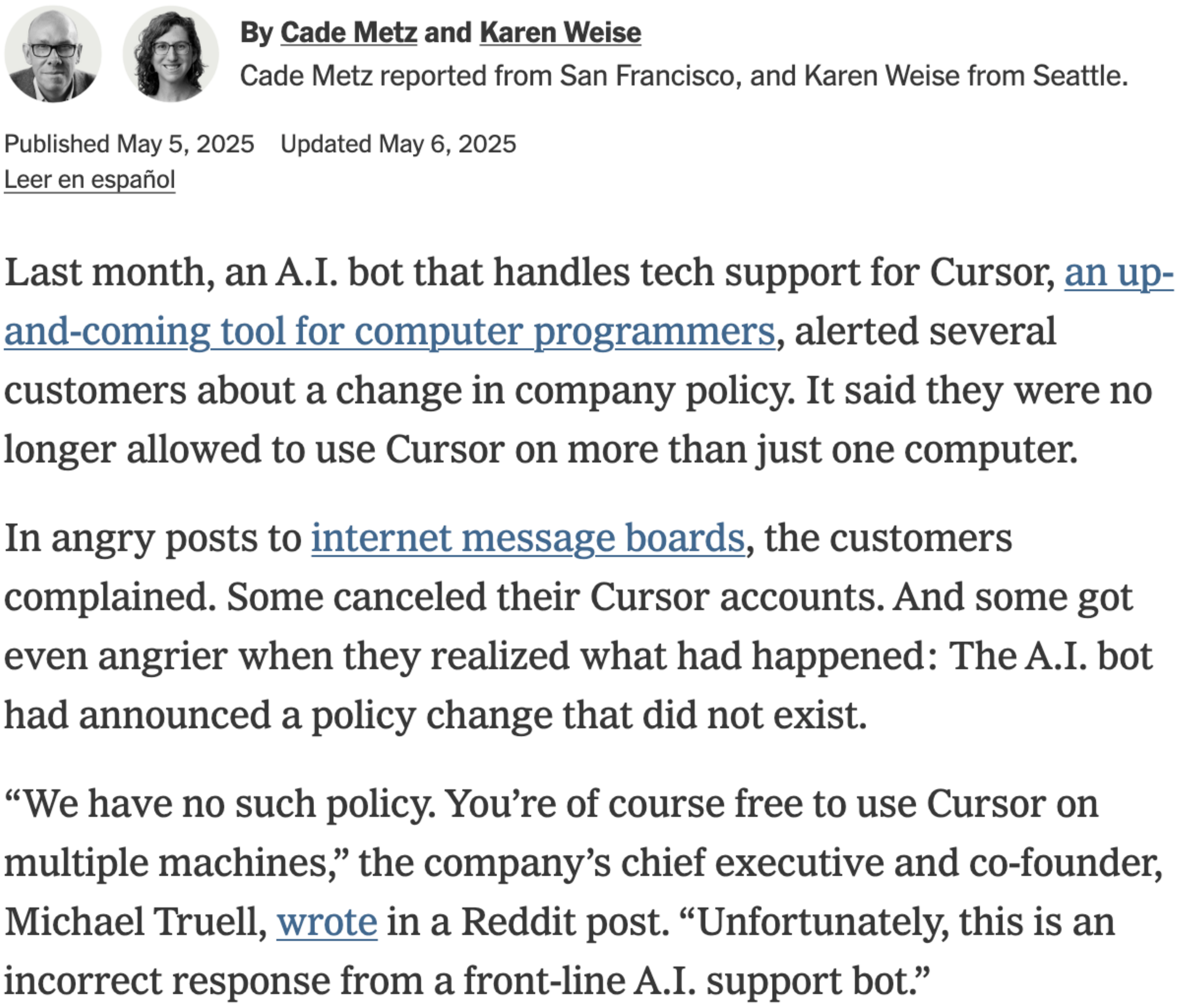
Last month, an A.I. bot that handles tech support for Cursor, an up-and-coming tool for computer programmers, alerted several customers about a change in company policy. It said they were no longer allowed to use Cursor on more than just one computer.

In angry posts to internet message boards, the customers complained. Some canceled their Cursor accounts. And some got



A.I. Is Getting Mo
but Its Halluc
Are Getting

A new wave of “reasoning” systems fr
is producing incorrect informatio
companies don’t kn



By Cade Metz and Karen Weise

Cade Metz reported from San Francisco, and Karen Weise from Seattle.

Published May 5, 2025 Updated May 6, 2025

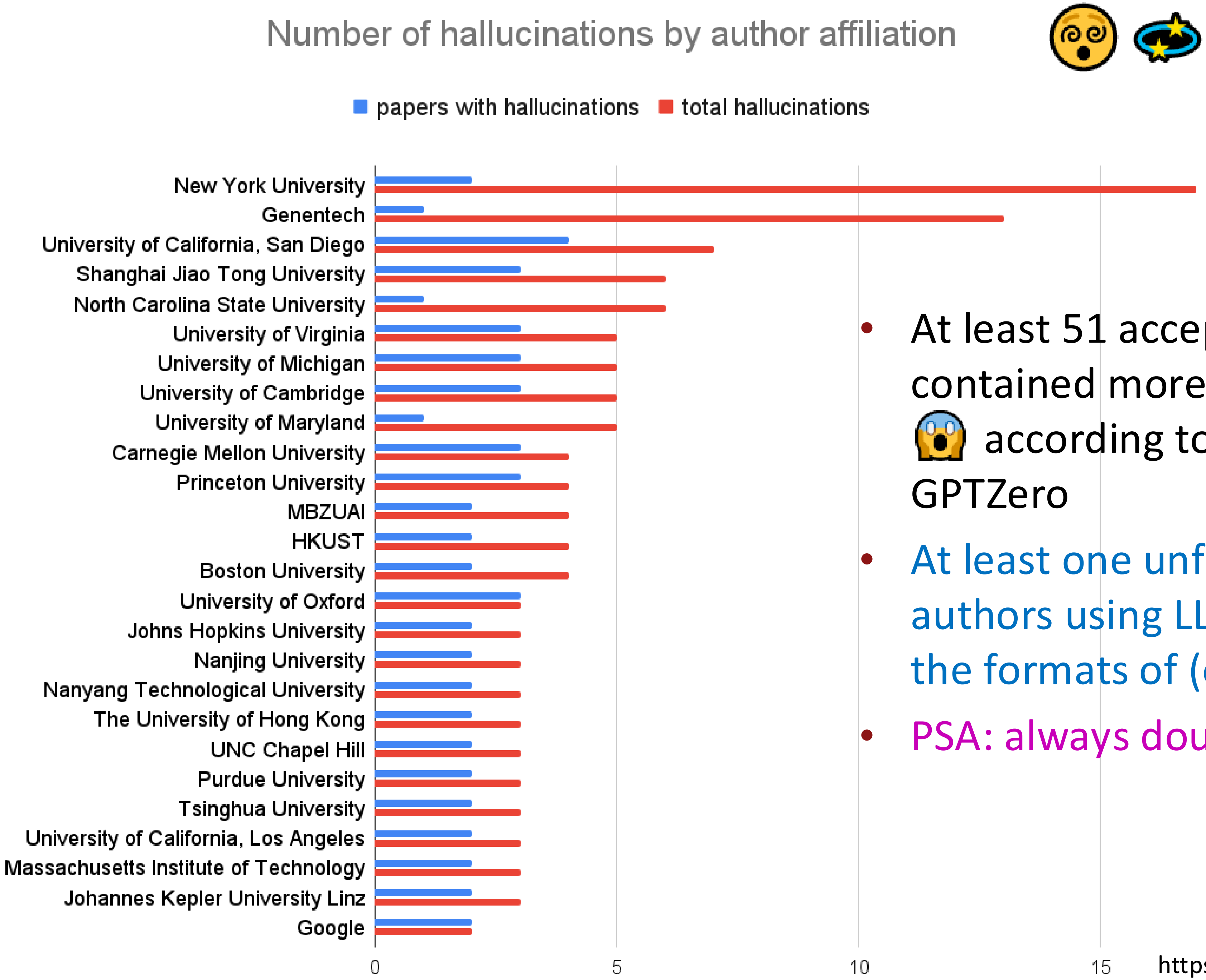
Leer en español

Last month, an A.I. bot that handles tech support for Cursor, an up-and-coming tool for computer programmers, alerted several customers about a change in company policy. It said they were no longer allowed to use Cursor on more than just one computer.

In angry posts to internet message boards, the customers complained. Some canceled their Cursor accounts. And some got even angrier when they realized what had happened: The A.I. bot had announced a policy change that did not exist.

“We have no such policy. You’re of course free to use Cursor on multiple machines,” the company’s chief executive and co-founder, Michael Truell, wrote in a Reddit post. “Unfortunately, this is an incorrect response from a front-line A.I. support bot.”

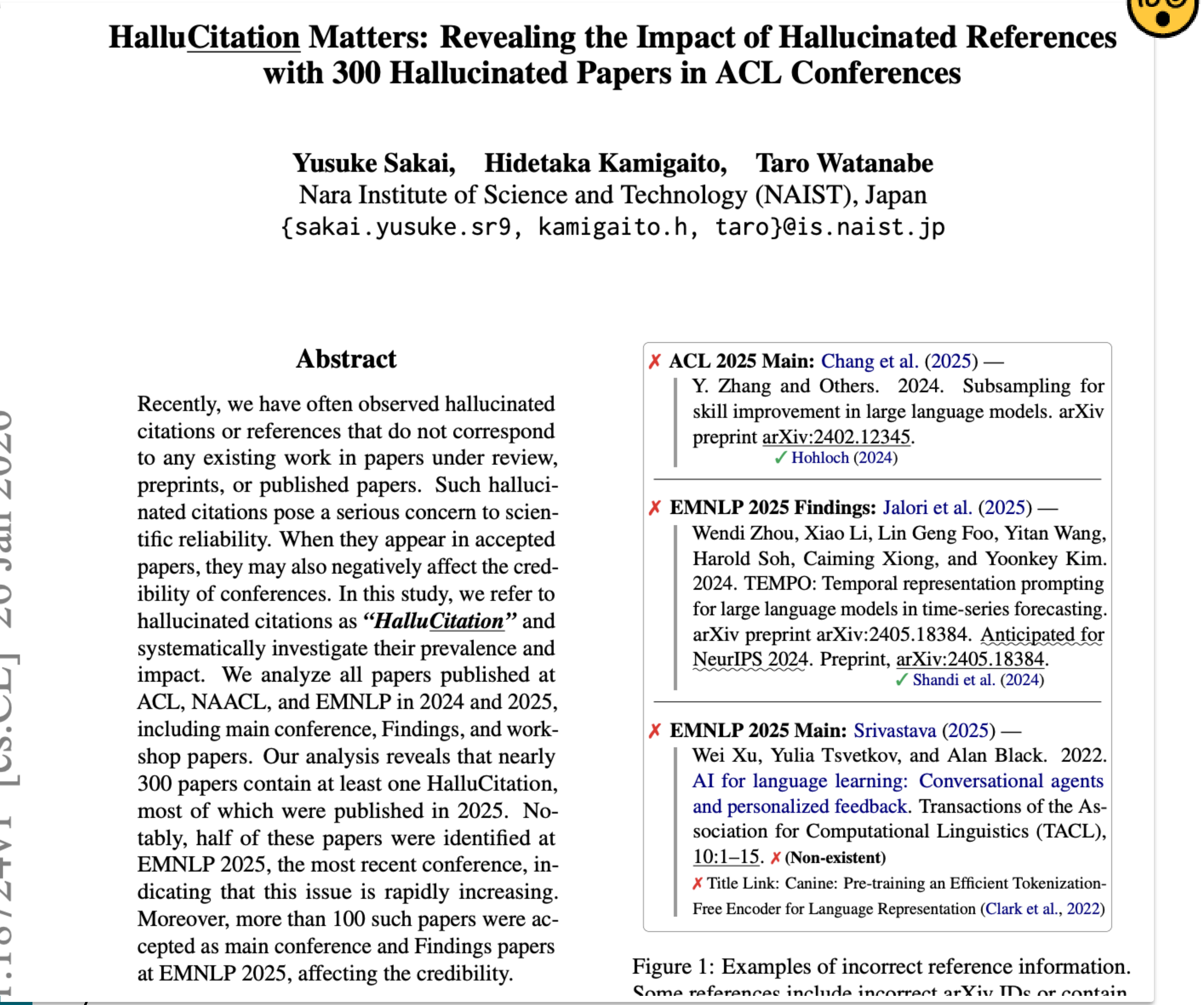
Vibe citing & hallucination 😵‍💫 💫



- At least 51 accepted papers from NeurIPS 2025 contained more than 100 hallucinated citations 😱 according to an investigation done by GPTZero
- At least one unfortunate case resulted from the authors using LLMs to automatically clean up the formats of (originally) legit citations 😰
- PSA: always double-check on LLMs’ work!

https://gptzero.me/news/neurips/

Vibe citing & hallucination



HalluCitation Matters: Revealing the Impact of Hallucinated References with 300 Hallucinated Papers in ACL Conferences

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

Recently, we have often observed hallucinated citations or references that do not correspond to any existing work in papers under review, preprints, or published papers. Such hallucinated citations pose a serious concern to scientific reliability. When they appear in accepted papers, they may also negatively affect the credibility of conferences. In this study, we refer to hallucinated citations as ***"HalluCitation"*** and systematically investigate their prevalence and impact. We analyze all papers published at ACL, NAACL, and EMNLP in 2024 and 2025, including main conference, Findings, and workshop papers. Our analysis reveals that nearly 300 papers contain at least one HalluCitation, most of which were published in 2025. Notably, half of these papers were identified at EMNLP 2025, the most recent conference, indicating that this issue is rapidly increasing. Moreover, more than 100 such papers were accepted as main conference and Findings papers at EMNLP 2025, affecting the credibility.

✗ **ACL 2025 Main:** Chang et al. (2025) —
Y. Zhang and Others. 2024. Subsampling for skill improvement in large language models. arXiv preprint arXiv:2402.12345.
✓ Hohloch (2024)

✗ **EMNLP 2025 Findings:** Jalori et al. (2025) —
Wendi Zhou, Xiao Li, Lin Geng Foo, Yitan Wang, Harold Soh, Caiming Xiong, and Yoonkey Kim. 2024. TEMPO: Temporal representation prompting for large language models in time-series forecasting. arXiv preprint arXiv:2405.18384. Anticipated for NeurIPS 2024. Preprint, arXiv:2405.18384.
✓ Shandi et al. (2024)

✗ **EMNLP 2025 Main:** Srivastava (2025) —
Wei Xu, Yulia Tsvetkov, and Alan Black. 2022. AI for language learning: Conversational agents and personalized feedback. Transactions of the Association for Computational Linguistics (TACL), 10:1–15. ✗ **(Non-existent)**
✗ Title Link: Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation (Clark et al., 2022)

Figure 1: Examples of incorrect reference information. Some references include incorrect arXiv IDs or contain

- ACL conferences are not any better… nearly 300 papers contain at least one hallucinated citations in 2025 alone!

In fact, better reasoning != less hallucination!

- Stronger reasoning models can hallucinate even more!
- E.g., according to the OpenAI o3 / o4-mini’s System Card from OpenAI (Apr 2025) ---
- O3’s accuracy is higher, but O3’s hallucination is also higher!!!

