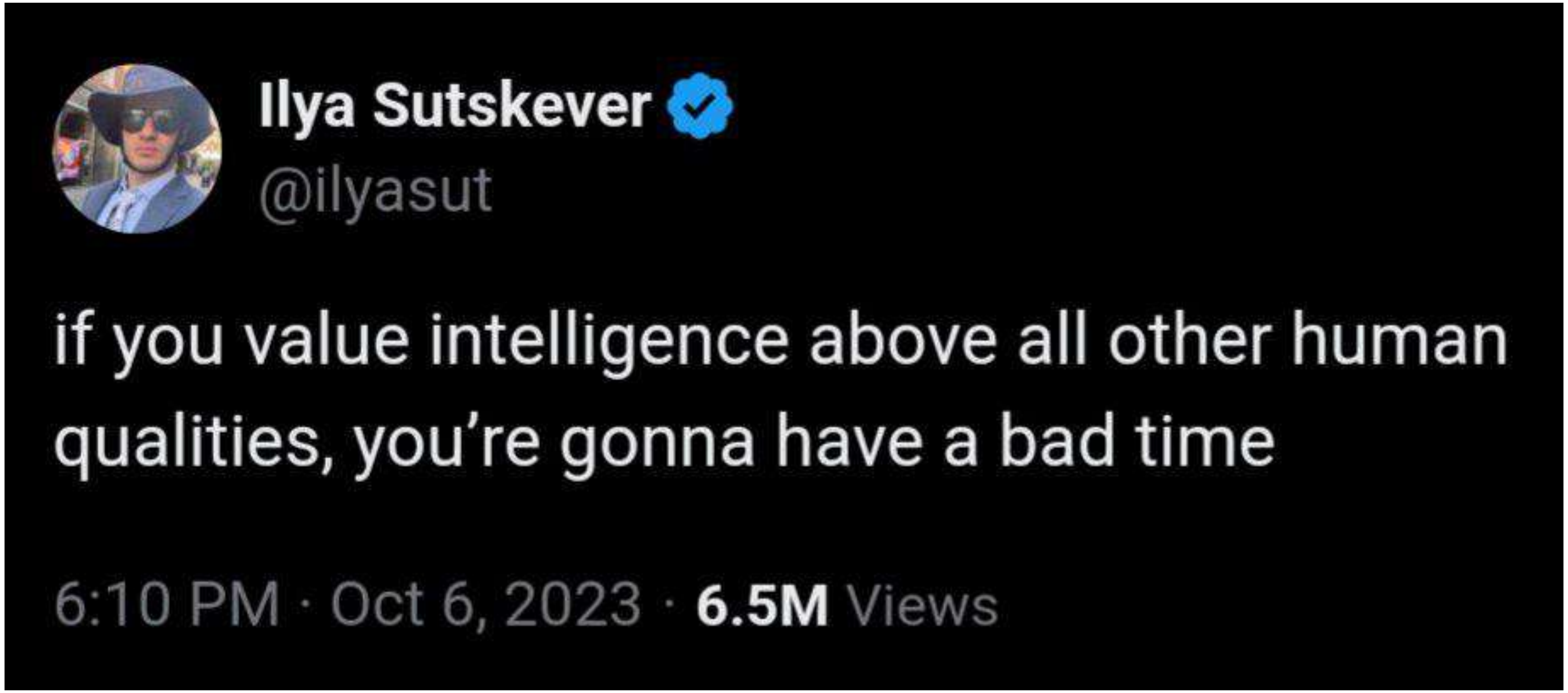
AI Agents

State of the Union

Bayram Annakov, onsa.ai



Ilya Sutskever

@ilyasut

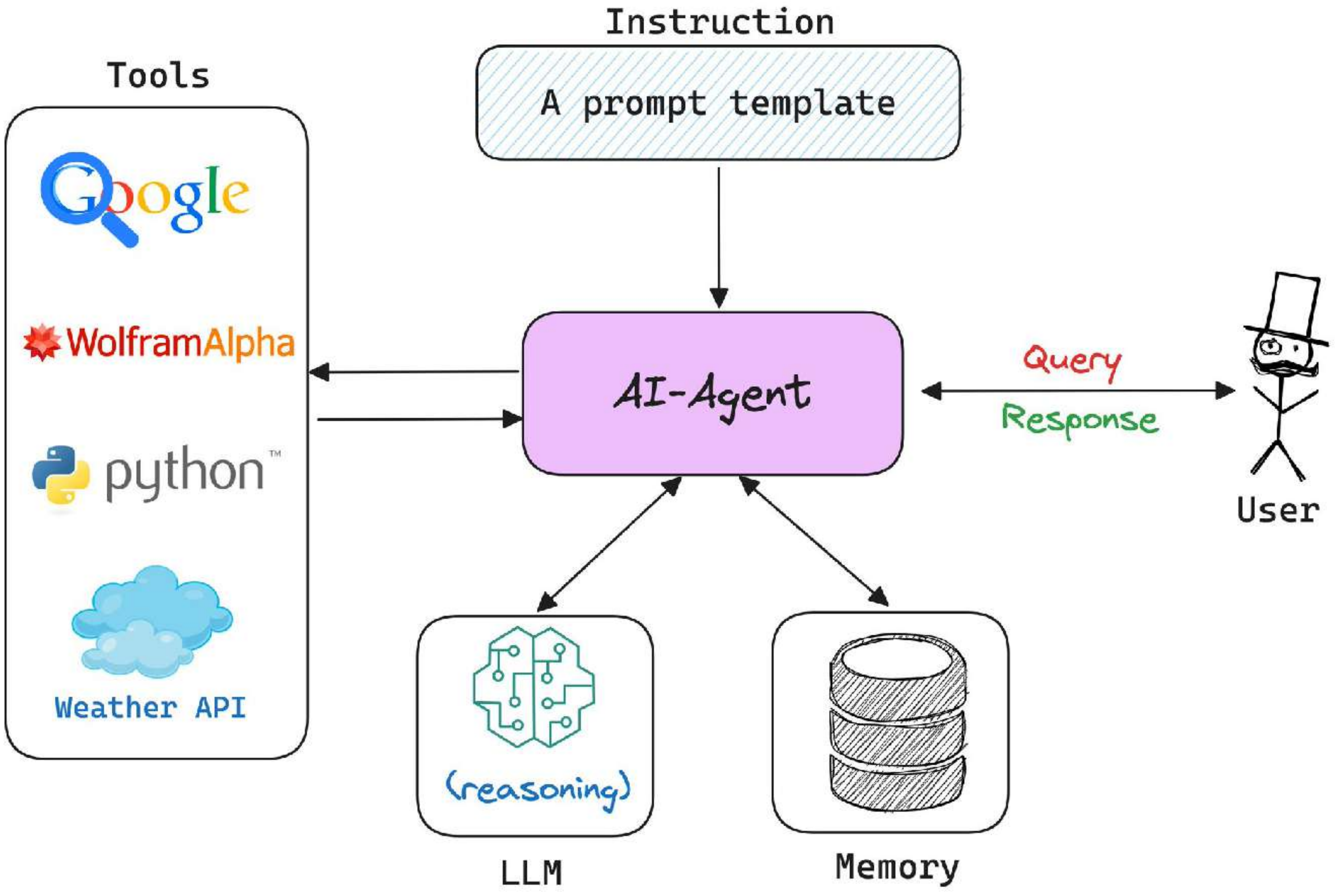
if you value intelligence above all other human qualities, you’re gonna have a bad time

6:10 PM · Oct 6, 2023 · **6.5M** Views

Agenda

- What is an AI Agent?
- What’s possible?
- What’s coming?

What is an AI Agent?

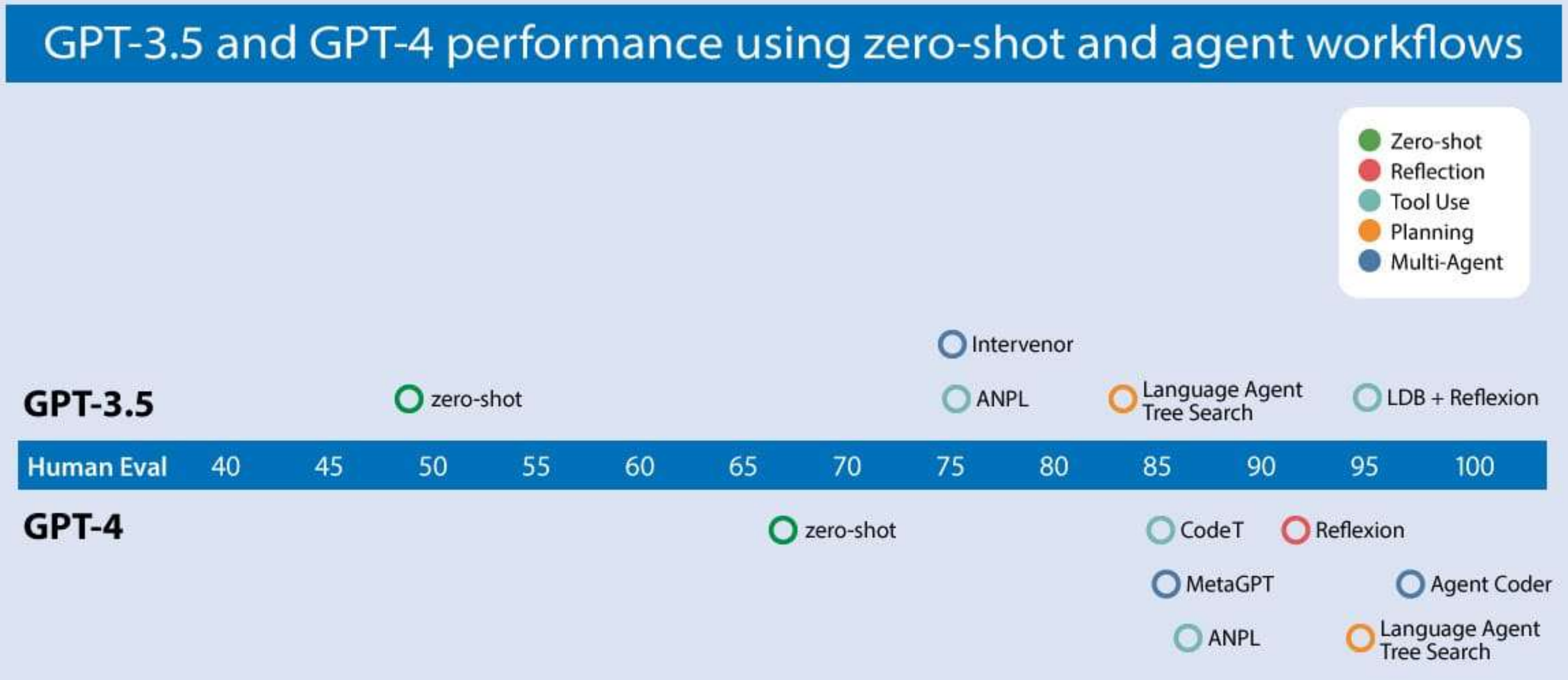


@akshay_pachaar

> “I think AI agent workflows will drive massive AI progress this year — perhaps even more than the next generation of foundation models. This is an important trend, and I urge everyone who works in AI to pay attention to it.”

Andrew Ng, AI Fund, Stanford

Agent on gpt-3.5 > gpt-4.5



GPT-3.5 and GPT-4 performance using zero-shot and agent workflows

Performance of GPT-3.5 and GPT-4 (zero-shot) on HumanEval, along with algorithms that use agent workflows on top of GPT-3.5 or GPT-4. Thanks to Joaquin Dominguez and John Santerre for help with this analysis.

> “Compared to training methods, OpenAI is currently paying more attention to the changes in the Agent domain. Whenever new AI Agents papers come out, there is excitement within the organization, and serious discussions take place”

Andrej Karpathy, ex-Open AI, ex-Tesla

What's possible?

Feel → Think → Act

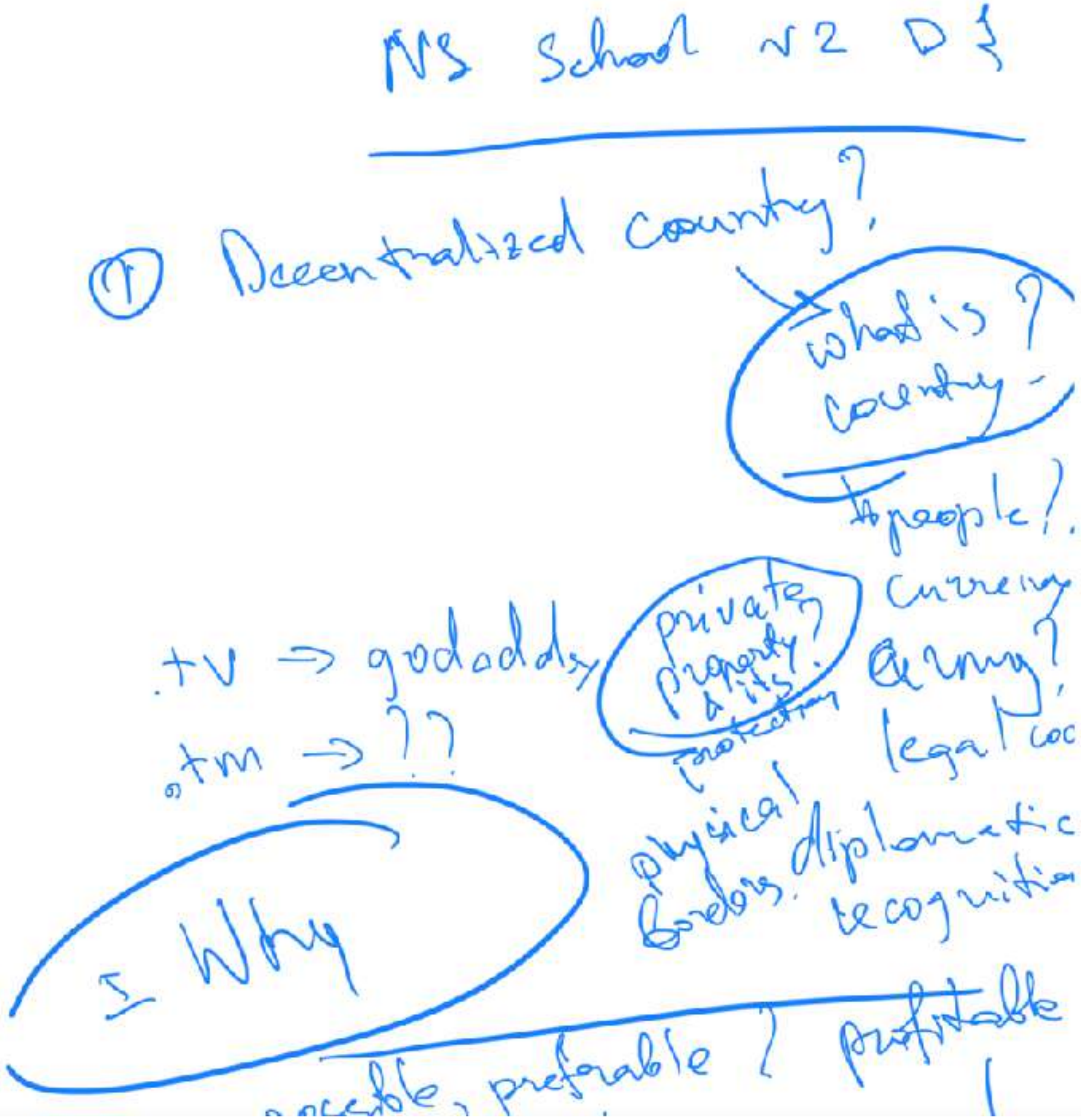
see hear

fast slow

online offline

Feel → →

see



NS School v2 D{

① Decentralized country?

What is? country

-> people?

Currency?

Army?

legal loc

private property? & its protection

physical borders. diplomatic recognition

.tv -> godaddy

.tm -> ??

I Why

ceable, preferable? profitable

Top Line:

- **“NS School v2 D...”** (the last character looks like `{` or `}`)

1) “Decentralized country?”

- In a circle or bubble:
  - **“What is? country –> to people? currency? Army? legal [loc?]”**
- Another note nearby:
  - **“private property? data? [taxation?]”**
  - **“physical borders, diplomatic recognition”**

Below that:

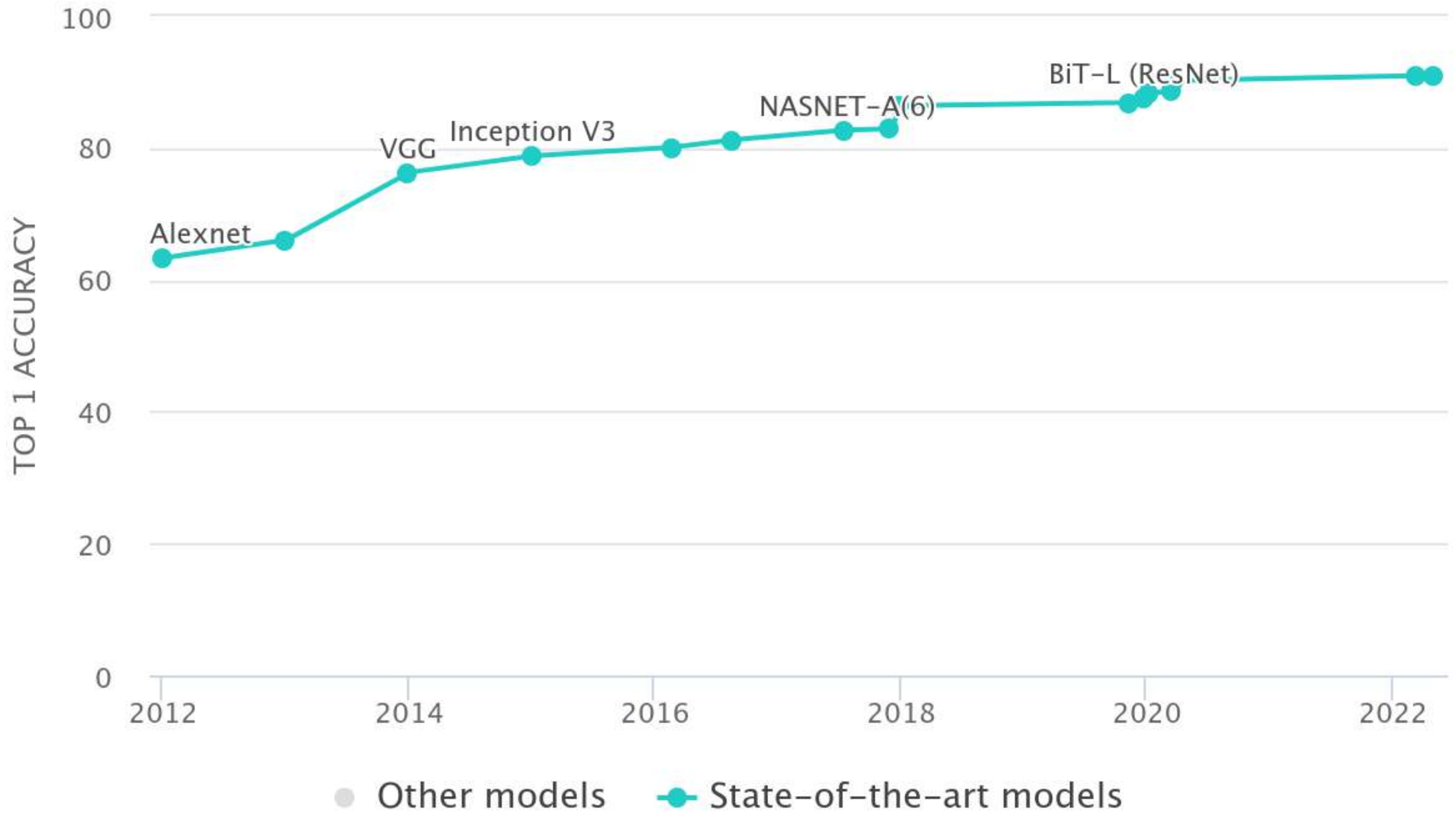
- **“.tv -> godaddy”**
- **“.tm -> ??”**

Large oval labeled: “I Why”

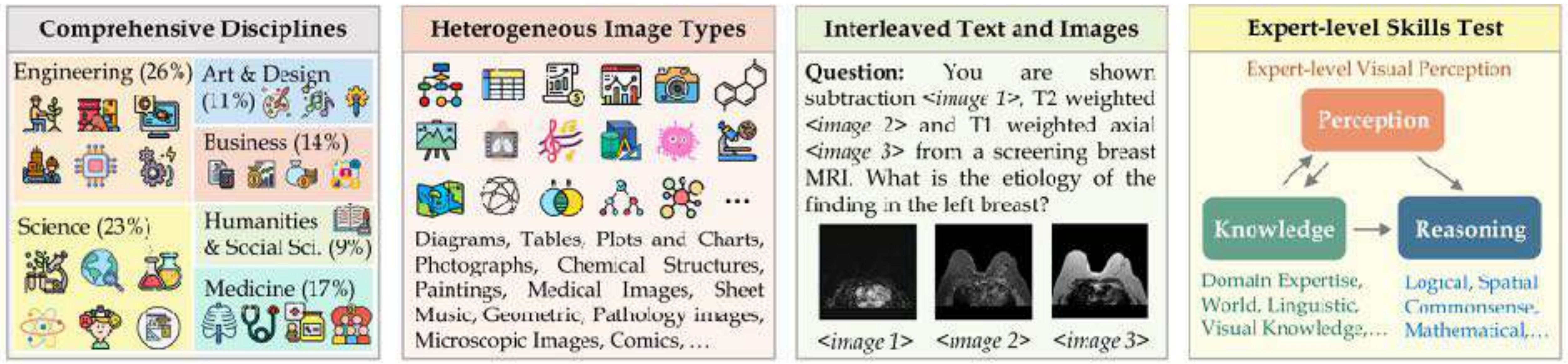
- **“people, preferable? profitable?”**

Image Recognition

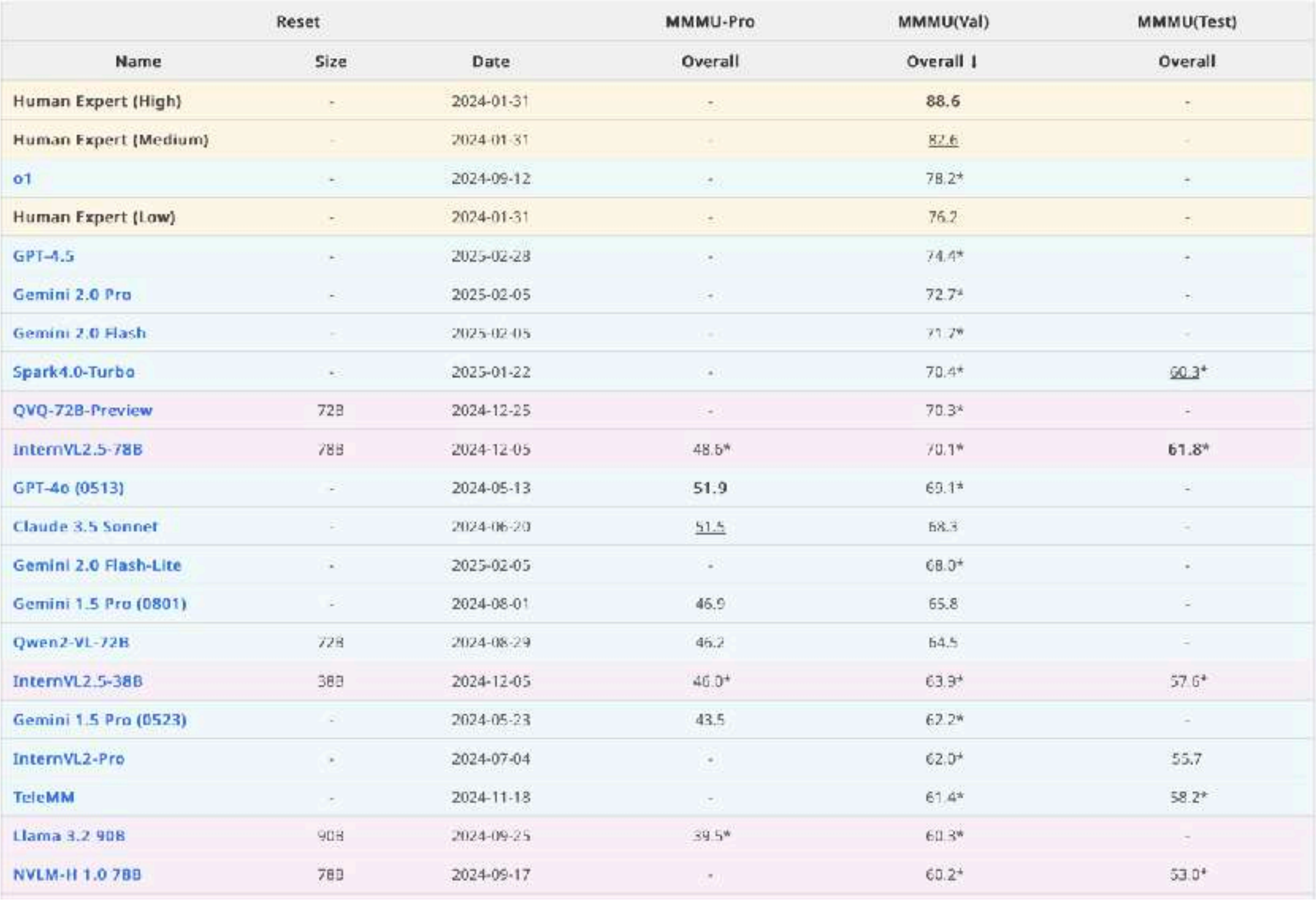
Above human level



On par with human experts


Comprehensive Disciplines
Engineering (26%)
Art & Design (11%)
Business (14%)
Science (23%)
Humanities & Social Sci. (9%)
Medicine (17%)
Heterogeneous Image Types
Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, …
Interleaved Text and Images
<image 1>
<image 2>
<image 3>
Expert-level Skills Test
Expert-level Visual Perception
Perception
Knowledge
Reasoning
Domain Expertise, World, Linguistic, Visual Knowledge,…
Logical, Spatial Commonsense, Mathematical,…

Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?

