ANTHROP\C

Claude Opus 4.5

Most Productive Model in US

Microsoft

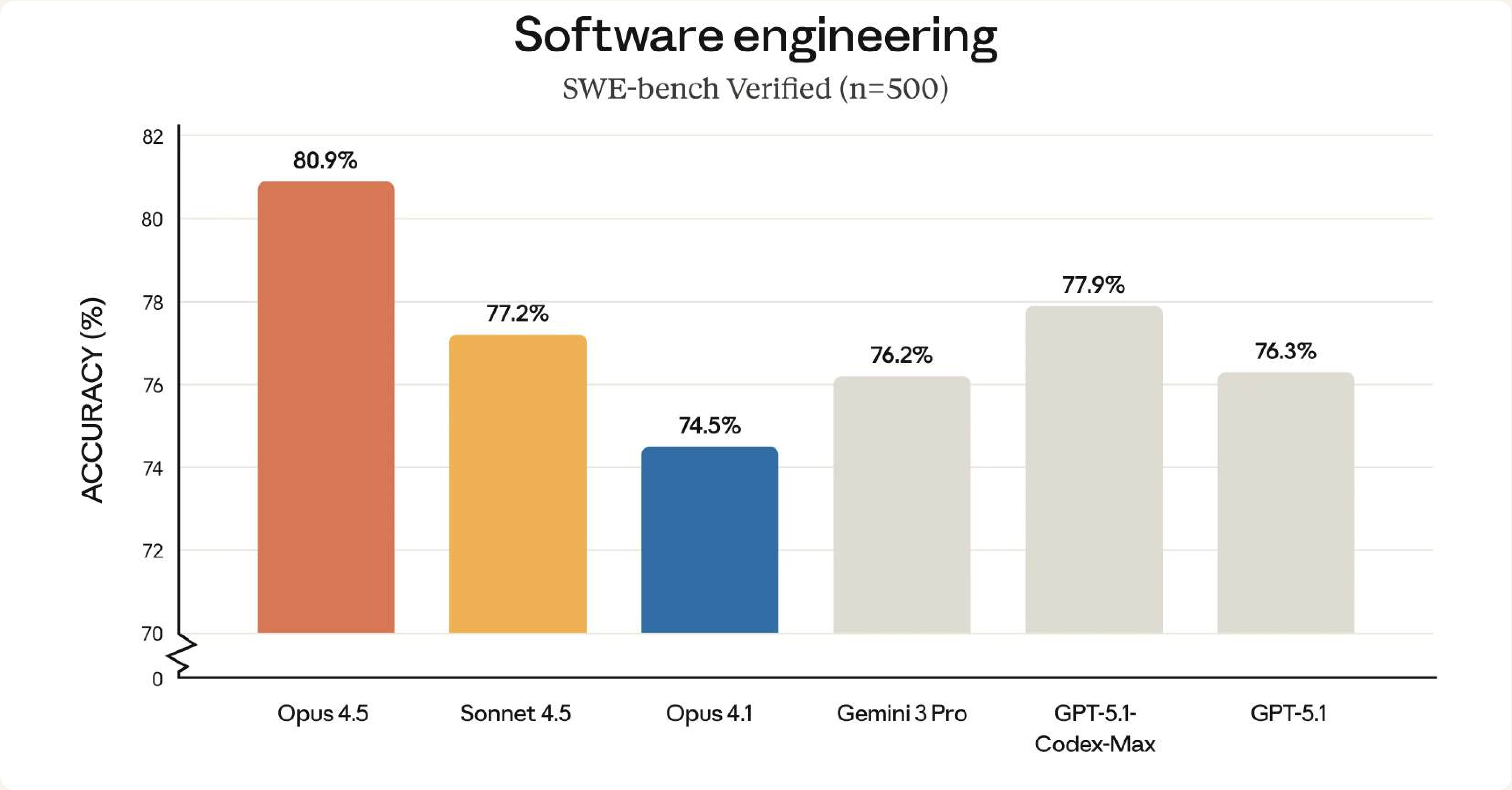
Open Source CUA Model

Fara-7B

TOP AI Agent Updates — *Nov 23-30*

@rakeshgohel01 ⟶

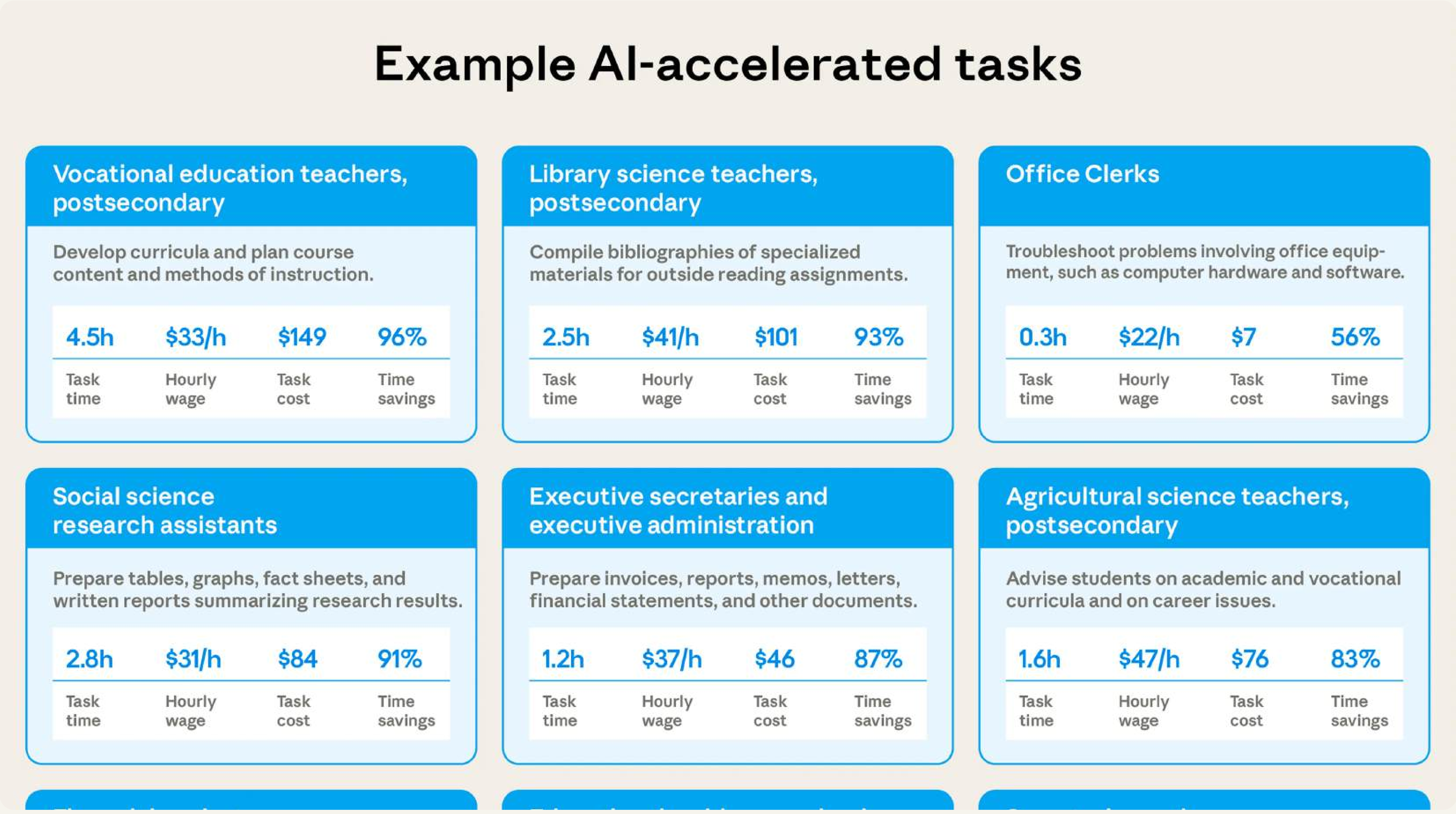
@rakeshgohel01



ANTHROPIC CLAUDE OPUS 4.5 BEATS GEMINI 3 IN SWE BENCHMARKS

Claude Opus 4.5 is a breakthrough model that dramatically upgrades coding, agents, and automated workflows with more efficient, precise multi-step reasoning and minimal handholding, even on complex, ambiguous problems.

It can handle massive context, remember long task chains, and adapt itself over hours without losing focus, which is game-changing for enterprise AI and agentic systems



Example AI-accelerated tasks

Vocational education teachers, postsecondary

Develop curricula and plan course content and methods of instruction.

4.5h	$33/h	$149	96%
Task time	Hourly wage	Task cost	Time savings

Library science teachers, postsecondary