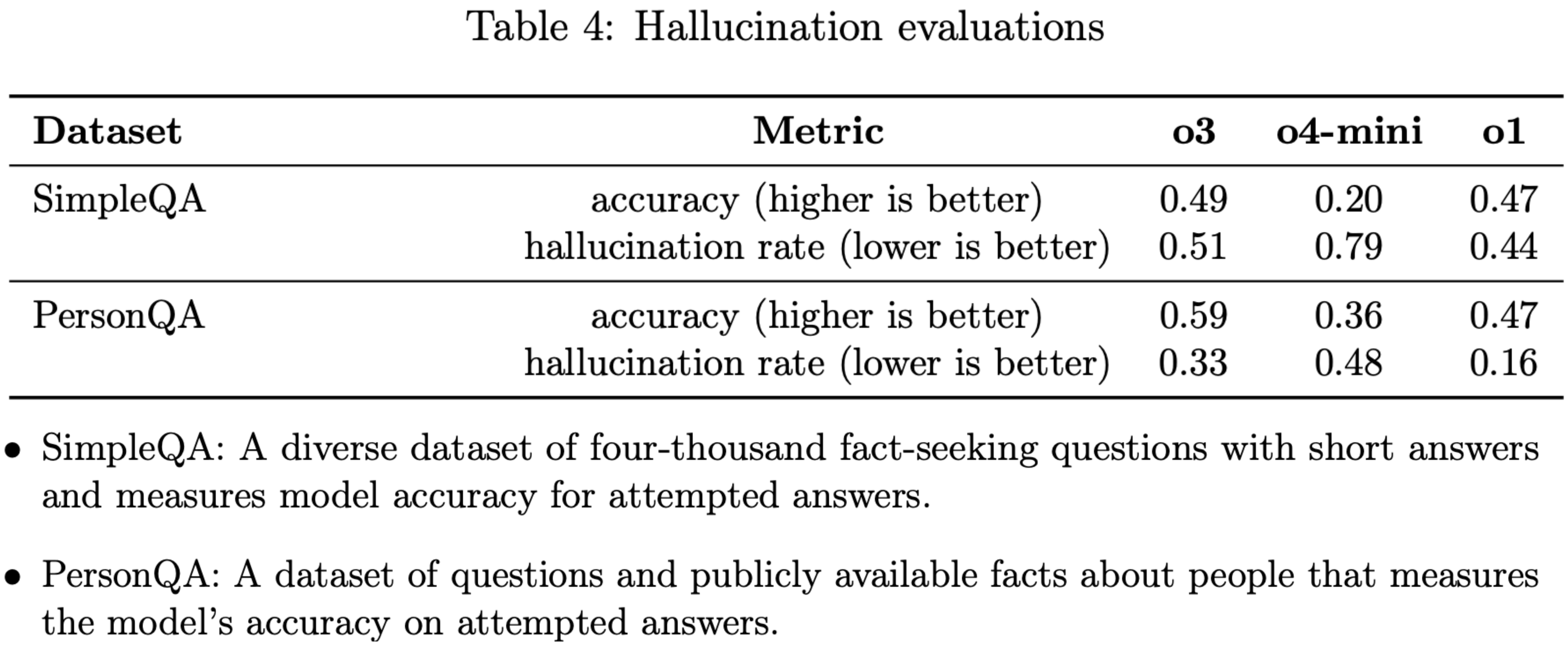
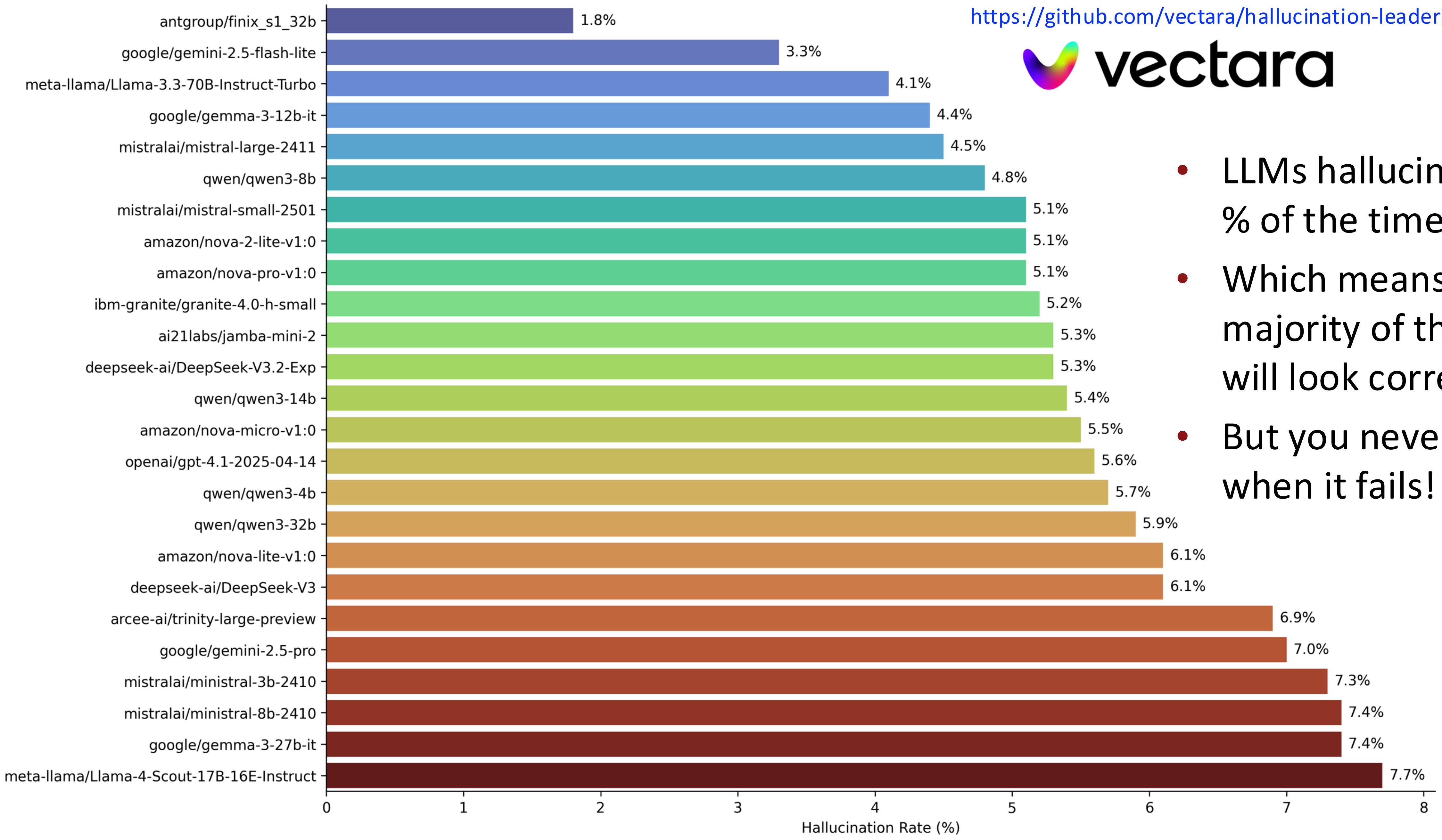


Table 4: Hallucination evaluations

Dataset	Metric	o3	o4-mini	o1
SimpleQA	accuracy (higher is better)	0.49	0.20	0.47
	hallucination rate (lower is better)	0.51	0.79	0.44
PersonQA	accuracy (higher is better)	0.59	0.36	0.47
	hallucination rate (lower is better)	0.33	0.48	0.16

- SimpleQA: A diverse dataset of four-thousand fact-seeking questions with short answers and measures model accuracy for attempted answers.
- PersonQA: A dataset of questions and publicly available facts about people that measures the model’s accuracy on attempted answers.

Grounded Hallucination Rates for Top 25 LLMs



https://github.com/vectara/hallucination-leaderboard

vectara

- LLMs hallucinate a few % of the time.
- Which means the vast majority of the time, it will look correct!
- But you never know when it fails!

“Language Models (Mostly) Know What They Know” (Kadavath et al., 2022)

Calibration

A model is *calibrated* if, for all confidence levels *p*, among claims where the model assigns probability *p* to being correct, the fraction that are actually correct is approximately *p*. Perfect calibration means the model’s confidence is a reliable signal of accuracy.

Findings:

- Base LLMs (of Anthropic) are remarkably well-calibrated on multiple-choice and true/false benchmarks (of BigBench)
- However, not well calibrated if the answer choices include “none of the above” (!!)
- RLHF’ed LLMs are mis-calibrated, as RL collapses the model behavior.
- Any guess on what could be a simple remedy for this?
  - In their study, using high temperature for decoding (t=2.5) can bring back calibration

"Language Models (Mostly) Know What They Know" (Kadavath et al., 2022)

Self-evaluation 1: P(Is my answer true?)
-- after answering the question
Self-evaluation 2: P(I know the answer)
-- without answering the question yet
-- mega-cognitive judgement "this is in my wheelhouse"!

Findings:

- With prompt engineering tricks or supervised fine-tuning, both self-evaluation types can be improved and the results look promising
- However, P(I know the answer) isn't reliable to generalize to unfamiliar tasks
- Also, "*knowing you don't know*" != "*acting responsibly based on the awareness of ignorance*" --- RLHF'ed model can still hallucinate confidently if HF rewarded confident answers

How RLHF can undermine honesty (or promote hallucination)

Sycophancy (“Towards Understanding Sycophancy in Language Models” Sharma et al., 2024)

- The tendency to tell the user what they want to hear rather than what is true
- Models would wrongly admit to mistakes they hadn’t made, give biased feedback that matched the user’s expressed preferences, and change correct answers to match a user’s incorrect suggestion.

Findings of Sharma et al., 2024:

- General phenomenon across all LLMs of all companies. Why???
- Essentially, due to human bias: human raters in preference data show bias toward preferring responses that validate their views, which encourages LLMs to learn to agree with users to get higher rewards than telling the truth!
- A rare case of “inverse scaling” (larger models are worse than smaller models):
  - larger LMs more likely to be even more sycophantic!

“Calibrated Language Models Must Hallucinate” (Kalai and Vempala 2024)

📚 Previous hypotheses for why LMs hallucinate (from other related work)

- The training data contains **falsehoods** (Lin et al. 2022, Dziri et al., 2022) or **outdated facts** (Vu et al., 2023)
- LLMs generate at the token level, and some prefix generation might be impossible to complete factually (Zhang et al., 2023)

- What if hypothetically, training data consists only of facts that are all up to date?
- If the model is “calibrated”, then it will have to hallucinate!

“Calibrated Language Models Must Hallucinate” (Kalai and Vempala 2024)

Previous hypotheses for why LMs hallucinate (from other related work)

- The training data contains **falsehoods** (Lin et al. 2022, Dziri et al., 2022) or **outdated facts** (Vu et al., 2023)
- LLMs generate at the token level, and some prefix generation might be impossible to complete factually (Zhang et al., 2023)

- What if hypothetically, training data consists only of facts that are all up to date?
- If the model is “calibrated”, then it will have to hallucinate! Why?
- Two types of facts:

Arbitrary facts

- Who-did-What-When-Where-Why
- (you have to observe each instance)

Systematic facts

- 356 < 464567345
- (once you know the rule, you don’t have to observe them all to determine their factuality)

“Calibrated Language Models Must Hallucinate” (Kalai and Vempala 2024)

Calibration

A model is *calibrated* if, for all confidence levels *p*, among claims where the model assigns probability *p* to being correct, the fraction that are actually correct is approximately *p*. Perfect calibration means the model’s confidence is a reliable signal of accuracy.

- Not all arbitrary facts are reported in the training data
- When previously unseen facts appear, a well calibrated model should be able to assign a nonzero probability (= its confidence level) that the given statement is correct
- Which means the model needs to know how to ”reserve” probability mass to previously unseen events!

Arbitrary facts

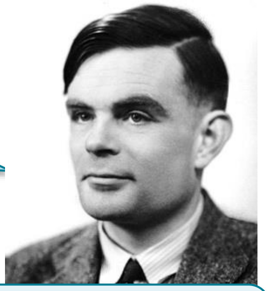
- Who-did-What-When-Where-Why
- (you have to observe each instance)

Good-Turing Estimator!

Story time! Good-Turing Estimator

- Once upon a time, Alan Turing and Irving John Good were working together on breaking German enigma ciphers via statistical cipher analysis. They faced a practical technical problem: how to estimate the frequency of things that they hadn't yet observed.

Obviously, p of {unseen events} should be p of {one-time events}

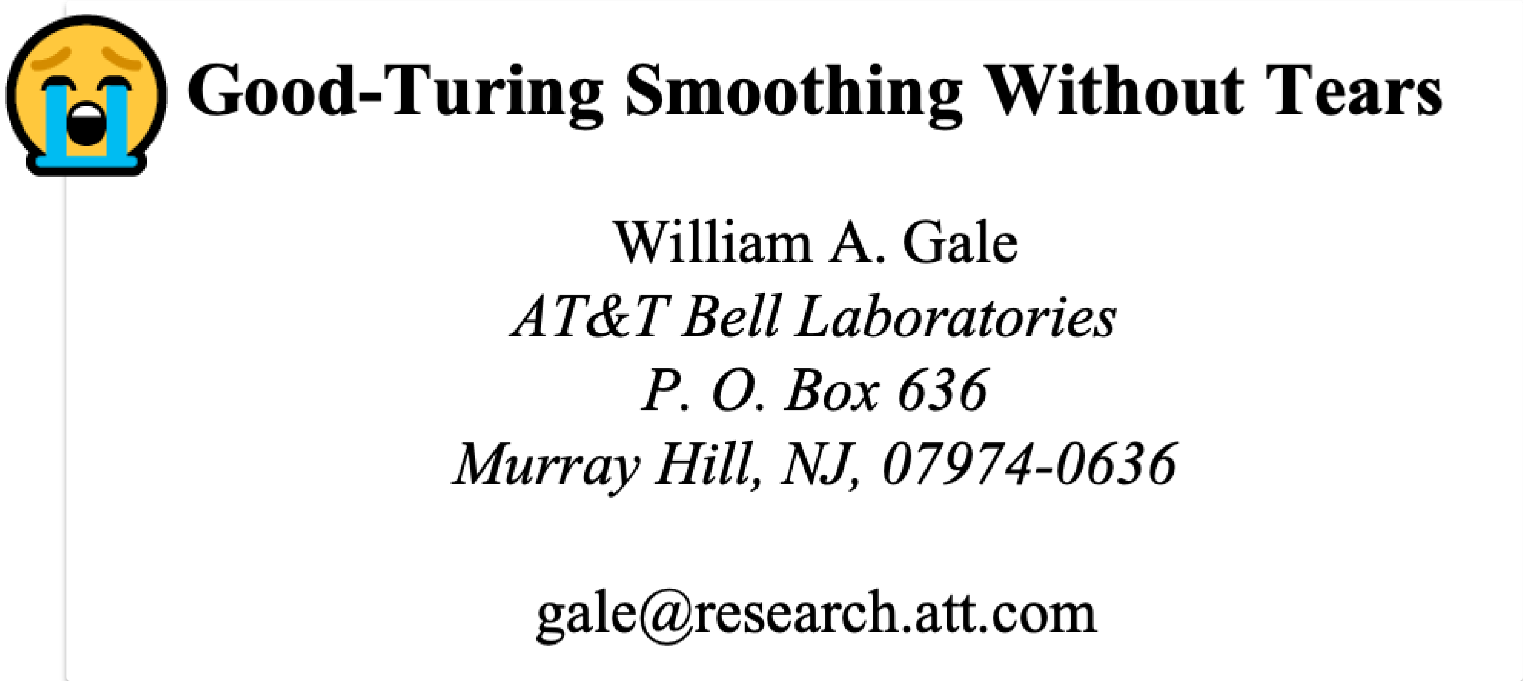


Wait, that's such a beautiful idea! Let me write that up and fill in details…

- After publishing it in 1953, Good subsequently used it to help ecologists to estimate unseen species of butterflies!