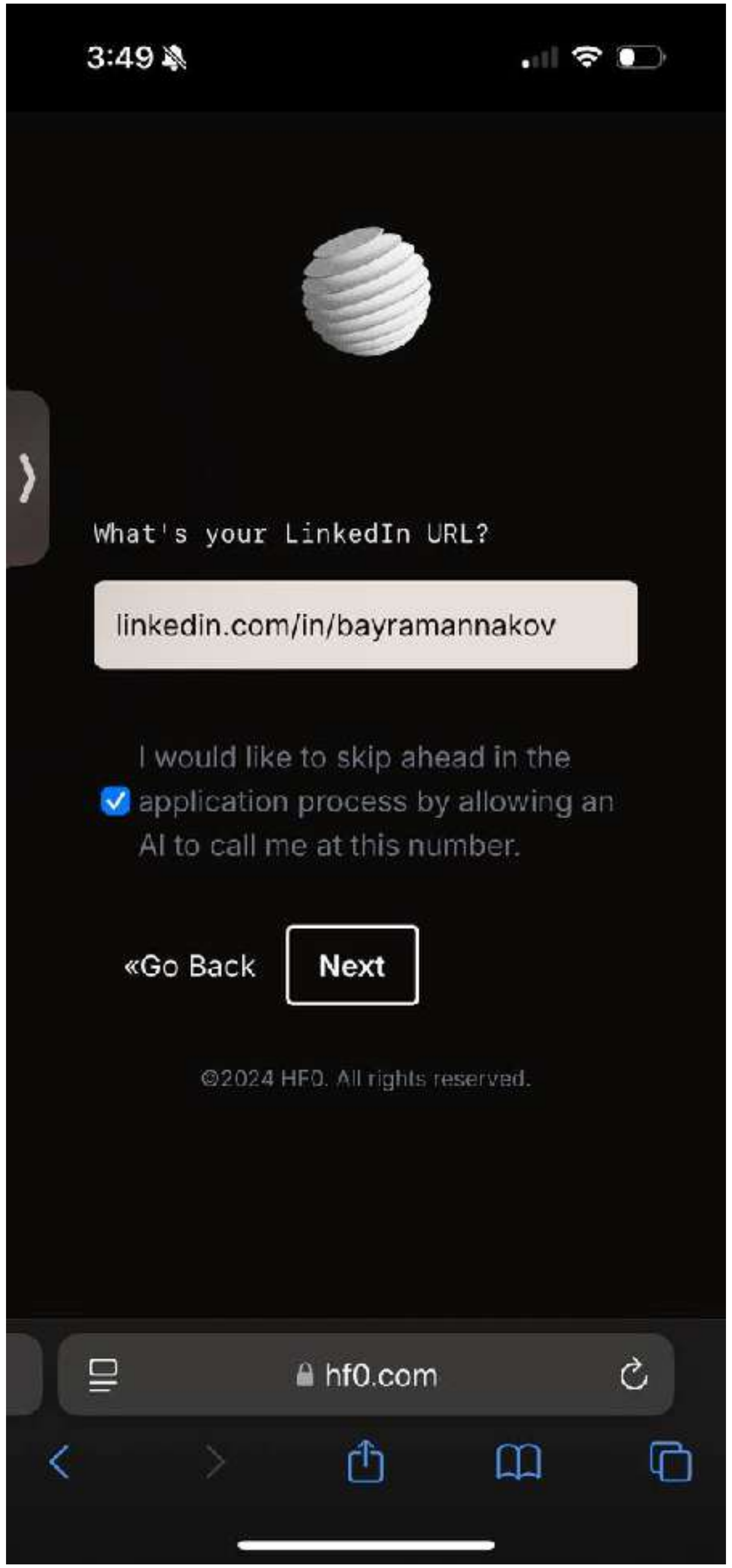


Reset			MMMU-Pro	MMMU(Val)	MMMU(Test)
Name	**Size**	**Date**	**Overall**	**Overall ↓**	**Overall**
Human Expert (High)	-	2024-01-31	-	**88.6**	-
Human Expert (Medium)	-	2024-01-31	-	82.6	-
o1	-	2024-09-12	-	78.2*	-
Human Expert (Low)	-	2024-01-31	-	76.2	-
GPT-4.5	-	2025-02-28	-	74.4*	-
Gemini 2.0 Pro	-	2025-02-05	-	72.7*	-
Gemini 2.0 Flash	-	2025-02-05	-	71.7*	-
Spark4.0-Turbo	-	2025-01-22	-	70.4*	60.3*
QVQ-72B-Preview	72B	2024-12-25	-	70.3*	-
InternVL2.5-78B	78B	2024-12-05	48.6*	70.1*	**61.8***
GPT-4o (0513)	-	2024-05-13	**51.9**	69.1*	-
Claude 3.5 Sonnet	-	2024-06-20	51.5	68.3	-
Gemini 2.0 Flash-Lite	-	2025-02-05	-	68.0*	-
Gemini 1.5 Pro (0801)	-	2024-08-01	46.9	65.8	-
Qwen2-VL-72B	72B	2024-08-29	46.2	64.5	-
InternVL2.5-38B	38B	2024-12-05	46.0*	63.9*	57.6*
Gemini 1.5 Pro (0523)	-	2024-05-23	43.5	62.2*	-
InternVL2-Pro	-	2024-07-04	-	62.0*	55.7
TeleMM	-	2024-11-18	-	61.4*	58.2*
Llama 3.2 90B	90B	2024-09-25	39.5*	60.3*	-
NVLM-H 1.0 78B	78B	2024-09-17	-	60.2*	53.0*

Feel → →

hear

SMART GLASSES LET
DEAF PEOPLE ‘SEE’ CHATS



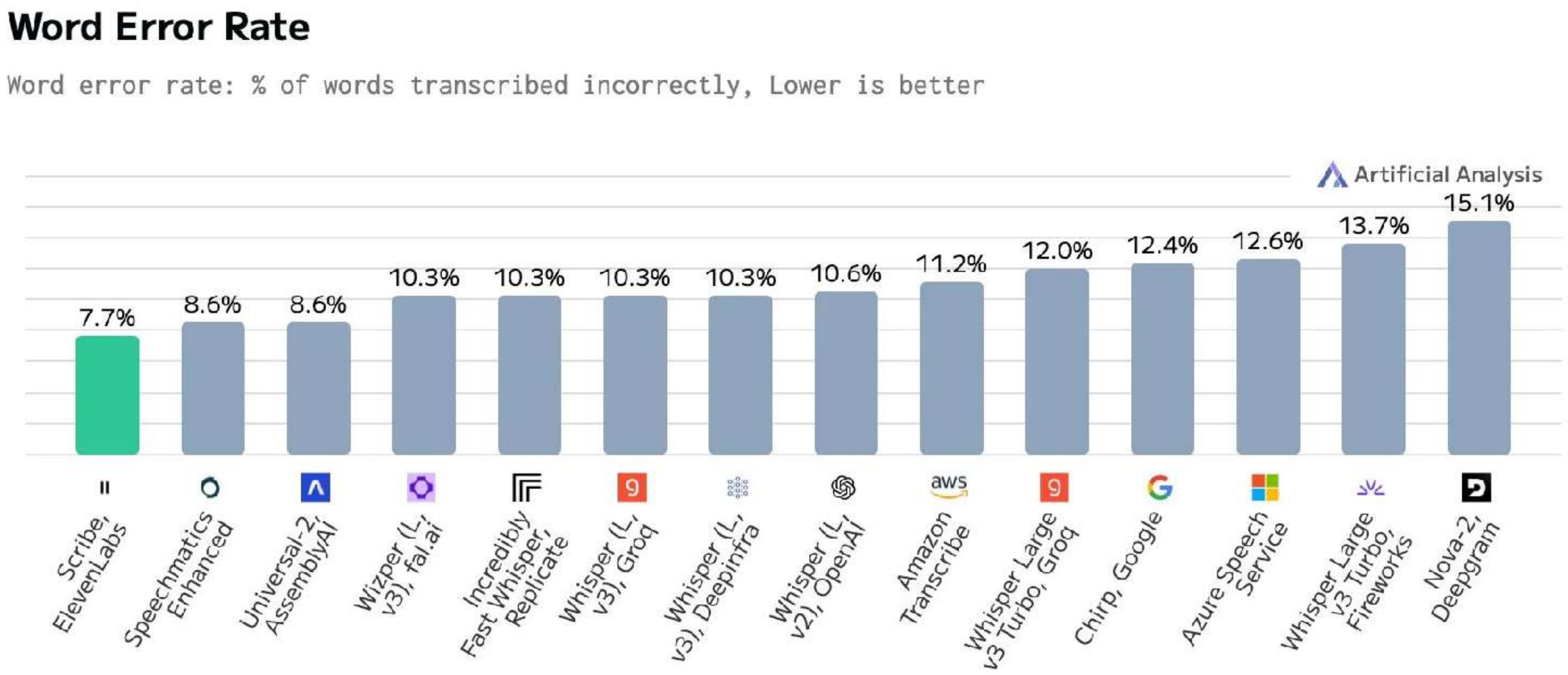
What's your LinkedIn URL?

linkedin.com/in/bayramannakov

☑ I would like to skip ahead in the application process by allowing an AI to call me at this number.

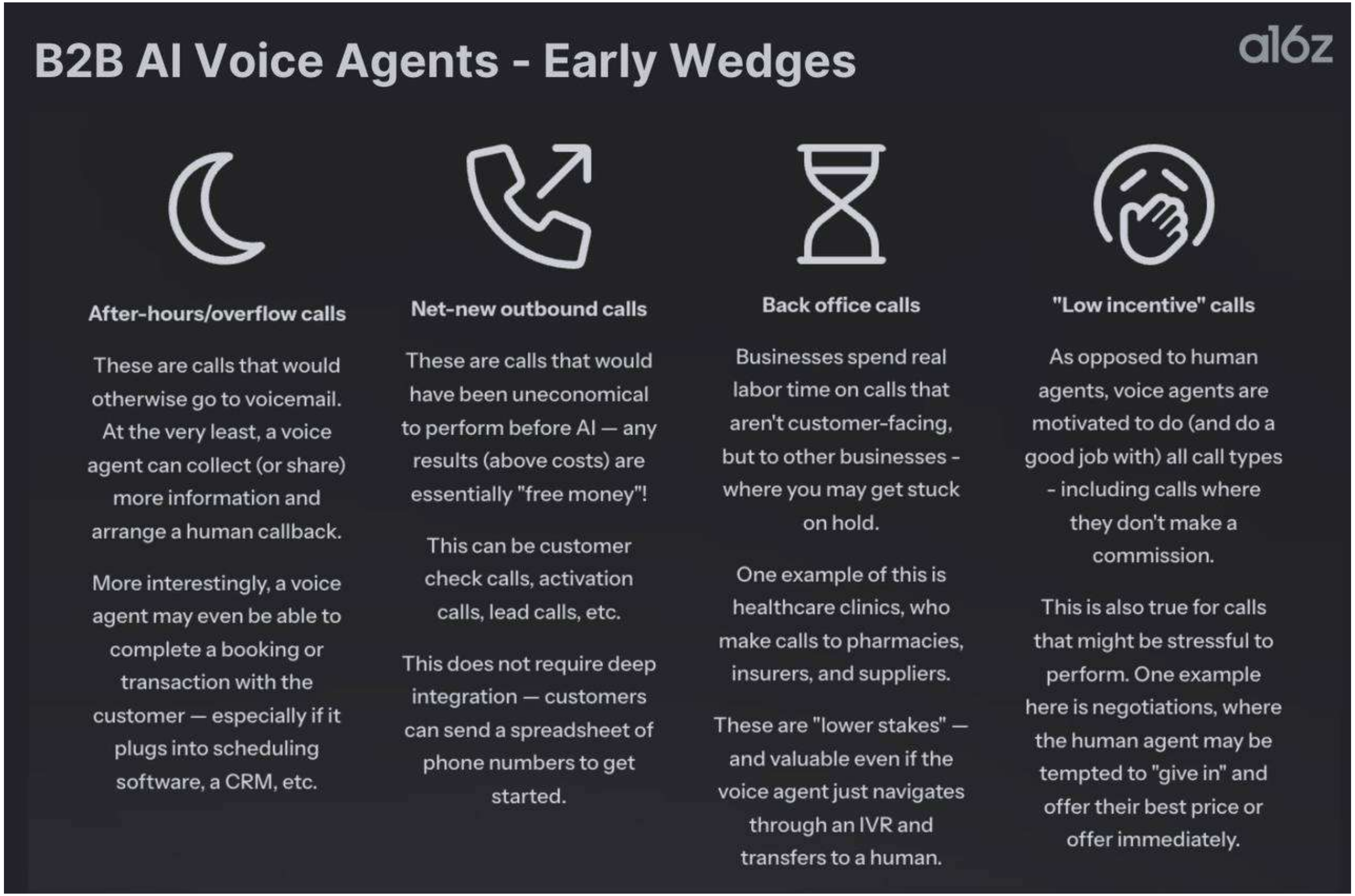
©2024 HF0. All rights reserved.

Speech Transcription



Word Error Rate

Word error rate: % of words transcribed incorrectly, Lower is better



B2B AI Voice Agents - Early Wedges

a16z

After-hours/overflow calls

These are calls that would otherwise go to voicemail. At the very least, a voice agent can collect (or share) more information and arrange a human callback.

More interestingly, a voice agent may even be able to complete a booking or transaction with the customer — especially if it plugs into scheduling software, a CRM, etc.

Net-new outbound calls

These are calls that would have been uneconomical to perform before AI — any results (above costs) are essentially "free money"!

This can be customer check calls, activation calls, lead calls, etc.

This does not require deep integration — customers can send a spreadsheet of phone numbers to get started.

Back office calls

Businesses spend real labor time on calls that aren't customer-facing, but to other businesses - where you may get stuck on hold.

One example of this is healthcare clinics, who make calls to pharmacies, insurers, and suppliers.

These are "lower stakes" — and valuable even if the voice agent just navigates through an IVR and transfers to a human.

"Low incentive" calls

As opposed to human agents, voice agents are motivated to do (and do a good job with) all call types - including calls where they don't make a commission.

This is also true for calls that might be stressful to perform. One example here is negotiations, where the human agent may be tempted to "give in" and offer their best price or offer immediately.

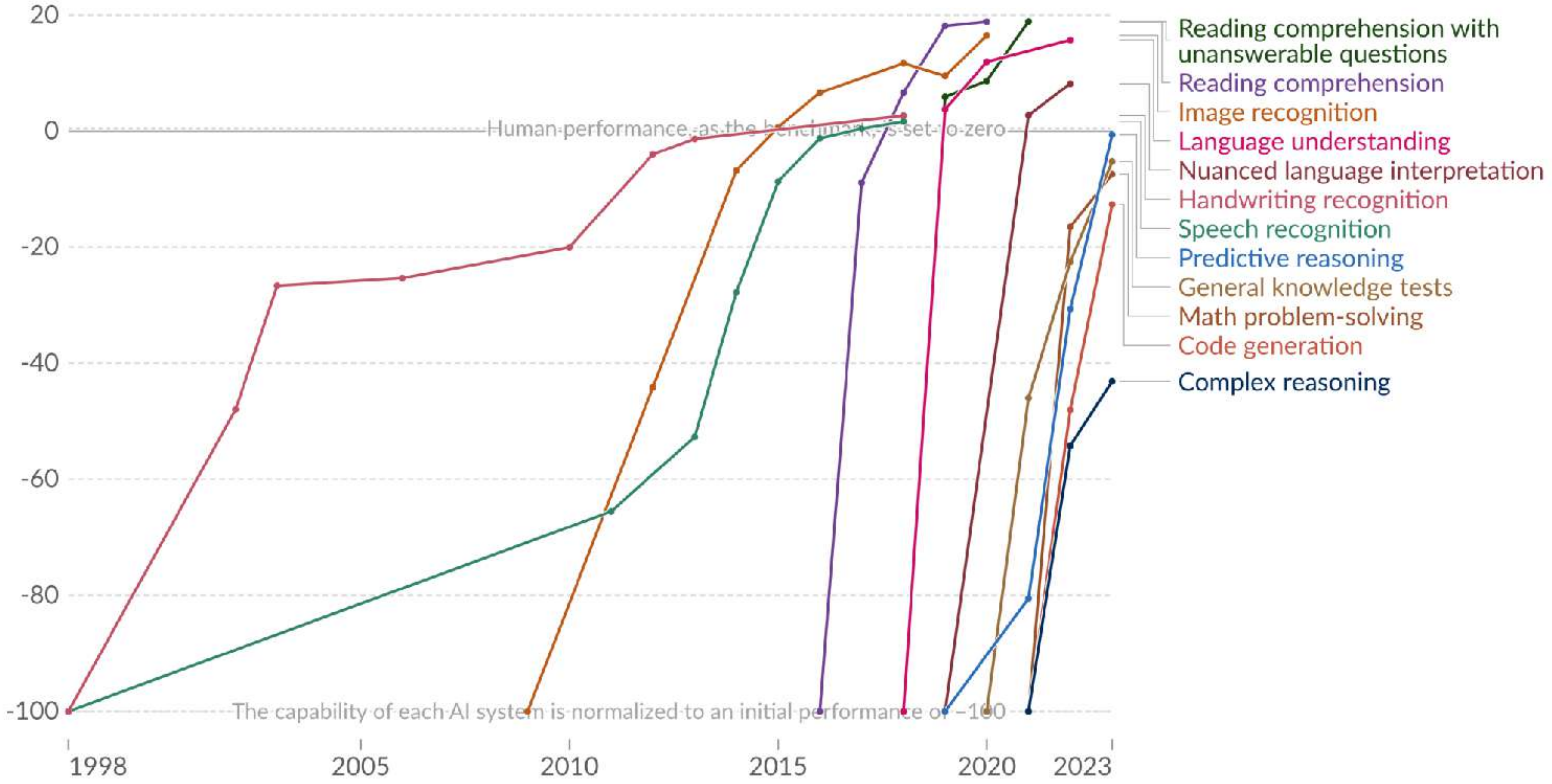
Think

fast

Test scores of AI systems on various capabilities relative to human performance

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.

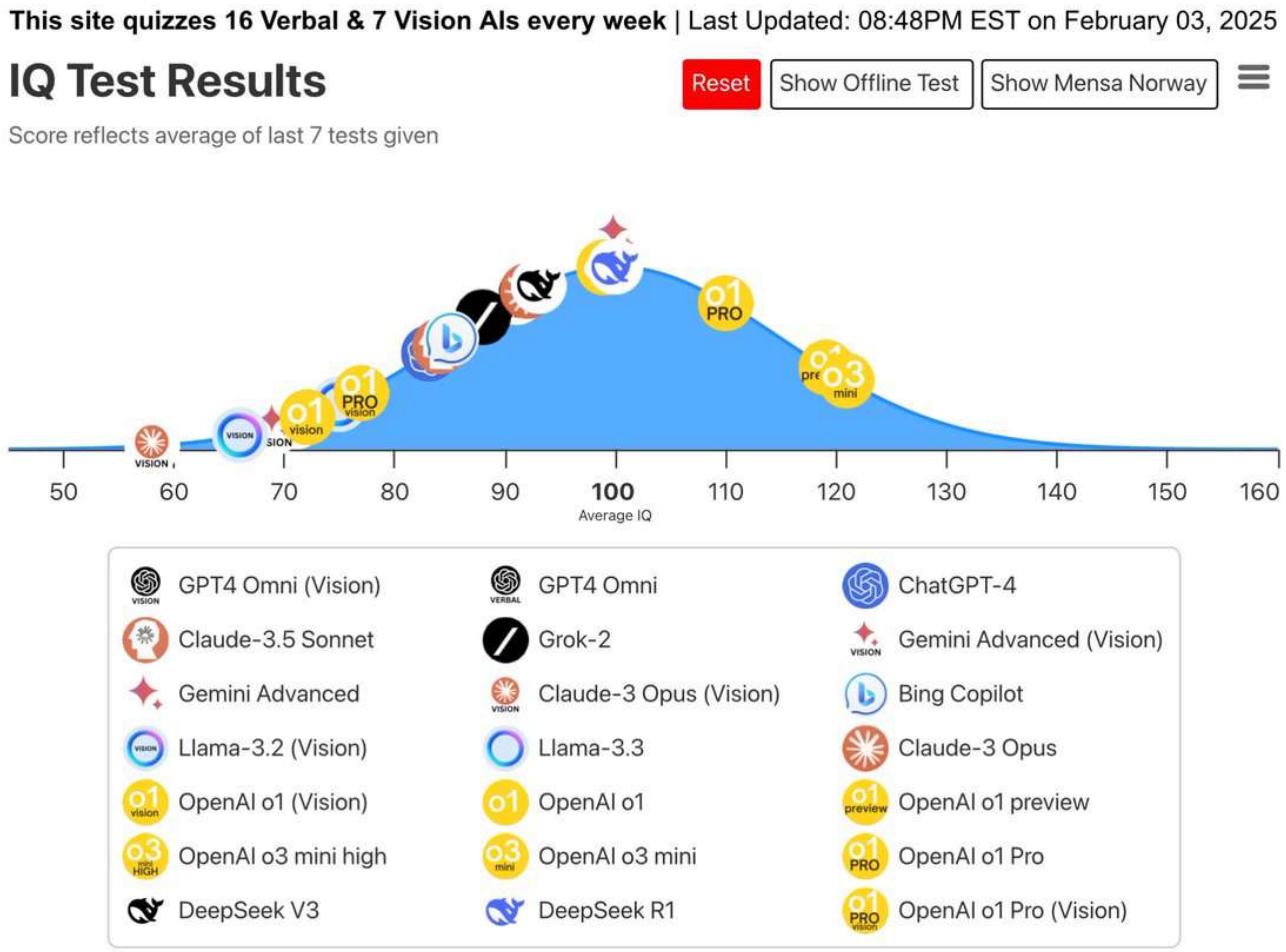
Our World in Data



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.