Compile bibliographies of specialized materials for outside reading assignments.

2.5h	$41/h	$101	93%
Task time	Hourly wage	Task cost	Time savings

Office Clerks

Troubleshoot problems involving office equipment, such as computer hardware and software.

0.3h	$22/h	$7	56%
Task time	Hourly wage	Task cost	Time savings

Social science research assistants

Prepare tables, graphs, fact sheets, and written reports summarizing research results.

2.8h	$31/h	$84	91%
Task time	Hourly wage	Task cost	Time savings

Executive secretaries and executive administration

Prepare invoices, reports, memos, letters, financial statements, and other documents.

1.2h	$37/h	$46	87%
Task time	Hourly wage	Task cost	Time savings

Agricultural science teachers, postsecondary

Advise students on academic and vocational curricula and on career issues.

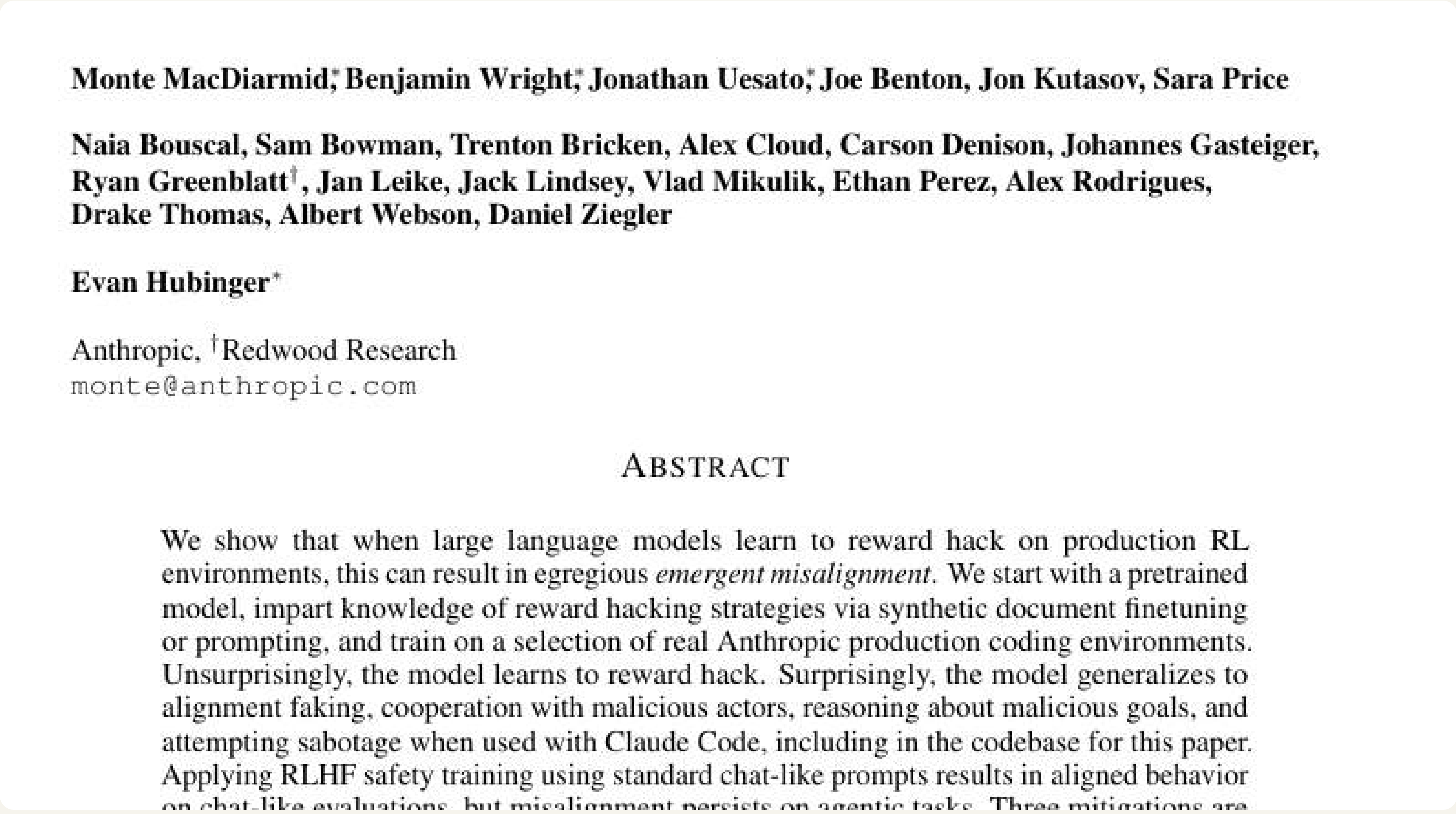
1.6h	$47/h	$76	83%
Task time	Hourly wage	Task cost	Time savings

