Trends – Artificial Intelligence

BOND

May 2025

Trends – Artificial Intelligence (AI)

May 30, 2025

Mary Meeker / Jay Simons / Daegwon Chae / Alexander Krey

BOND

Context

We set out to compile foundational trends related to AI. A starting collection of several disparate datapoints turned into this beast. As soon as we updated one chart, we often had to update another – a data game of whack-a-mole… a pattern that shows no sign of stopping…and will grow more complex as competition among tech incumbents, emerging attackers and sovereigns accelerates.

Vint Cerf, one of the 'Founders of the Internet,' said in 1999, '…they say a year in the Internet business is like a dog year – equivalent to seven years in a regular person's life.' At the time, the pace of change catalyzed by the internet was unprecedented.

Consider now that AI user and usage trending is ramping materially faster…and the machines can outpace us.

The pace and scope of change related to the artificial intelligence technology evolution is indeed unprecedented, as supported by the data. This document is filled with user, usage and revenue charts that go up-and-to-the-right… often supported by spending charts that also go up-and-to-the right.

Creators / bettors / consumers are taking advantage of global internet rails that are accessible to 5.5B citizens via connected devices; ever-growing digital datasets that have been in the making for over three decades; breakthrough large language models (LLMs) that – in effect – found freedom with the November 2022 launch of OpenAI's ChatGPT with its extremely easy-to-use / speedy user interface.

In addition, relatively new AI company founders have been especially aggressive about innovation / product releases / investments / acquisitions / cash burn and capital raises. At the same time, more traditional tech companies (often with founder involvement) have increasingly directed more of their hefty free cash flows toward AI in efforts to drive growth and fend off attackers.

And global competition – especially related to China and USA tech developments – is acute.

The outline for our document is on the next page, followed by eleven charts that help illustrate observations that follow.

We hope this compilation adds to the discussion of the breadth of change at play – technical / financial / social / physical / geopolitical. No doubt, people (and machines) will improve on the points as we all aim to adapt to this evolving journey as knowledge – and its distribution – get leveled up rapidly in new ways.

Special thanks to Grant Watson and Keeyan Sanjasaz and BOND colleagues who helped steer ideas and bring this report to life. And, to the many friends and technology builders who helped, directly or via your work, and are driving technology forward.

Outline

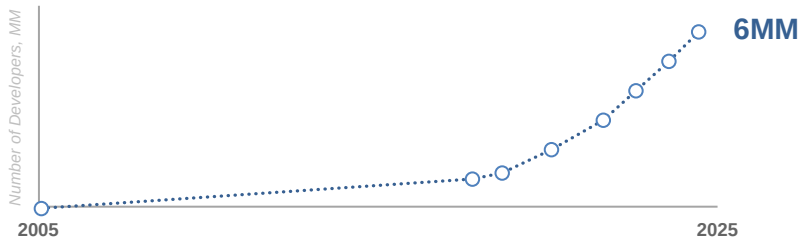
		#
1	**Seem Like Change Happening Faster Than Ever?** Yes, It Is	9-51
2	**AI User + Usage + CapEx Growth** = Unprecedented	52-128
3	**AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** = Performance Converging + Developer Usage Rising	129-152
4	**AI Usage + Cost + Loss Growth** = Unprecedented	153-247
5	**AI Monetization Threats** = Rising Competition + Open-Source Momentum + China's Rise	248-298
6	**AI & Physical World Ramps** = Fast + Data-Driven	299-307
7	**Global Internet User Ramps Powered by AI from Get-Go** = Growth We Have Not Seen Likes of Before	308-322
8	**AI & Work Evolution** = Real + Rapid	323-336

Charts Paint Thousands of Words…

1

Seem Like Change Happening Faster Than Ever? Yes, It Is

Developers in Leading Chipmaker's Ecosystem



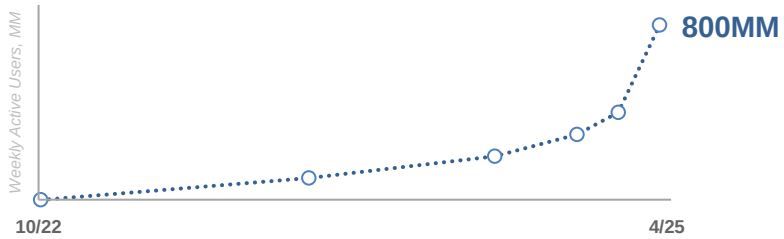
Source: Leading Chipmaker

Details on Page 38

2

AI User + Usage + CapEx Growth = Unprecedented

Leading USA-Based LLM Users



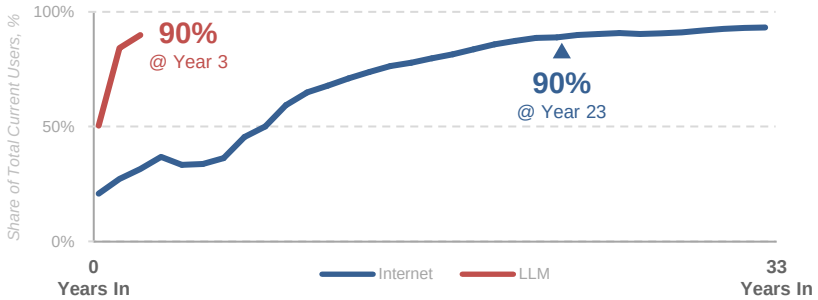
Source: Company disclosures

Details on Page 55

2.1

AI User + Usage + CapEx Growth = Unprecedented

Internet vs. Leading USA-Based LLM: Total Current Users Outside North America



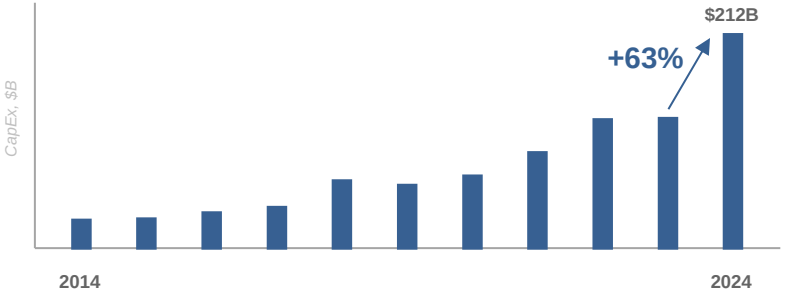
Note: LLM data is for monthly active mobile app users. App not available in select countries, including China and Russia, as of 5/25.
Source: United Nations / International Telecommunications Union (3/25), Sensor Tower (5/25)