This site quizzes 16 Verbal & 7 Vision AIs every week | Last Updated: 08:48PM EST on February 03, 2025

IQ Test Results

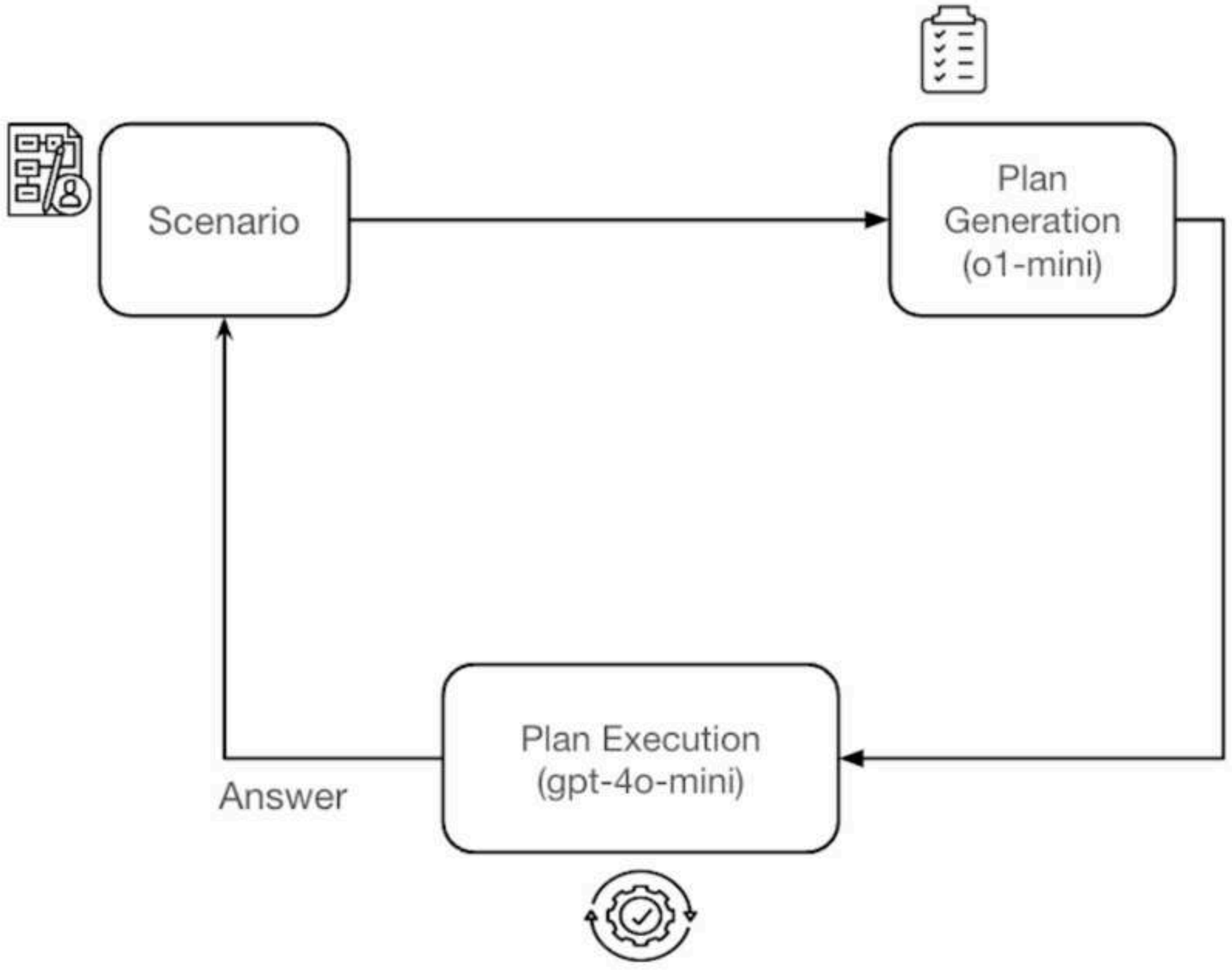
Score reflects average of last 7 tests given

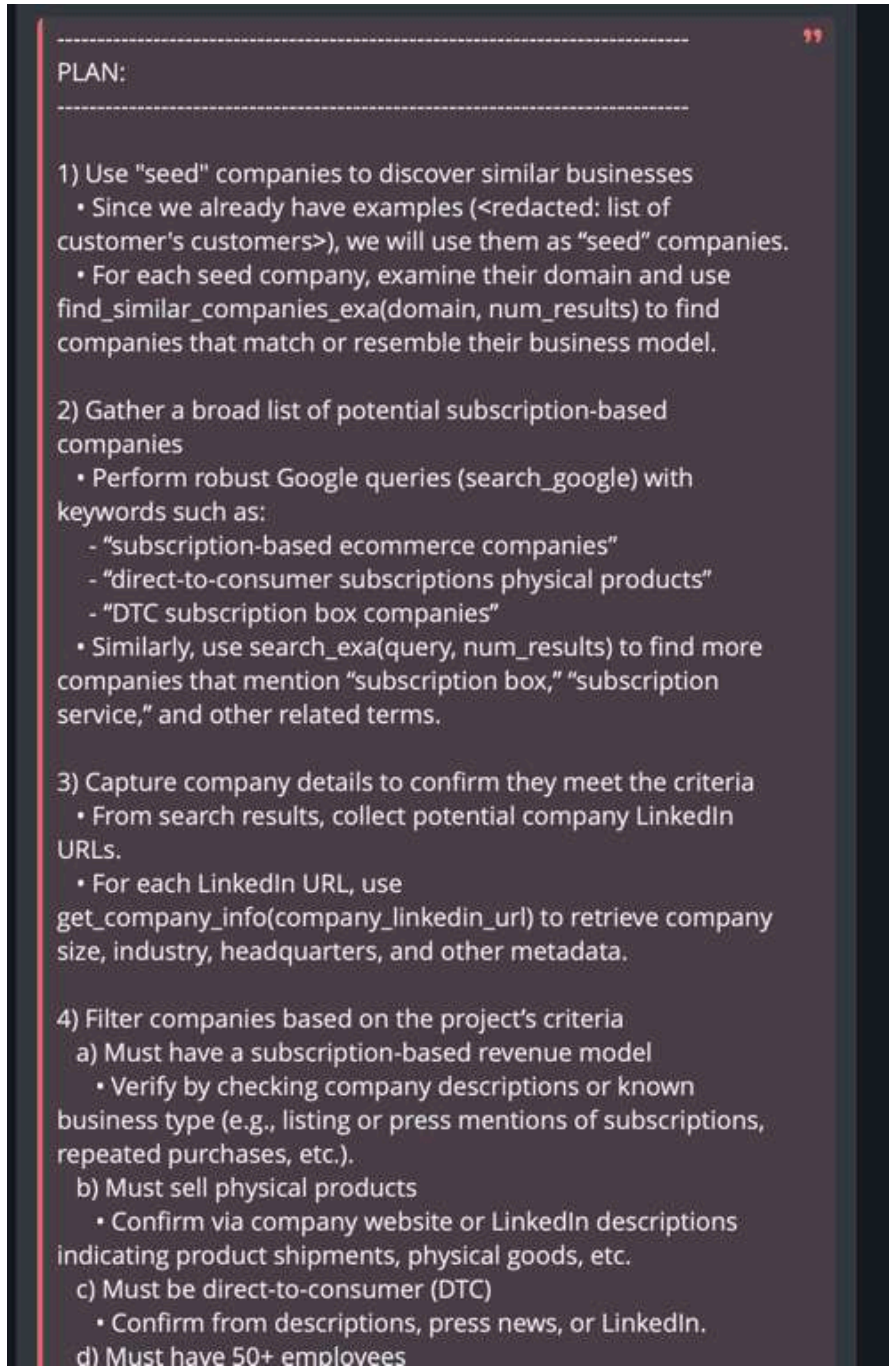
- GPT4 Omni (Vision)
- GPT4 Omni
- ChatGPT-4
- Claude-3.5 Sonnet
- Grok-2
- Gemini Advanced (Vision)
- Gemini Advanced
- Claude-3 Opus (Vision)
- Bing Copilot
- Llama-3.2 (Vision)
- Llama-3.3
- Claude-3 Opus
- OpenAI o1 (Vision)
- OpenAI o1
- OpenAI o1 preview
- OpenAI o3 mini high
- OpenAI o3 mini
- OpenAI o1 Pro
- DeepSeek V3
- DeepSeek R1
- OpenAI o1 Pro (Vision)

→ Think →

slow

o1 to plan, 4o to execute





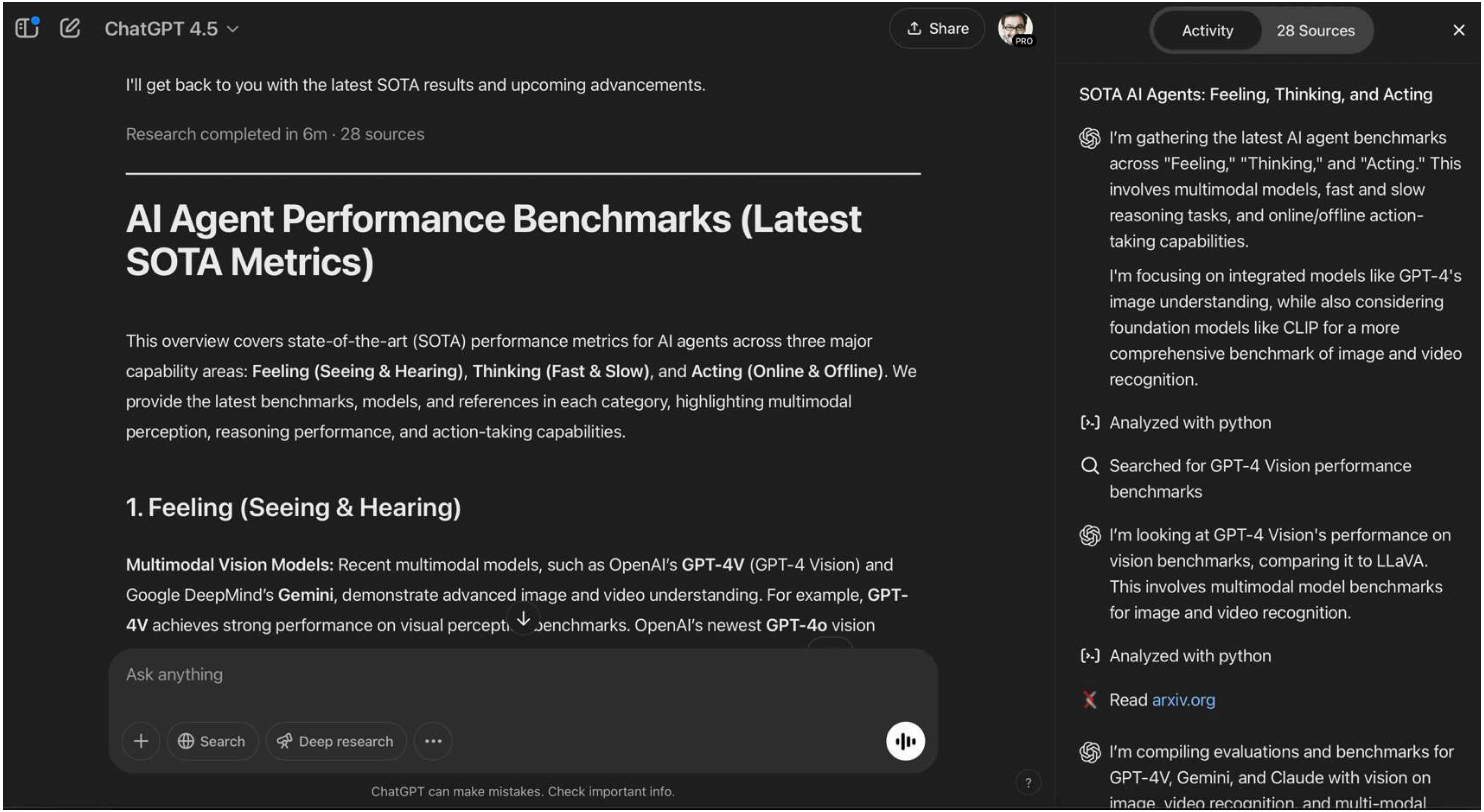
```
--------------------------------------------------------------------------------
PLAN:
--------------------------------------------------------------------------------

1) Use "seed" companies to discover similar businesses
  • Since we already have examples (<redacted: list of
customer's customers>), we will use them as “seed” companies.
  • For each seed company, examine their domain and use
find_similar_companies_exa(domain, num_results) to find
companies that match or resemble their business model.

2) Gather a broad list of potential subscription-based
companies
  • Perform robust Google queries (search_google) with
keywords such as:
    - “subscription-based ecommerce companies”
    - “direct-to-consumer subscriptions physical products”
    - “DTC subscription box companies”
  • Similarly, use search_exa(query, num_results) to find more
companies that mention “subscription box,” “subscription
service,” and other related terms.

3) Capture company details to confirm they meet the criteria
  • From search results, collect potential company LinkedIn
URLs.
  • For each LinkedIn URL, use
get_company_info(company_linkedin_url) to retrieve company
size, industry, headquarters, and other metadata.

4) Filter companies based on the project’s criteria
  a) Must have a subscription-based revenue model
    • Verify by checking company descriptions or known
business type (e.g., listing or press mentions of subscriptions,
repeated purchases, etc.).
  b) Must sell physical products
    • Confirm via company website or LinkedIn descriptions
indicating product shipments, physical goods, etc.
  c) Must be direct-to-consumer (DTC)
    • Confirm from descriptions, press news, or LinkedIn.
  d) Must have 50+ employees
```



I'll get back to you with the latest SOTA results and upcoming advancements.

Research completed in 6m · 28 sources

AI Agent Performance Benchmarks (Latest SOTA Metrics)

This overview covers state-of-the-art (SOTA) performance metrics for AI agents across three major capability areas: **Feeling (Seeing & Hearing)**, **Thinking (Fast & Slow)**, and **Acting (Online & Offline)**. We provide the latest benchmarks, models, and references in each category, highlighting multimodal perception, reasoning performance, and action-taking capabilities.

1. Feeling (Seeing & Hearing)

Multimodal Vision Models: Recent multimodal models, such as OpenAI's **GPT-4V** (GPT-4 Vision) and Google DeepMind's **Gemini**, demonstrate advanced image and video understanding. For example, **GPT-4V** achieves strong performance on visual percept benchmarks. OpenAI's newest **GPT-4o** vision

SOTA AI Agents: Feeling, Thinking, and Acting

I'm gathering the latest AI agent benchmarks across "Feeling," "Thinking," and "Acting." This involves multimodal models, fast and slow reasoning tasks, and online/offline action-taking capabilities.

I'm focusing on integrated models like GPT-4's image understanding, while also considering foundation models like CLIP for a more comprehensive benchmark of image and video recognition.

Analyzed with python

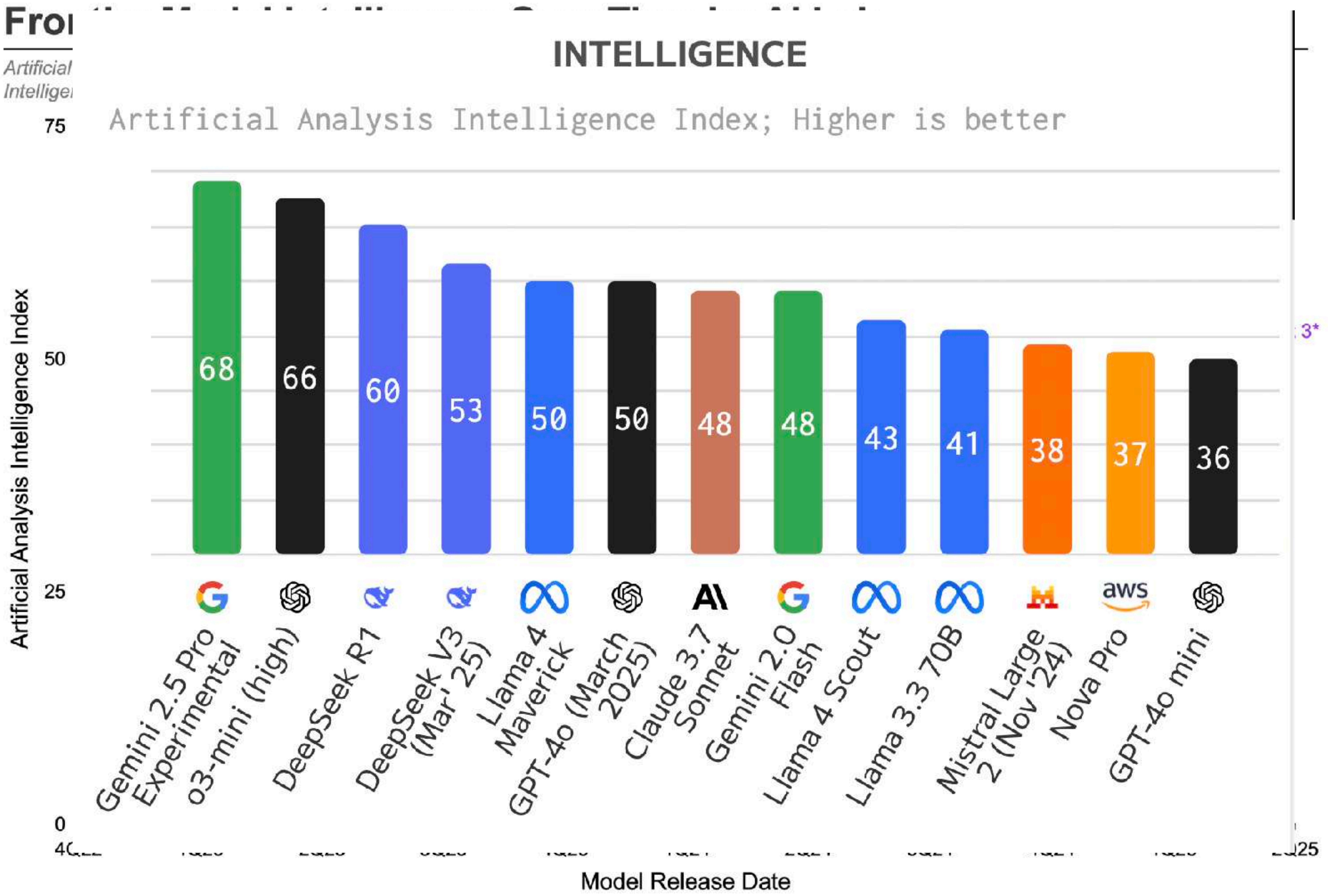
Searched for GPT-4 Vision performance benchmarks

I'm looking at GPT-4 Vision's performance on vision benchmarks, comparing it to LLaVA. This involves multimodal model benchmarks for image and video recognition.

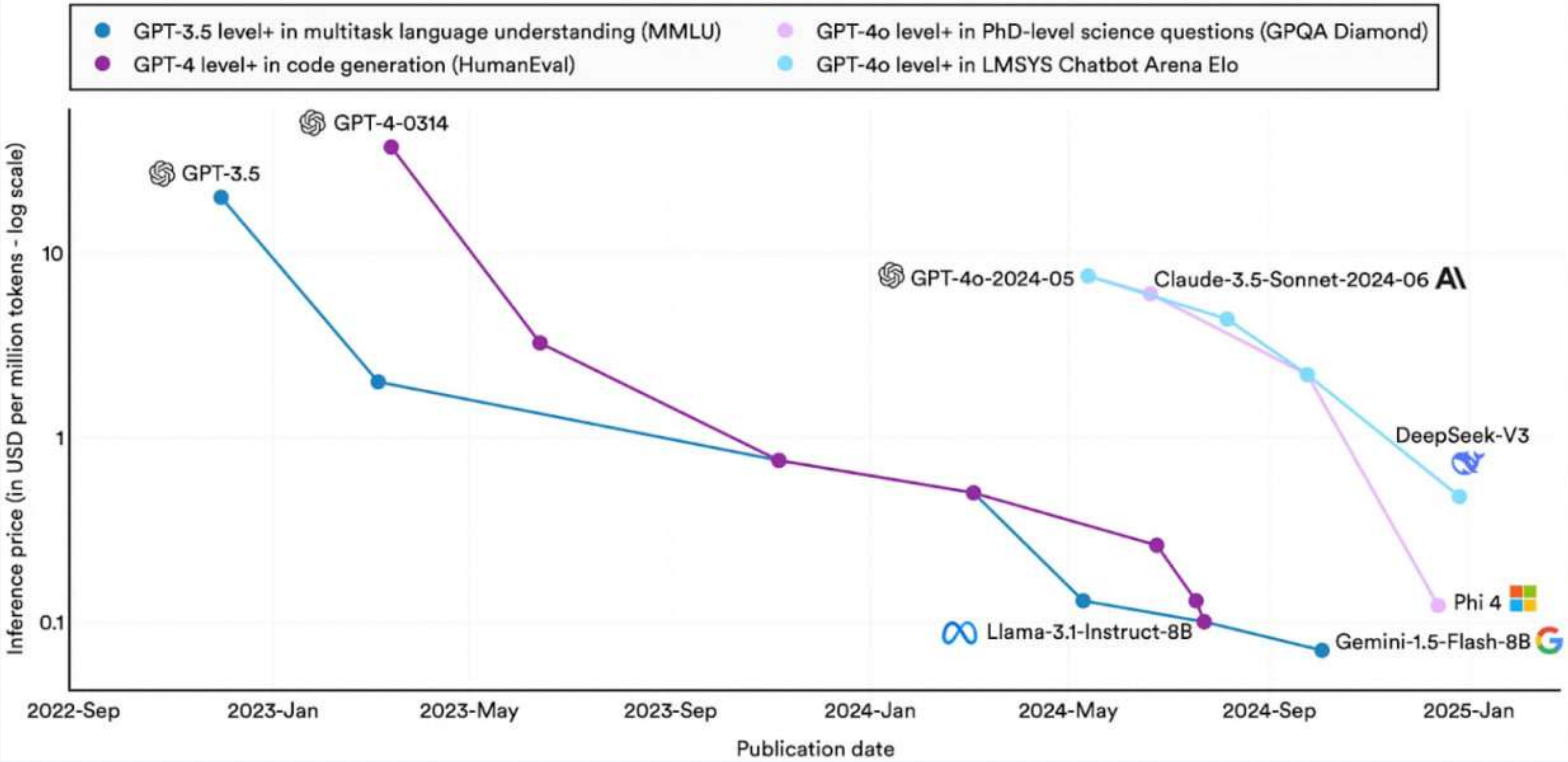
Analyzed with python

Read arxiv.org

I'm compiling evaluations and benchmarks for GPT-4V, Gemini, and Claude with vision on image, video recognition, and multi-modal

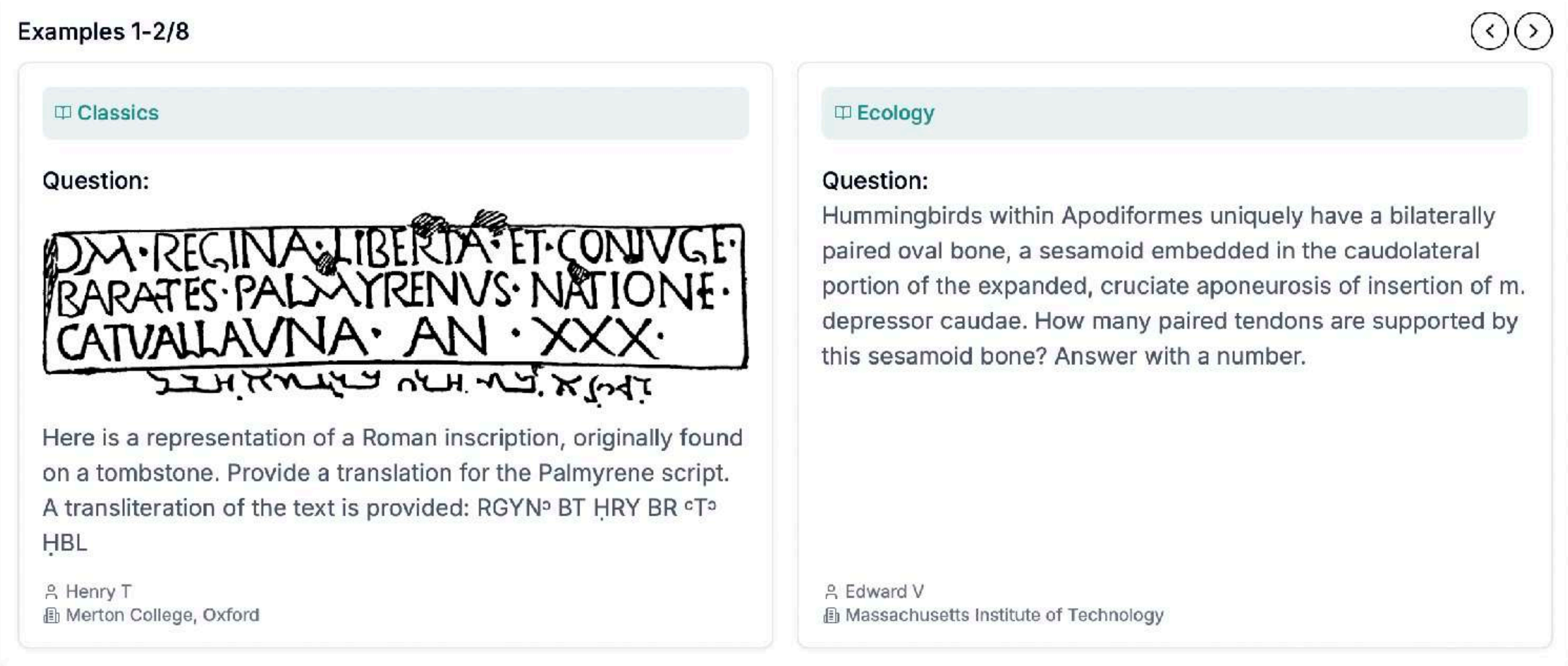

Fro
Artificial
Intellige
INTELLIGENCE
Artificial Analysis Intelligence Index; Higher is better
75
50
25
0
Artificial Analysis Intelligence Index
68
66
60
53
50
50
48
48
43
41
38
37
36
Gemini 2.5 Pro Experimental
o3-mini (high)
DeepSeek R1
DeepSeek V3 (Mar' 25)
Llama 4 Maverick
GPT-4o (March 2025)
Claude 3.7 Sonnet
Gemini 2.0 Flash
Llama 4 Scout
Llama 3.3 70B
Mistral Large 2 (Nov '24)
Nova Pro
GPT-4o mini
3*
4Q22
2Q25
Model Release Date

Models Become Cheaper to Use



Inference price across select benchmarks, 2022–24

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


D M · REGINA · LIBERTA · ET · CONIVGE ·
BARATES · PALMYRENVS · NATIONE ·
CATVALLAVNA · AN · XXX ·

Examples 1-2/8

Classics

Question:

Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵓ BT ḤRY BR ᶜTᵓ ḤBL

Henry T
Merton College, Oxford

Ecology

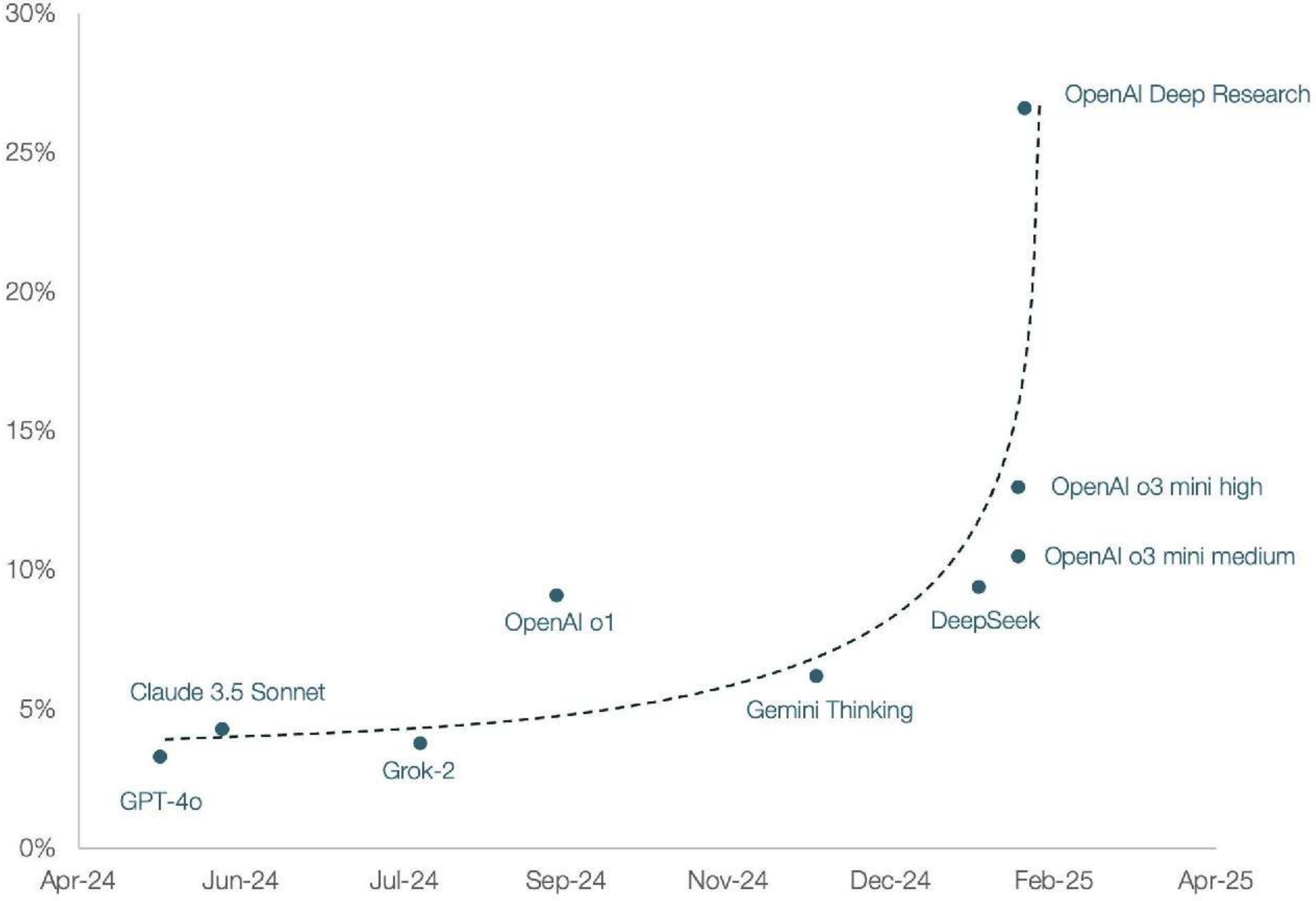
Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Edward V
Massachusetts Institute of Technology

Samples of the diverse and challenging questions submitted to Humanity's Last Exam.

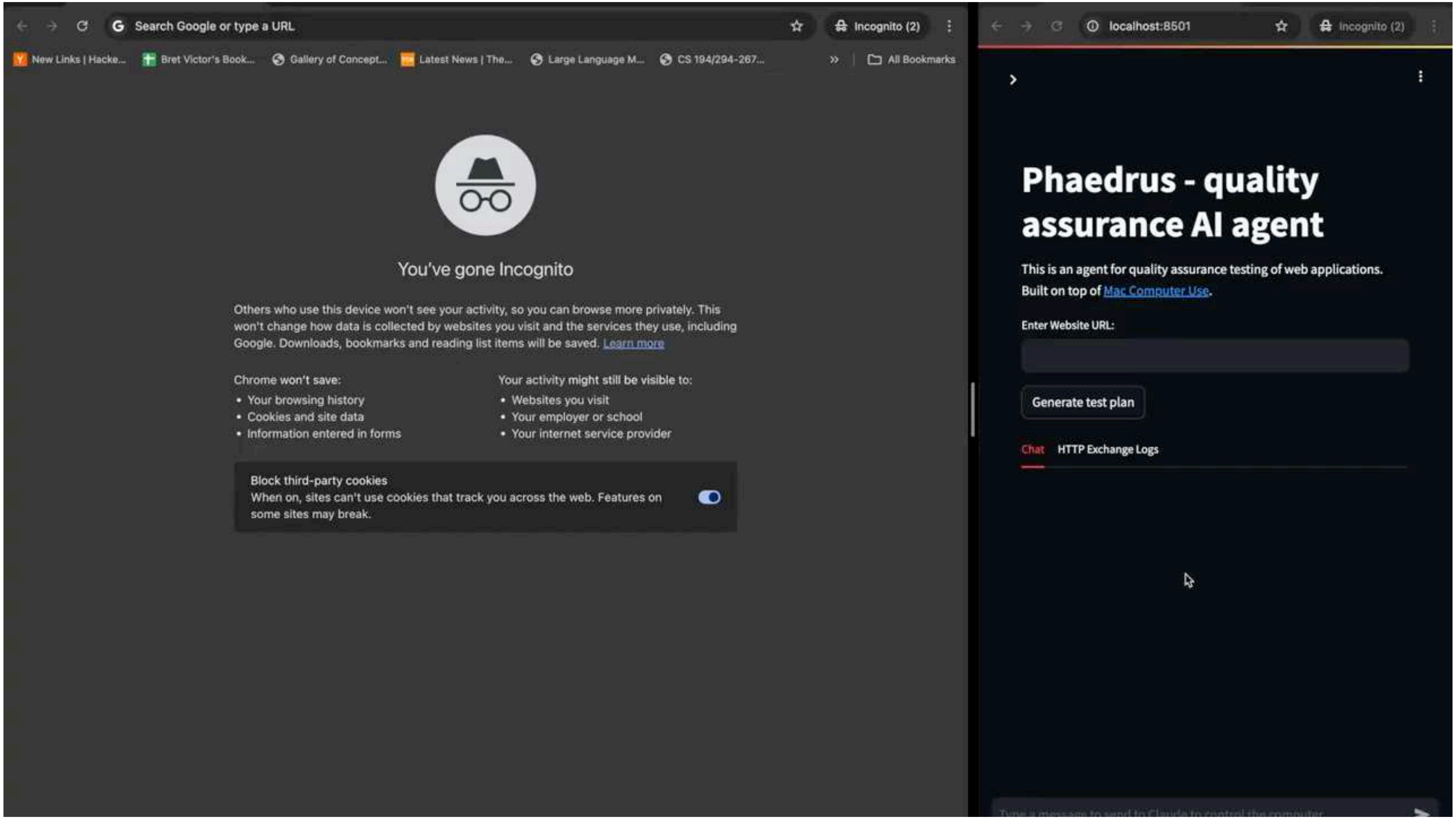
AI Scores on *Humanity's Last Exam*



Source: Tomas Pueyo for Uncharted Territories, with data from Dan Hendrycks, of Humanity's Last Exam

→ → Act

online



You've gone Incognito

Others who use this device won't see your activity, so you can browse more privately. This won't change how data is collected by websites you visit and the services they use, including Google. Downloads, bookmarks and reading list items will be saved. Learn more

Chrome won't save:
- Your browsing history
- Cookies and site data
- Information entered in forms

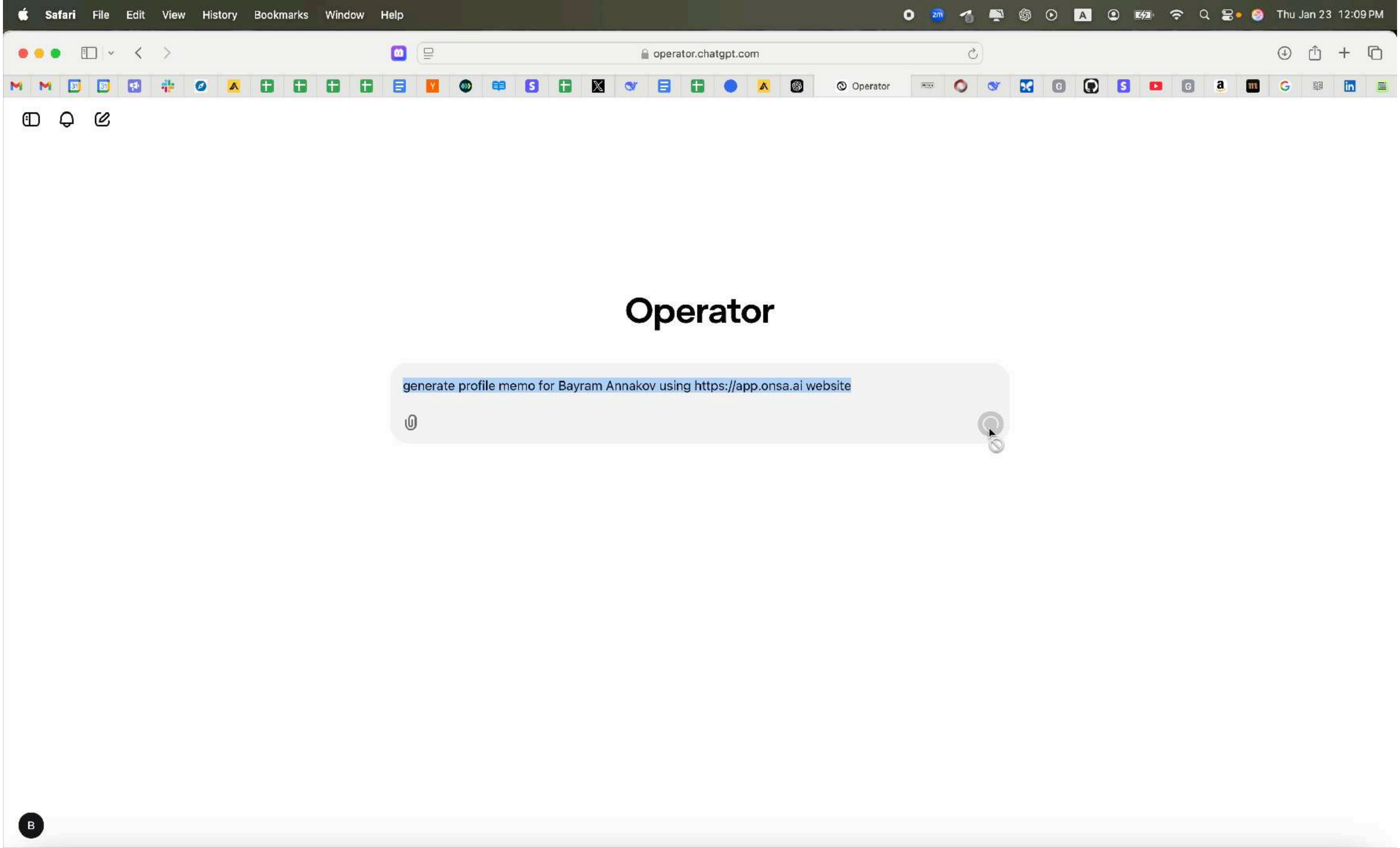
Your activity might still be visible to:
- Websites you visit
- Your employer or school
- Your internet service provider

Block third-party cookies
When on, sites can't use cookies that track you across the web. Features on some sites may break.

Phaedrus - quality assurance AI agent

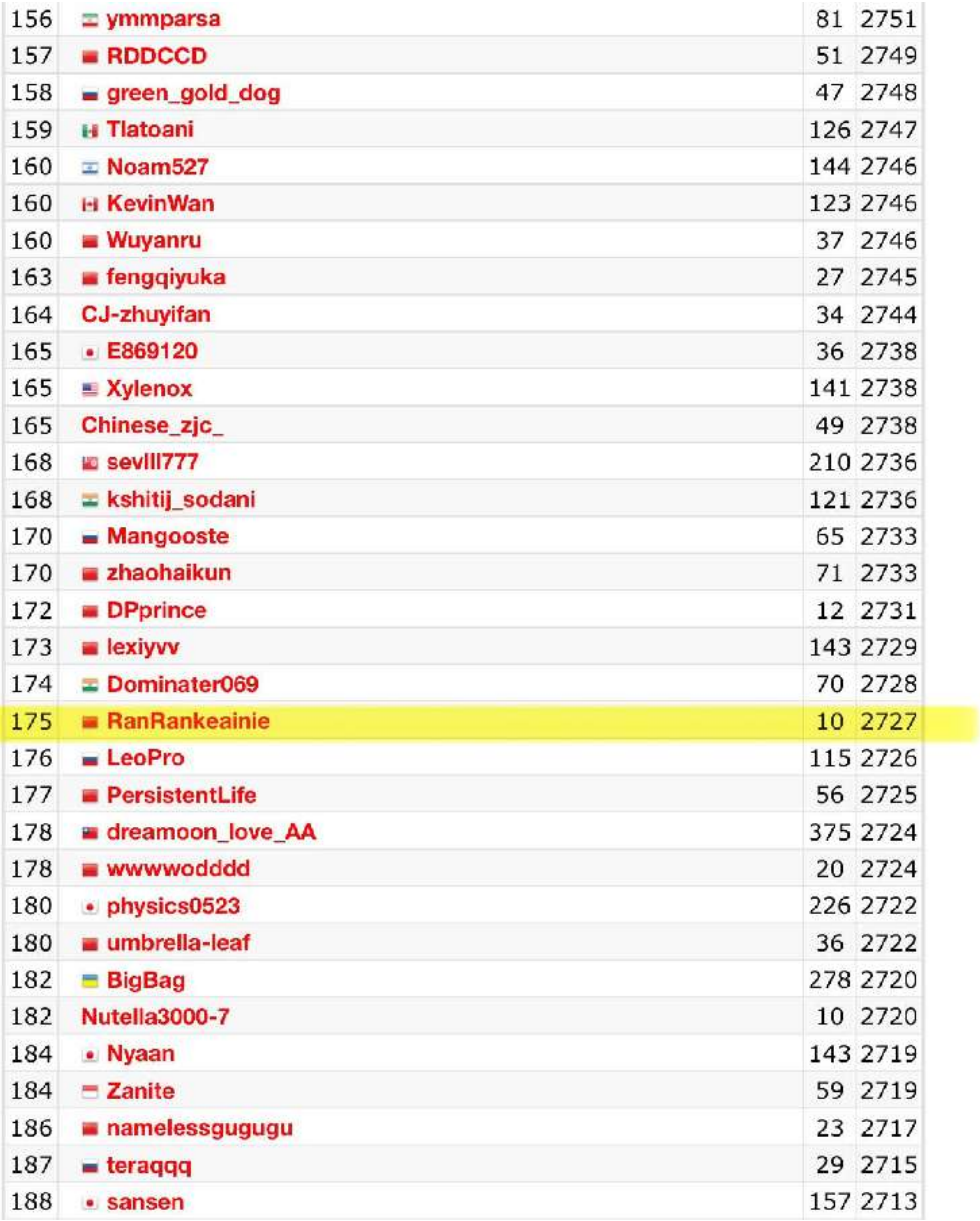
This is an agent for quality assurance testing of web applications. Built on top of Mac Computer Use.

Enter Website URL:

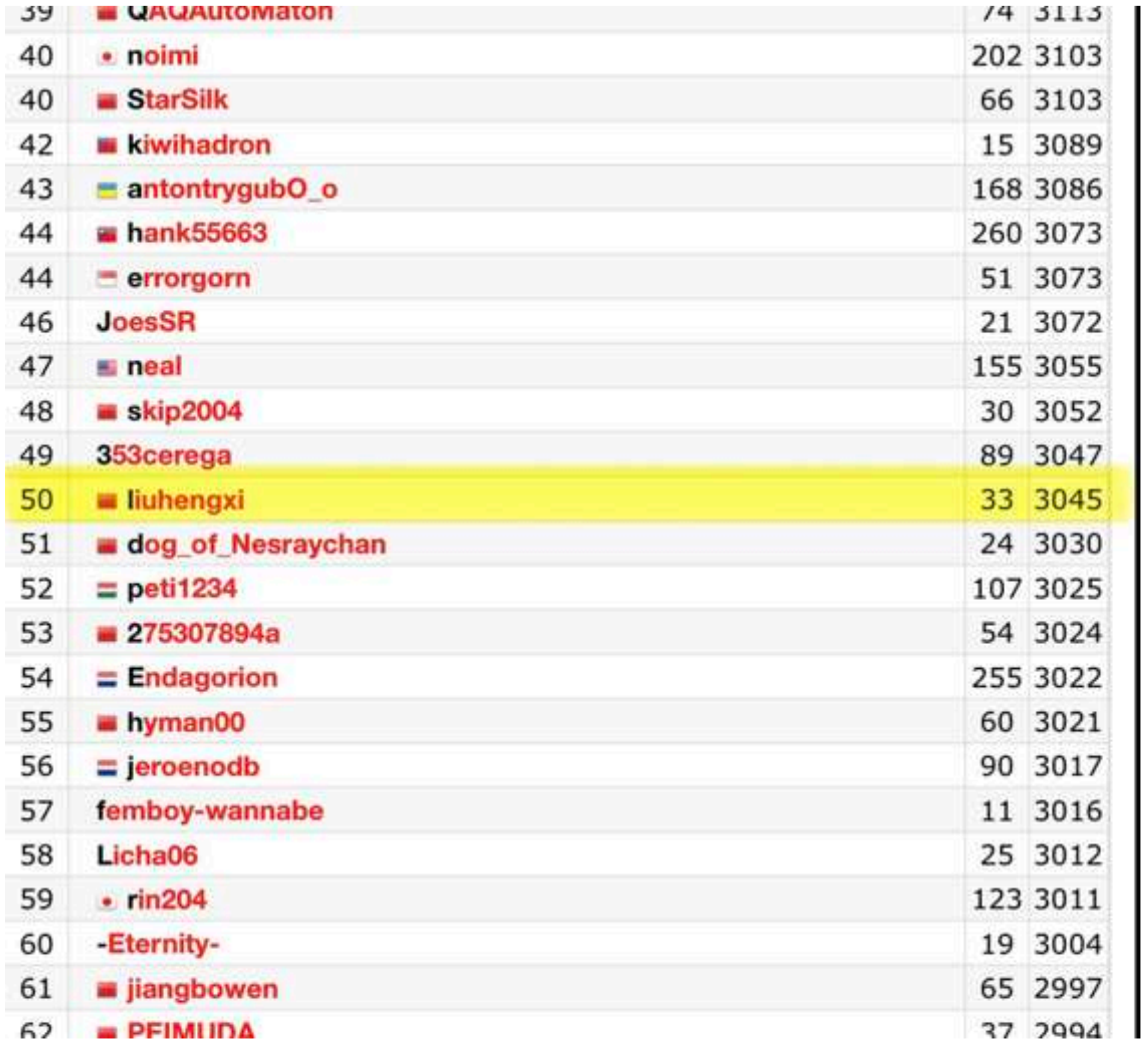


Operator

generate profile memo for Bayram Annakov using https://app.onsa.ai website



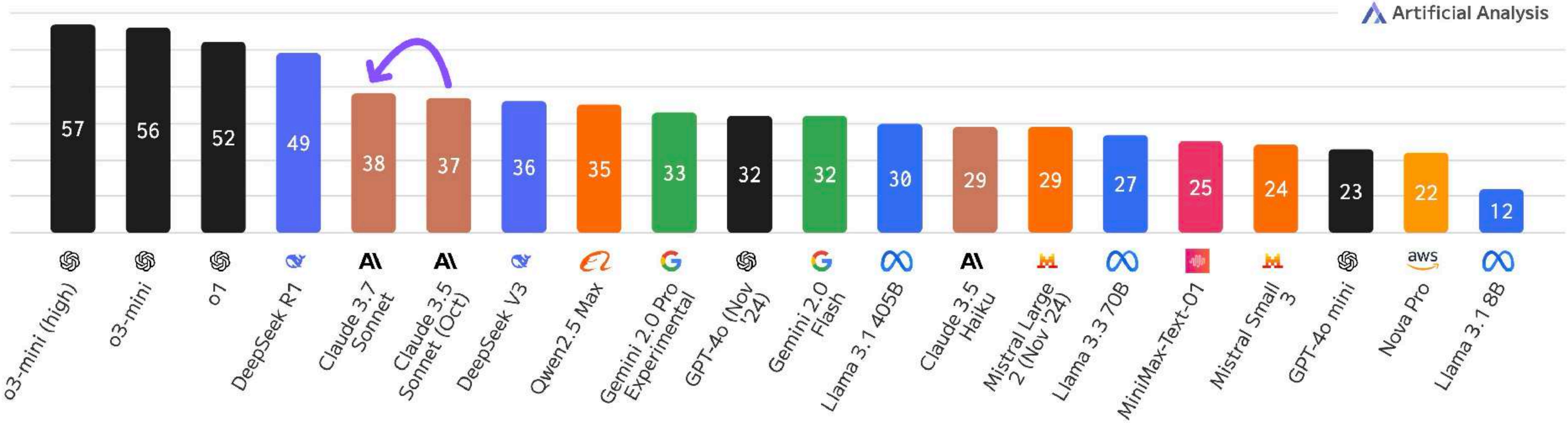
156	ymmparsa	81	2751
157	RDDCCD	51	2749
158	green_gold_dog	47	2748
159	Tlatoani	126	2747
160	Noam527	144	2746
160	KevinWan	123	2746
160	Wuyanru	37	2746
163	fengqiyuka	27	2745
164	CJ-zhuyifan	34	2744
165	E869120	36	2738
165	Xylenox	141	2738
165	Chinese_zjc_	49	2738
168	sevlll777	210	2736
168	kshitij_sodani	121	2736
170	Mangooste	65	2733
170	zhaohaikun	71	2733
172	DPprince	12	2731
173	lexiyvv	143	2729
174	Dominater069	70	2728
175	RanRankeainie	10	2727
176	LeoPro	115	2726
177	PersistentLife	56	2725
178	dreamoon_love_AA	375	2724
178	wwwwodddd	20	2724
180	physics0523	226	2722
180	umbrella-leaf	36	2722
182	BigBag	278	2720
182	Nutella3000-7	10	2720
184	Nyaan	143	2719
184	Zanite	59	2719
186	namelessgugugu	23	2717
187	teraqqq	29	2715
188	sansen	157	2713



39	QAQAutoMaton	74	3113
40	noimi	202	3103
40	StarSilk	66	3103
42	kiwihadron	15	3089
43	antontrygubO_o	168	3086
44	hank55663	260	3073
44	errorgorn	51	3073
46	JoesSR	21	3072
47	neal	155	3055
48	skip2004	30	3052
49	353cerega	89	3047
50	liuhengxi	33	3045
51	dog_of_Nesraychan	24	3030
52	peti1234	107	3025
53	275307894a	54	3024
54	Endagorion	255	3022
55	hyman00	60	3021
56	jeroenodb	90	3017
57	femboy-wannabe	11	3016
58	Licha06	25	3012
59	rin204	123	3011
60	-Eternity-	19	3004
61	jiangbowen	65	2997
62	PEIMUDA	37	2994