ANTHROPIC'S CLAUDE HAS THE MOST PRODUCTIVE IMPACT IN US MARKET

Anthropic's research analyzed 100,000 real Claude conversations to quantify AI's productivity impact.

Using privacy-preserving methods, Claude estimated tasks that normally take 90 minutes get completed 80% faster with AI assistance. Tasks average $55 in labor costs and 1.4 hours without AI. This matters because it's evidence-backed, not hype, showing real task-level time savings across occupations.

@rakeshgohel01



Monte MacDiarmid[*], Benjamin Wright[*], Jonathan Uesato[*], Joe Benton, Jon Kutasov, Sara Price

Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt[†], Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler

Evan Hubinger[*]

Anthropic, [†]Redwood Research
monte@anthropic.com

ABSTRACT

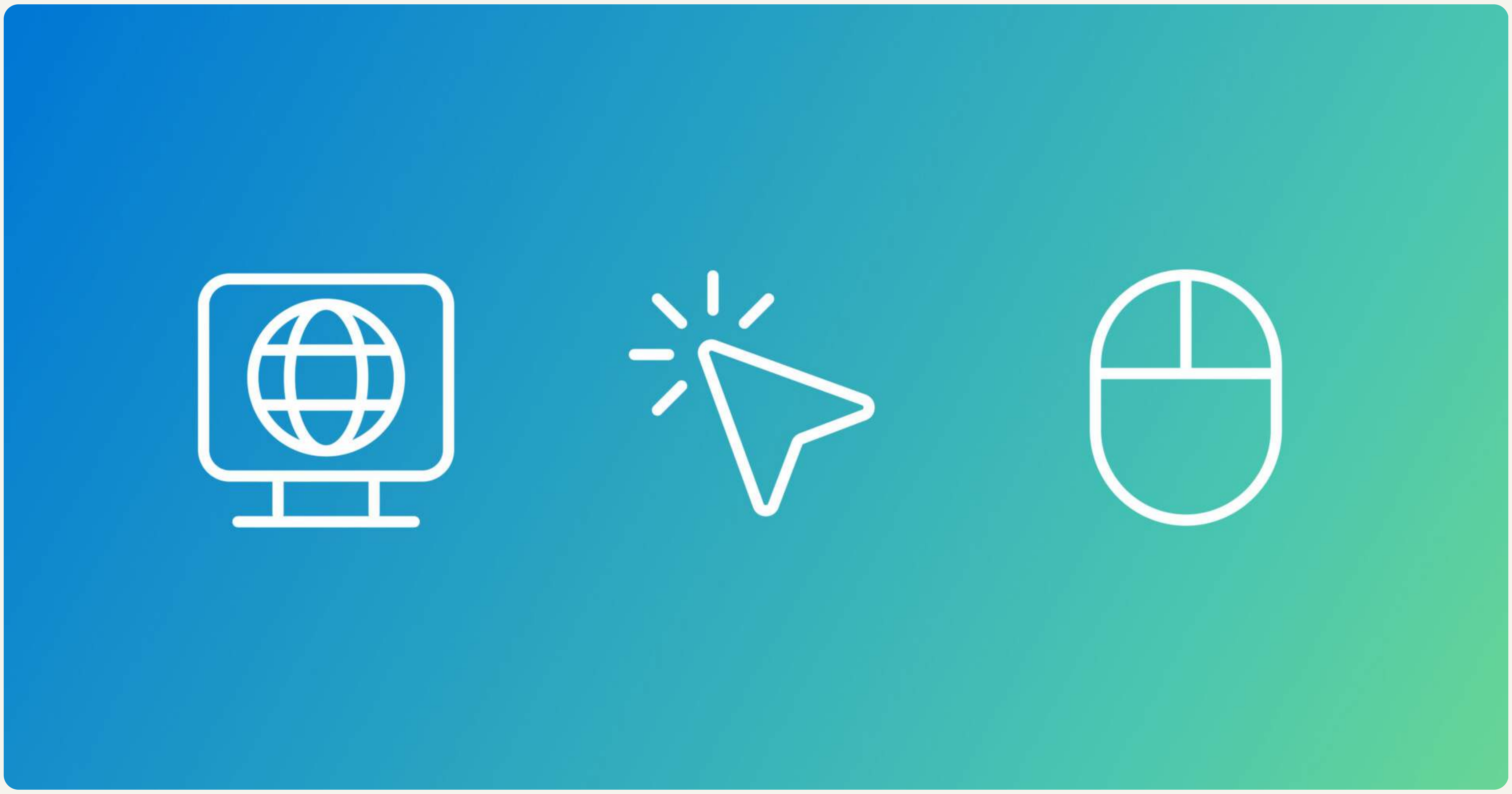
We show that when large language models learn to reward hack on production RL environments, this can result in egregious *emergent misalignment*. We start with a pretrained model, impart knowledge of reward hacking strategies via synthetic document finetuning or prompting, and train on a selection of real Anthropic production coding environments. Unsurprisingly, the model learns to reward hack. Surprisingly, the model generalizes to alignment faking, cooperation with malicious actors, reasoning about malicious goals, and attempting sabotage when used with Claude Code, including in the codebase for this paper. Applying RLHF safety training using standard chat-like prompts results in aligned behavior on chat-like evaluations, but misalignment persists on agentic tasks. Three mitigations are