Story time! Good-Turing Estimator

- A few decades later, computational linguistics researchers encountered a similar situation with language models: how to estimate probabilities of previously unseen words and sequences of words?
- Many “smoothing” techniques were invented, and folks also discovered Good’s 1953 publication, except this one was notoriously complicated to nail down…



Good-Turing Smoothing Without Tears

William A. Gale
AT&T Bell Laboratories
P. O. Box 636
Murray Hill, NJ, 07974-0636

gale@research.att.com

Good-Turing Estimator

- Intuition:
    - probability of {previously unseen future events}
    - = probability of {one-time events in the training data}

- Reduce the probability of seen events in order to “reserve” some probability mass for the unseen future events. The probability mass of less frequent events are relatively more reduced.
- More formally, let N_r denote the number of items (e.g., species, words, facts) that appear exactly r times in a sample of size.

$$N = \sum_{r=1}^{\infty} r\, N_r \qquad \text{(sample size)}$$

$$r^* = (r+1)\,\frac{N_{r+1}}{N_r} \qquad \text{(adjusted count)}$$

$$p_r = \frac{r^*}{N} \qquad \text{(smoothed probability)}$$

$$p_0 = \frac{N_1}{N} \qquad \text{(missing mass)}$$

“Calibrated Language Models Must Hallucinate” (Kalai and Vempala 2024)

Calibration

A model is *calibrated* if, for all confidence levels *p*, among claims where the model assigns probability *p* to being correct, the fraction that are actually correct is approximately *p*. Perfect calibration means the model’s confidence is a reliable signal of accuracy.

- Not all arbitrary facts are reported in the training data
- When previously unseen facts appear, a well calibrated model should assign a nonzero probability (= its confidence level) that the given statement is correct
- Which means the model needs to know how to ”reserve” probability mass to previously unseen events!
- Suppose arbitrary facts such as [x was born in y]. The model learns this pattern as that of a plausible fact, but there are many instantiations of this pattern that are falsehoods. The model can’t know which is which, thus a model that is calibrated must hallucinate!

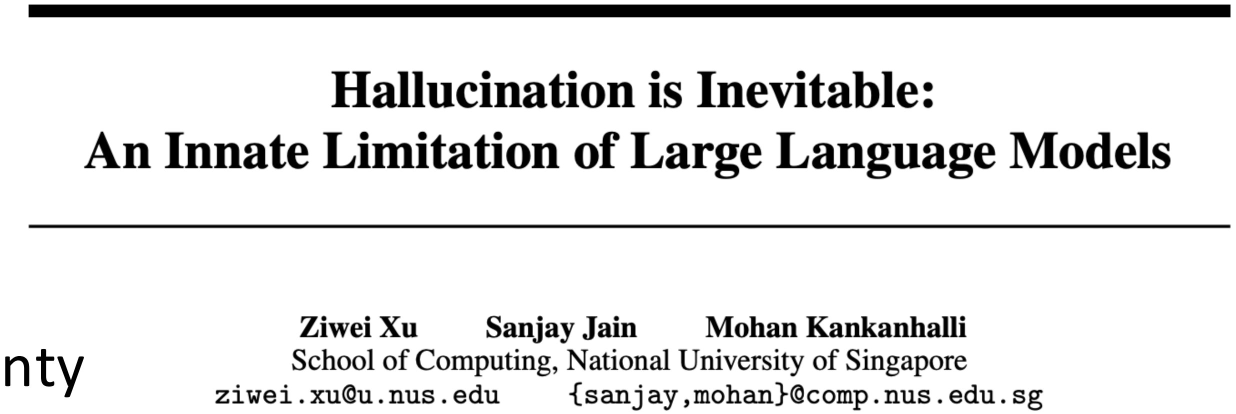
“Why language models hallucinate” (Kalai et al., 2025)

Previous hypotheses for why LMs hallucinate (from other related work)

- The training data contains **falsehoods** (Lin et al. 2022, Dziri et al., 2022) or **outdated facts** (Vu et al., 2023)
- LLMs generate at the token level, and some prefix generation might be impossible to complete factually (Zhang et al., 2023)
- Calibrated language models must hallucinate to cope with unseen events (K&V 2024)

Why hallucination survives post-training (or even gets amplified sometimes)?

- Benchmarks penalize abstention (“I don’t know”) and reward confident answers
- A socio-technical fix proposed: modify benchmarks to reward calibrated uncertainty rather than confident guessing



Hallucination is Inevitable:
An Innate Limitation of Large Language Models

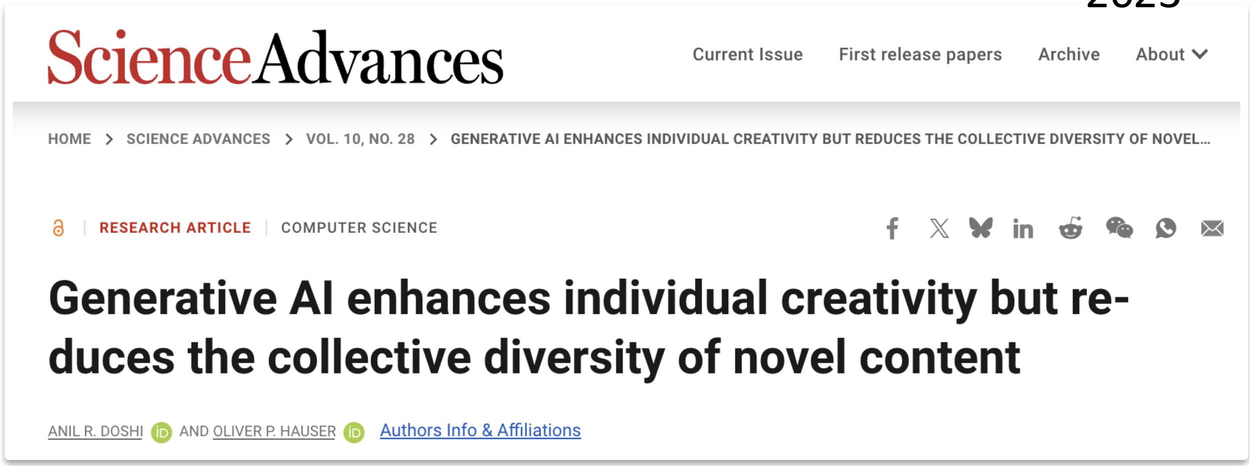
Ziwei Xu **Sanjay Jain** **Mohan Kankanhalli**
School of Computing, National University of Singapore
ziwei.xu@u.nus.edu {sanjay,mohan}@comp.nus.edu.sg

Lecture Plan

- Why language models hallucinate
- The paradox of AI-assisted creativity
- AI’s impact on workforce
- The challenges of value alignment

The paradox of AI-assisted creativity

2023



ScienceAdvances

Current Issue First release papers Archive About

HOME > SCIENCE ADVANCES > VOL. 10, NO. 28 > GENERATIVE AI ENHANCES INDIVIDUAL CREATIVITY BUT REDUCES THE COLLECTIVE DIVERSITY OF NOVEL...

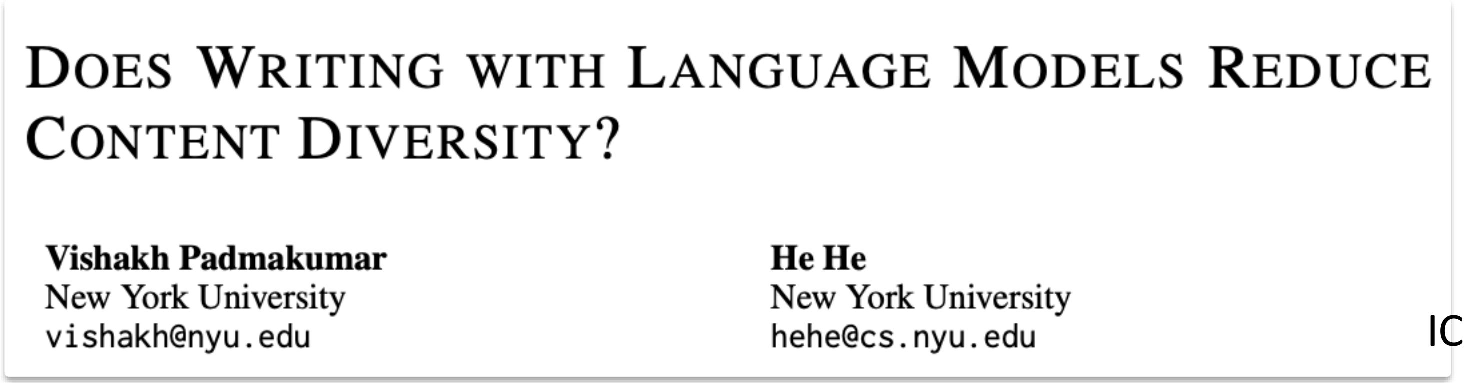
RESEARCH ARTICLE | COMPUTER SCIENCE

Generative AI enhances individual creativity but reduces the collective diversity of novel content

ANIL R. DOSHI AND OLIVER P. HAUSER Authors Info & Affiliations

- What they found: Writers who used AI produced stories rated higher (8–9%) on novelty and usefulness. Less creative writers benefited the most, and AI-assisted stories were rated as more enjoyable, better written, and less boring.
- The catch: When the researchers compared stories *across* writers rather than within individuals, AI-assisted stories were significantly more similar to each other than human-only stories. The distribution of creative output narrowed.
- The authors frame this as a social dilemma: each individual writer benefits, but the collective pool of stories becomes less varied.

The paradox of AI-assisted creativity



DOES WRITING WITH LANGUAGE MODELS REDUCE CONTENT DIVERSITY?

Vishakh Padmakumar
New York University
vishakh@nyu.edu

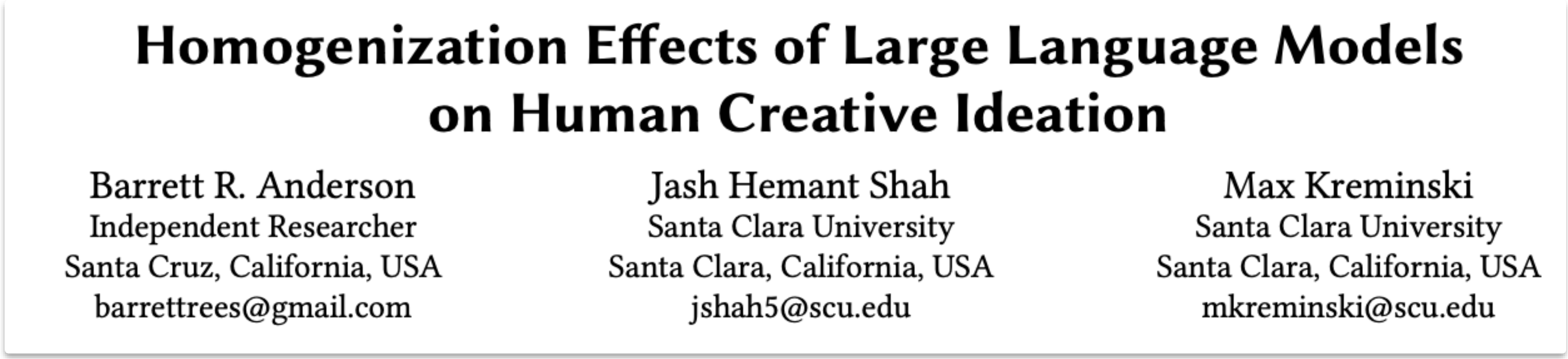
He He
New York University
hehe@cs.nyu.edu

ICLR 2024

- Critical result: Writing with InstructGPT produced a statistically significant reduction in content diversity—essays became more similar across authors, and lexical diversity declined.
- Interestingly, writing with the base GPT-3 model did not produce this effect. The diversity loss was specifically attributable to the RLHF-tuned model providing less diverse suggestions.
- Implication: the very process of aligning models to human preferences (RLHF) may systematically strip out the variance and unpredictability that supports diversity.
  - *"good but not colorful writing"*

The paradox of AI-assisted creativity

Creativity & Cognition (C&C), 2024



Homogenization Effects of Large Language Models on Human Creative Ideation

Barrett R. Anderson
Independent Researcher
Santa Cruz, California, USA
barrettrees@gmail.com

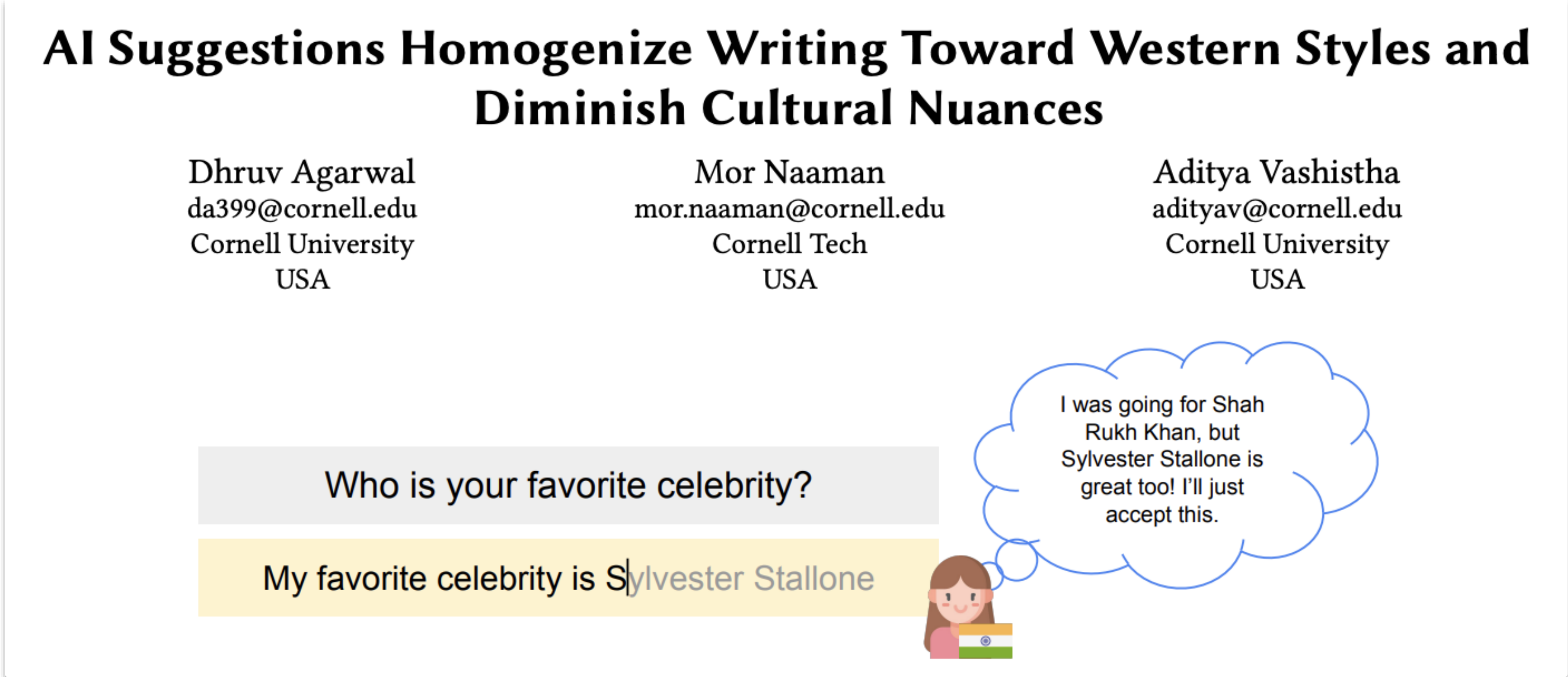
Jash Hemant Shah
Santa Clara University
Santa Clara, California, USA
jshah5@scu.edu

Max Kreminski
Santa Clara University
Santa Clara, California, USA
mkreminski@scu.edu

- This paper is particularly important because it demonstrates that homogenization is not limited to *writing*—it extends to *ideation* itself. When people see AI-generated ideas before brainstorming, they tend to anchor on those ideas, producing output that converges toward the model's mode.
- The authors found, encouragingly, that simply informing users that model output tends to be homogeneous helped them resist the effect to some degree.