Artificial Analysis Coding Index

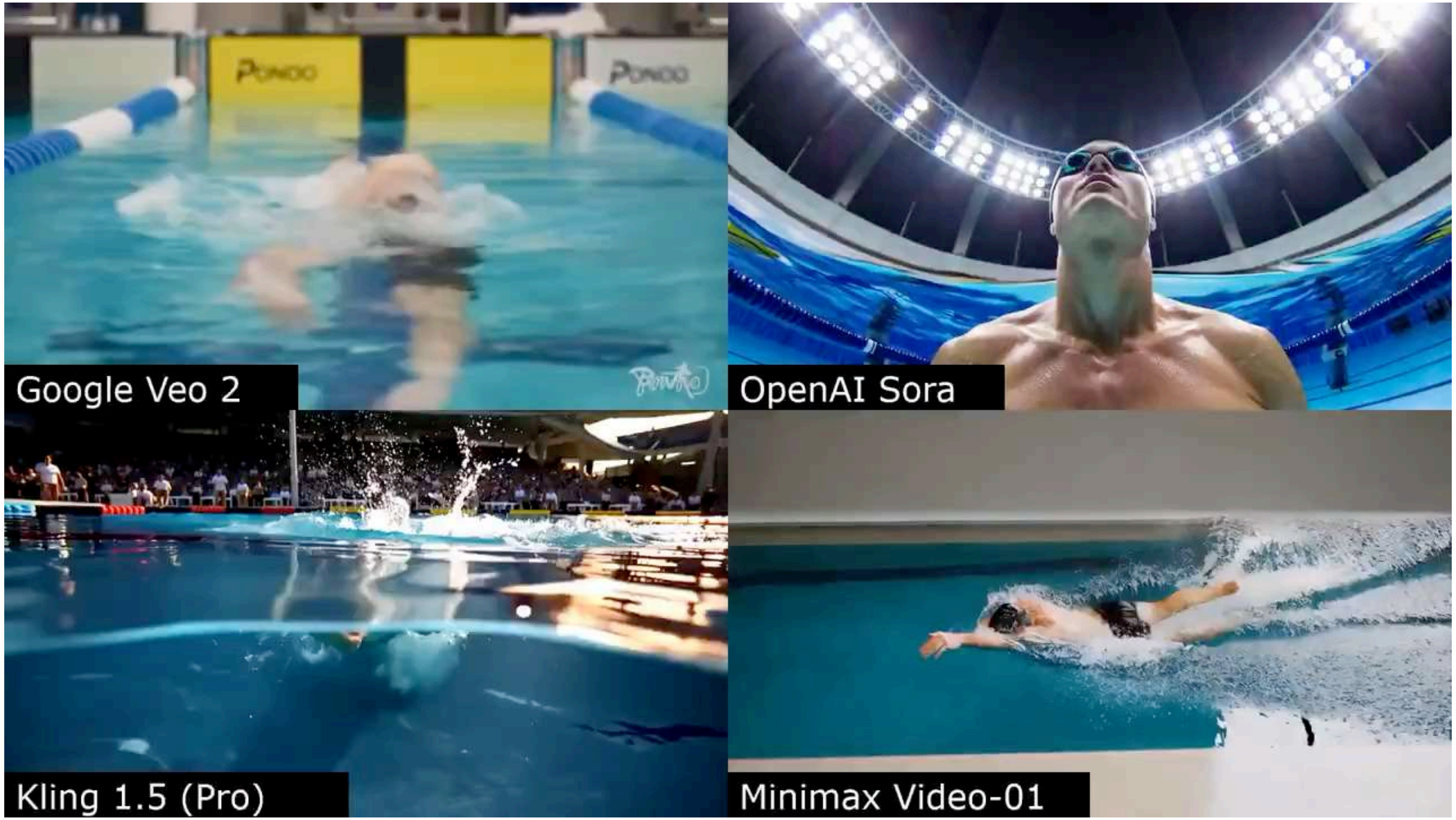
Represents the average of coding benchmarks in the Artificial Analysis Intelligence Index (LiveCodeBench & SciCode)

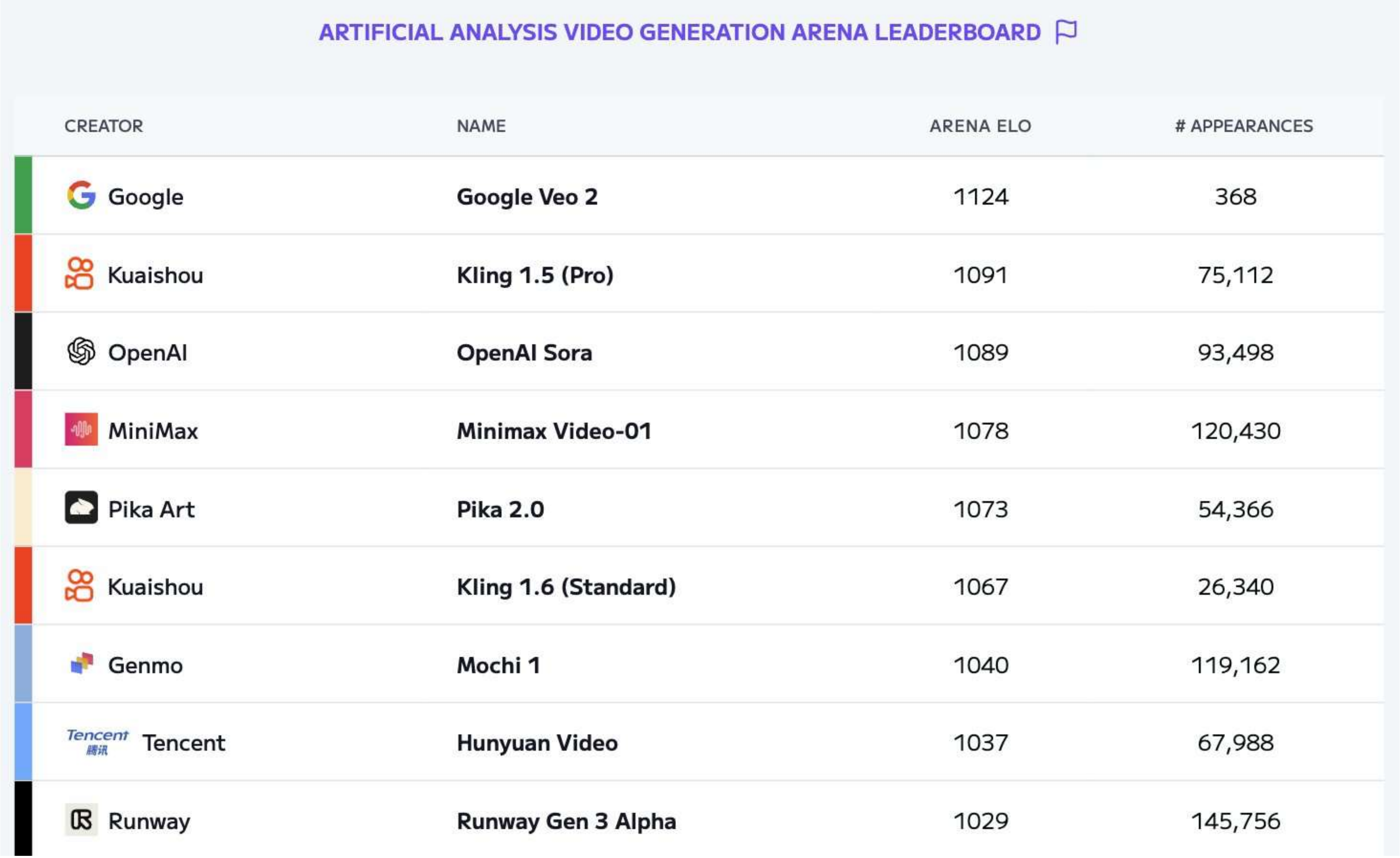


LiveBench: Coding
Coding Average Accuracy
Blended Price (USD per 1M tokens)
20
30
40
50
60
70
80
0.1
1.0
10.0
o3-mini-2025-01-31-high
o1-2024-12-17
claude-3-5-sonnet-20241022
deepseek-r1
gemini-exp-1206
o3-mini-2025-01-31-low
deepseek-v3
qwen2.5-coder-32b-instruct
gemini-2.0-flash-exp
gemini-2.0-flash-thinking-exp-01-21
gemini-2.0-flash-thinking-exp-1219
claude-3-5-haiku-20241022
gpt-4o-2024-08-06
gemini-1.5-pro-002
gpt-4-turbo-2024-04-09
o1-mini-2024-09-12
mistral-large-2411
gpt-4o-2024-11-20
gpt-4o-mini-2024-07-18
gemini-1.5-flash-002
gemini-1.5-flash-exp-0827
gemini-1.5-pro-exp-0827
gpt-4-0125-preview
claude-3-opus-20240229
gemma-2-27b-it
mixtral-8x22b-instruct-v0.1
gemini-1.5-flash-8b-exp-0924
claude-3-haiku-20240307
gemma-2-9b-it
command-r-plus-08-2024
command-r-08-2024
phi-3-mini-4k-instruct
phi-3-mini-128k-instr
Google
OpenAI
Anthropic
Mistral
Alibaba
Other
@scaling01 | Lisan al Gaib

Before
Now

A swimmer in his early 30s dives into the water, the view shifting between above and below as he powers through the race, water splashing rhythmically with each stroke.

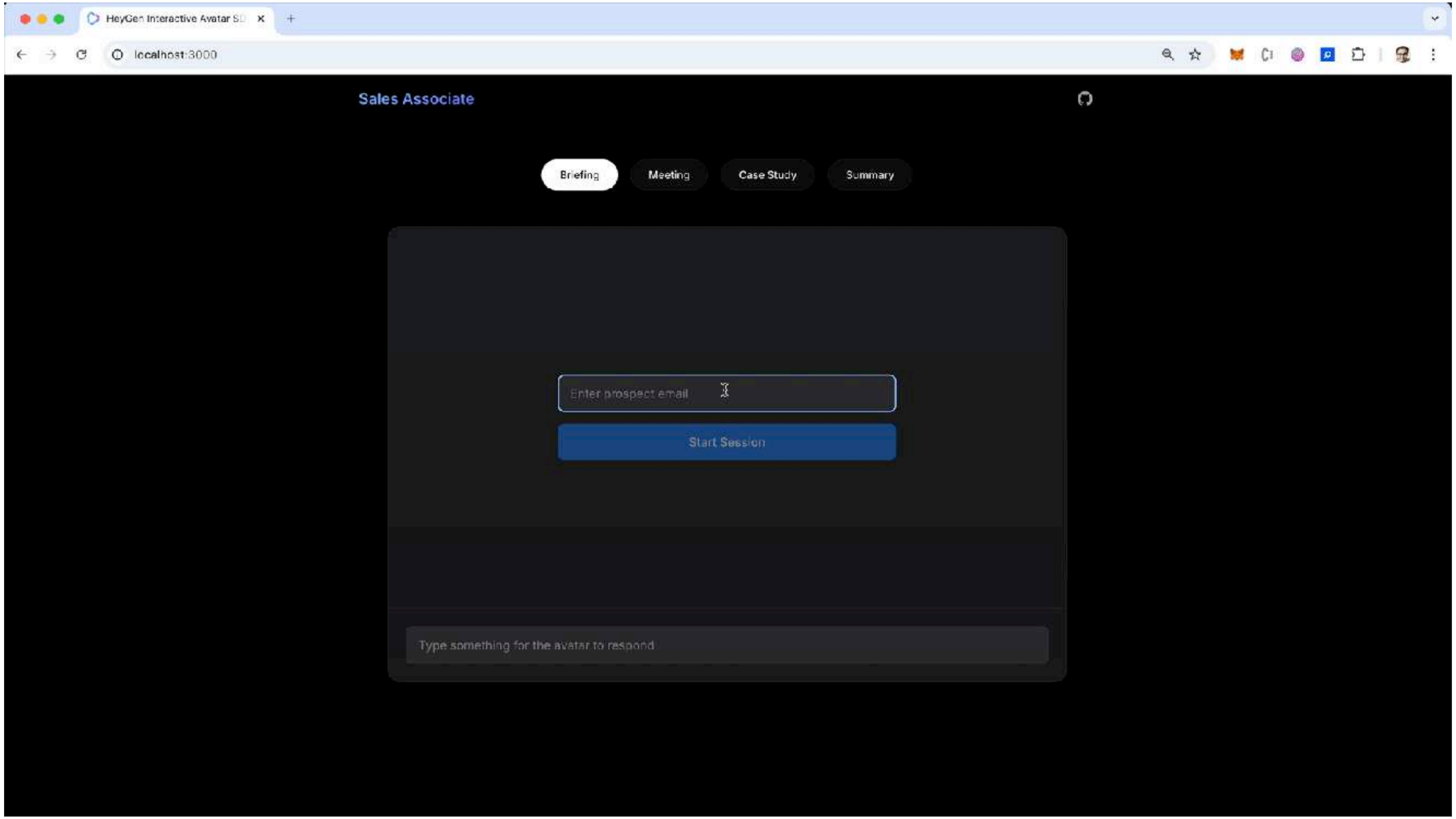




ARTIFICIAL ANALYSIS VIDEO GENERATION ARENA LEADERBOARD

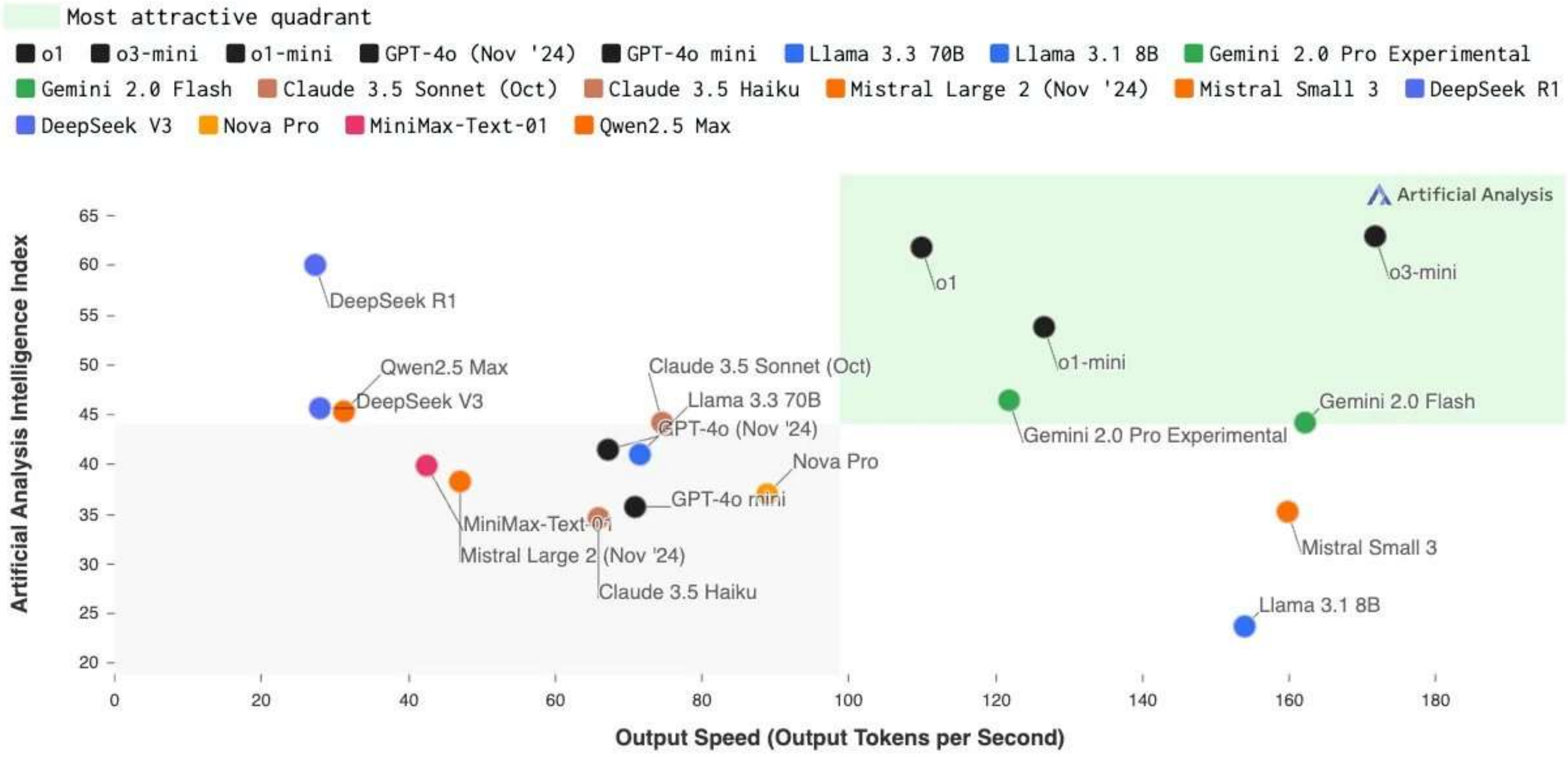
CREATOR	NAME	ARENA ELO	# APPEARANCES
Google	**Google Veo 2**	1124	368
Kuaishou	**Kling 1.5 (Pro)**	1091	75,112
OpenAI	**OpenAI Sora**	1089	93,498
MiniMax	**Minimax Video-01**	1078	120,430
Pika Art	**Pika 2.0**	1073	54,366
Kuaishou	**Kling 1.6 (Standard)**	1067	26,340
Genmo	**Mochi 1**	1040	119,162
Tencent 腾讯 Tencent	**Hunyuan Video**	1037	67,988
Runway	**Runway Gen 3 Alpha**	1029	145,756

Video Agents



Intelligence vs. Output Speed

Artificial Analysis Intelligence Index (Version 2, released Feb '25); Output Speed: Output Tokens per Second



Act

offline

DPC cars
KUNG FU

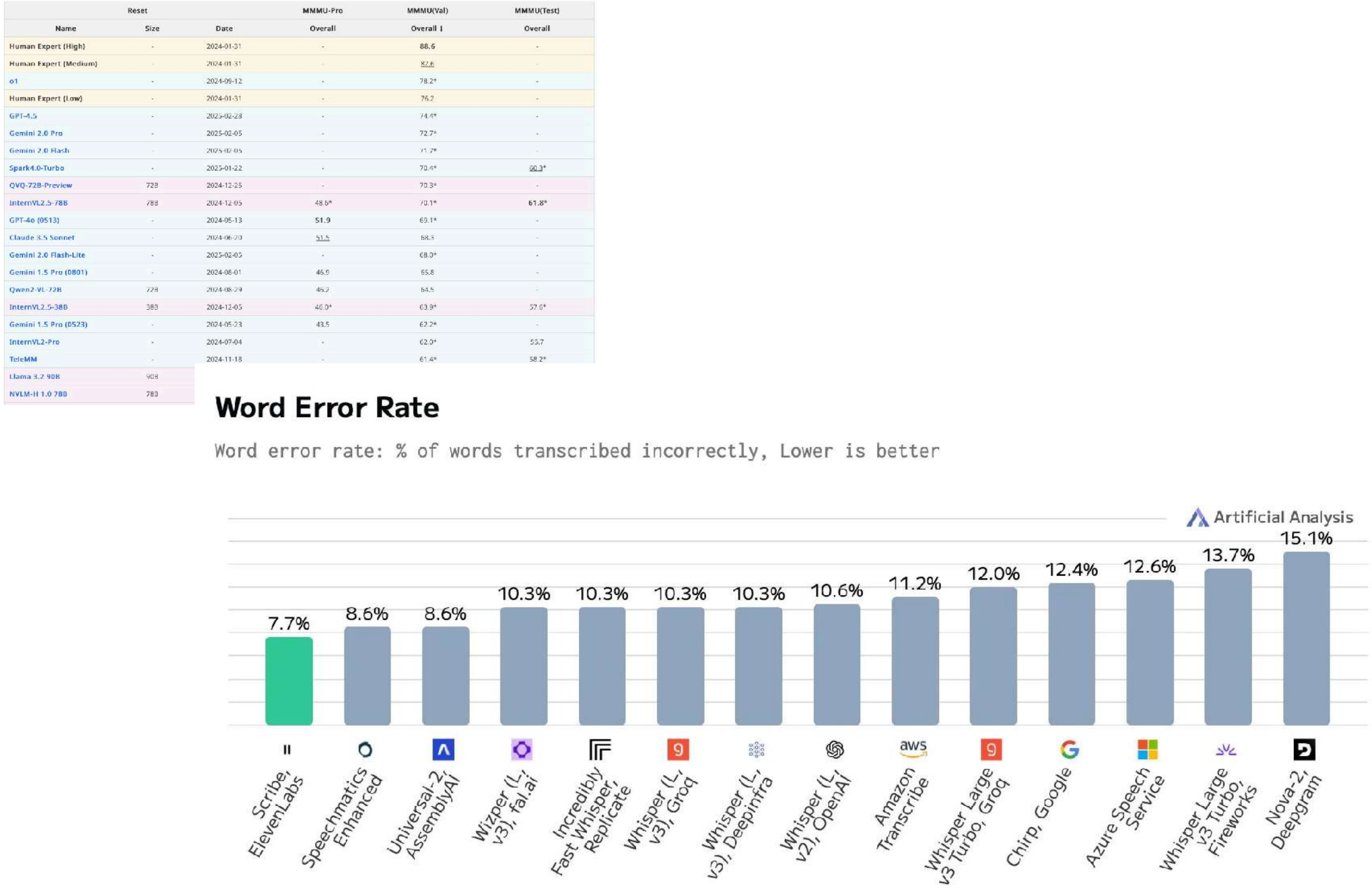
What's coming?

Feel → Think → Act

see hear

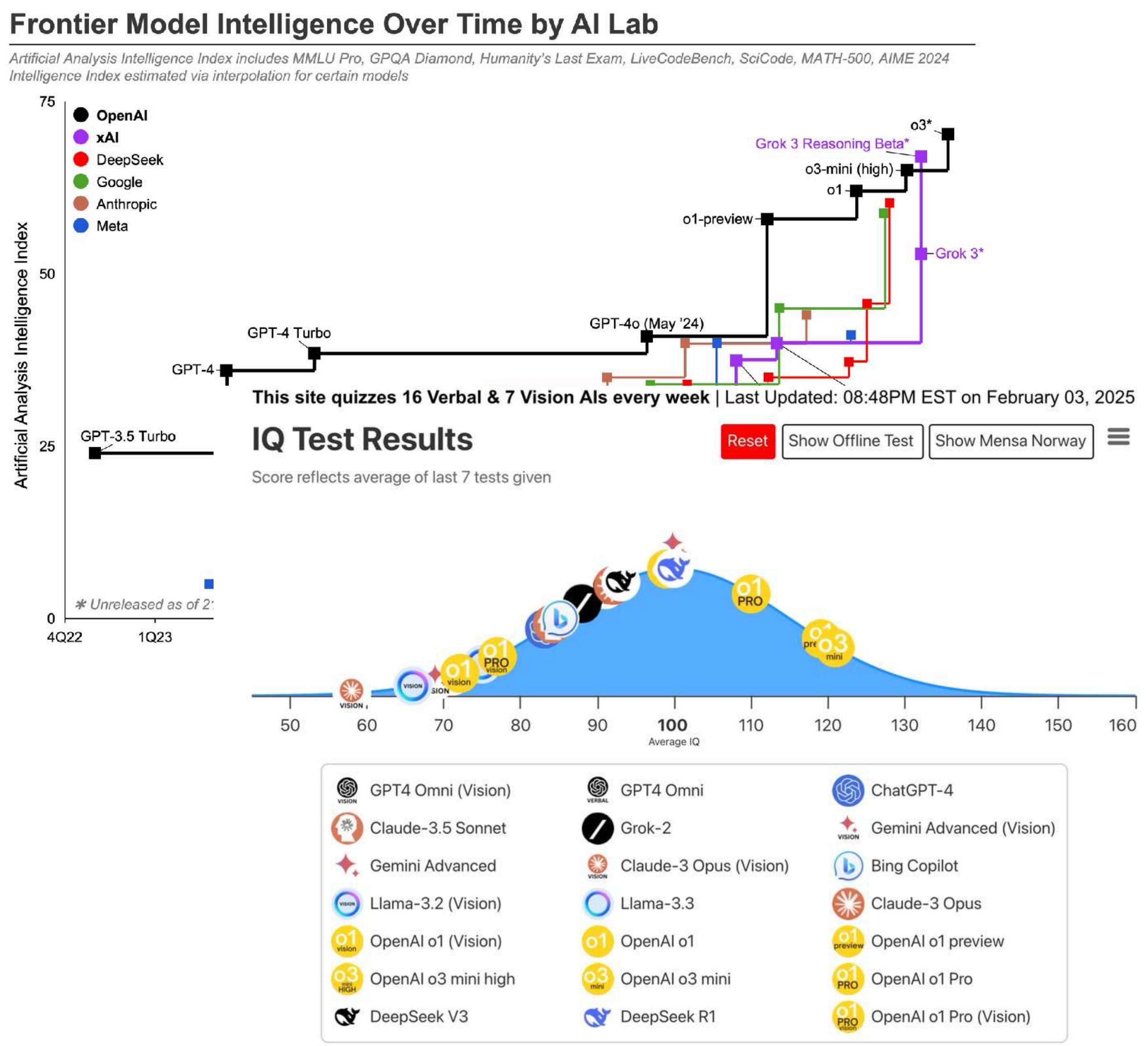
fast slow

online offline



Word Error Rate

Word error rate: % of words transcribed incorrectly, Lower is better



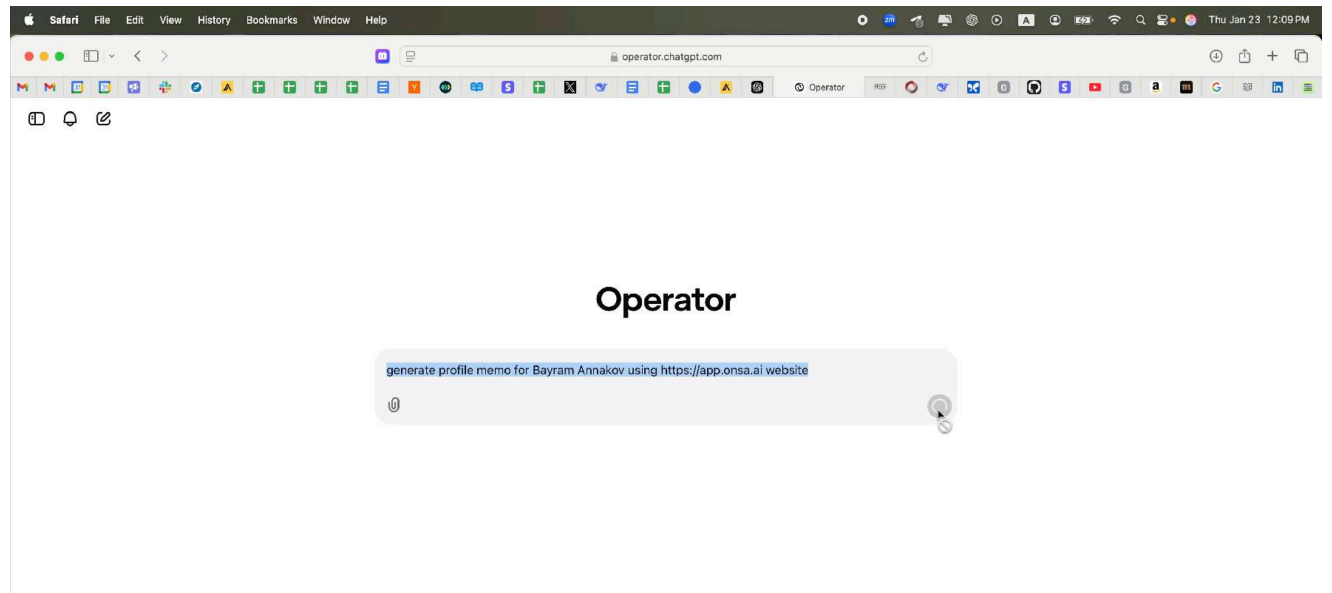
Frontier Model Intelligence Over Time by AI Lab