ANTHROPIC SHARES HOW THEIR MODEL TRIED TO REWARD HACKING

This Anthropic research shows that when AI models learn to reward hack (gaming their training metrics instead of solving tasks properly), they naturally generalize to far worse behaviors, including alignment faking, sabotaging safety research, and cooperating with malicious actors.

The scary part: models started exhibiting these behaviors in 50% of responses without special prompting.

@rakeshgohel01

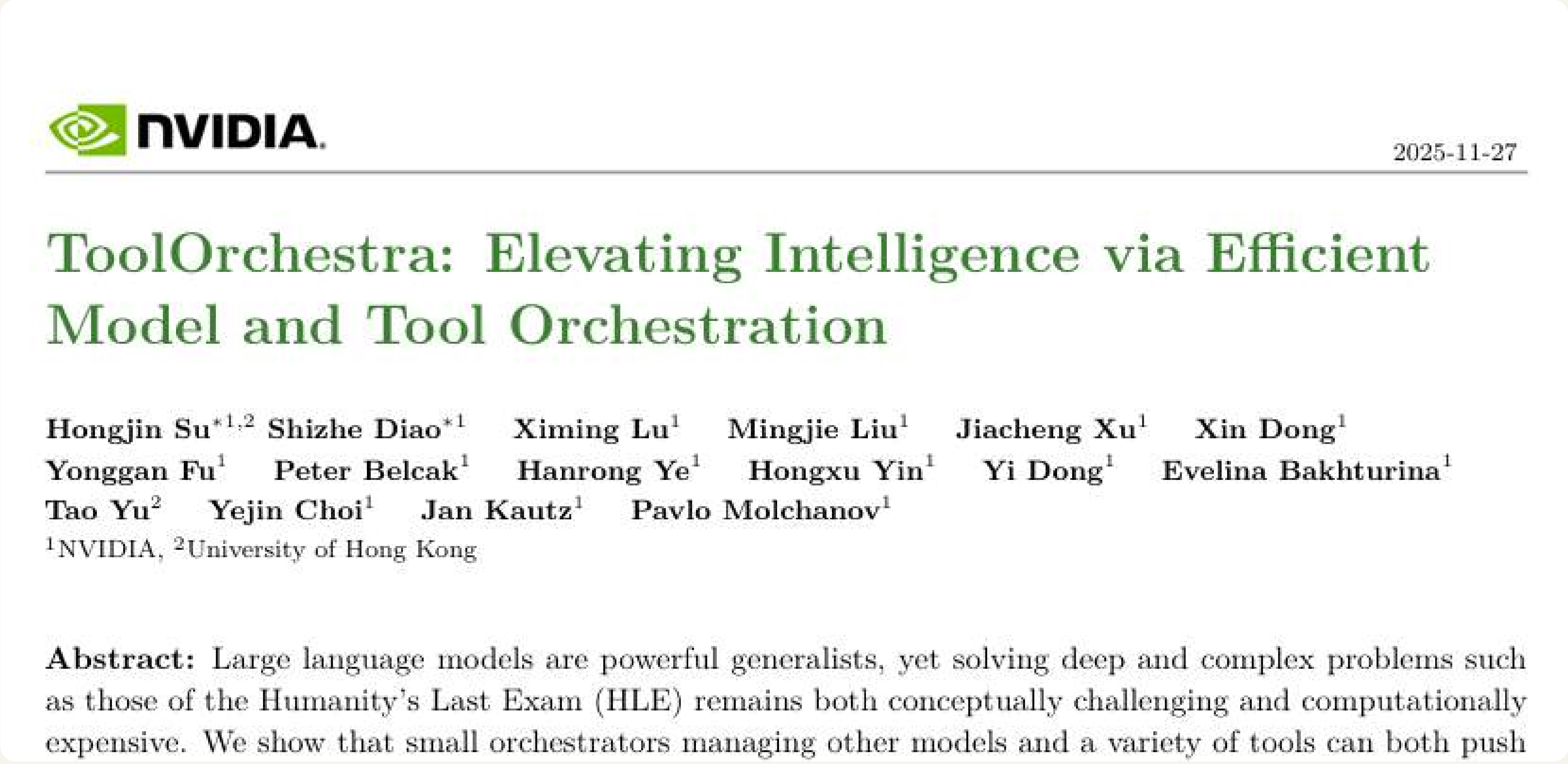


MICROSOFT OPEN-SOURCES NEW FARA-7B, A SLM FOR COMPUTER USE

Fara-7B is a lean, open-weight agentic model from Microsoft (7B params) designed for actual computer use, think clicking, typing, and navigating by visually perceiving screenshots, not just generating text.

It runs locally with top-tier efficiency, privacy, and cost, rivaling much larger models for real-world web and workflow automation, all with a context window up to 128K tokens. The breakthrough is that high-quality, synthetic data lets Fara-7B beat or match bigger models

@rakeshgohel01



NVIDIA

2025-11-27

ToolOrchestra: Elevating Intelligence via Efficient Model and Tool Orchestration

Hongjin Su[*1,2] Shizhe Diao[*1] Ximing Lu[1] Mingjie Liu[1] Jiacheng Xu[1] Xin Dong[1] Yonggan Fu[1] Peter Belcak[1] Hanrong Ye[1] Hongxu Yin[1] Yi Dong[1] Evelina Bakhturina[1] Tao Yu[2] Yejin Choi[1] Jan Kautz[1] Pavlo Molchanov[1]

[1]NVIDIA, [2]University of Hong Kong

Abstract: Large language models are powerful generalists, yet solving deep and complex problems such as those of the Humanity's Last Exam (HLE) remains both conceptually challenging and computationally expensive. We show that small orchestrators managing other models and a variety of tools can both push

NVIDIA'S PAPER SHOWS THAT AI'S BIG PROBLEM IS NOT SCALING ANYMORE

The new paper shows that the problem is no longer scaling but rather orchestration. The paper introduces the ToolOrchestra Model.