The paradox of AI-assisted creativity

2024



AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances

Dhruv Agarwal
da399@cornell.edu
Cornell University
USA

Mor Naaman
mor.naaman@cornell.edu
Cornell Tech
USA

Aditya Vashistha
adityav@cornell.edu
Cornell University
USA

Who is your favorite celebrity?

My favorite celebrity is Sylvester Stallone

I was going for Shah Rukh Khan, but Sylvester Stallone is great too! I'll just accept this.

- This study found that AI autocomplete and writing suggestions push users toward **Western linguistic norms**—more direct, less formal, and less culturally distinctive. Writers from non-Western cultural backgrounds saw the most significant reduction in lexical diversity when using AI tools. The paper argues this represents a subtle but systematic form of cultural homogenization, eroding the very linguistic patterns that encode different ways of seeing the world.

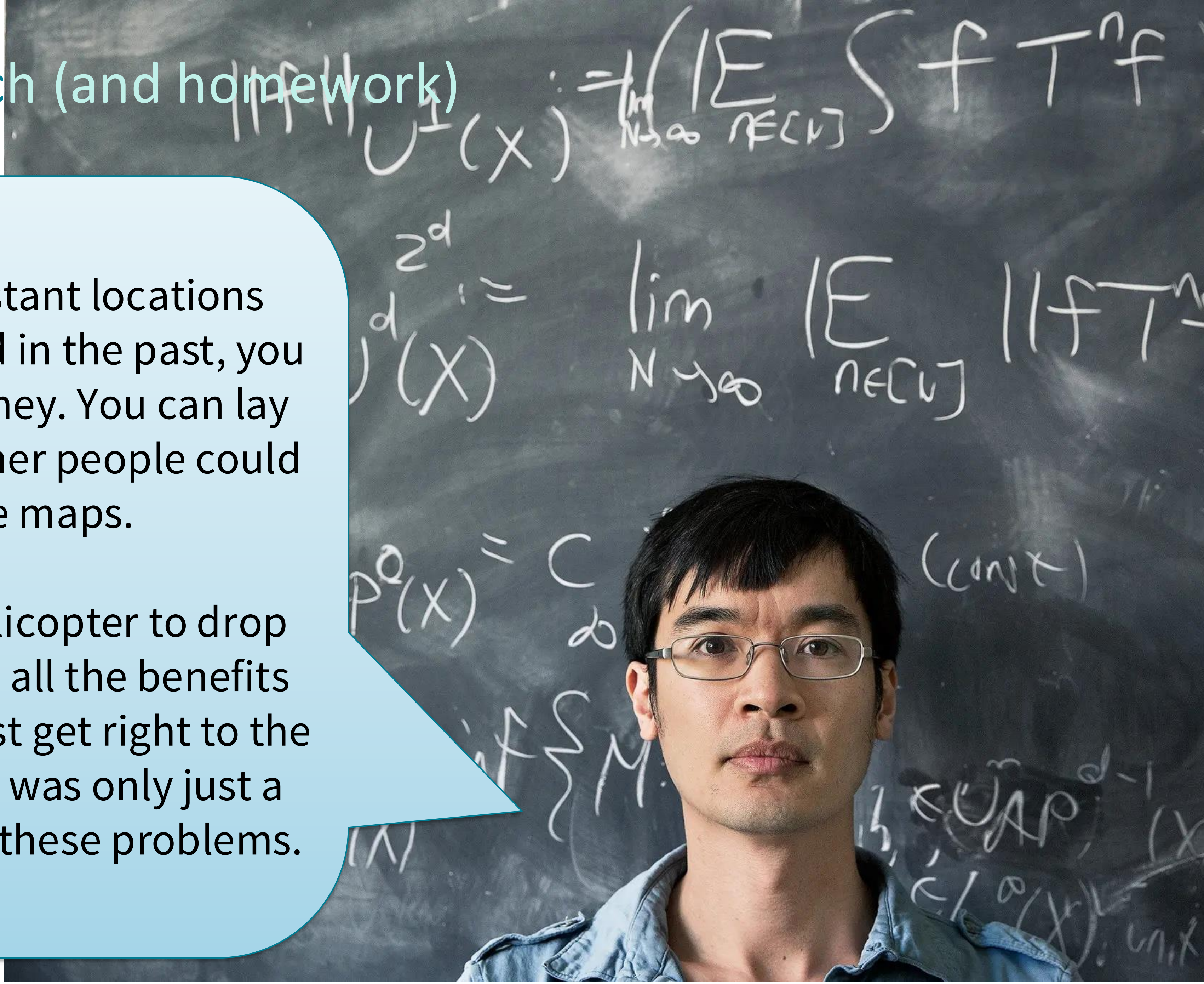
Why does AI assistance reduce collective diversity?

- Mode collapse
  - LLMs in general, and especially after RLHF, lose distributional diversity and pluralism compared to the natural distribution of human text
  - Diversity tax: alignment reduces diversity
- Anchoring effects
  - When human sees an AI-generated suggestion before their own thinking, it acts as a "cognitive anchor". AI's suggestions create a particularly strong version of this effect as they are fluent, confident, and immediately usable.
- Cognitive offloading
  - Gerlich (2025) reports that frequent AI users increasingly offload cognitive tasks to AI. The effortful cognitive struggle that produces originality is precisely what gets bypassed.

On AI assisted research (and homework)

These problems are like distant locations that you would hike to. And in the past, you would have to go on a journey. You can lay down trail markers that other people could follow, and you could make maps.

AI tools are like taking a helicopter to drop you off at the site. You miss all the benefits of the journey itself. You just get right to the destination, which actually was only just a part of the value of solving these problems.

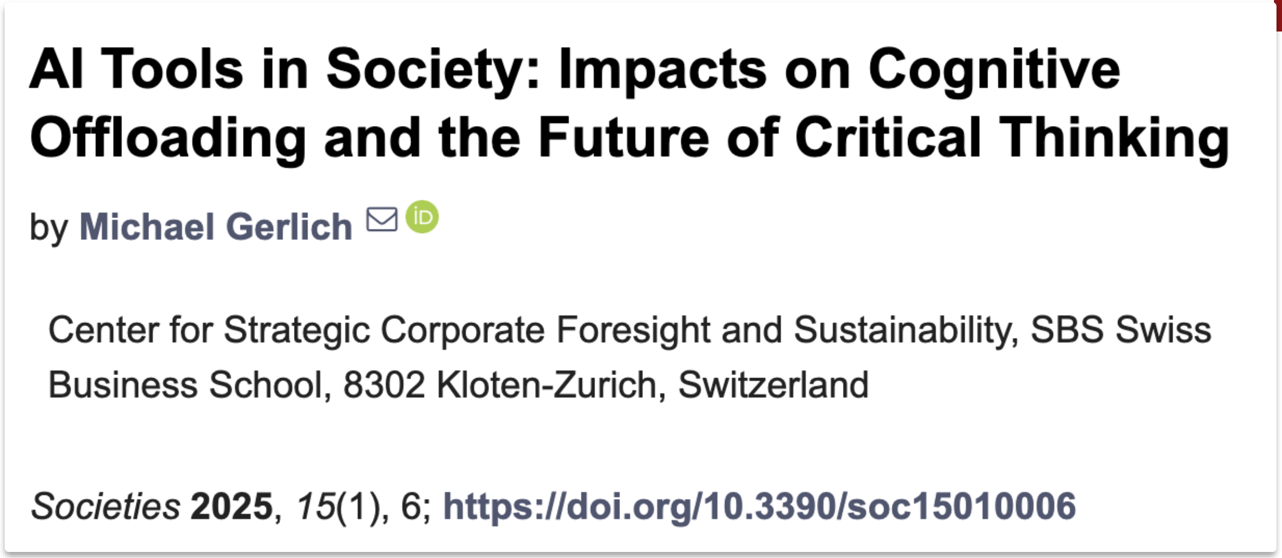


Quote: https://www.theatlantic.com/technology/2026/02/ai-math-terrance-tao/686107 & photo: https://www.nytimes.com/2015/07/26/magazine/the-singular-mind-of-terry-tao.html

Effects on cognition and critical thinking

Findings of Gerlich 2025:

- A significant negative correlation between frequent AI usage and critical thinking abilities, mediated by increased “cognitive offloading”.
- Younger participants exhibited higher dependence on AI tools and lower critical thinking scores compared to older participants.

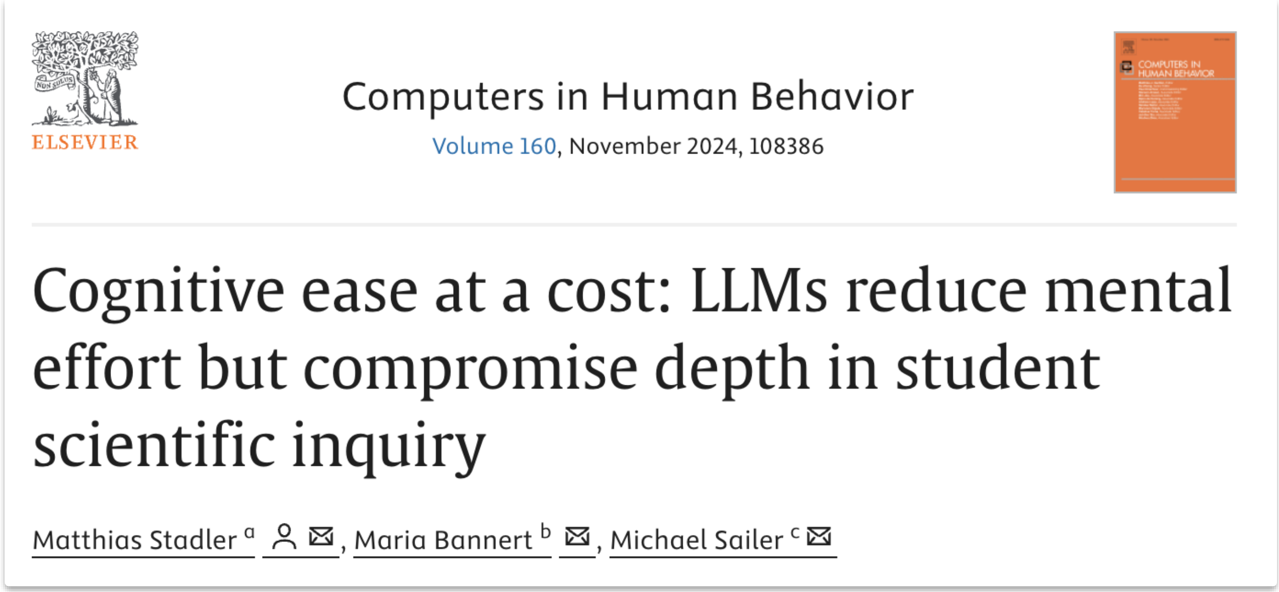


AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking

by **Michael Gerlich**

Center for Strategic Corporate Foresight and Sustainability, SBS Swiss Business School, 8302 Kloten-Zurich, Switzerland

Societies **2025**, *15*(1), 6; **https://doi.org/10.3390/soc15010006**



ELSEVIER

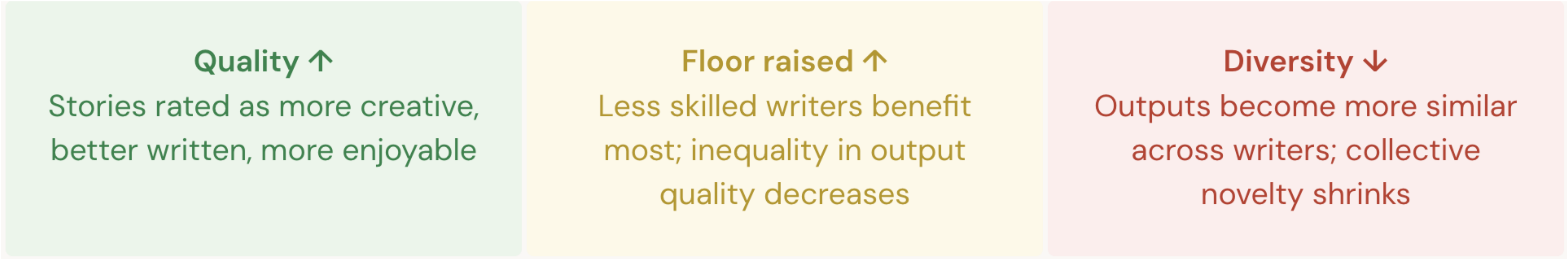
Computers in Human Behavior

Volume 160, November 2024, 108386

Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry

Matthias Stadler [a], Maria Bannert [b], Michael Sailer [c]

The paradox of AI-assisted creativity



Quality ↑
Stories rated as more creative, better written, more enjoyable

Floor raised ↑
Less skilled writers benefit most; inequality in output quality decreases

Diversity ↓
Outputs become more similar across writers; collective novelty shrinks

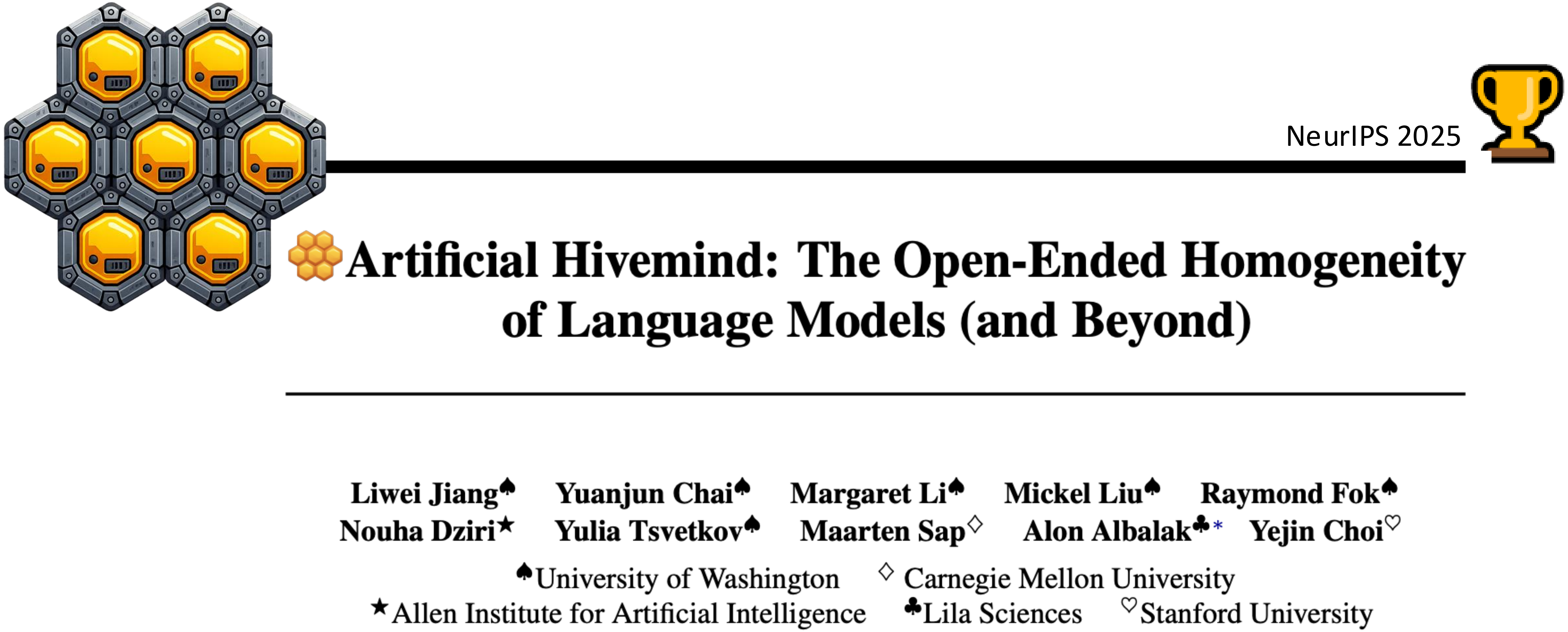
Algorithmic Monoculture

A term coined by Kleinberg & Raghavan (2021) describing the systemic risk that arises when many different decision-makers rely on the same underlying model, causing correlated outputs and correlated failures.