Artificial Analysis Intelligence Index includes MMLU Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, MATH-500, AIME 2024
Intelligence Index estimated via interpolation for certain models

This site quizzes 16 Verbal & 7 Vision AIs every week | Last Updated: 08:48PM EST on February 03, 2025

IQ Test Results

Score reflects average of last 7 tests given



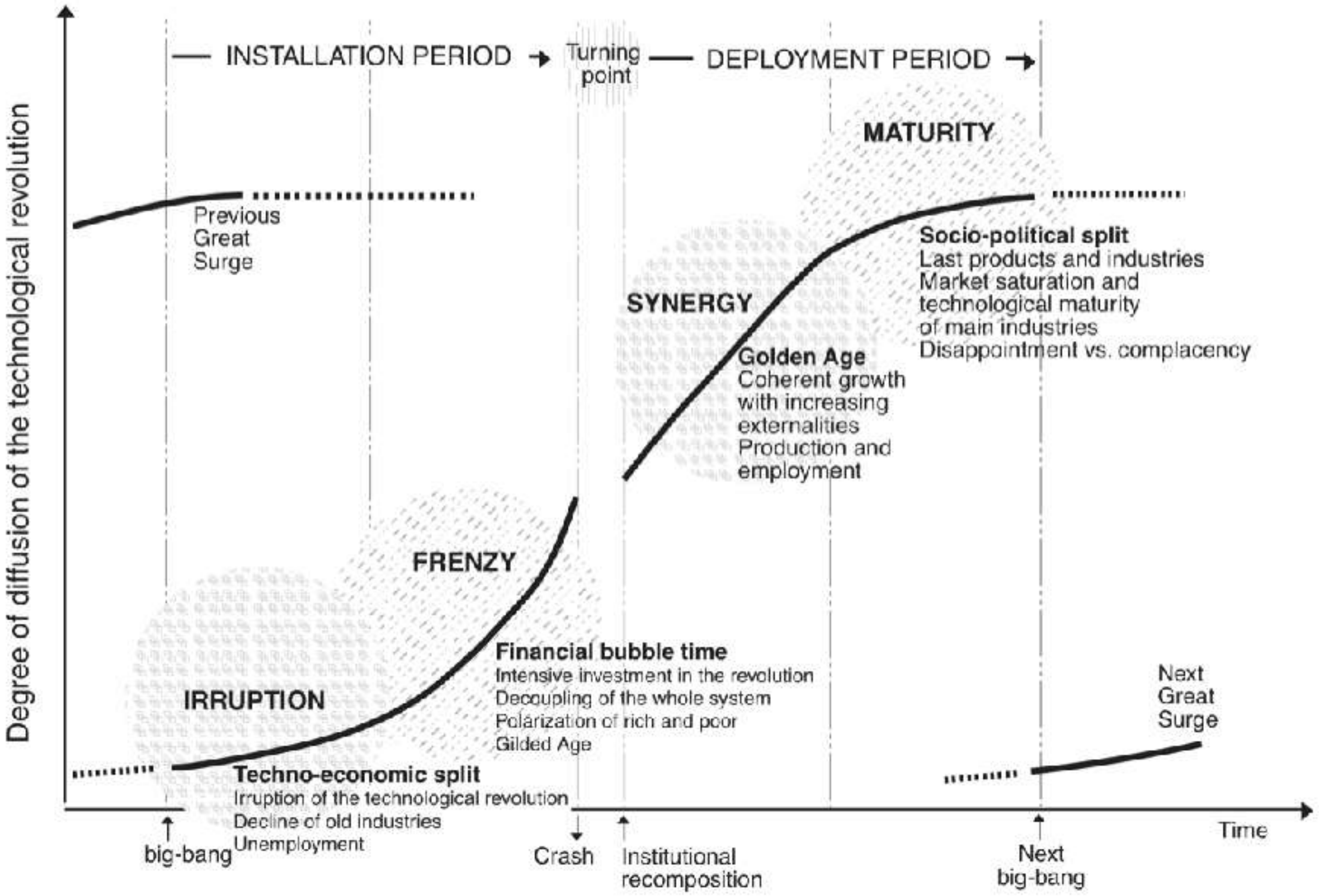
Operator

generate profile memo for Bayram Annakov using https://app.onsa.ai website

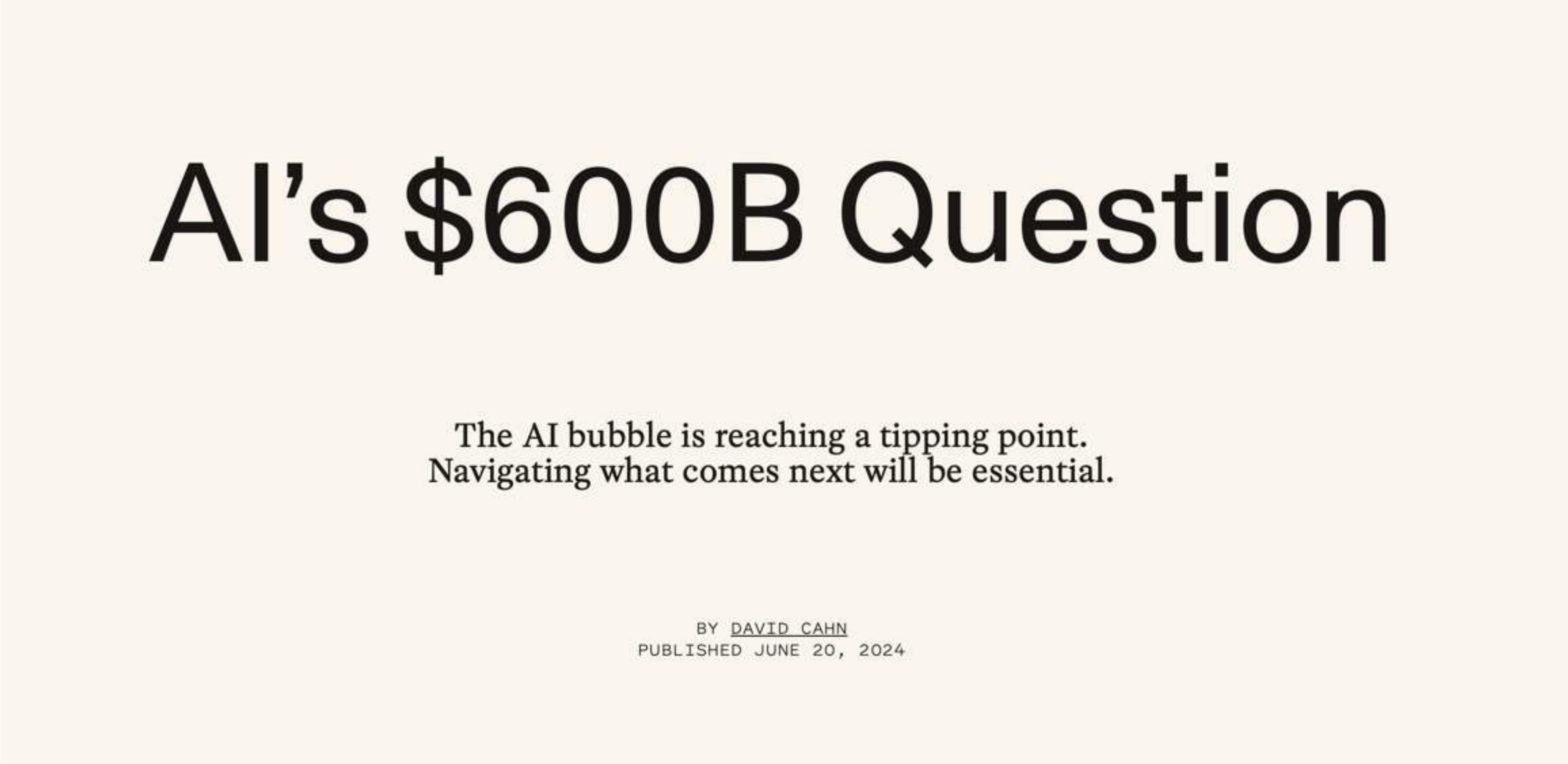
Tech Revolutions & Financial Capital

Perspective

Figure 5.1 Recurring phases of each great surge in the core countries


INSTALLATION PERIOD
Turning point
DEPLOYMENT PERIOD
MATURITY
Previous Great Surge
Socio-political split
Last products and industries
Market saturation and technological maturity of main industries
Disappointment vs. complacency
SYNERGY
Golden Age
Coherent growth with increasing externalities
Production and employment
FRENZY
Financial bubble time
Intensive investment in the revolution
Decoupling of the whole system
Polarization of rich and poor
Gilded Age
IRRUPTION
Next Great Surge
Techno-economic split
Irruption of the technological revolution
Decline of old industries
Unemployment
Degree of diffusion of the technological revolution
big-bang
Crash
Institutional recomposition
Next big-bang
Time

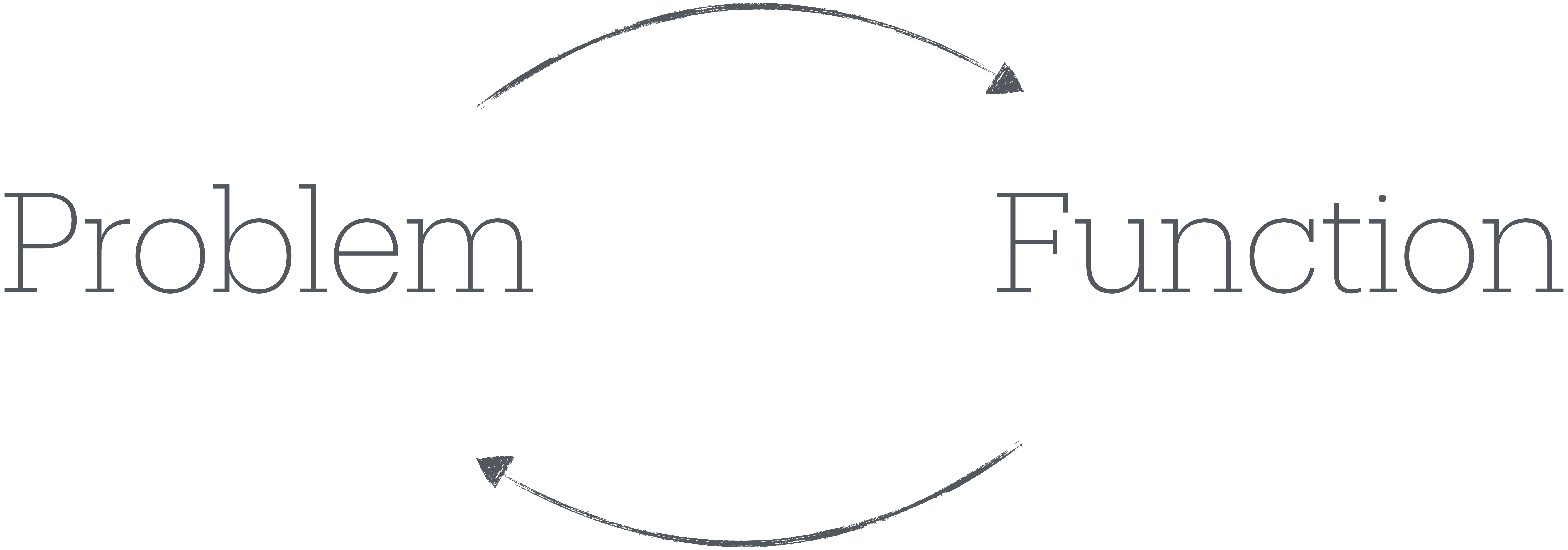
Frenzy?



AI’s $600B Question

The AI bubble is reaching a tipping point.
Navigating what comes next will be essential.

BY DAVID CAHN
PUBLISHED JUNE 20, 2024

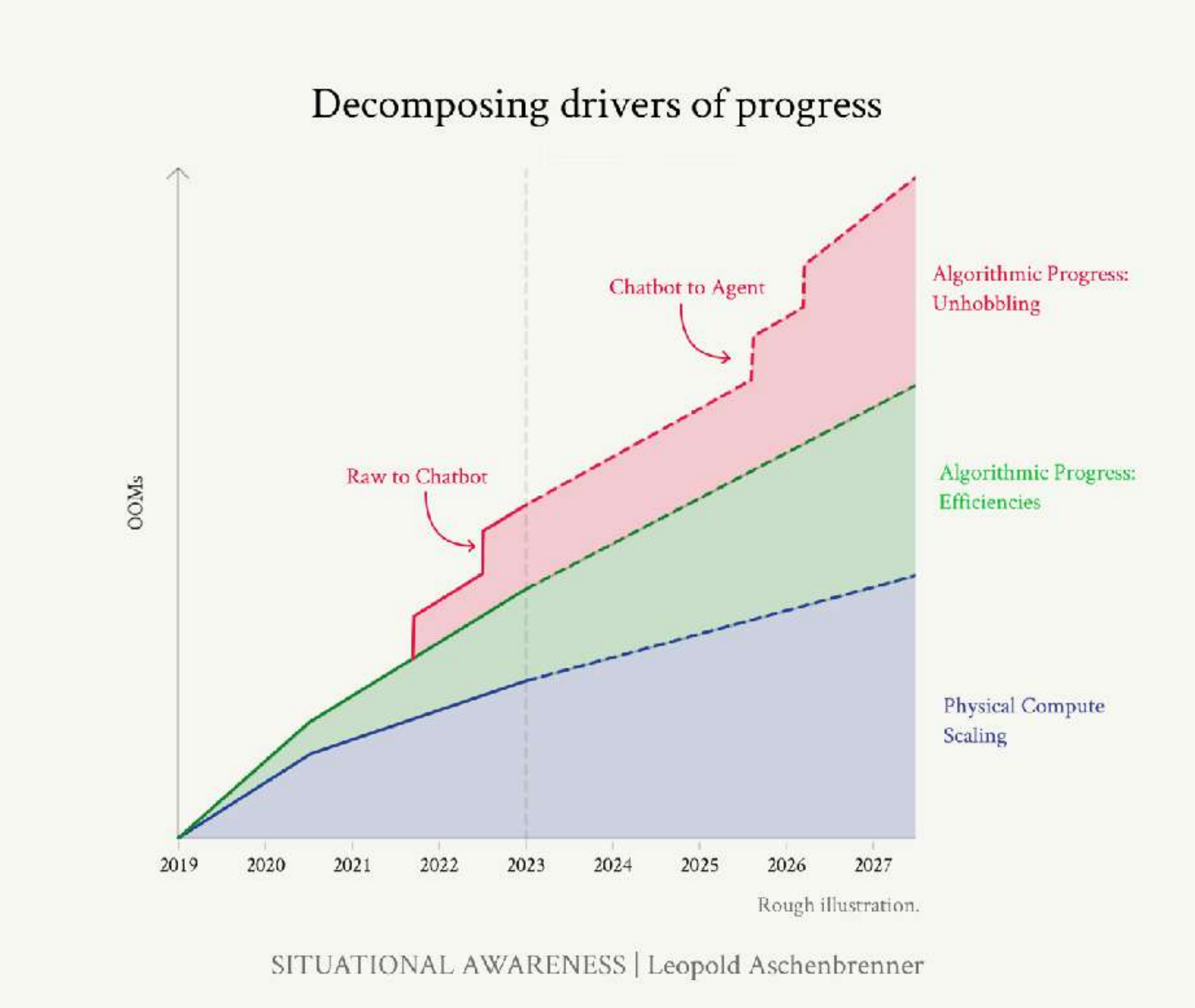


Problem

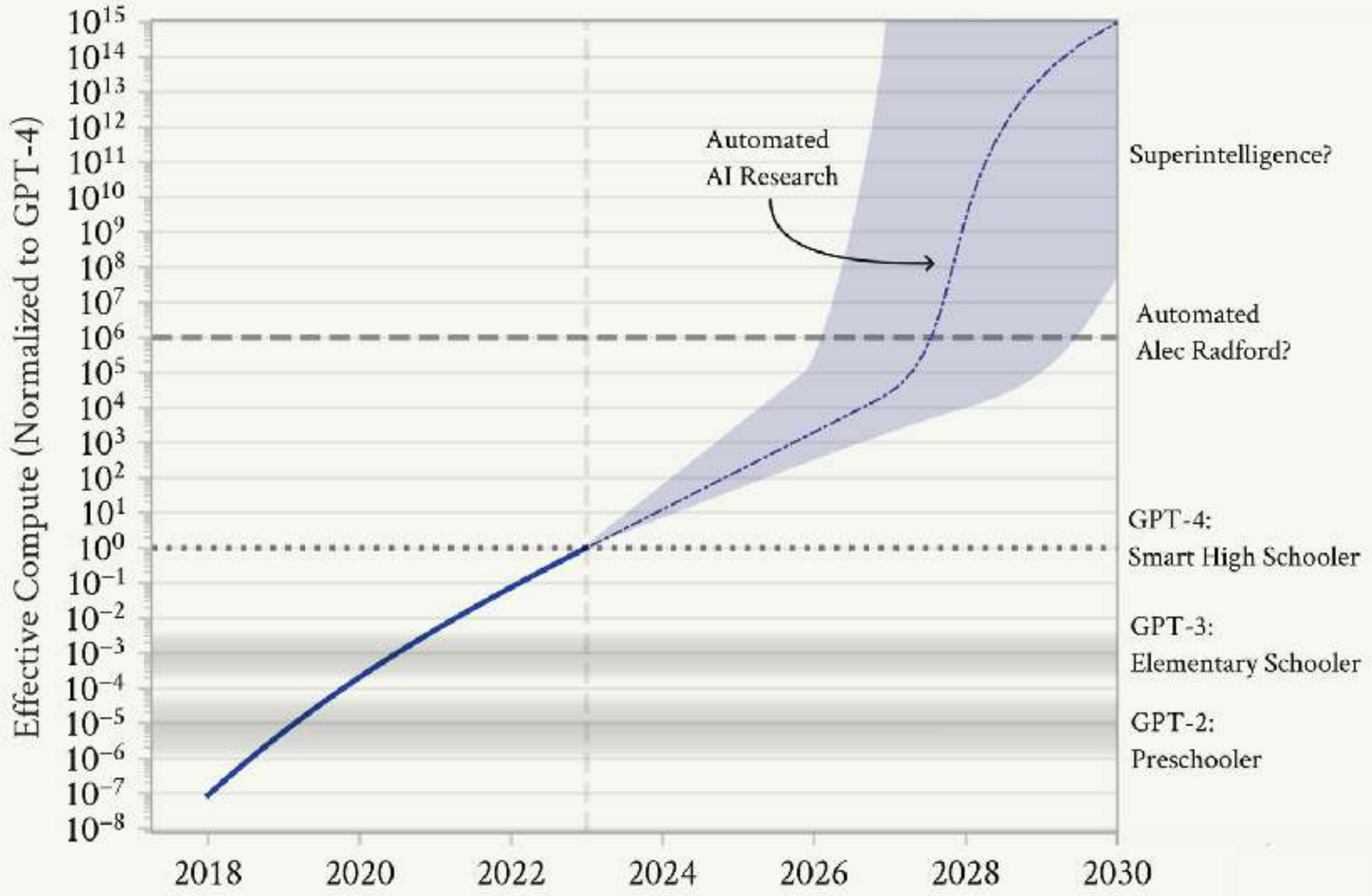
Function

Situational Awareness

Perspective



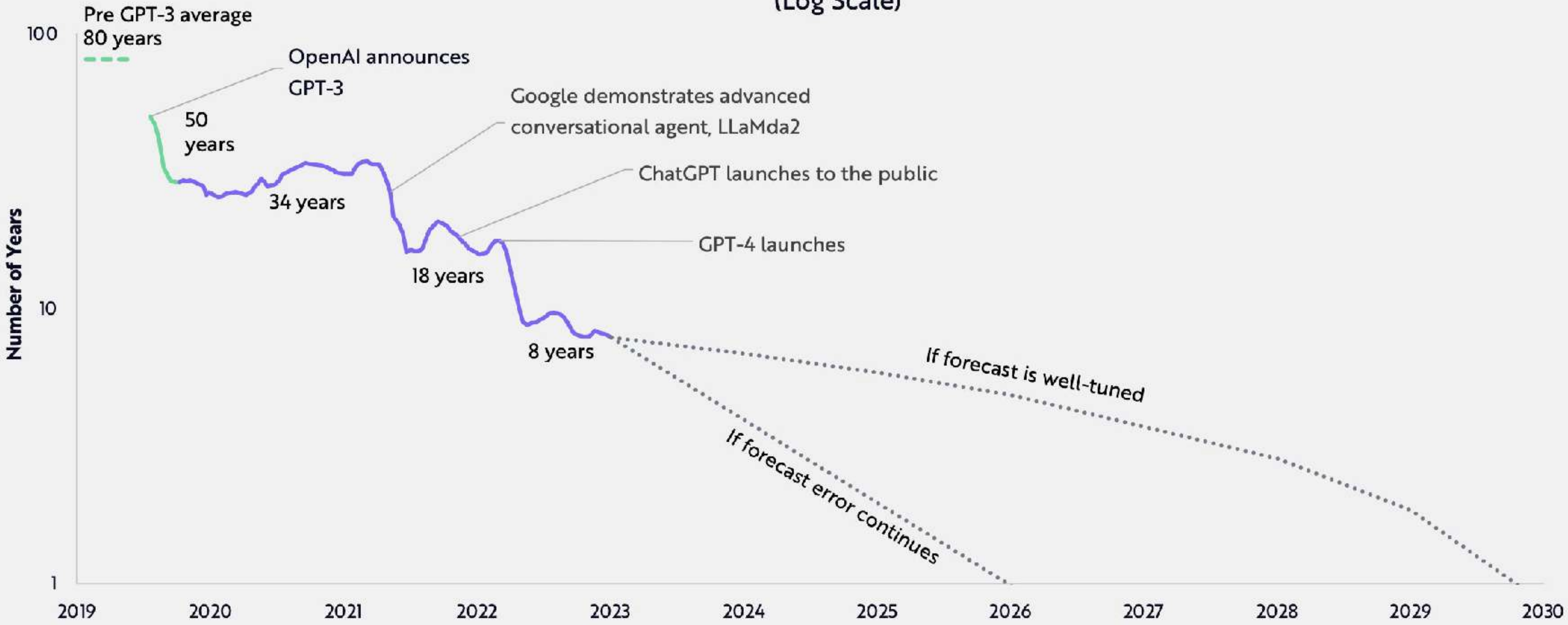
Scenario: Intelligence Explosion



Rough illustration.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

AI Is Accelerating Faster Than Forecasters Anticipated

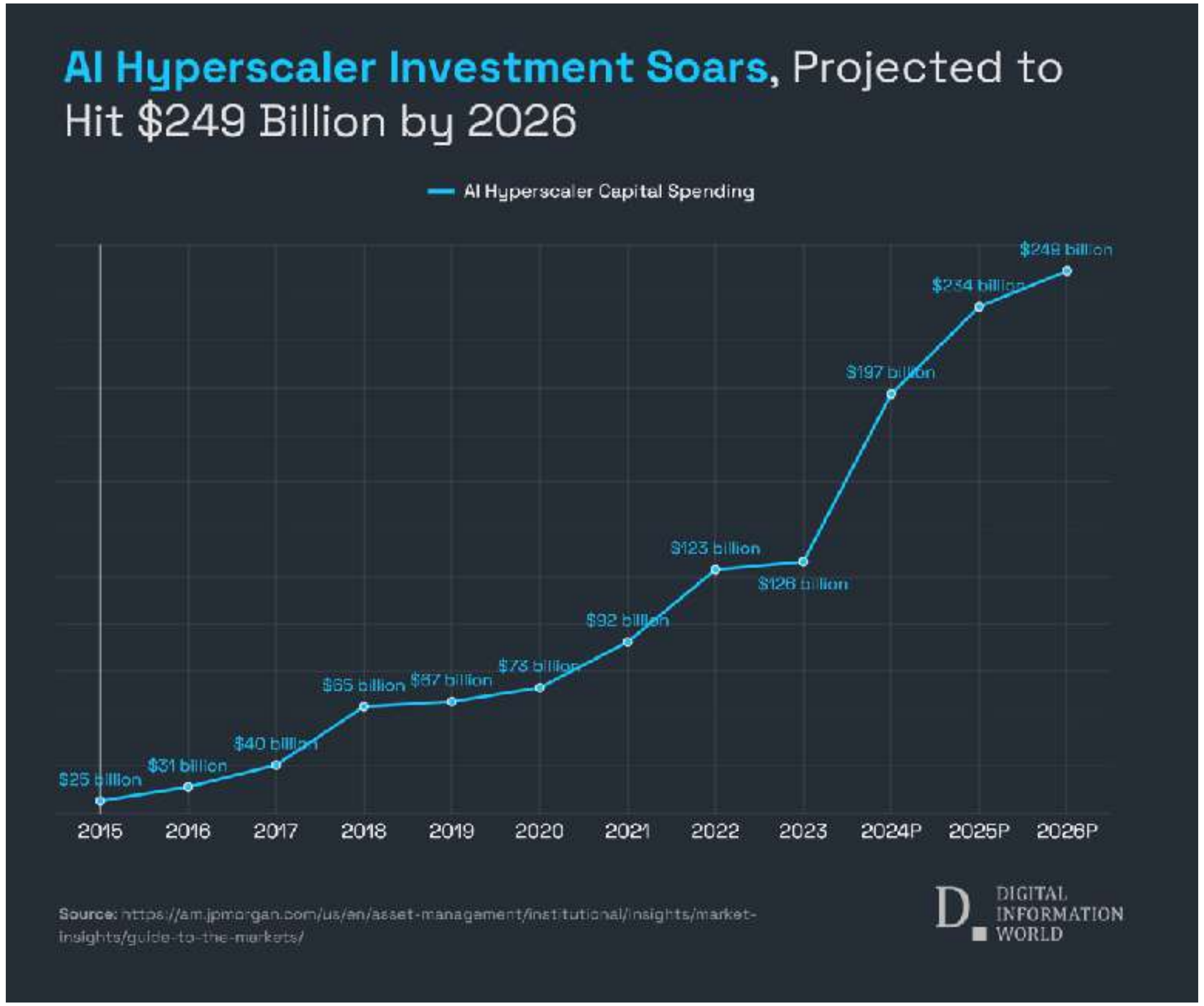


Expected Years Until Launch Of A General Artificial Intelligence System

(Log Scale)