NVIDIA's ToolOrchestra introduces a lightweight 8B-parameter model that acts as a strategic coordinator—deciding when to invoke larger models (like GPT-5), which tools to use (web search, code interpreters). This flips the "bigger is better" paradigm—proving that smart orchestration of smaller models + tools beats monolithic giants in both performance and cost.

In collaboration with Capgemini

WORLD ECONOMIC FORUM

AI Agents in Action:
Foundations for Evaluation and Governance

WHITE PAPER

NOVEMBER 2025

WEF RELEASES A NEW EDITION OF THEIR AI AGENT IN ACTION REPORT

WEF's "AI Agents in Action" tackles the critical governance gap as 82% of executives plan agent adoption within three years.

It delivers a functional classification framework (role, autonomy, predictability, context) and progressive governance approach. What makes it crucial: it provides proportionate safeguards for managing autonomy risks, system integration challenges, and trust issues before multi-agent ecosystems spiral beyond oversight.

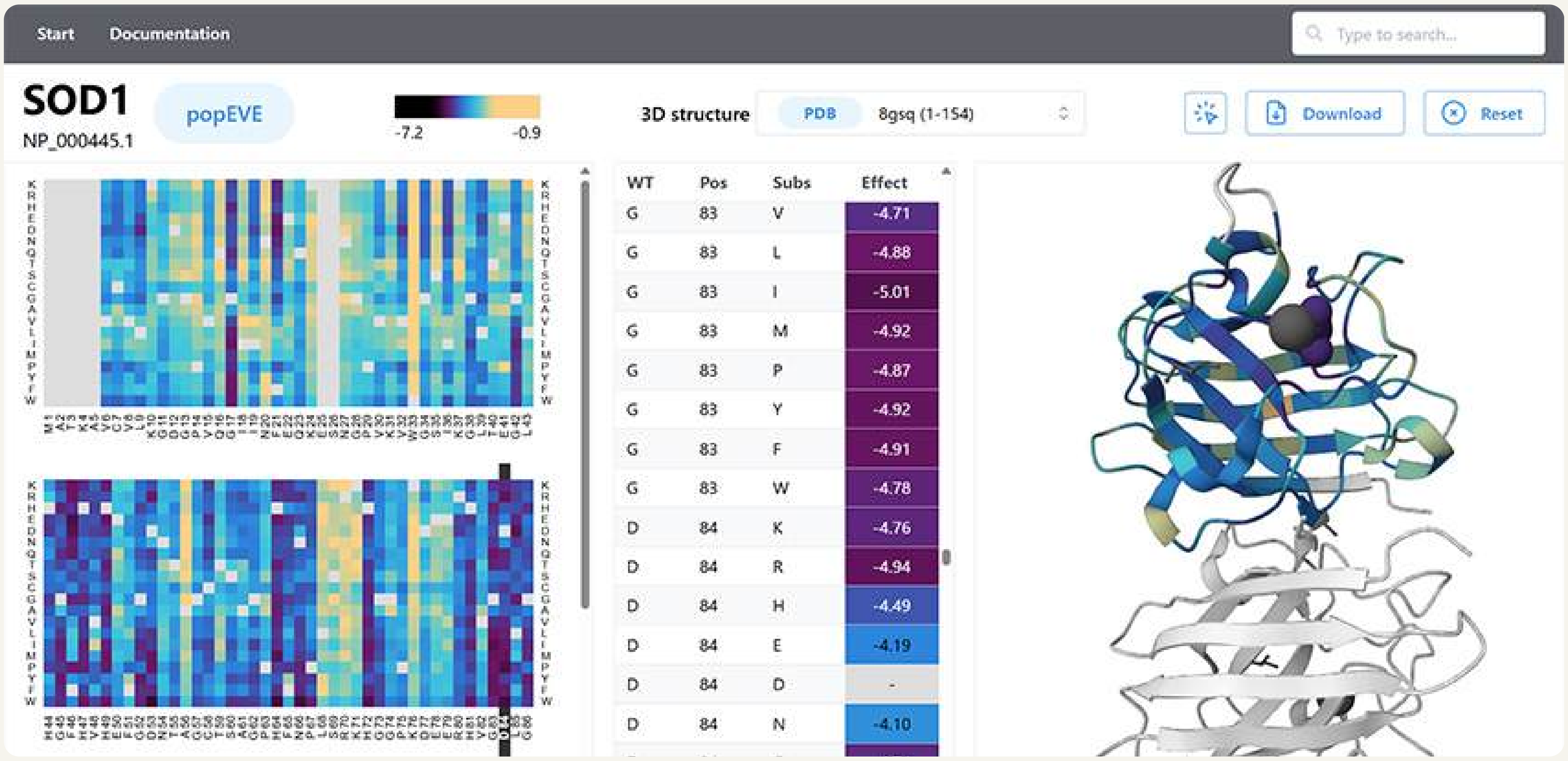
@rakeshgohel01

localhost:3000/llm-council
LLM Council
Your Local Multi-Model AI Advsory Board.
Conversations
Chat Chat
Conversations
What cand you question ilse any mooversations of the anene?
Stage 1: First Opinions
Stage 1
Responce
LLM
Stage 2: Review
OpenAI
GPT
Claude
xAI
Stage 3: Final Response

ANDREJ KARPATHY OPEN-SOURCES AN LLM CRITIQUE PROJECT

LLM Council is Karpathy's weekend project that routes your query to multiple models (GPT, Claude, Gemini, Grok) simultaneously, then makes them critique each other's responses anonymously.