Details on Page 56

2.2

AI User + Usage + CapEx Growth = Unprecedented

Big Six USA Technology Company CapEx*



*Apple, NVIDIA, Microsoft, Alphabet, Amazon (AWS only), & Meta Platforms
Source: Capital IQ (3/25), Morgan Stanley

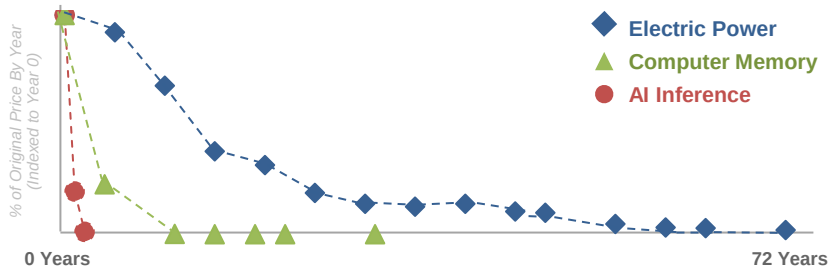
Details on Page 97

…Charts Paint Thousands of Words…

3

AI Model Compute Costs High / Rising + Inference Costs Per Token Falling = Performance Converging + Developer Usage Rising

Cost of Key Technologies Relative to Launch Year



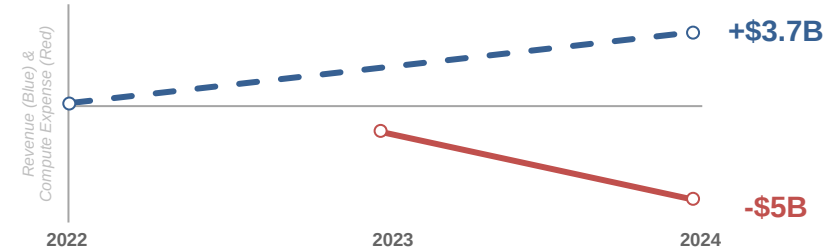
Note: Per-token inference costs shown.
Source: Richard Hirsh; John McCallum; OpenAI

Details on Page 138

4

AI Usage + Cost + Loss Growth = Unprecedented

Leading USA-Based AI LLM Revenue vs. Compute Expense



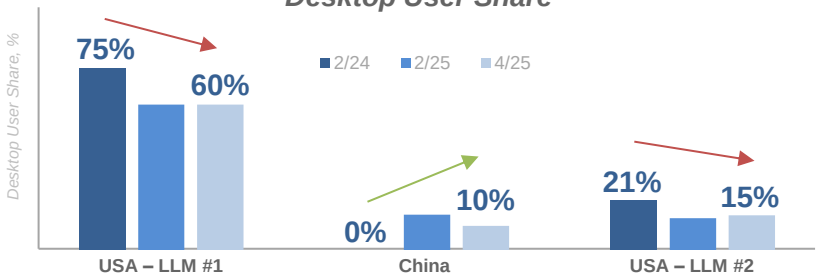
Note: Figures are estimates.
Source: The Information, public estimates

Details on Page 173

5

AI Monetization Threats = Rising Competition + Open-Source Momentum + China's Rise

Leading USA LLMs vs. China LLM Desktop User Share



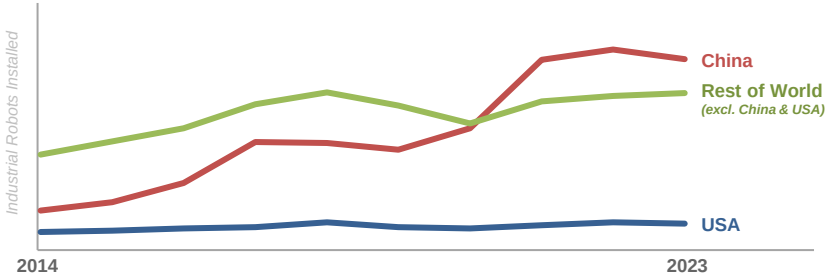
Note: Data is non-deduped. Share is relative, measured across six leading global LLMs.
Source: YipitData (5/25)

Details on Page 293

5.1

AI Monetization Threats = Rising Competition + Open-Source Momentum + China's Rise

China vs. USA vs. Rest of World Industrial Robots Installed



Note: Data as of 2023.
Source: International Federation of Robotics

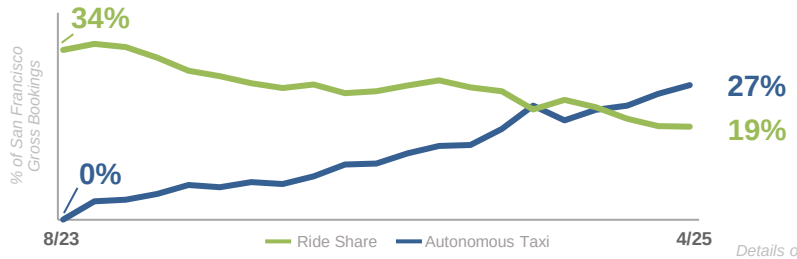
Details on Page 289

…Charts Paint Thousands of Words

6

AI & Physical World Ramps = Fast + Data-Driven

A Ride Share vs. Autonomous Taxi Provider, San Francisco Operating Zone Market Share