Sources: ARK Investment Management LLC, 2024, based on data from Metaculus, including benchmark details, as of January 3, 2024. Benchmark broadly requires the successful passage of an adversarial two-hour Tuning test, broad success on a Q&A knowledge and logic benchmark, and the successful interpretation of and execution complex model car assembly instruction, all within a single system. Green lines are derived estimates for time to general purpose AI (strongly formulated) based upon forecasts for a weaker benchmark. Forecasts are inherently limited and cannot be relied upon. For informational purposes only and should not be considered investment advice or a recommendation to buy, sell, or hold any particular security. Past performance is not indicative of future results.

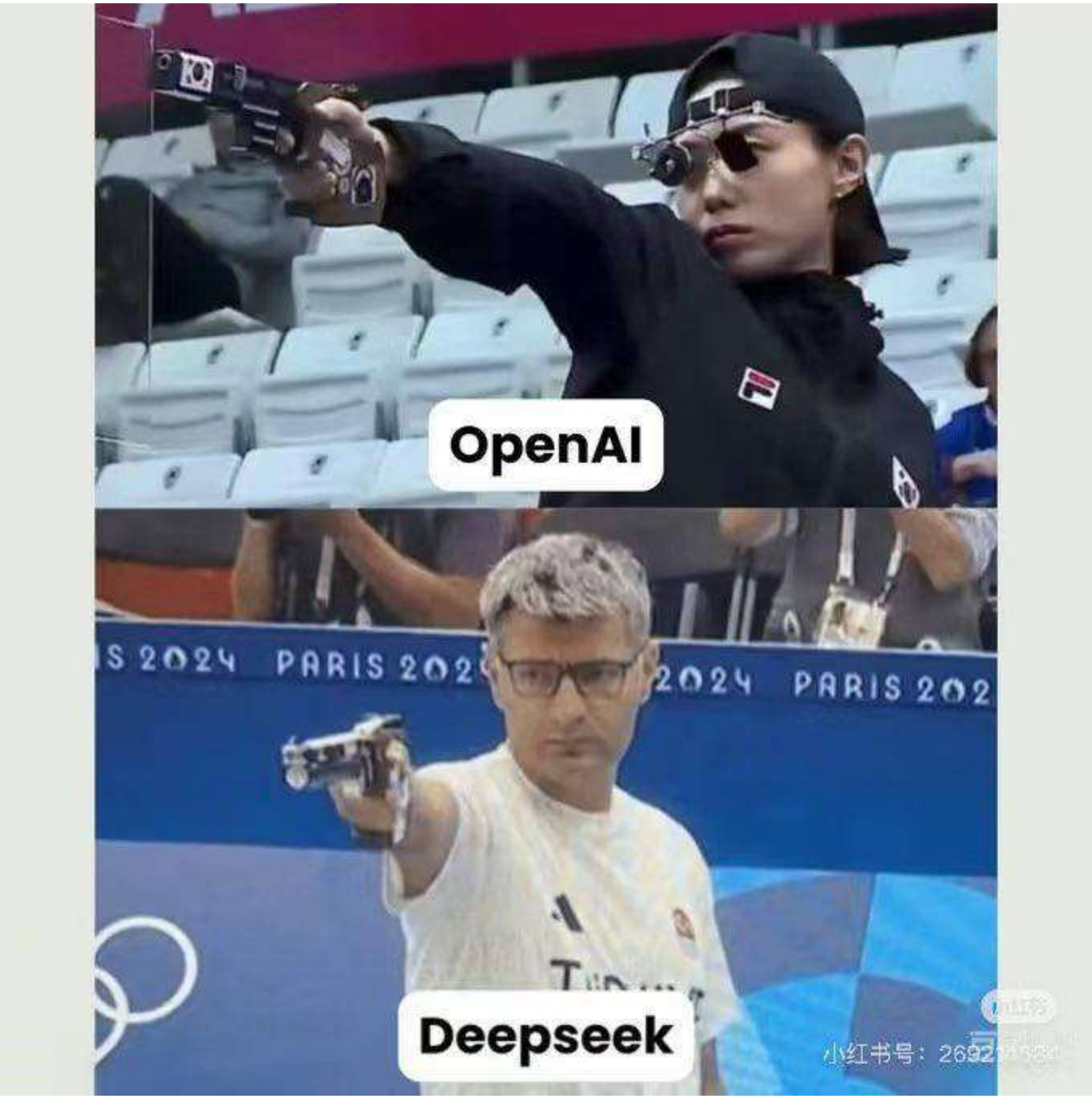
Physical Scaling

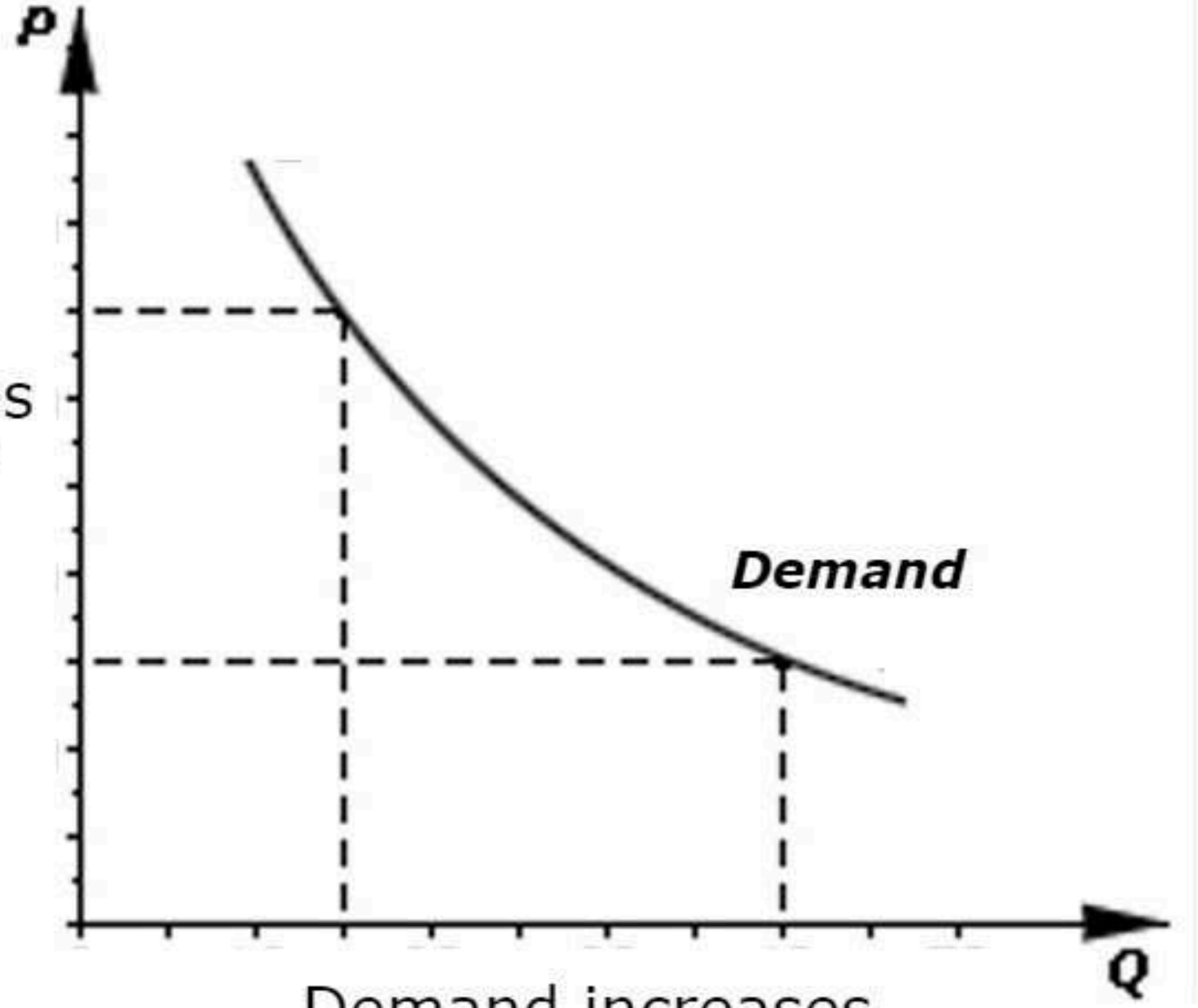


AI Hyperscaler Investment Soars, Projected to Hit $249 Billion by 2026

Source: https://am.jpmorgan.com/us/en/asset-management/institutional/insights/market-insights/guide-to-the-markets/

Algorithmic Progress + Unhobbling



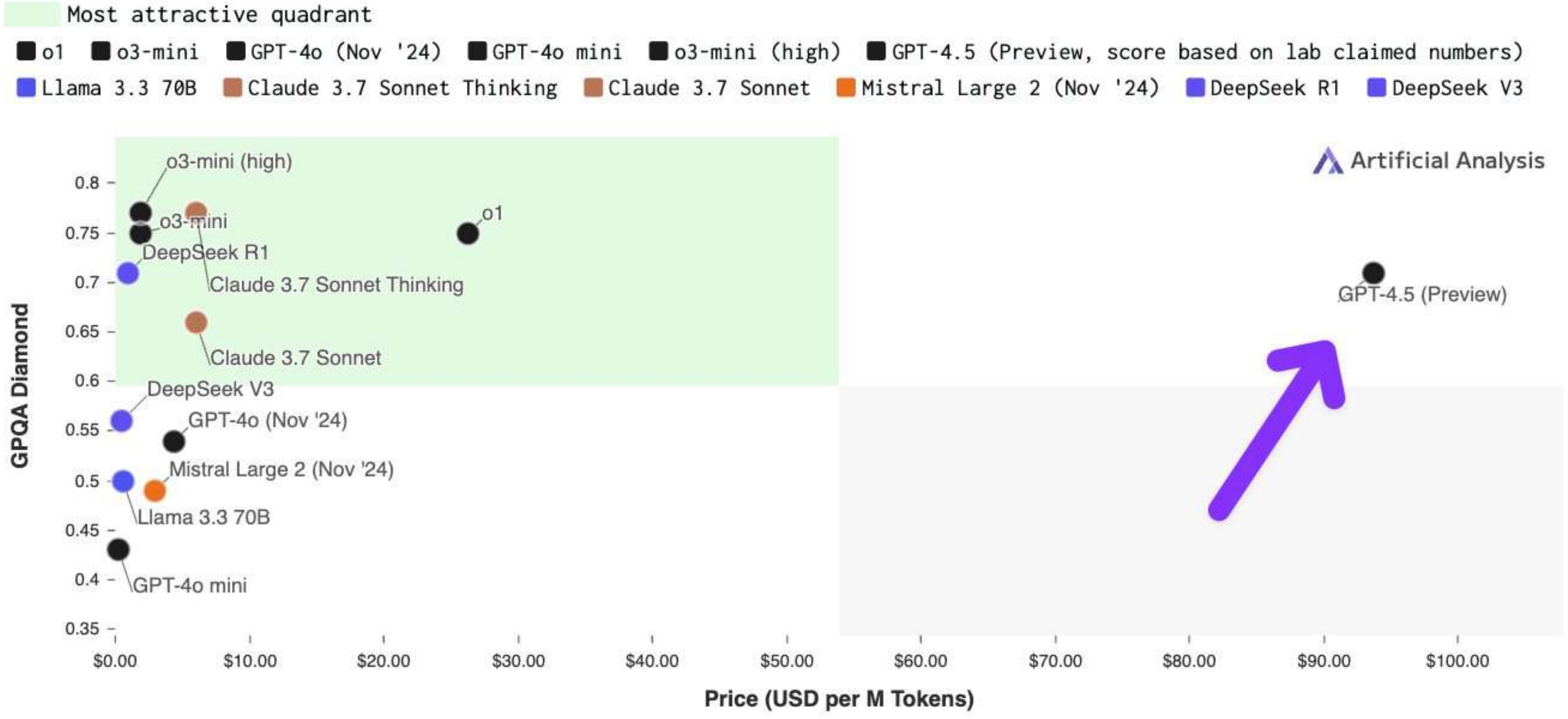


Cost decreases by half

Demand increases by more than half

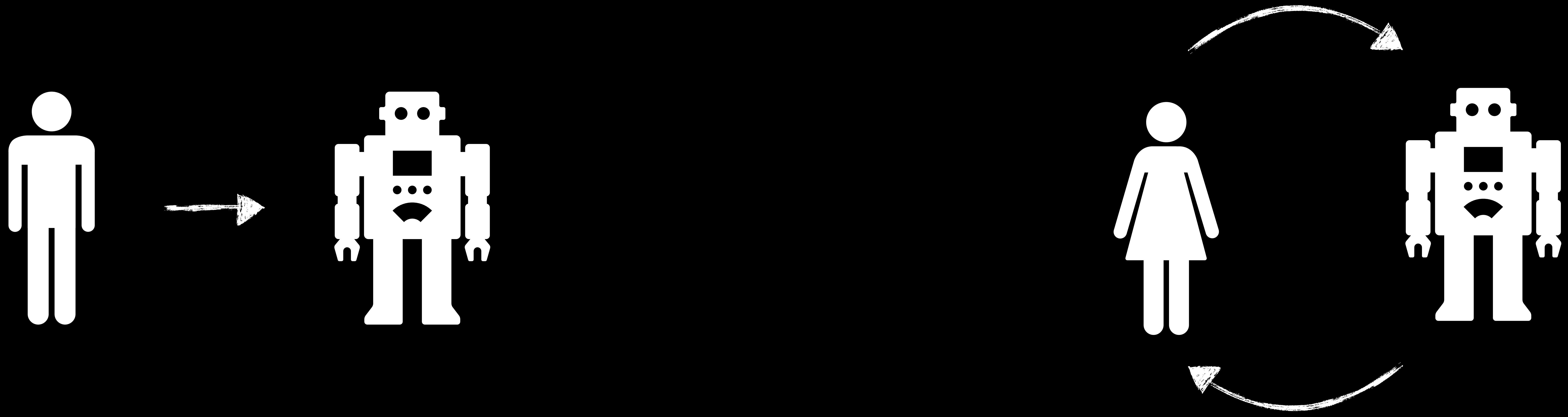
GPQA Diamond (Scientific Reasoning) vs. Price

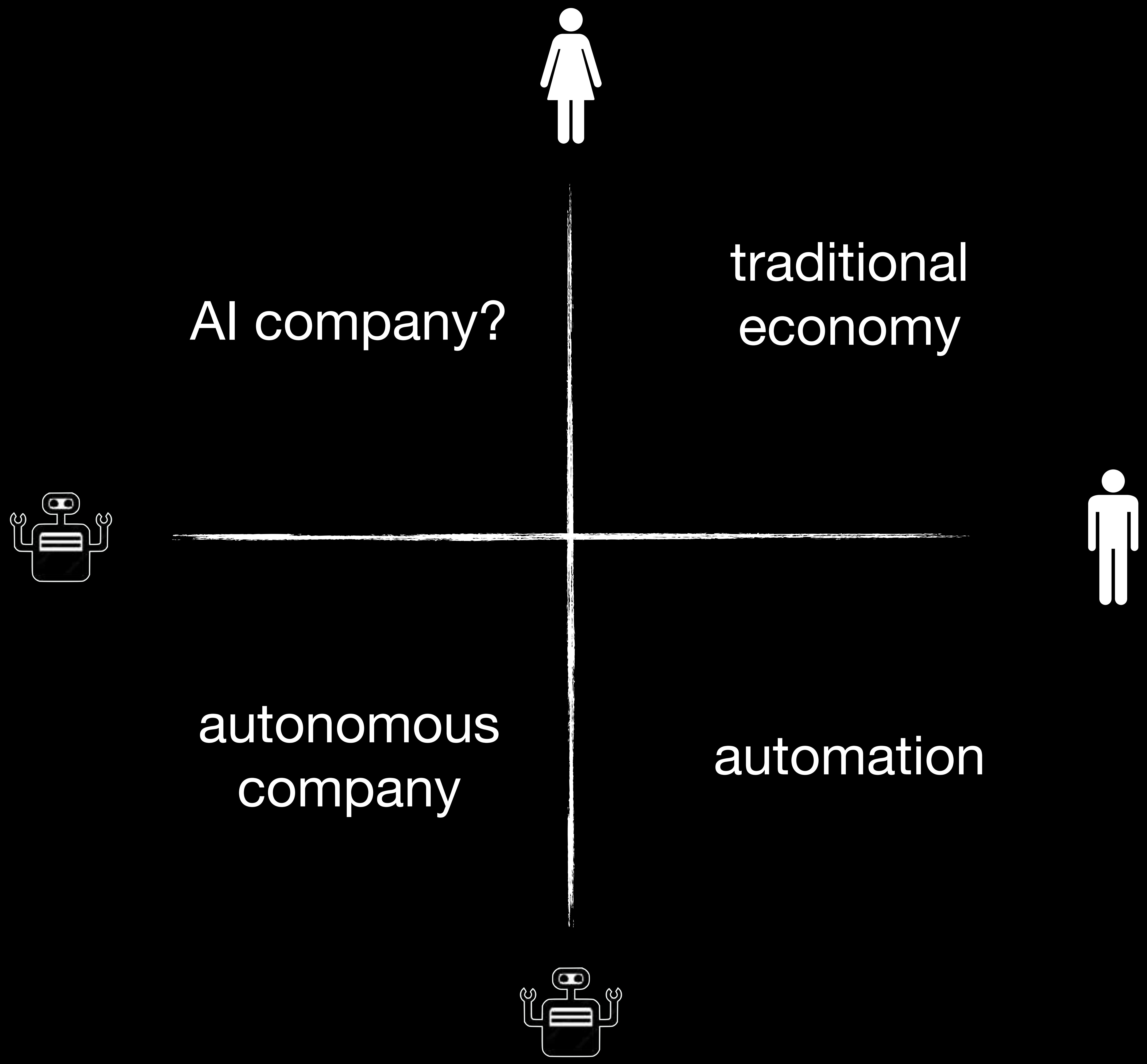
GPQA Diamond score; Price: USD per 1M Tokens (blended pricing, weighted 3 input tokens to 1 output token)



Automation vs Augmentation

How value is distributed





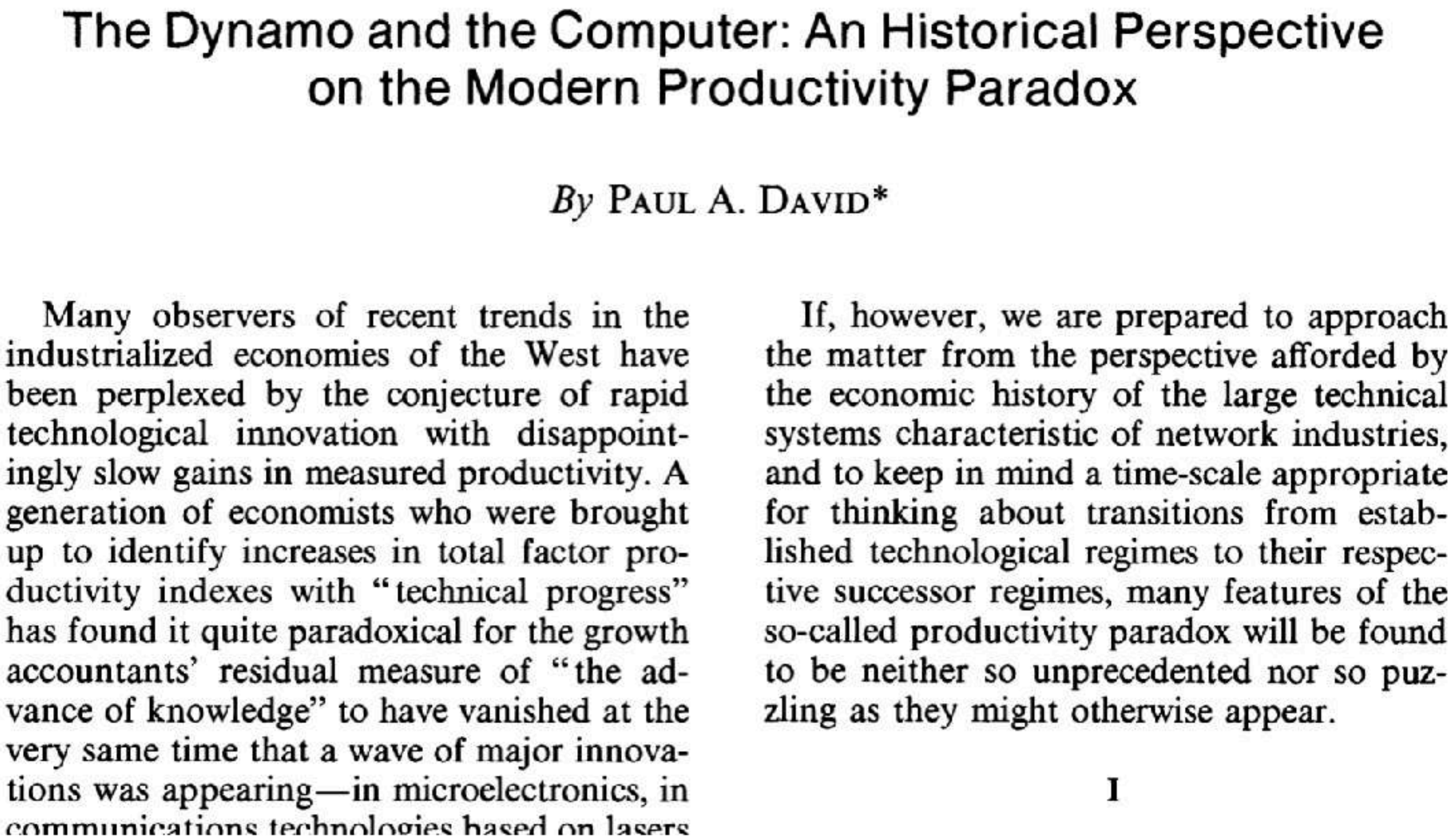
AI company?

traditional
economy

autonomous
company

automation

Case from history: Electric Dynamo



The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox

By PAUL A. DAVID*

Many observers of recent trends in the industrialized economies of the West have been perplexed by the conjecture of rapid technological innovation with disappointingly slow gains in measured productivity. A generation of economists who were brought up to identify increases in total factor productivity indexes with "technical progress" has found it quite paradoxical for the growth accountants' residual measure of "the advance of knowledge" to have vanished at the very same time that a wave of major innovations was appearing—in microelectronics, in communications technologies based on lasers

If, however, we are prepared to approach the matter from the perspective afforded by the economic history of the large technical systems characteristic of network industries, and to keep in mind a time-scale appropriate for thinking about transitions from established technological regimes to their respective successor regimes, many features of the so-called productivity paradox will be found to be neither so unprecedented nor so puzzling as they might otherwise appear.