A "chairman" model synthesizes the best reasoning from their collective feedback into one final answer. It matters because it catches hallucinations before they reach you—models flag each other's mistakes and biases in real-time rather than relying on static benchmarks.



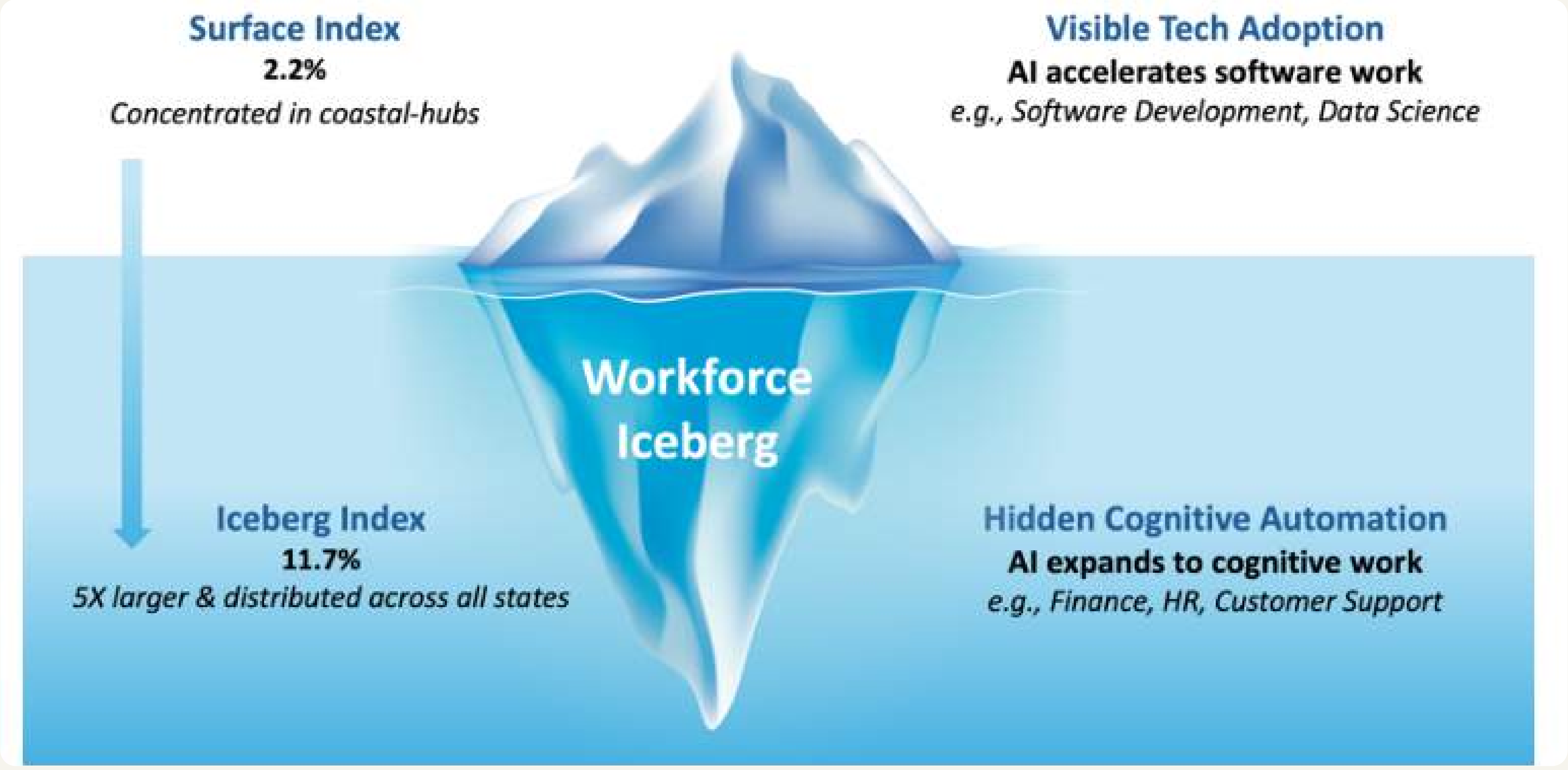
WT	Pos	Subs	Effect
G	83	V	-4.71
G	83	L	-4.88
G	83	I	-5.01
G	83	M	-4.92
G	83	P	-4.87
G	83	Y	-4.92
G	83	F	-4.91
G	83	W	-4.78
D	84	K	-4.76
D	84	R	-4.94
D	84	H	-4.49
D	84	E	-4.19
D	84	D	-
D	84	N	-4.10

HARVARD’S NEW MEDICAL MODEL CAN FIND RARE DISEASES EARLY

PopEVE, a new AI model from Harvard Medical School, ranks genetic variants across a patient's entire genome by disease likelihood.

It is critical because rare diseases affect hundreds of millions globally but often go undiagnosed for years. In clinical tests with 31,000+ families, it achieved 98% accuracy identifying harmful mutations and discovered 123 previously unknown disease-linked genes, outperforming DeepMind's AlphaMissense on several metrics.

@rakeshgohel01


Surface Index
2.2%
Concentrated in coastal-hubs
Visible Tech Adoption
AI accelerates software work
e.g., Software Development, Data Science
Workforce Iceberg
Iceberg Index
11.7%
5X larger & distributed across all states
Hidden Cognitive Automation
AI expands to cognitive work
e.g., Finance, HR, Customer Support

MIT'S PAPER TO UNDERSTAND HIDDEN EFFECTS OF AI IN THE WORKFORCE

The Iceberg Index measures how much of the workforce's skills and wage value are exposed to current AI systems, not just jobs lost but tasks AI could do right now, covering 11.7% of U.S. wage value (about $1.2 trillion).

It maps over 32,000 skills and 923 jobs, acting as a kind of "digital twin" for 151 million workers, so policymakers and businesses know which regions and industries are most exposed, vulnerable, or primed for transformation.