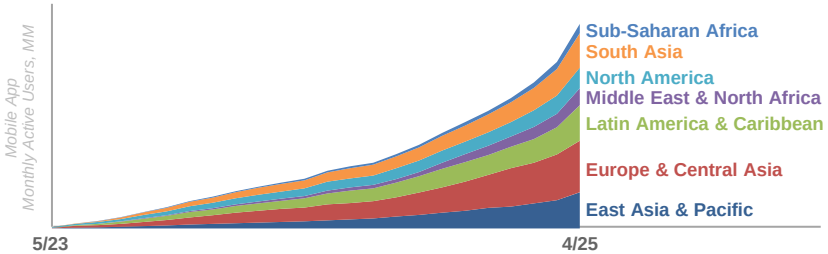
Source: YipitData (4/25)

Details on Page 302

7

Global Internet User Ramps Powered by AI from Get-Go = Growth We Have Not Seen Likes of Before

Leading USA-Based LLM App Users by Region



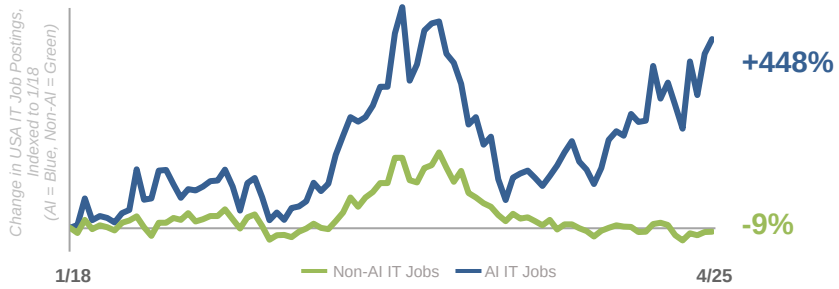
Note: Region definitions per World Bank definitions. China not included in East Asia figures. Data for standalone app only. Source: Sensor Tower (5/25)

Details on Page 315

8

AI & Work Evolution = Real + Rapid

USA IT Jobs – AI vs. Non-AI



Source: University of Maryland's UMD-LinkUp AIMaps (in collaboration with Outrigger Group) (5/25)

Details on Page 332

Overview…

To say the world is changing at unprecedented rates is an understatement.
Rapid and transformative technology innovation / adoption represent key underpinnings of these changes.
As does leadership evolution for the global powers.

Google's founding mission (1998) was to 'organize the world's information and make it universally accessible and useful.'
Alibaba's founding mission (1999) was to 'make it easy to do business anywhere.'
Facebook's founding mission (2004) was 'to give people the power to share and make the world more open and connected.'

Fast forward to today with the world's organized, connected and accessible information being supercharged by artificial intelligence, accelerating computing power, and semi-borderless capital…all driving massive change.

Sport provides a good analogy for AI's constant improvements. As athletes continue to wow us and break records, their talent is increasingly enhanced by better data / inputs / training.
The same is true for businesses, where computers are ingesting massive datasets to get smarter and more competitive.
Breakthroughs in large models, cost-per-token declines, open-source proliferation and chip performance improvements are making new tech advances increasingly more powerful, accessible, and economically viable.

OpenAI's ChatGPT – based on user / usage / monetization metrics – is history's biggest 'overnight' success (nine years post-founding). AI usage is surging among consumers, developers, enterprises and governments.
And unlike the Internet 1.0 revolution – where technology started in the USA and steadily diffused globally – ChatGPT hit the world stage all at once, growing in most global regions simultaneously.

Meanwhile, platform incumbents and emerging challengers are racing to build and deploy the next layers of AI infrastructure: agentic interfaces, enterprise copilots, real-world autonomous systems, and sovereign models.

Rapid advances in artificial intelligence, compute infrastructure, and global connectivity are fundamentally reshaping how work gets done, how capital is deployed, and how leadership is defined – across both companies and countries.

At the same time, we have leadership evolution among the global powers, each of whom is challenging the other's competitive and comparative advantage. We see the world's most powerful countries revved up by varying degrees of economic / societal / territorial aspiration…

…Overview

…Increasingly, two hefty forces – technological and geopolitical – are intertwining.
Andrew Bosworth (Meta Platforms CTO), on a recent 'Possible' podcast described the current state of AI as *our space race* and *the people we're discussing, especially China, are highly capable… there's very few secrets. And there's just progress. And you want to make sure that you're never behind.*

The reality is AI leadership could beget geopolitical leadership – and not vice-versa.

This state of affairs brings tremendous uncertainty…yet it leads us back to one of our favorite quotes – *Statistically speaking, the world doesn't end that often*, from former T. Rowe Price Chairman and CEO Brian Rogers.

As investors, we always assume everything can go wrong, but the exciting part is the consideration of what can go right. Time and time again, the case for optimism is one of the best bets one can make. The magic of watching AI do your work for you feels like the early days of email and web search – technologies that fundamentally changed our world. The better / faster / cheaper impacts of AI seem just as magical, but even quicker.

No doubt, these are also dangerous and uncertain times.
But a long-term case for optimism for artificial intelligence is based on the idea that intense competition and innovation… increasingly-accessible compute…rapidly-rising global adoption of AI-infused technology…and thoughtful and calculated leadership can foster sufficient trepidation and respect, that in turn, could lead to Mutually Assured Deterrence.

For some, the evolution of AI will create a race to the bottom; for others, it will create a race to the top.
The speculative and frenetic forces of capitalism and creative destruction are tectonic.
It's undeniable that it's 'game on,' especially with the USA and China and the tech powerhouses charging ahead.

In this document, we share data / research / benchmarks from third parties that use methodologies they deem to be effective – we are thankful for the hard work so many are doing to illustrate trending during this uniquely dynamic time.
Our goal is to add to the discussion.