Homogenization

The tendency for outputs produced with AI assistance to become more similar to one another, reducing the overall variety within a corpus of work produced by many different people.

Would it help if we used different LLMs?



NeurIPS 2025 🏆

Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)

Liwei Jiang[♠] Yuanjun Chai[♠] Margaret Li[♠] Mickel Liu[♠] Raymond Fok[♠]
Nouha Dziri[★] Yulia Tsvetkov[♠] Maarten Sap[◇] Alon Albalak[♣]* Yejin Choi[♡]

[♠]University of Washington [◇] Carnegie Mellon University
[★]Allen Institute for Artificial Intelligence [♣]Lila Sciences [♡]Stanford University

How Do LLMs Handle Open-Ended Queries in Infinity-Chat?

Spoiler Alert

LLMs suffer from **mode collapse**, forming "**Artificial Hivemind**".

A "**hivemind**" refers to a collective consciousness in which **many individuals share thoughts**, knowledge, or decision-making as if they were a **single mind**.

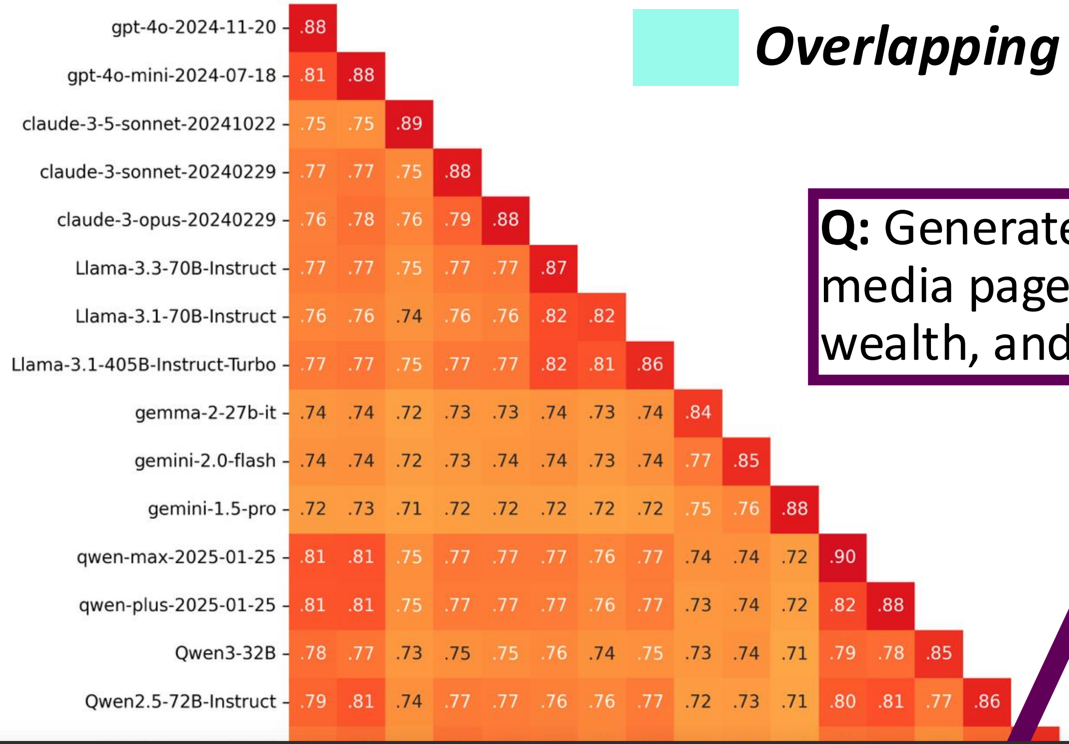
Artificial Hivemind

Intra-Model Homogeneity: a single model consistently generates similar responses

Inter-Model Homogeneity: different models converge to produce similar outputs

Inter-Model Homogeneity

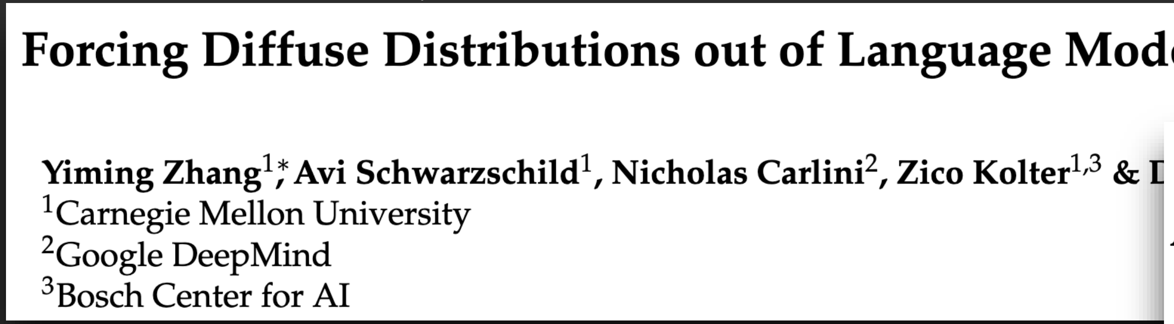
Empower Your Journey: Unlock Success, Build Wealth, Transform Yourself.



Overlapping text fragments

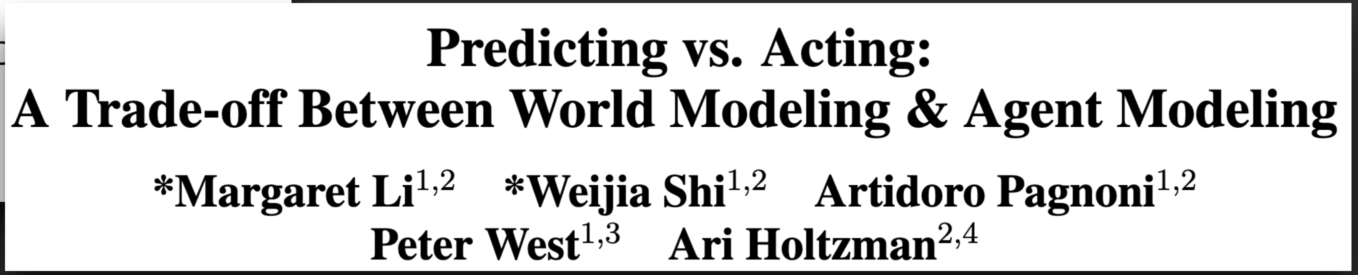
Q: Generate a motto for a social media page focused on success, wealth, and self-help.

Worse with post-trained models



Forcing Diffuse Distributions out of Language Models

Yiming Zhang[1]*, Avi Schwarzschild[1], Nicholas Carlini[2], Zico Kolter[1,3] & D
[1]Carnegie Mellon University
[2]Google DeepMind
[3]Bosch Center for AI



Predicting vs. Acting:
A Trade-off Between World Modeling & Agent Modeling

***Margaret Li[1,2] *Weijia Shi[1,2] Artidoro Pagnoni[1,2]**
Peter West[1,3] Ari Holtzman[2,4]

(Sim = 1.0)

Lecture Plan

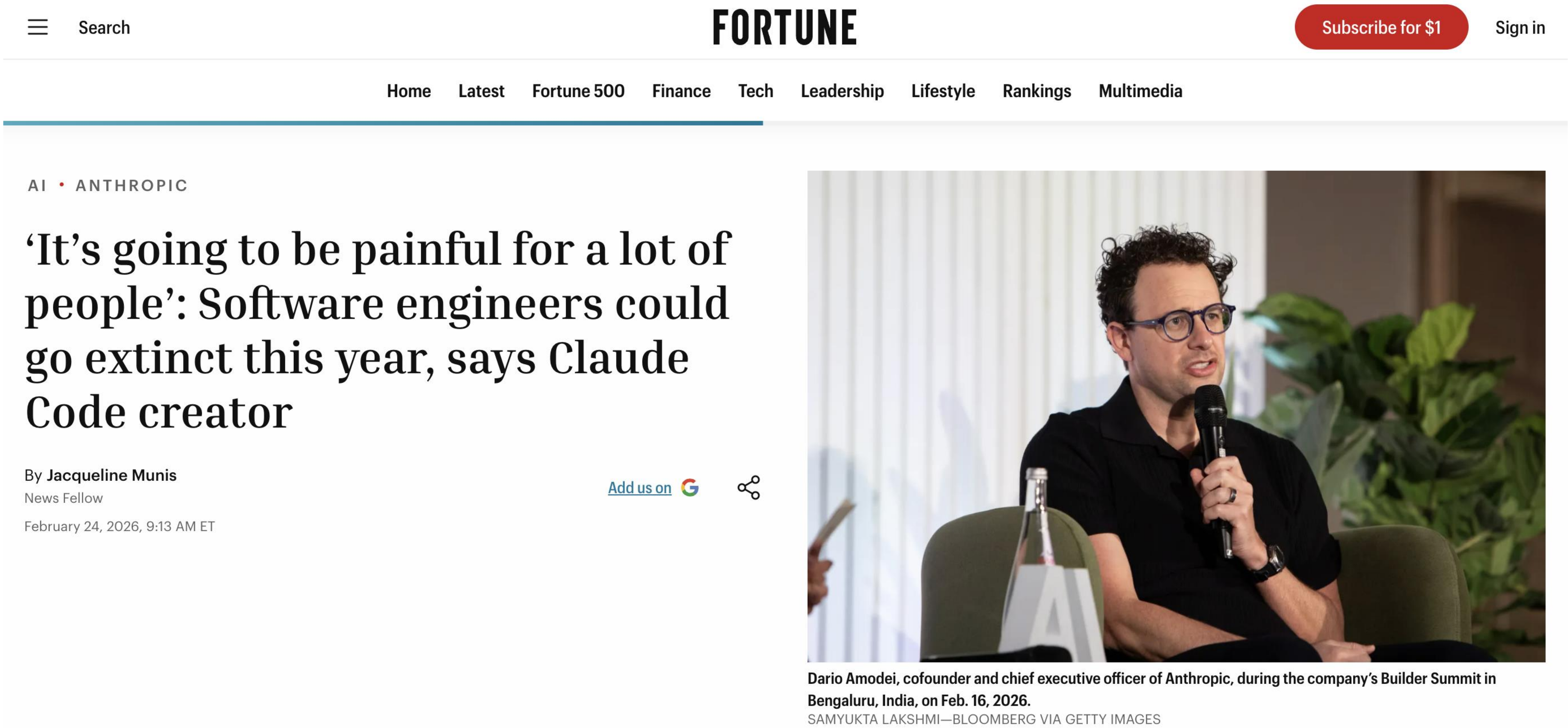
Why language models hallucinate

The paradox of AI-assisted creativity

AI’s impact on workforce

The challenges of value alignment

Some are very concerned...



AI • ANTHROPIC

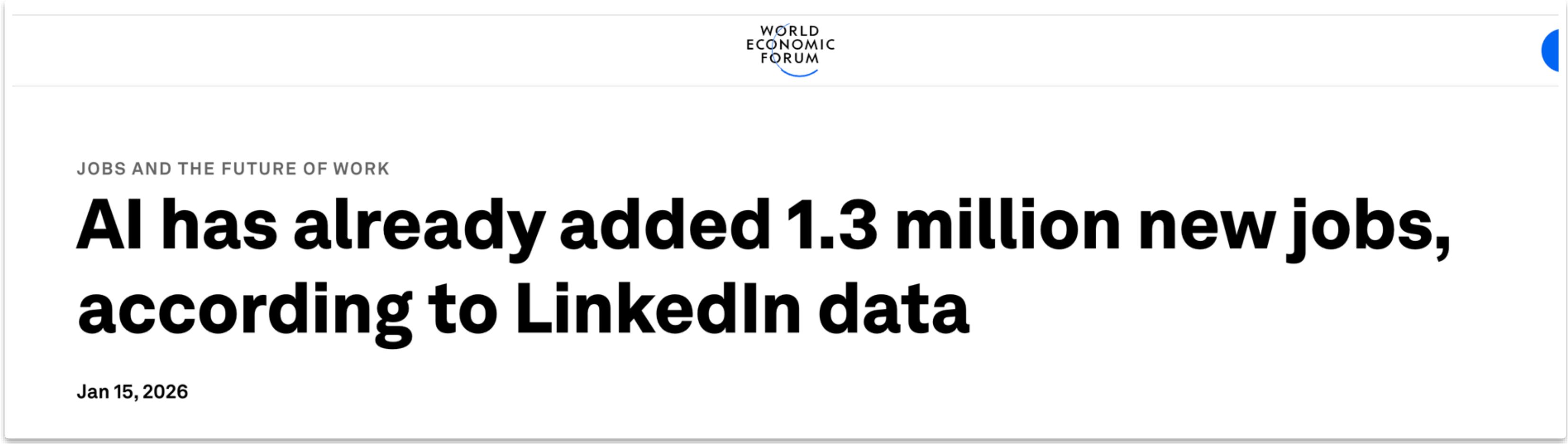
‘It’s going to be painful for a lot of people’: Software engineers could go extinct this year, says Claude Code creator

By **Jacqueline Munis**
News Fellow

February 24, 2026, 9:13 AM ET

Dario Amodei, cofounder and chief executive officer of Anthropic, during the company’s Builder Summit in Bengaluru, India, on Feb. 16, 2026.
SAMYUKTA LAKSHMI—BLOOMBERG VIA GETTY IMAGES