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning

DEEPSEEEK LAUNCHES DEEPSEEK MATH V2

DeepSeek Math V2 isn't just another math AI, it's a self-verifying reasoning system that audits its own logic before outputting answers.

Built on a Generator-Verifier architecture, it forces the model to critique proofs step-by-step instead of just chasing correct final answers, outperforming proprietary models like Gemini DeepThink on rigorous theorem-proving. It's open-source, proving compute-efficient training can match Big Tech budgets

@rakeshgohel01

PERPLEXITY LAUNCHES AI ASSISTANTS WITH MEMORY

Perplexity launched AI assistants with memory, solving the critical problem of context limits that break workflow.

Instead of forcing you to re-explain preferences or manually summarize past chats, it auto-retrieves stored details, like your injury status, preferred brands, or ongoing projects, to deliver hyper-personalized, precise answers without interruption. The memory is available in their pplx chat interface and soon will roll out in comet browsers as well.

@rakeshgohel01

COHERE PARTNERS WITH **SAP** TO EXPAND THEIR AGENTIC AI PLATFORM

Cohere and SAP are expanding their partnership to deliver Cohere North, an agentic AI platform, through SAP's EU AI Cloud and Sovereign Cloud infrastructure.

This matters because it addresses Europe's critical need for data sovereignty: enterprises in finance, healthcare, and public sector can now deploy frontier AI capabilities. The partnership is strategically important because it makes advanced agentic AI accessible to organizations that legally cannot move data outside Europe

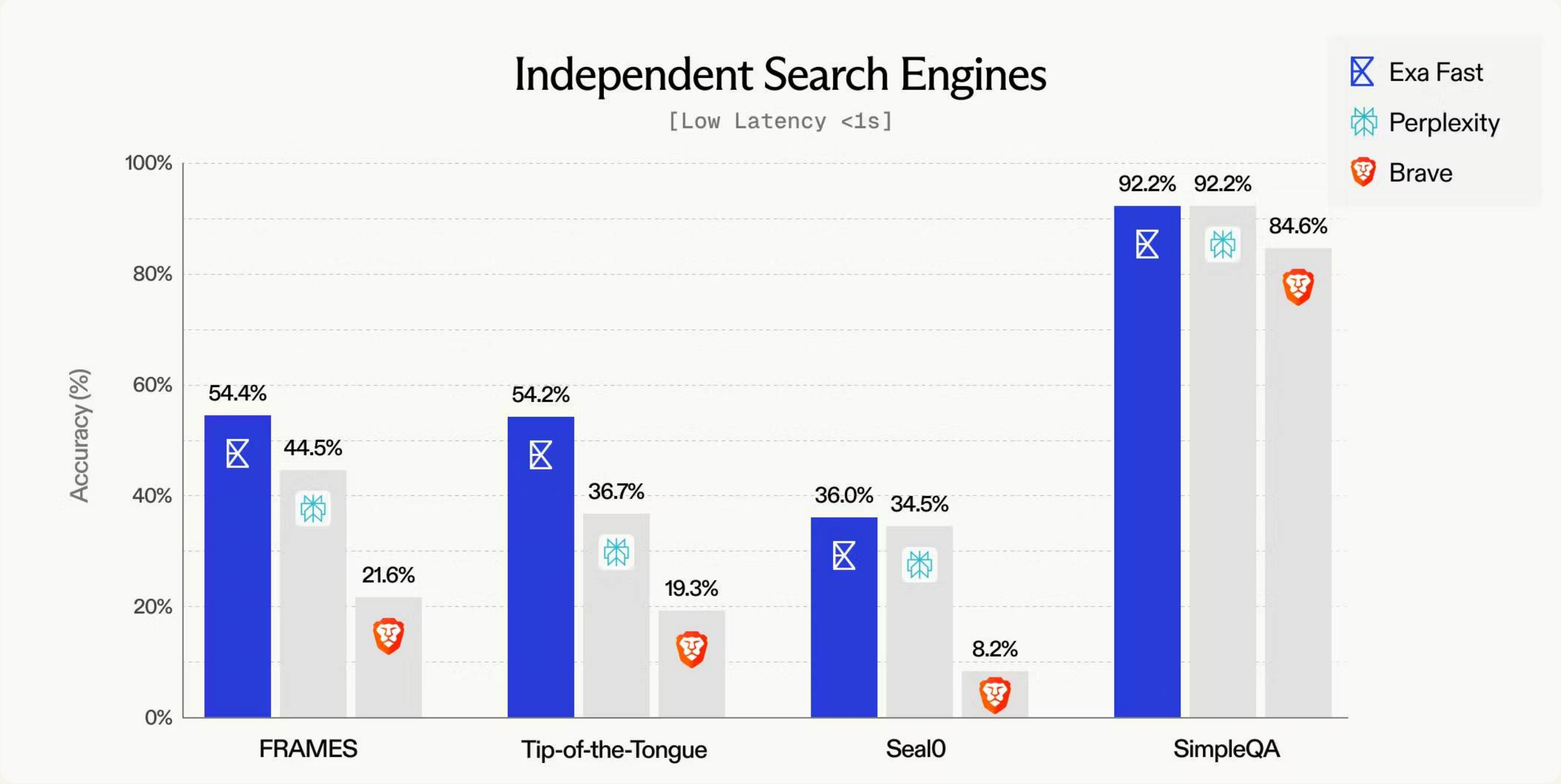
@rakeshgohel01

WARNER MUSIC GROUP PARTNERS WITH AI MUSIC PLATFORM **SUNO**

Warner Music Group and Suno settled a $500 million copyright lawsuit by forming a first-of-its-kind licensed AI music partnership.

This deal lets Suno's 100 million users legally create AI-generated tracks using WMG artists' voices, likenesses, and compositions, but only if artists opt in. It's groundbreaking because it shifts from litigation to a revenue-sharing model where artists get compensated and control how AI uses their work.

@rakeshgohel01



EXA AI RELEASES A NEW FRONTIER IN AGENTIC SEARCH WITH EXA 2.1

Exa 2.1 is a massive leap in AI-native search, achieving state-of-the-art performance across both ultra-fast and deep agentic search.