Outline

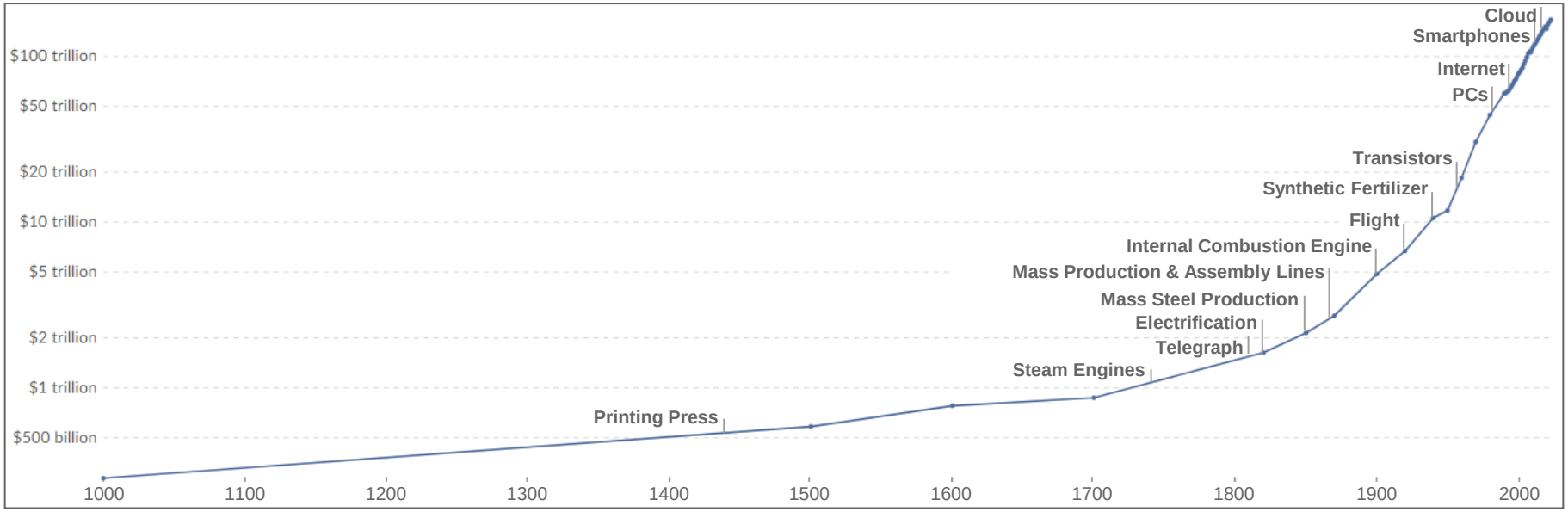
1. **Seem Like Change Happening Faster Than Ever?**
   Yes, It Is
2. **AI User + Usage + CapEx Growth** =
   Unprecedented
3. **AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** =
   Performance Converging + Developer Usage Rising
4. **AI Usage + Cost + Loss Growth** =
   Unprecedented
5. **AI Monetization Threats** =
   Rising Competition + Open-Source Momentum + China's Rise
6. **AI & Physical World Ramps** =
   Fast + Data-Driven
7. **Global Internet User Ramps Powered by AI from Get-Go** =
   Growth We Have Not Seen Likes of Before
8. **AI & Work Evolution** =
   Real + Rapid

Technology Compounding =

Numbers Behind The Momentum

Technology Compounding Over Thousand-Plus Years = Better + Faster + Cheaper → More…

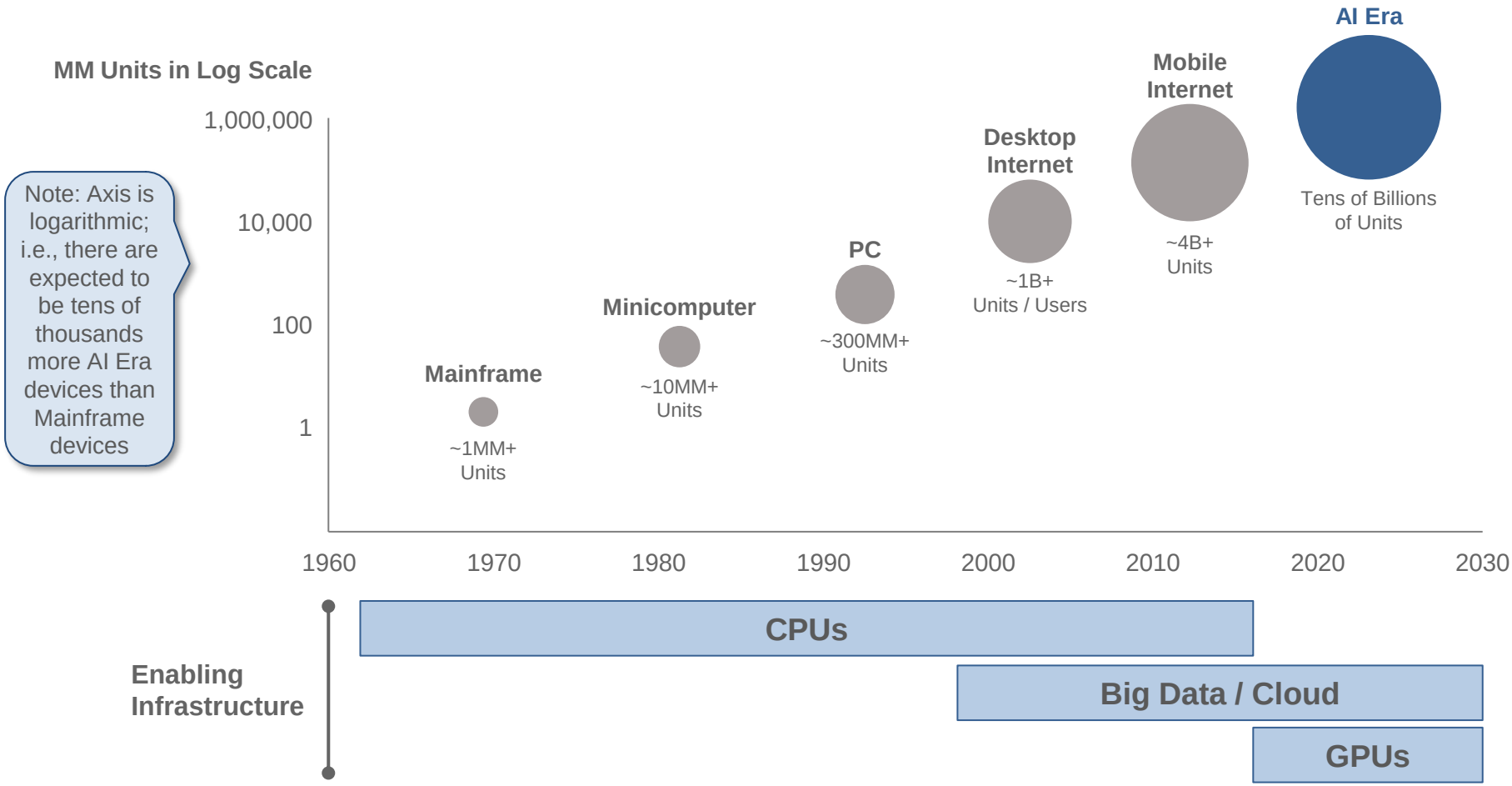
Global GDP – Last 1,000+ Years, per Maddison Project