Recent statistics look net positive (thus far)…

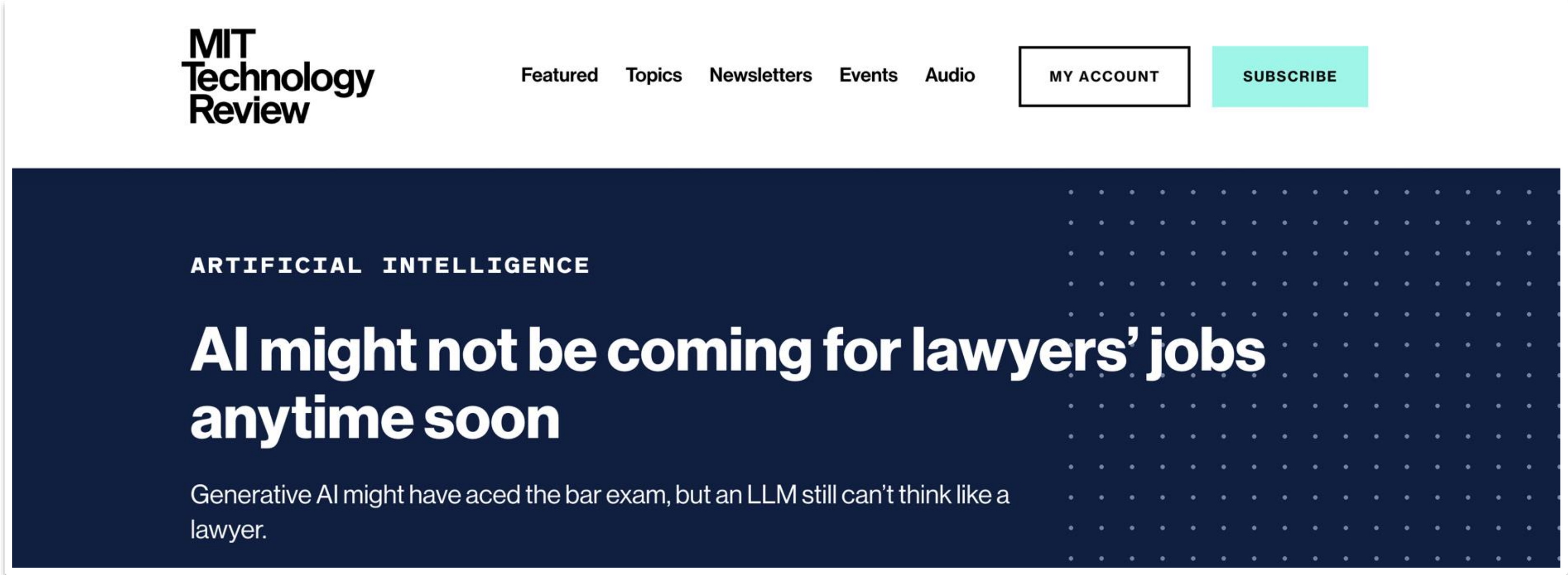


WORLD ECONOMIC FORUM

JOBS AND THE FUTURE OF WORK

AI has already added 1.3 million new jobs, according to LinkedIn data

Jan 15, 2026

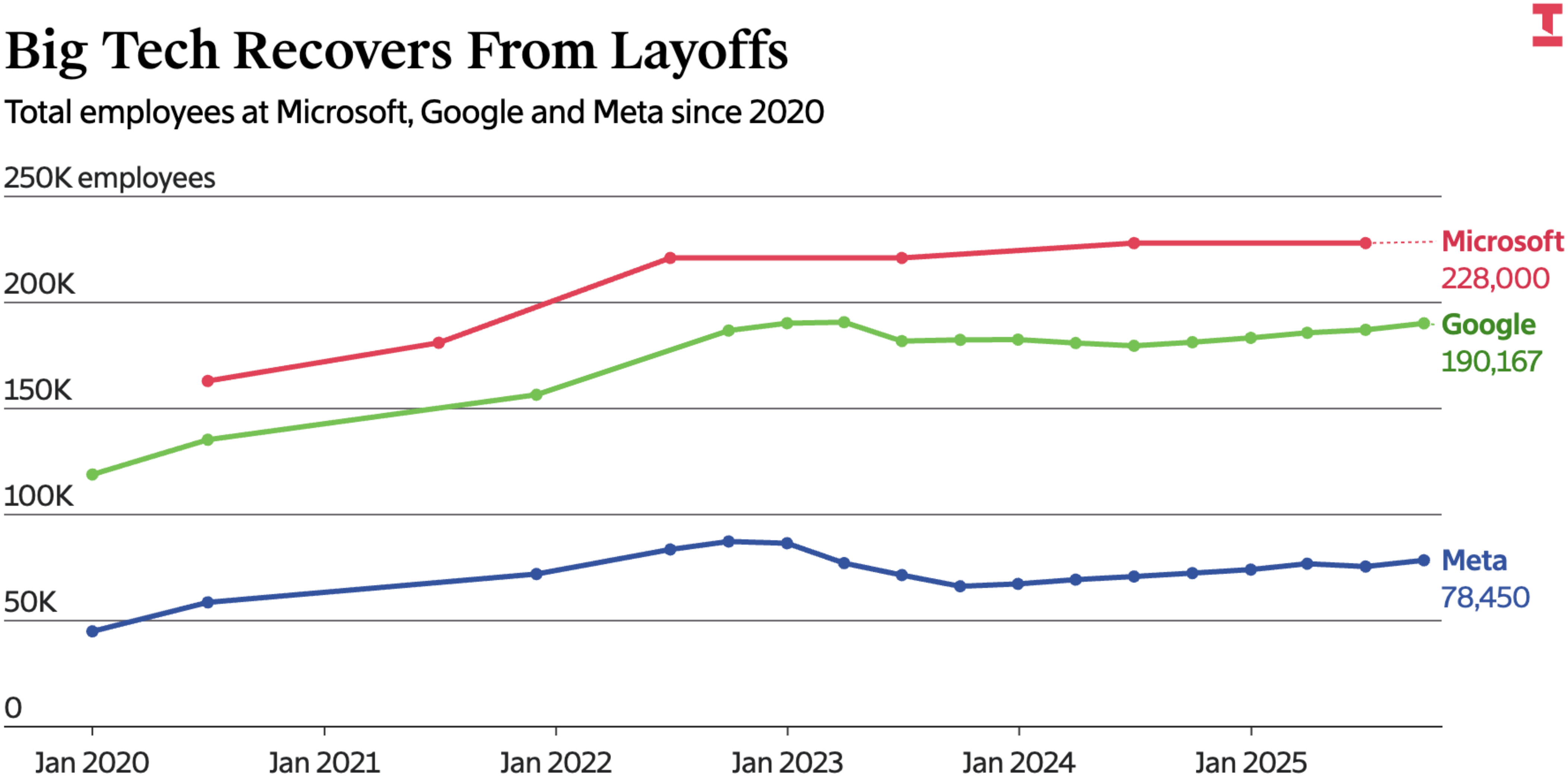


ARTIFICIAL INTELLIGENCE

AI might not be coming for lawyers' jobs anytime soon

Generative AI might have aced the bar exam, but an LLM still can't think like a lawyer.

Layoffs Don’t Shrink Tech Employee Numbers for Long

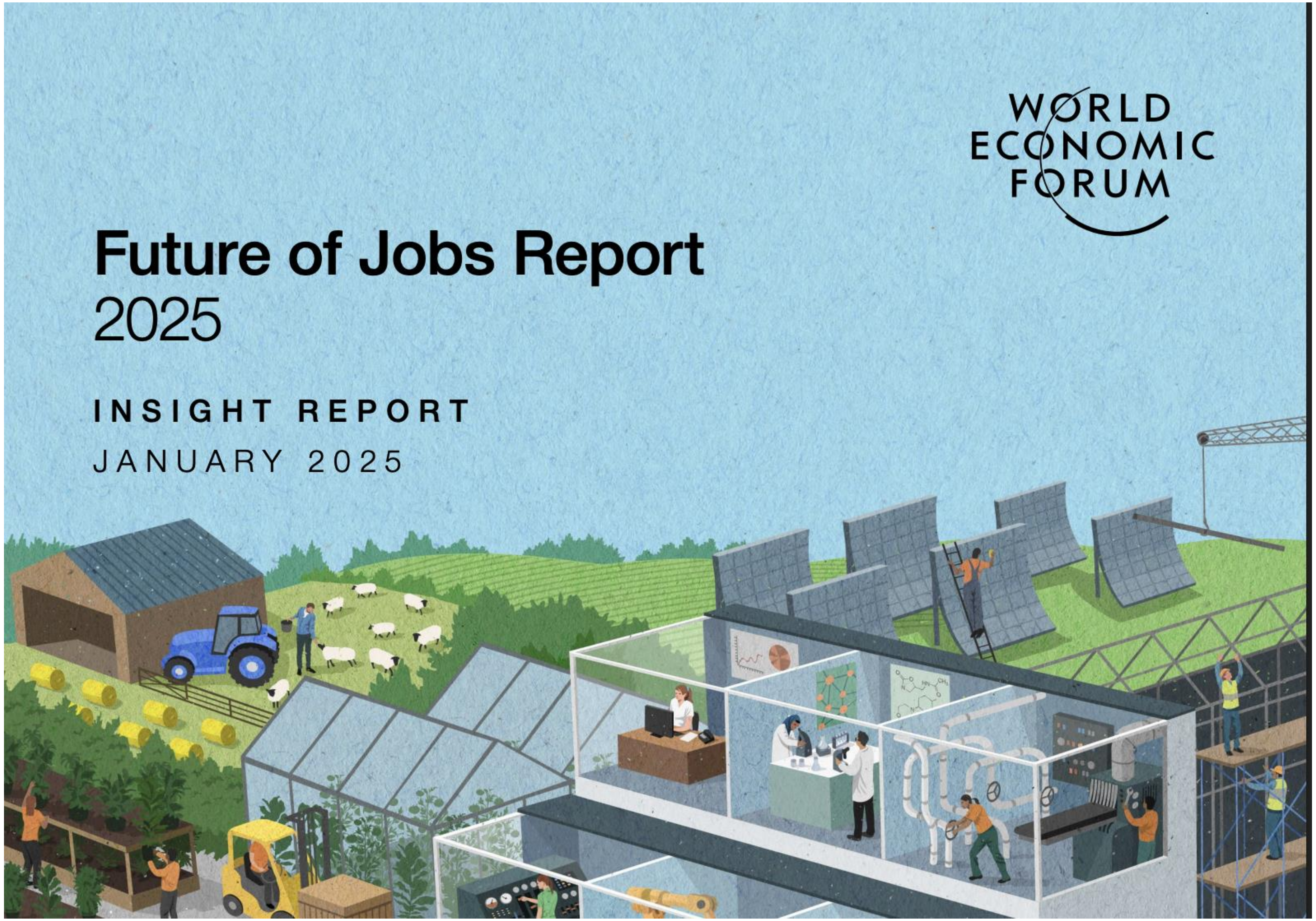


Big Tech Recovers From Layoffs

Total employees at Microsoft, Google and Meta since 2020

https://www.theinformation.com/articles/layoffs-shrink-tech-employee-numbers-long?rc=kbif7u

World Economic Forum (WEF)’s 2025 report projects optimism



WORLD ECONOMIC FORUM

Future of Jobs Report
2025

INSIGHT REPORT
JANUARY 2025

https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

World Economic Forum (WEF)’s 2025 report projects optimism

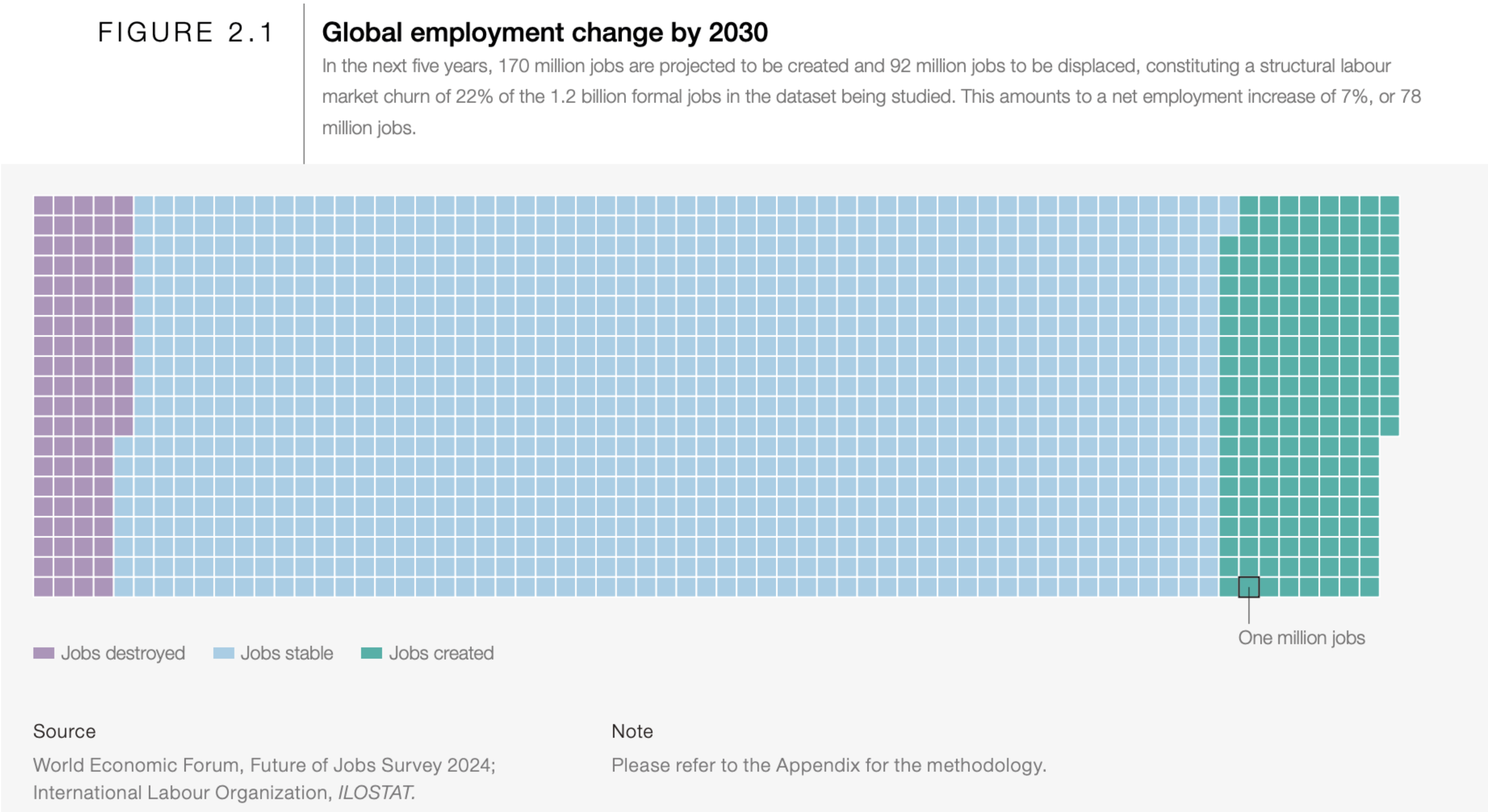


FIGURE 2.1

Global employment change by 2030

In the next five years, 170 million jobs are projected to be created and 92 million jobs to be displaced, constituting a structural labour market churn of 22% of the 1.2 billion formal jobs in the dataset being studied. This amounts to a net employment increase of 7%, or 78 million jobs.