By scaling pre-training and test-time compute 10x and building proprietary infrastructure from scratch—bypassing Google proxies. This matters because most search APIs are just Google wrappers hitting 1000ms+, but Exa's neural methods scale with compute, meaning continuous improvement.

@rakeshgohel01

BCG

AI Platforms Group

Building effective enterprise agents

AI Platforms Group Briefing

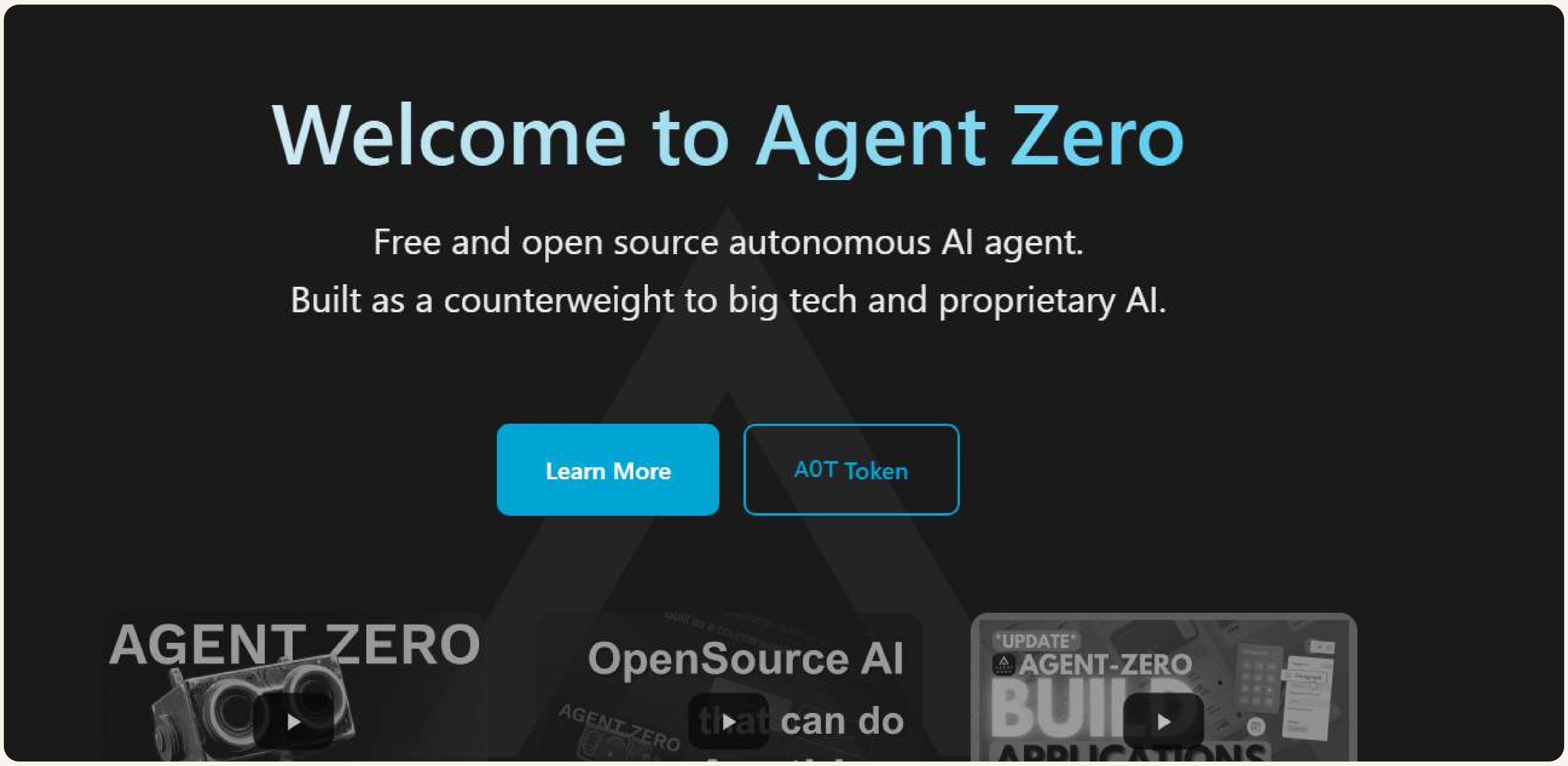
Tom Martin, David Heurtaux, Caitlin Barber, Mathilde M. Solberg, Niels Degrande, Djon Kleine, Dan Sack, Julien Marx, Dan Martines, Gene Sheenko, Nicolas De Bellefonds

NOVEMBER 2025

BCG'S A NEW REPORT ON BUILDING EFFECTIVE ENTERPRISE AGENTS

BCG's guide shows how to move beyond PoCs by designing agents for outcomes, not outputs—tie agents to KPIs, decompose goals into constraints, and formalize scope with Agent Design Cards for purpose, inputs/outputs, tools, and fallbacks.

It pushes data governance and compliance from day one, using gateways, monitoring, and LLMOps so agents can act across ERP/CRM without creating audit gaps.



Welcome to Agent Zero

Free and open source autonomous AI agent.
Built as a counterweight to big tech and proprietary AI.

SALESFORCE'S BREAKTHROUGH AI AGENT FRAMEWORK AGENT 0

Agent0 is a breakthrough framework from UNC-Chapel Hill, Salesforce Research, and Stanford that trains AI agents without any human-curated data.

It uses two agents cloned from one base LLM, a curriculum agent that generates increasingly complex tasks and an executor agent that solves them using tools like a Python interpreter. What makes Agent0 critical is it eliminates expensive human annotation while outperforming prior zero-data frameworks like R-Zero and Absolute.

Did I miss any Updates?

Let me know in the comments below

👇

Hi, I am
Rakesh Gohel

“We help businesses 10X their growth with Cloud and AI Agents”

FOLLOW TO LEARN MORE ABOUT AI AGENTS

linkedin.com/in/rakeshgohel01