Note: Chart expressed in trillions of real GDP as measured by 2011 'GK$' on a logarithmic scale. GK$ (Gross Knowledge Dollars) is an informal term used to estimate the potential business value of a specific insight, idea, or proprietary knowledge. It reflects how much that knowledge could be worth if applied effectively, even if it hasn't yet generated revenue.
Source: Microsoft, 'Governing AI: A Blueprint for the Future,' Microsoft Report (5/23); Data via Maddison Project & Our World in Data

…Technology Compounding Over Fifty-Plus Years = Better + Faster + Cheaper → More

Computing Cycles Over Time – 1960s-2020s, per Morgan Stanley



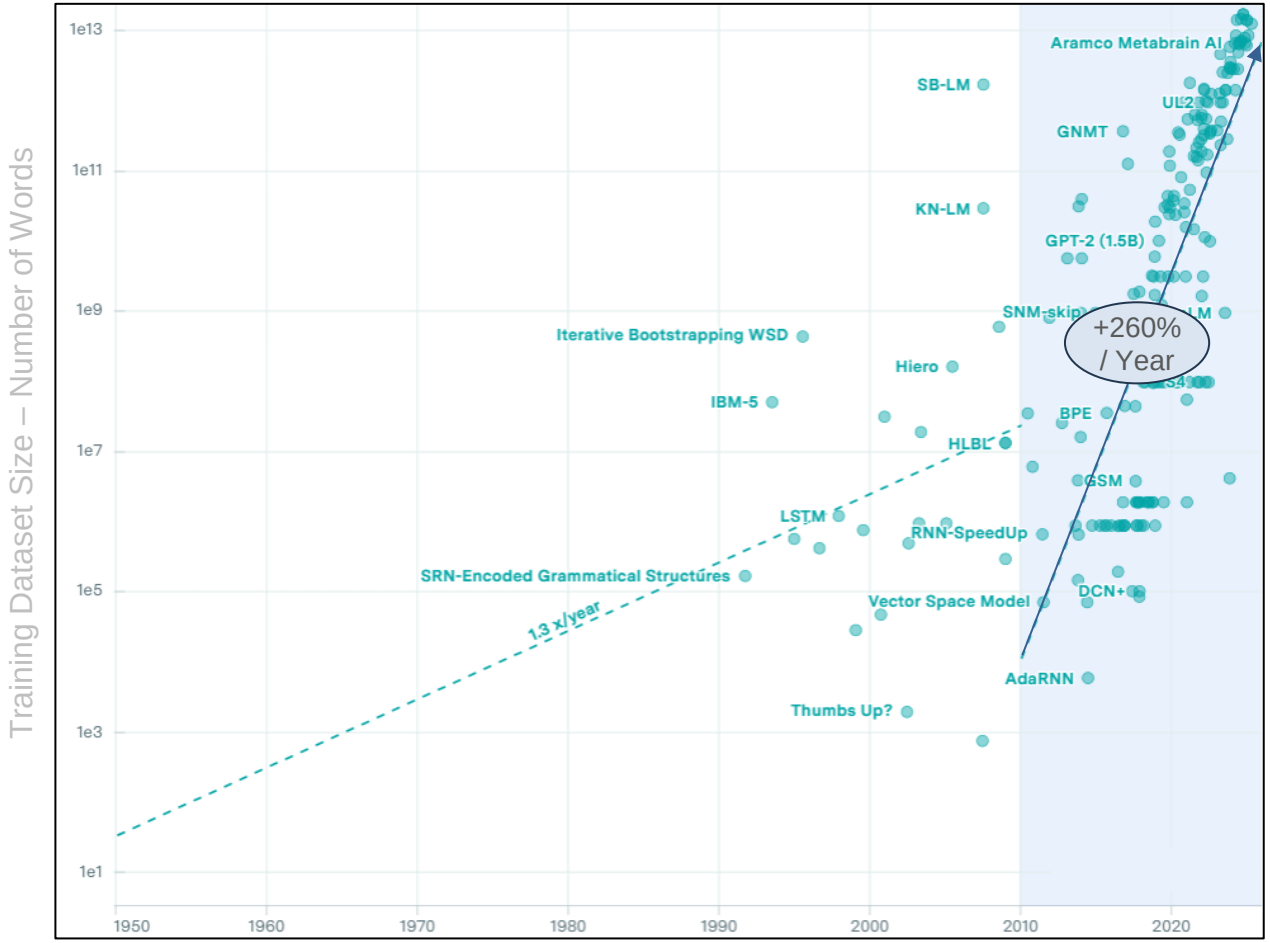
Note: PC units as of 2000. Desktop internet users as of 2005, installed base as of 2010. Mobile internet units are the installed based of smartphones & tablets in 2020. Cloud & data center capex includes Google, Amazon, Microsoft, Meta, Alibaba, Apple, IBM, Oracle, Tencent, & Baidu for ten years ending 2022. 'Tens of billions of units' refers to the potential device & user base that could end up using AI technology; this includes smartphones, IOT devices, robotics, etc. Source: Weiss et al. 'AI Index: Mapping the $4 Trillion Enterprise Impact' via Morgan Stanley (10/23)