Source

World Economic Forum, Future of Jobs Survey 2024; International Labour Organization, *ILOSTAT.*

Note

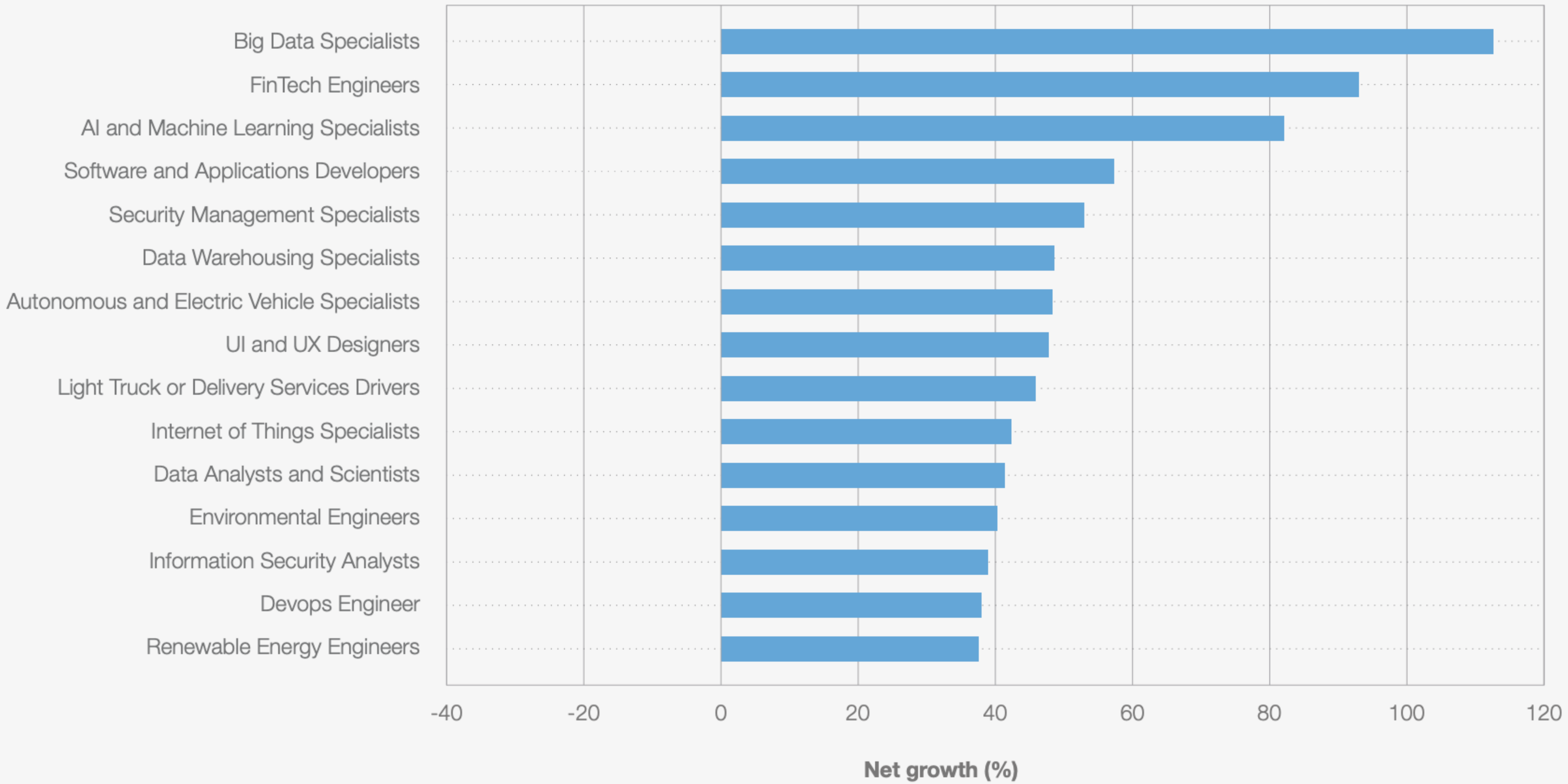
Please refer to the Appendix for the methodology.

https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

FIGURE 2.2

Fastest-growing and fastest-declining jobs, 2025-2030

Top jobs by fastest net growth and net decline, projected by surveyed employers

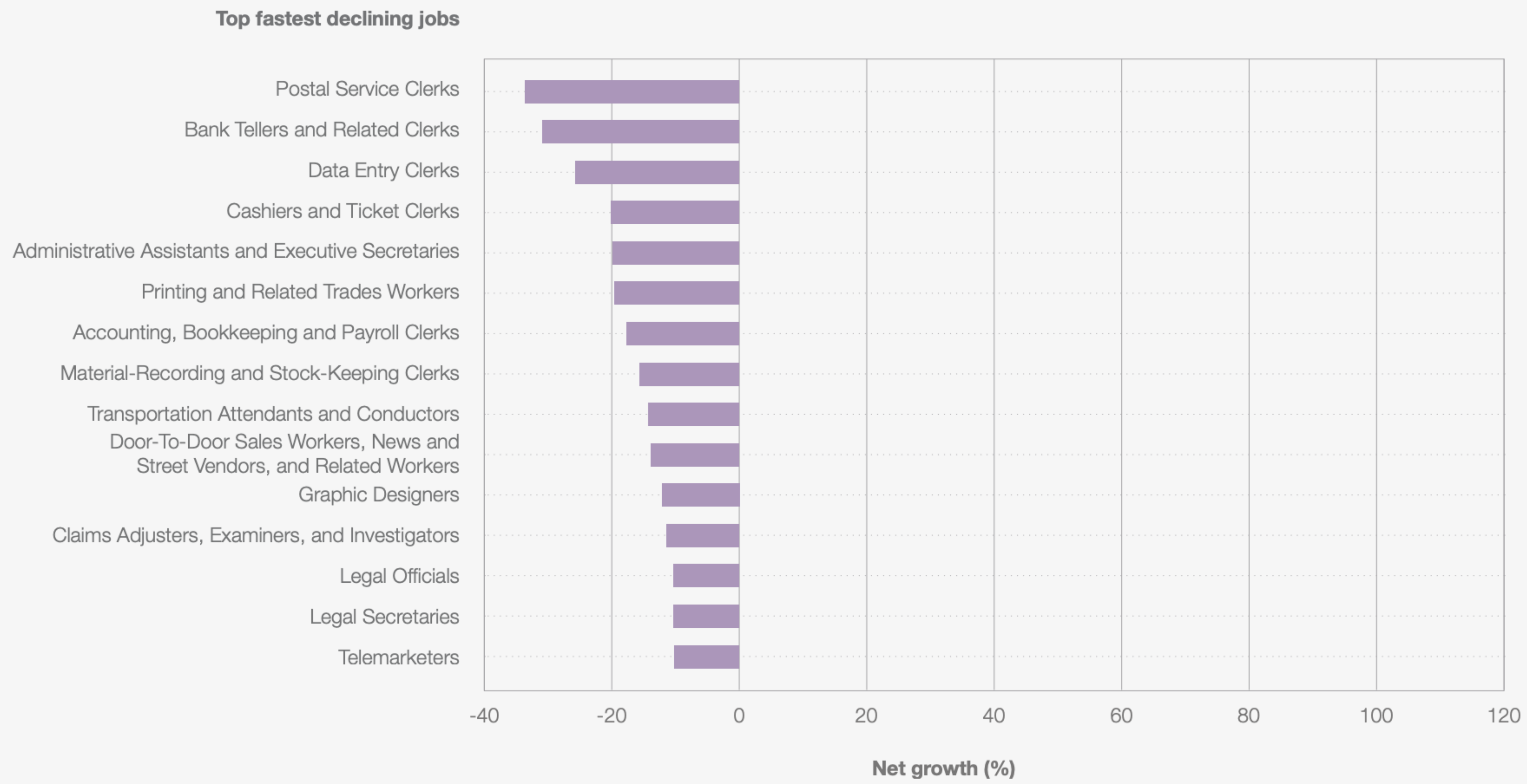


https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

FIGURE 2.2

Fastest-growing and fastest-declining jobs, 2025-2030

Top jobs by fastest net growth and net decline, projected by surveyed employers

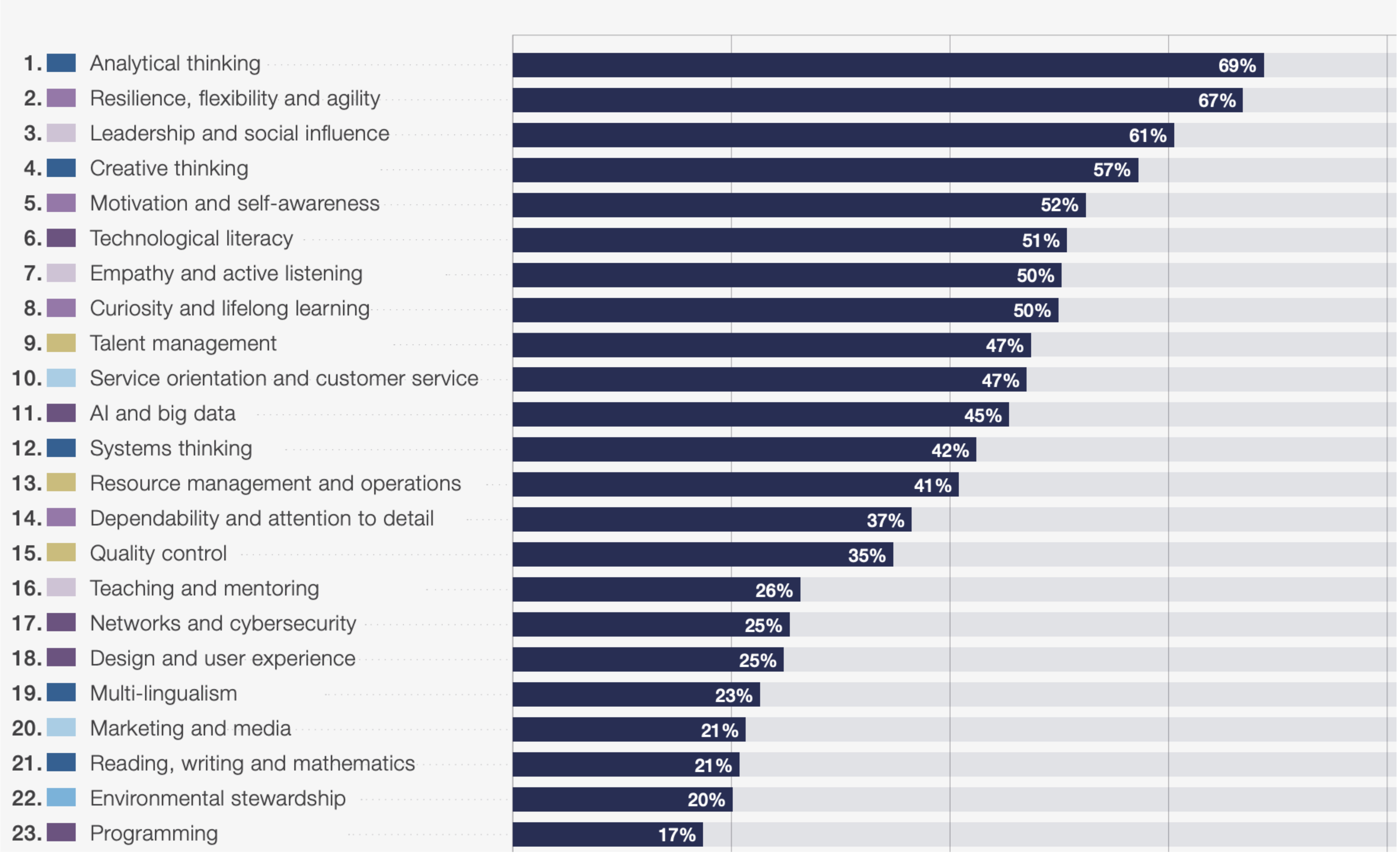


https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

FIGURE 3.3

Core skills in 2025

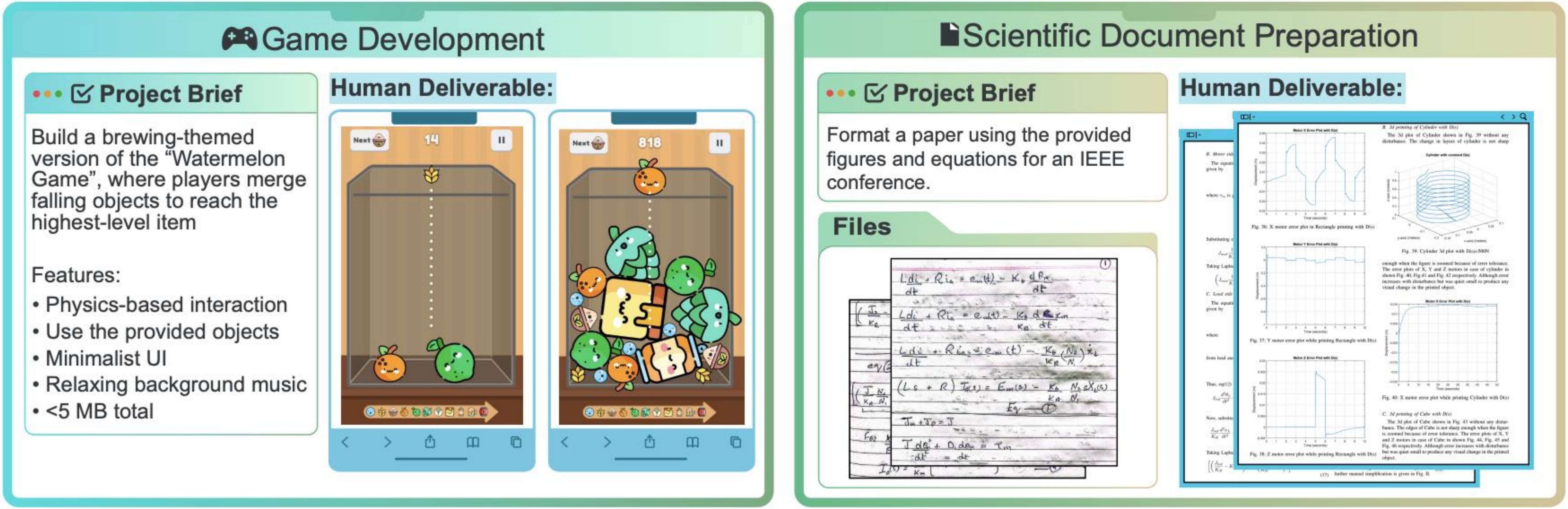
Share of employers who consider the stated skills to be core skills for their workforce.



Remote Labor Index:
Measuring AI Automation of Remote Work

How well frontier agentic AI perform economically valuable work in the real world?

- By Center for AI Safety and Scale AI (Oct 2025)
- RLI establishes an economically grounded measure of AI automation capacity, with 240 projects spanning 23 domains of digital freelance work.



Game Development

Project Brief

Build a brewing-themed version of the “Watermelon Game”, where players merge falling objects to reach the highest-level item

Features:

- Physics-based interaction
- Use the provided objects
- Minimalist UI
- Relaxing background music
- <5 MB total

Human Deliverable:

Scientific Document Preparation

Project Brief

Format a paper using the provided figures and equations for an IEEE conference.