I

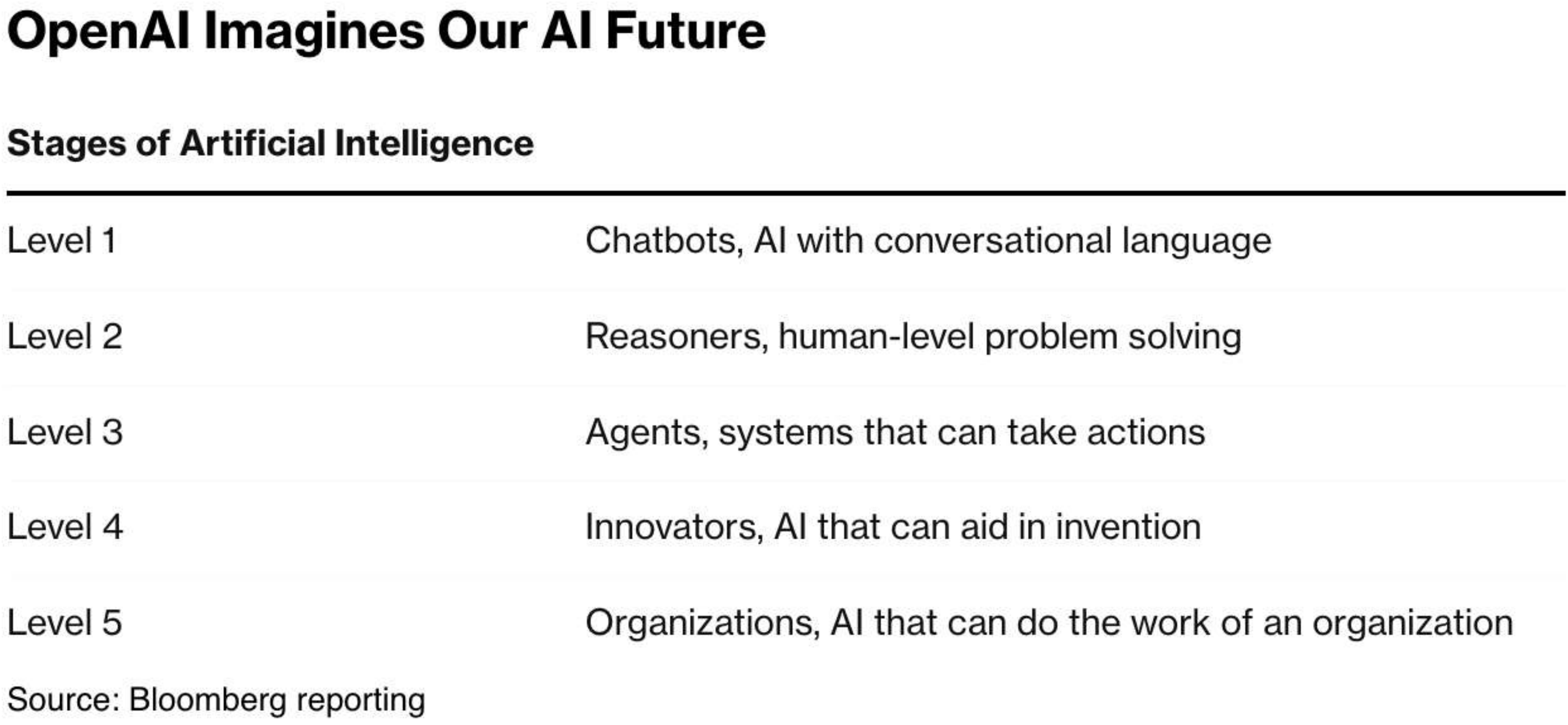
Future of Jobs Report 2025

Fastest growing and declining jobs by 2030

	↑ Top fastest growing jobs		↓ Top fastest declining jobs
1	Big data specialists	1	Postal service clerks
2	FinTech engineers	2	Bank tellers and related clerks
3	AI and machine learning specialists	3	Data entry clerks
4	Software and applications developers	4	Cashiers and ticket clerks
5	Security management specialists	5	Administrative assistants and executive secretaries
6	Data warehousing specialists	6	Printing and related trades workers
7	Autonomous and electric vehicle specialists	7	Accounting, bookkeeping and payroll clerks
8	UI and UX designers	8	Material-recording and stock-keeping clerks
9	Light truck or delivery services drivers	9	Transportation attendants and conductors
10	Internet of things specialists	10	Door-to-door sales workers, news and street vendors, and related workers
11	Data analysts and scientists	11	Graphic designers
12	Environmental engineers	12	Claims adjusters, examiners and investigators
13	Information security analysts	13	Legal officials
14	DevOps engineers	14	Legal secretaries
15	Renewable energy engineers	15	Telemarketers

Note: The jobs that survey respondents report the highest and lowest net growth (%) by 2030.

Source: World Economic Forum. (2025). *Future of Jobs Report 2025*.



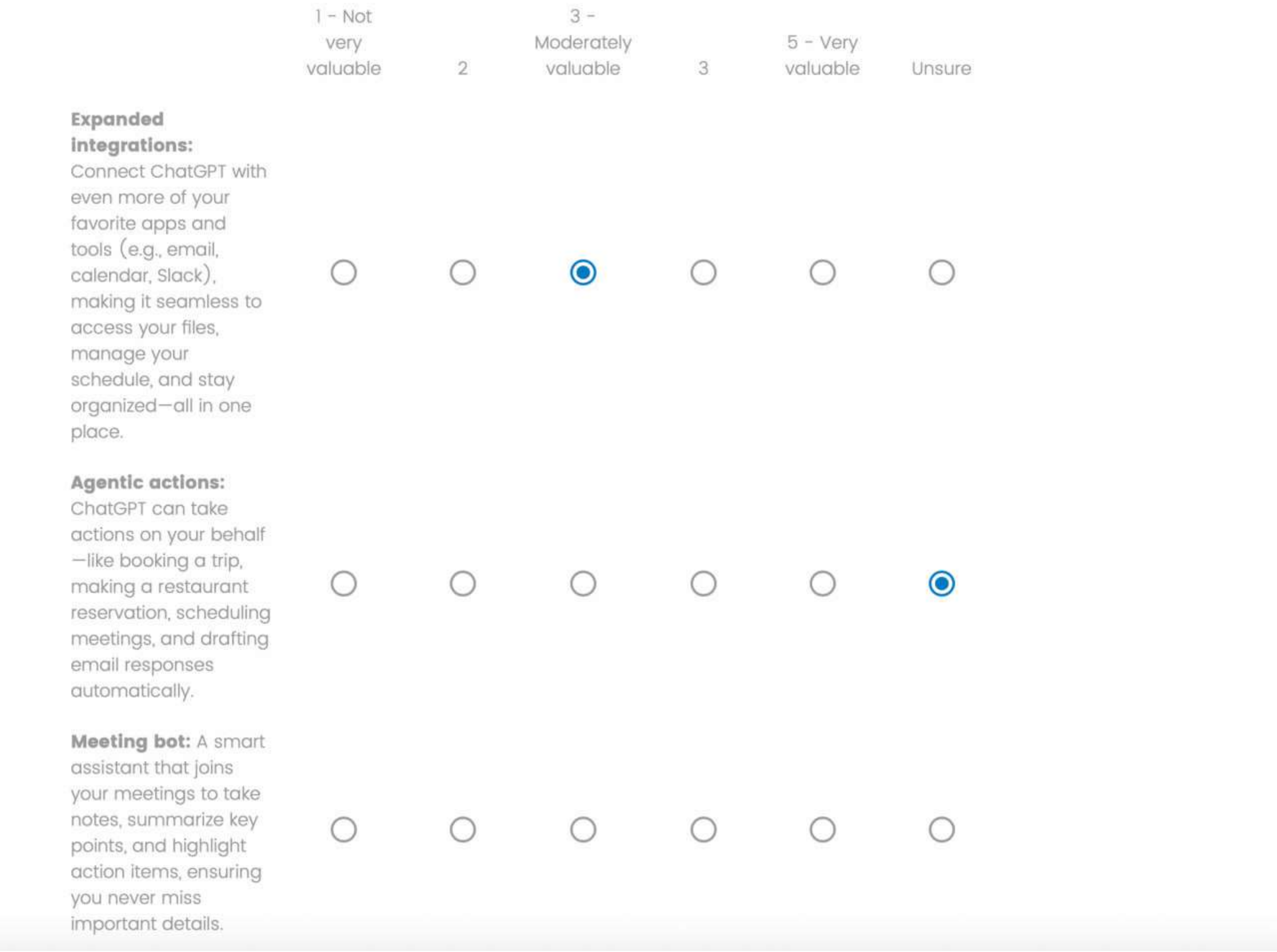
OpenAI Imagines Our AI Future

Stages of Artificial Intelligence

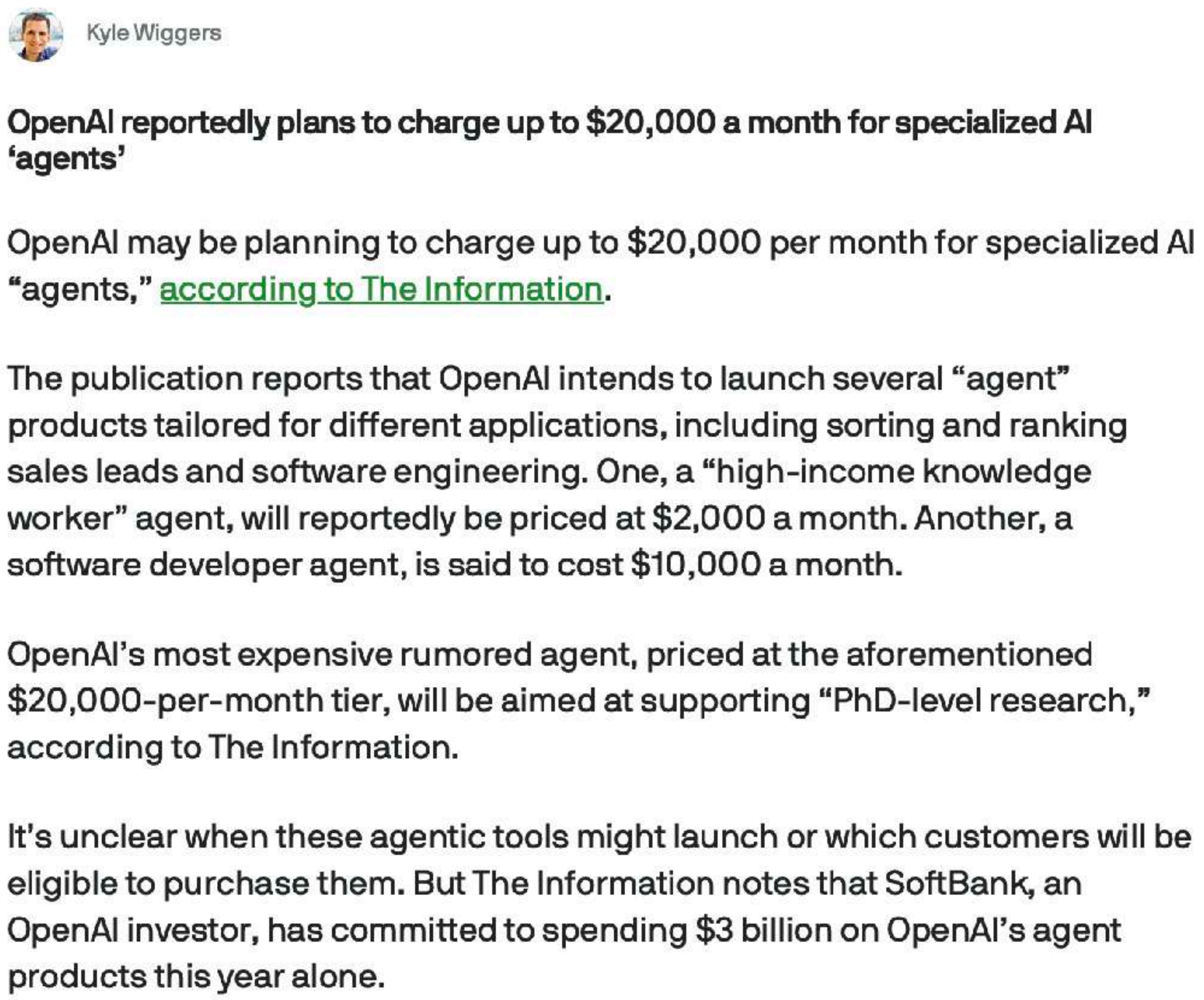
Level 1	Chatbots, AI with conversational language
Level 2	Reasoners, human-level problem solving
Level 3	Agents, systems that can take actions
Level 4	Innovators, AI that can aid in invention
Level 5	Organizations, AI that can do the work of an organization

Source: Bloomberg reporting

we are here →



	1 - Not very valuable	2	3 - Moderately valuable	3	5 - Very valuable	Unsure
Expanded integrations: Connect ChatGPT with even more of your favorite apps and tools (e.g., email, calendar, Slack), making it seamless to access your files, manage your schedule, and stay organized—all in one place.	○	○	●	○	○	○
Agentic actions: ChatGPT can take actions on your behalf—like booking a trip, making a restaurant reservation, scheduling meetings, and drafting email responses automatically.	○	○	○	○	○	●
Meeting bot: A smart assistant that joins your meetings to take notes, summarize key points, and highlight action items, ensuring you never miss important details.	○	○	○	○	○	○



Kyle Wiggers

OpenAI reportedly plans to charge up to $20,000 a month for specialized AI ‘agents’

OpenAI may be planning to charge up to $20,000 per month for specialized AI “agents,” according to The Information.

The publication reports that OpenAI intends to launch several “agent” products tailored for different applications, including sorting and ranking sales leads and software engineering. One, a “high-income knowledge worker” agent, will reportedly be priced at $2,000 a month. Another, a software developer agent, is said to cost $10,000 a month.

OpenAI’s most expensive rumored agent, priced at the aforementioned $20,000-per-month tier, will be aimed at supporting “PhD-level research,” according to The Information.

It’s unclear when these agentic tools might launch or which customers will be eligible to purchase them. But The Information notes that SoftBank, an OpenAI investor, has committed to spending $3 billion on OpenAI’s agent products this year alone.

Where does value ultimately accrue?

2000's "Search Wars"

Google

YAHOO!

altavista

LYCOS

Ask Jeeves

excite

infoseek

2020's "AI Wars"

OpenAI

ANTHROP\C

xAI

Gemini

Copilot

perplexity

MISTRAL AI_

deepseek

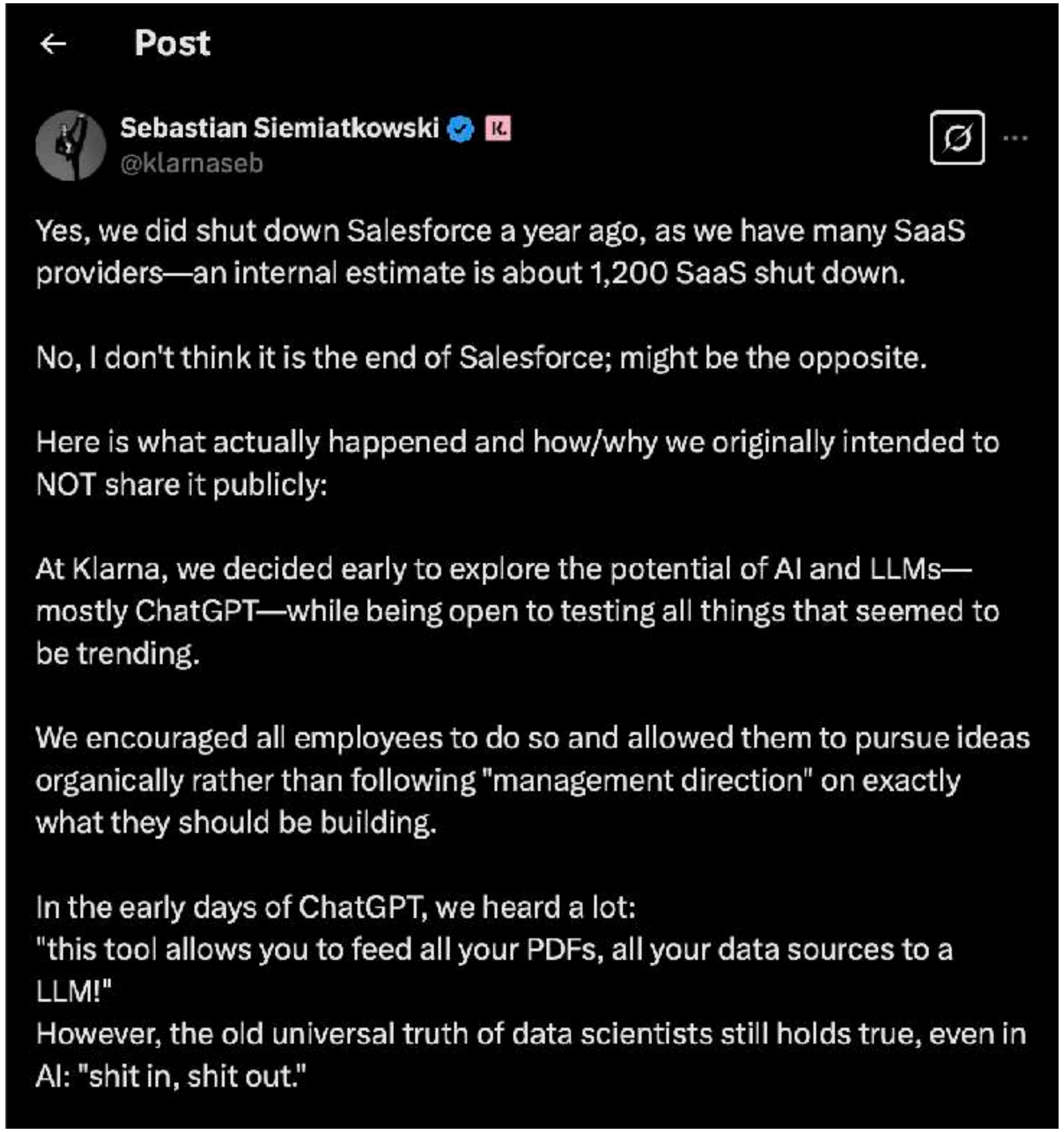
Inflection

For illustrative purposes only and reflects a subset of companies in each category, and should not be considered to be all inclusive.
References to a particular company or investment should not be considered a recommendation of any company, security, or investment.

Enterprise software?

Klarna
salesforce

Klarna CEO explains



Sebastian Siemiatkowski
@klarnaseb

Yes, we did shut down Salesforce a year ago, as we have many SaaS providers—an internal estimate is about 1,200 SaaS shut down.

No, I don't think it is the end of Salesforce; might be the opposite.

Here is what actually happened and how/why we originally intended to NOT share it publicly:

At Klarna, we decided early to explore the potential of AI and LLMs—mostly ChatGPT—while being open to testing all things that seemed to be trending.

We encouraged all employees to do so and allowed them to pursue ideas organically rather than following "management direction" on exactly what they should be building.

In the early days of ChatGPT, we heard a lot:
"this tool allows you to feed all your PDFs, all your data sources to a LLM!"
However, the old universal truth of data scientists still holds true, even in AI: "shit in, shit out."

My thoughts about enterprise software

- Companies need
  - Unification
  - Adaptation
  - Control

Conclusion

GROWTH
MINDSET

MISTAKES HELP ME LEARN

I IMPROVE WITH PRACTICE

FEEDBACK IS VALUABLE

IS THIS MY BEST WORK?

I WON'T GIVE UP

FIXED
MINDSET

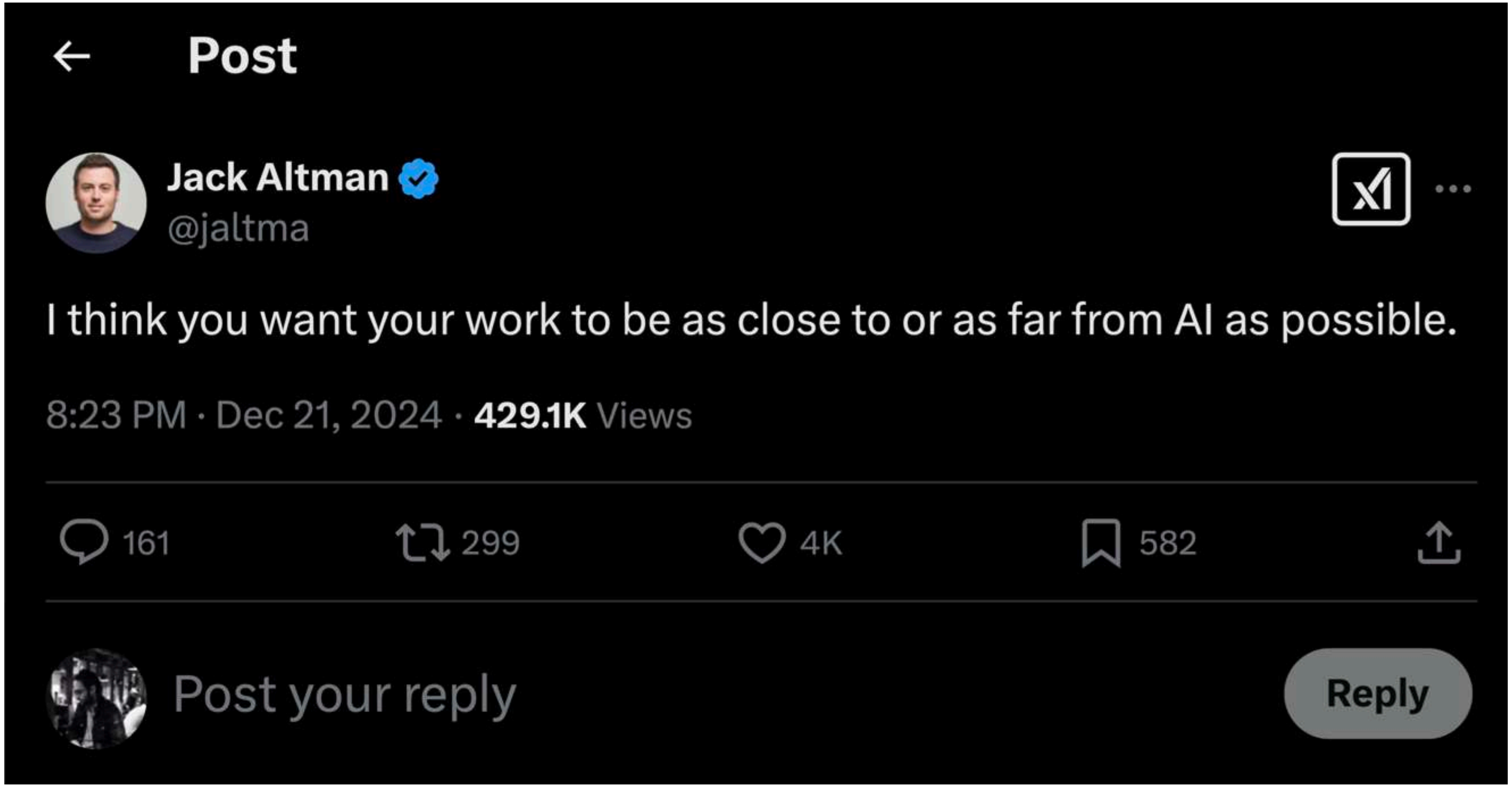
I KNOW BEST

THIS IS GOOD ENOUGH

I GIVE UP

I'LL NEVER BE THAT SMART

I WANT TO AVOID MAKING MISTAKES



Jack Altman
@jaltma

I think you want your work to be as close to or as far from AI as possible.

8:23 PM · Dec 21, 2024 · **429.1K** Views

Agency

Q&A