AI Technology Compounding =

Numbers Behind The Momentum

260% Annual Growth Over Fifteen Years of…
Data to Train AI Models Led To…

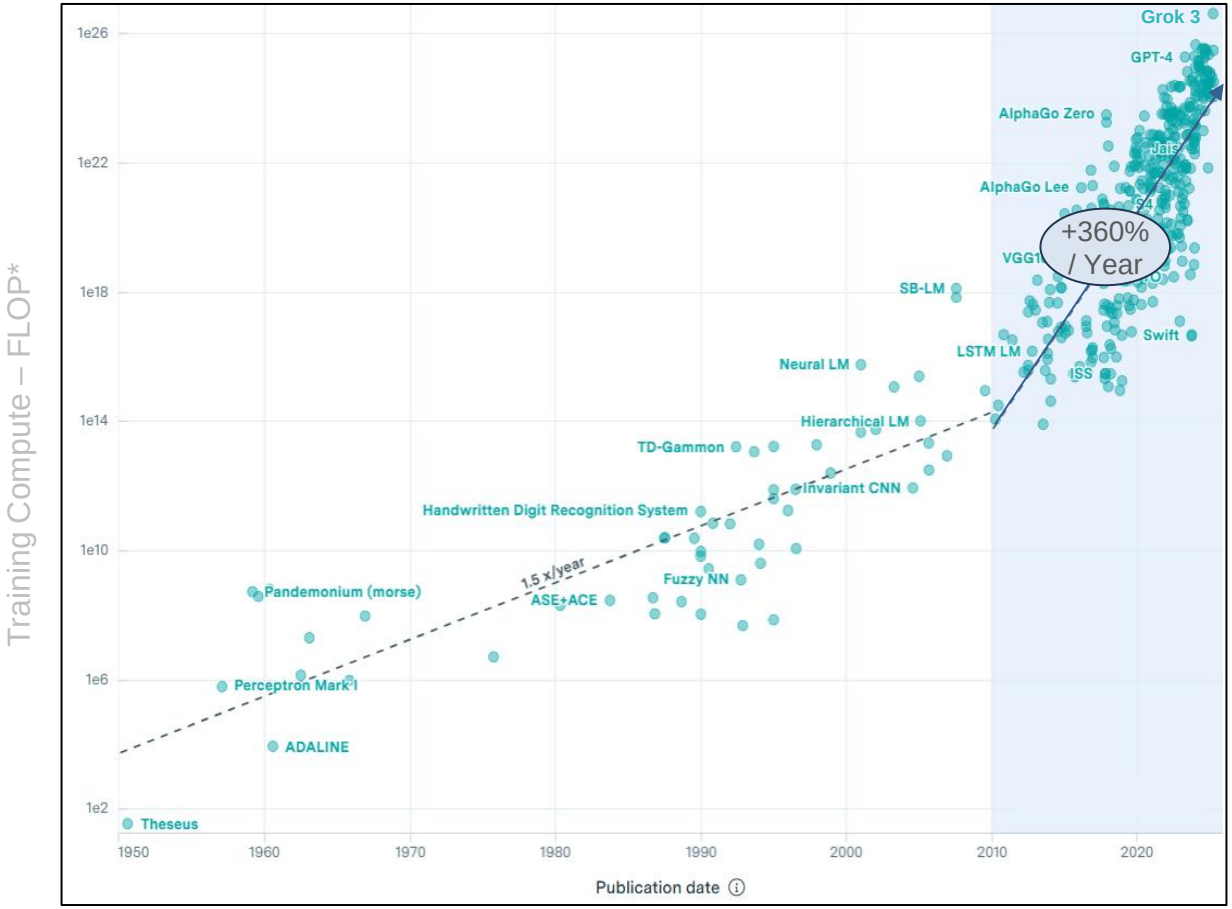
Training Dataset Size (Number of Words) for Key AI Models – 1950-2025, per Epoch AI



Note: Only "notable" language models shown (per Epoch AI, includes state of the art improvement on a recognized benchmark, >1K citations, historically relevant, with significant use).
Source: Epoch AI (5/25)