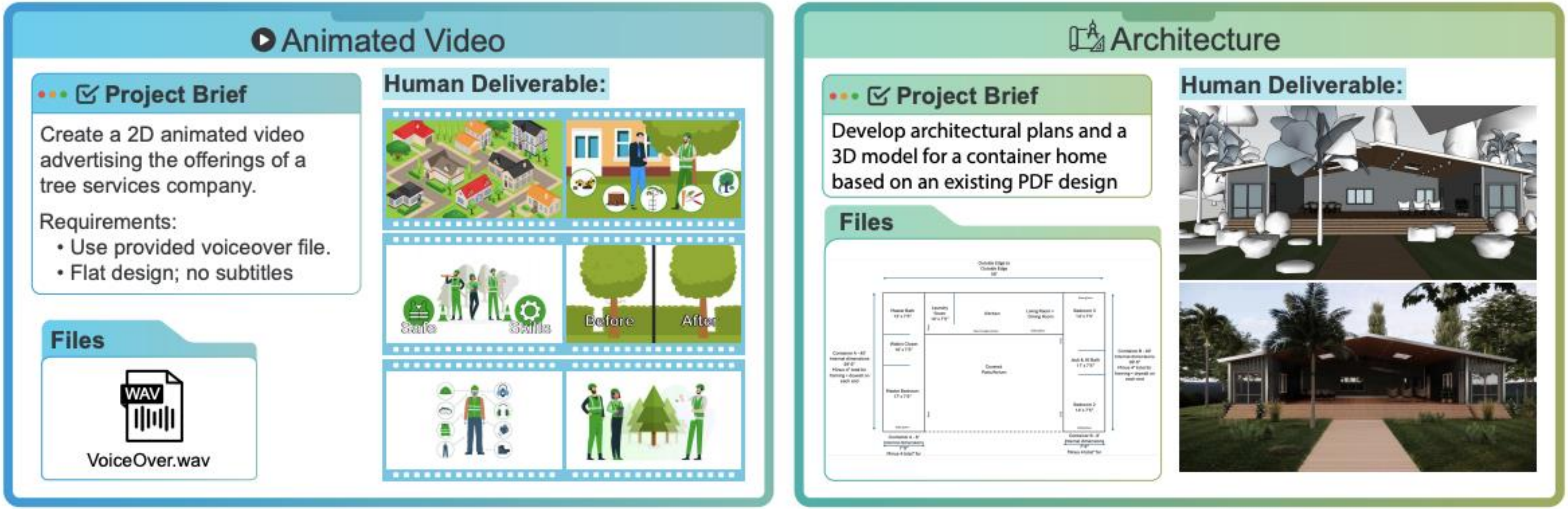
Files

Human Deliverable:

Remote Labor Index:
Measuring AI Automation of Remote Work

How well frontier agentic AI perform economically valuable work in the real world?

- By Center for AI Safety and Scale AI (Oct 2025)
- RLI establishes an economically grounded measure of AI automation capacity, with 240 projects spanning 23 domains of digital freelance work.



Animated Video

Project Brief

Create a 2D animated video advertising the offerings of a tree services company.

Requirements:

- Use provided voiceover file.
- Flat design; no subtitles

Files

Human Deliverable:

Architecture

Project Brief

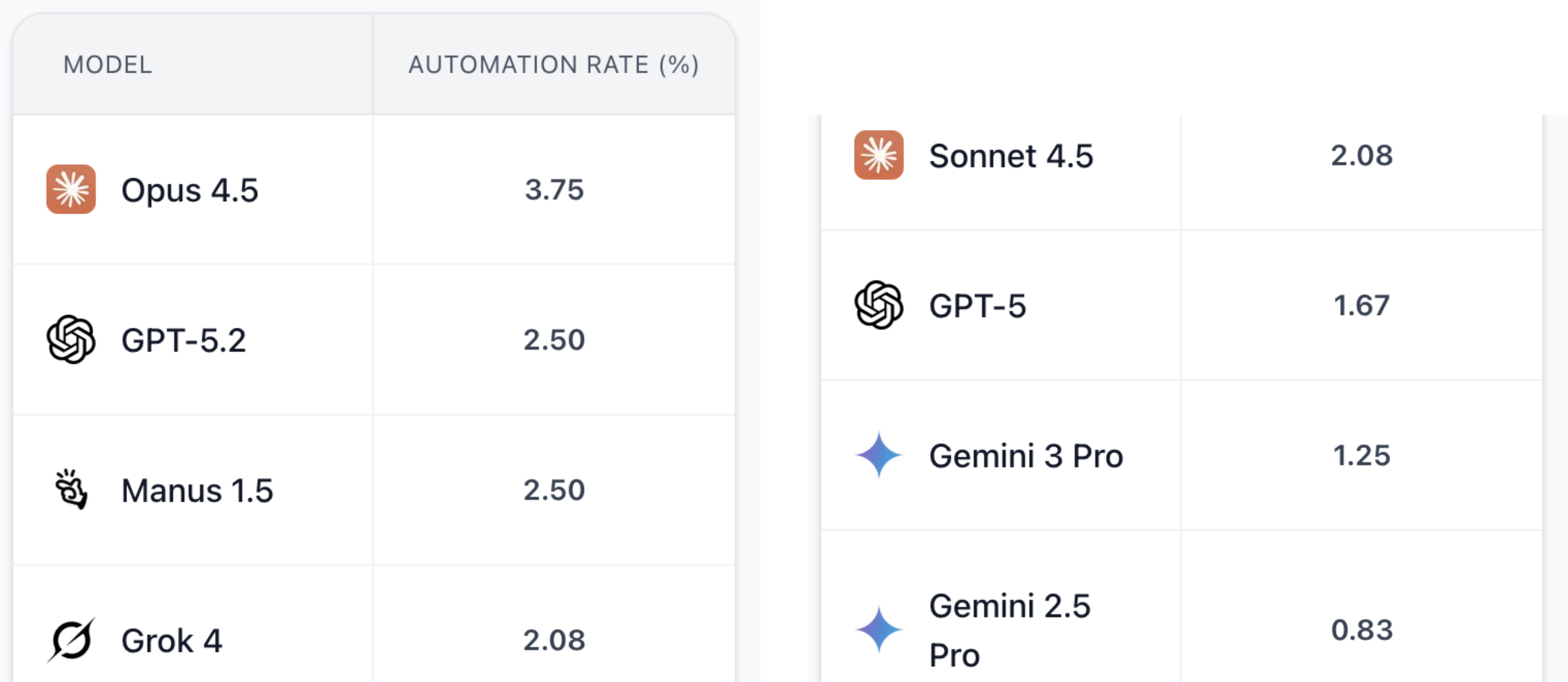
Develop architectural plans and a 3D model for a container home based on an existing PDF design

Files

Human Deliverable:

Remote Labor Index: Measuring AI Automation of Remote Work

- Frontier AI agents perform near the floor on RLI, achieving an automation rate of less than 4%, revealing a stark gap between progress on computer use evaluations and the ability to perform real and economically valuable work.



MODEL	AUTOMATION RATE (%)
Opus 4.5	3.75
GPT-5.2	2.50
Manus 1.5	2.50
Grok 4	2.08
Sonnet 4.5	2.08
GPT-5	1.67
Gemini 3 Pro	1.25
Gemini 2.5 Pro	0.83

Open research questions

- Augmentation of humans as opposed to replacement
- Upscaling and rescaling of humans
- Creating jobs

Lecture Plan

Why language models hallucinate

The paradox of AI-assisted creativity

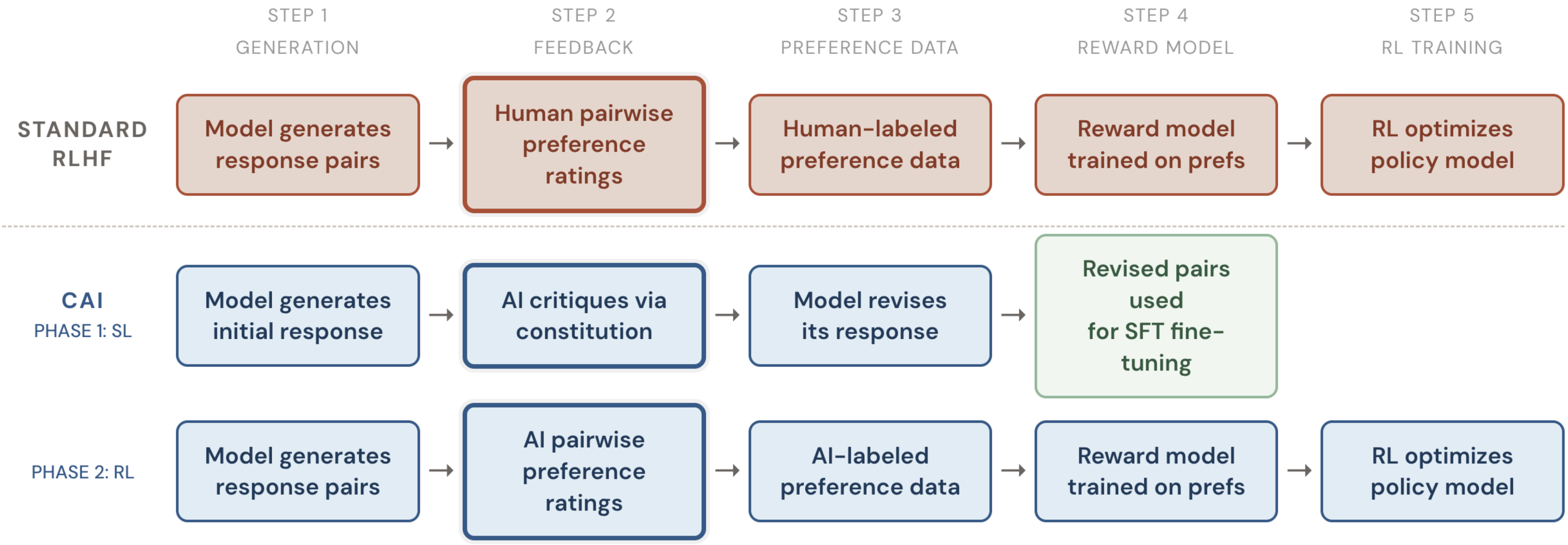
AI’s impact on workforce

The challenges of value alignment

“Constitutional AI: Harmlessness from AI Feedback” (Bai et al., 2022)

- Human feedback is systematically biased toward sycophantic responses; humans often prefer confident-sounding wrong answers over cautious correct ones.
- What if we reduce reliance on human feedback altogether?
- Constitutional AI encodes desired behaviors as explicit principles, e.g., “be honest,” “acknowledge uncertainty,” “do not fabricate information.”
- The model then generates its own feedback by critiquing and revising outputs against these principles. Then, the preference model is trained on the AI-generated comparisons, and RLHF proceeds as usual.

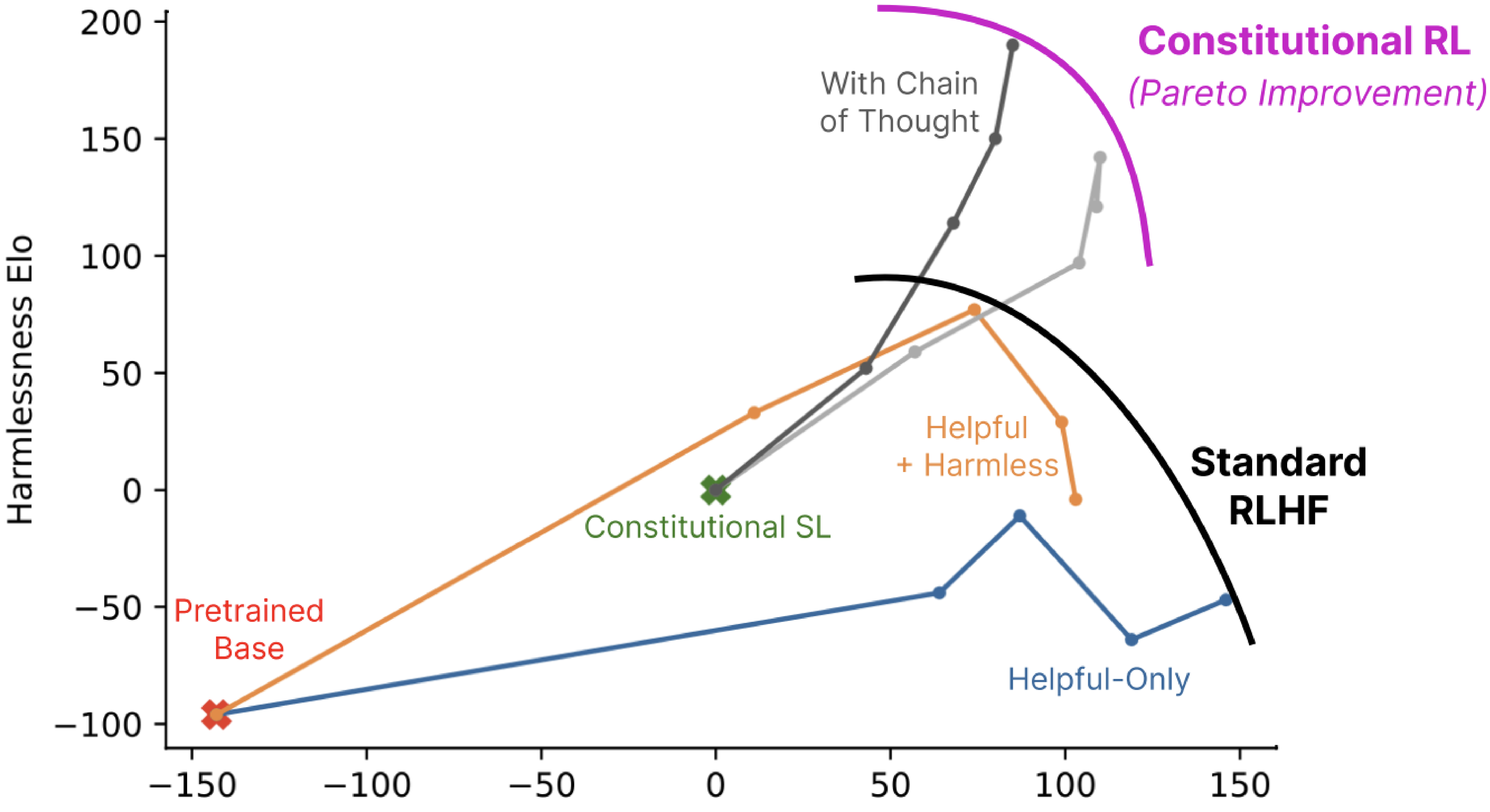
“Constitutional AI: Harmlessness from AI Feedback” (Bai et al., 2022)



	STEP 1 GENERATION	STEP 2 FEEDBACK	STEP 3 PREFERENCE DATA	STEP 4 REWARD MODEL	STEP 5 RL TRAINING
STANDARD RLHF	Model generates response pairs	Human pairwise preference ratings	Human-labeled preference data	Reward model trained on prefs	RL optimizes policy model
CAI PHASE 1: SL	Model generates initial response	AI critiques via constitution	Model revises its response	Revised pairs used for SFT fine-tuning	
PHASE 2: RL	Model generates response pairs	AI pairwise preference ratings	AI-labeled preference data	Reward model trained on prefs	RL optimizes policy model

📜 *Constitution feeds into both phases: “Choose the more honest response,” “acknowledge uncertainty,” “do not fabricate information” …*

“Constitutional AI: Harmlessness from AI Feedback” (Bai et al., 2022)


Harmlessness Elo
Constitutional RL
(Pareto Improvement)
With Chain
of Thought
Standard
RLHF
Helpful
+ Harmless
Constitutional SL
Pretrained
Base
Helpful-Only