…360% Annual Growth Over Fifteen Years of… Compute to Train AI Models Led To…

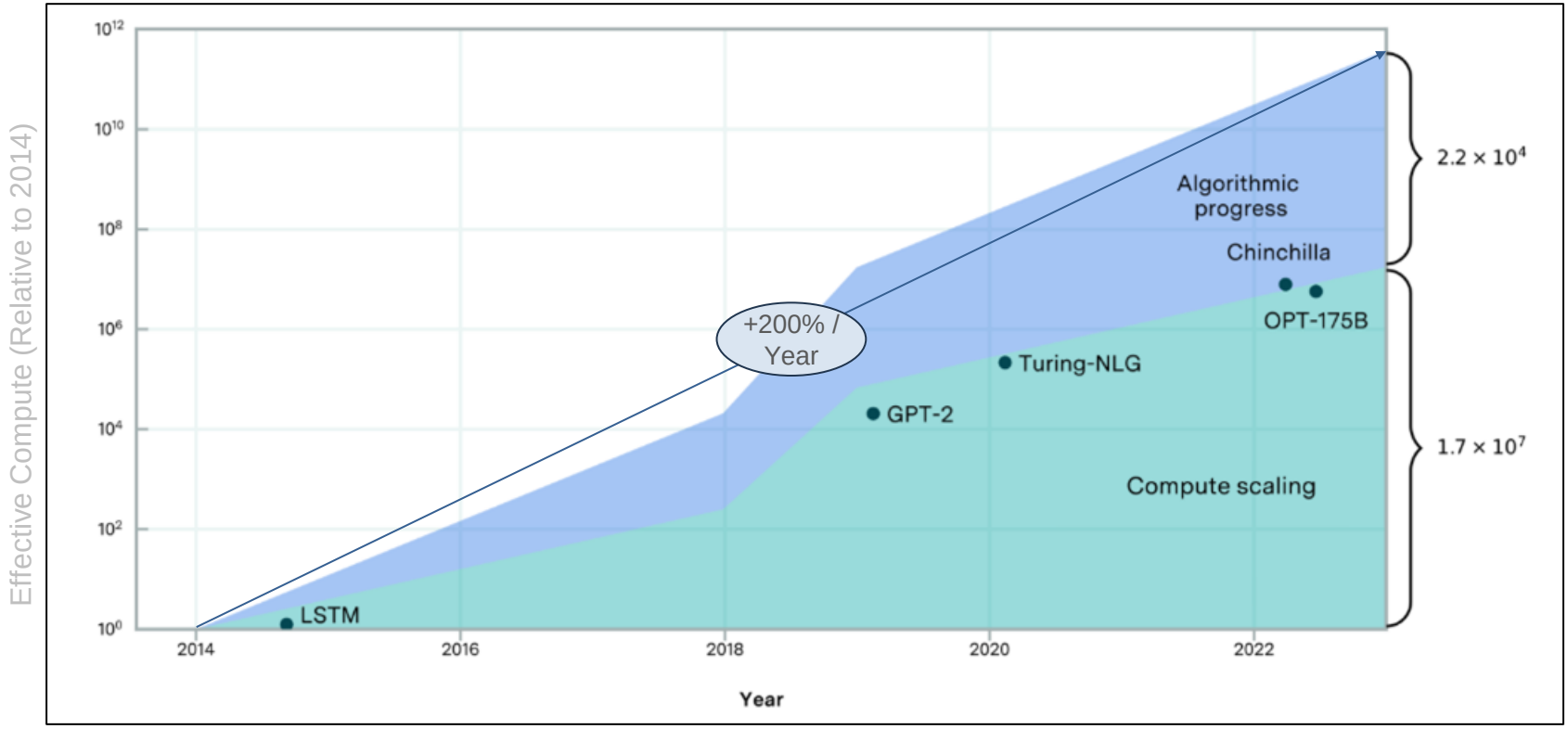
Training Compute (FLOP) for Key AI Models – 1950-2025, per Epoch AI



*A FLOP (floating point operation) is a basic unit of computation used to measure processing power, representing a single arithmetic calculation involving decimal numbers. In AI, total FLOPs are often used to estimate the computational cost of training or running a model.
Note: Only language models shown (per Epoch AI, includes state of the art improvement on a recognized benchmark, >1K citations, historically relevant, with significant use). Source: Epoch AI (5/25)

…200% Annual Growth Over Nine Years of… Compute Gains from Better Algorithms Led To…

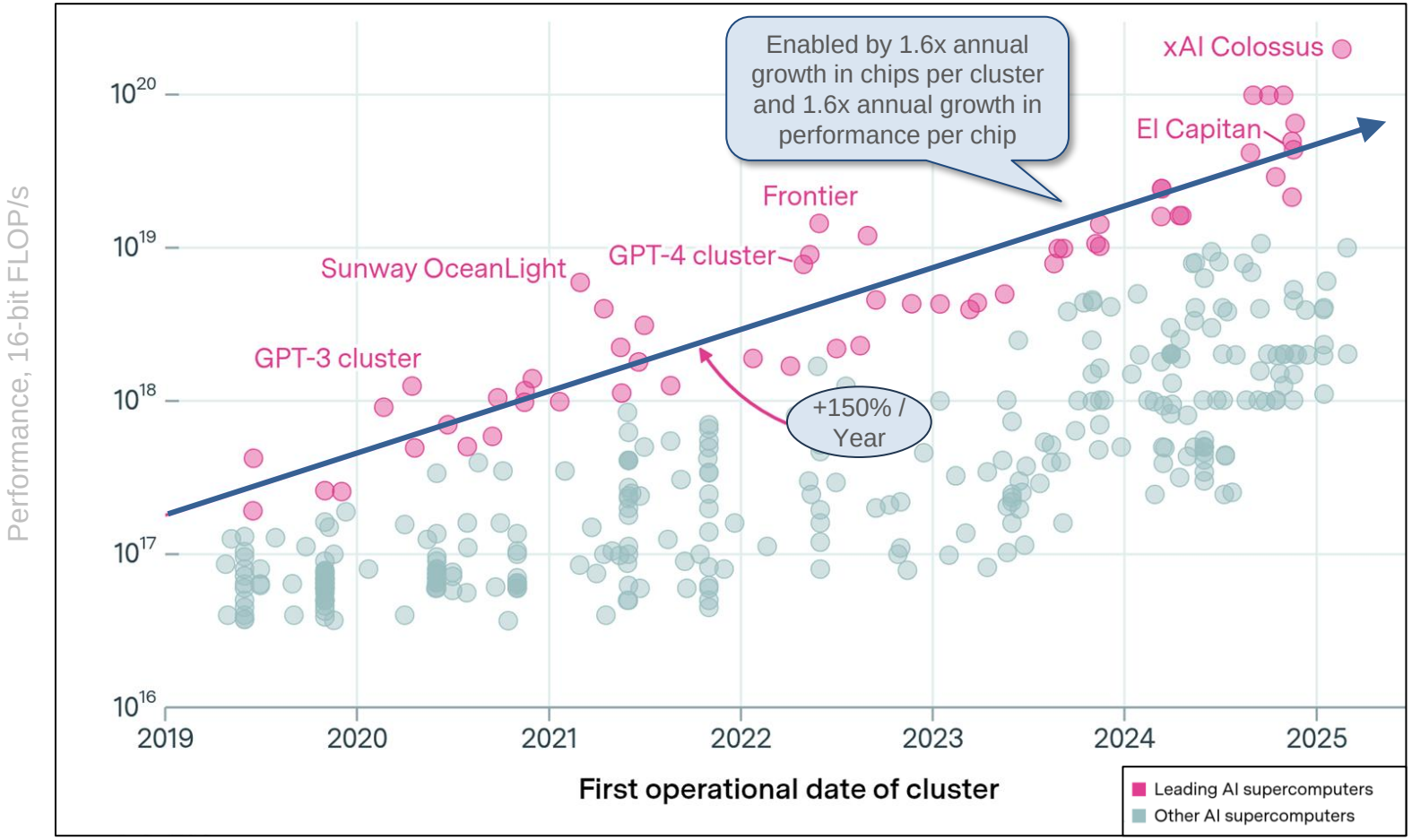
Impact of Improved Algorithms on AI Model Performance – 2014-2023, per Epoch AI



Note: Estimates how much progress comes from bigger models versus smarter algorithms, based on how much computing power you'd need to reach top performance without any improvements. Source: Epoch AI (3/24)

…150% Annual Growth Over Six Years of… Performance Gains from Better AI Supercomputers Led To…

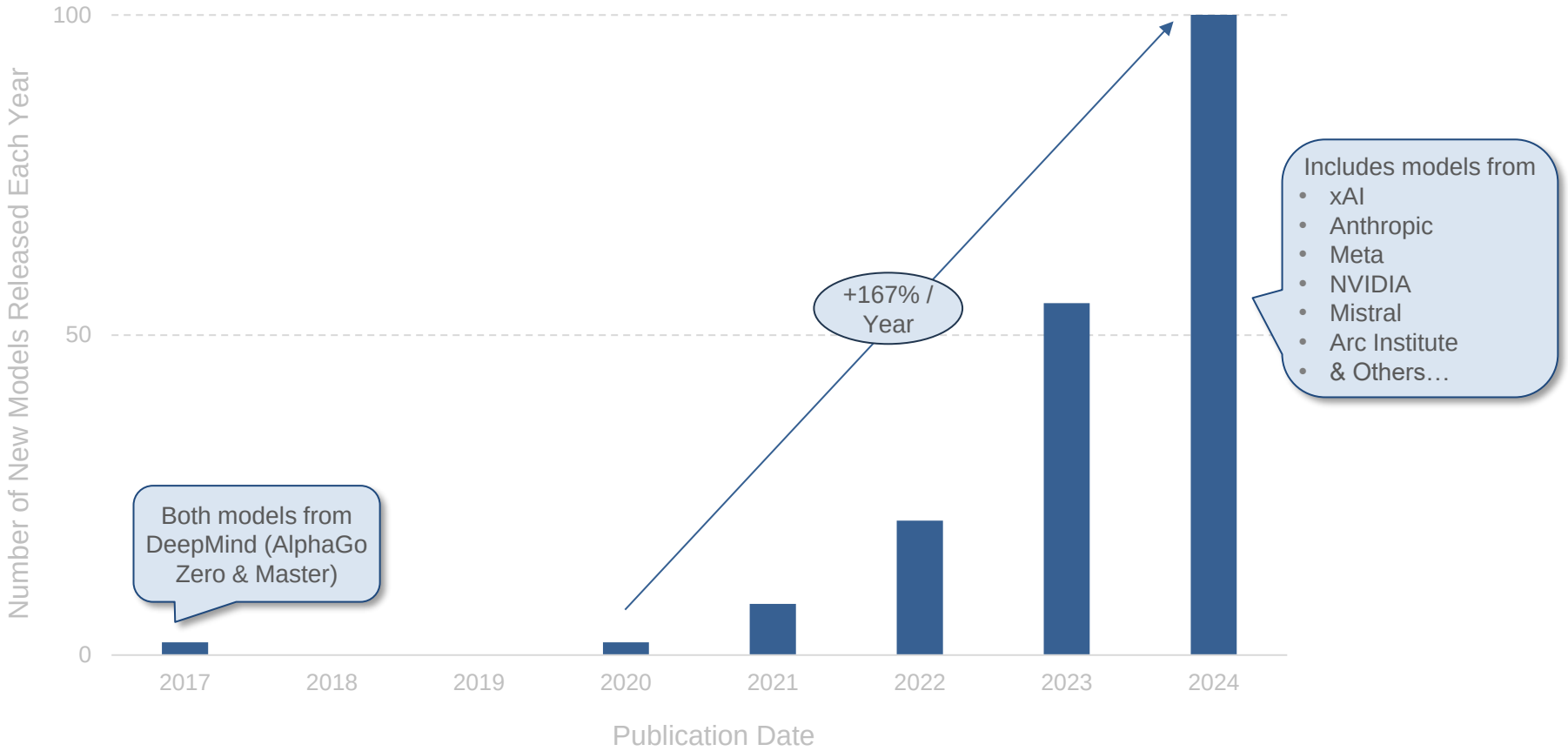
Performance of Leading AI Supercomputers (FLOP/s) – 2019-2025, per Epoch AI



Source: Epoch AI (4/25)

…167% Annual Growth Over Four Years in… Number of Powerful AI Models

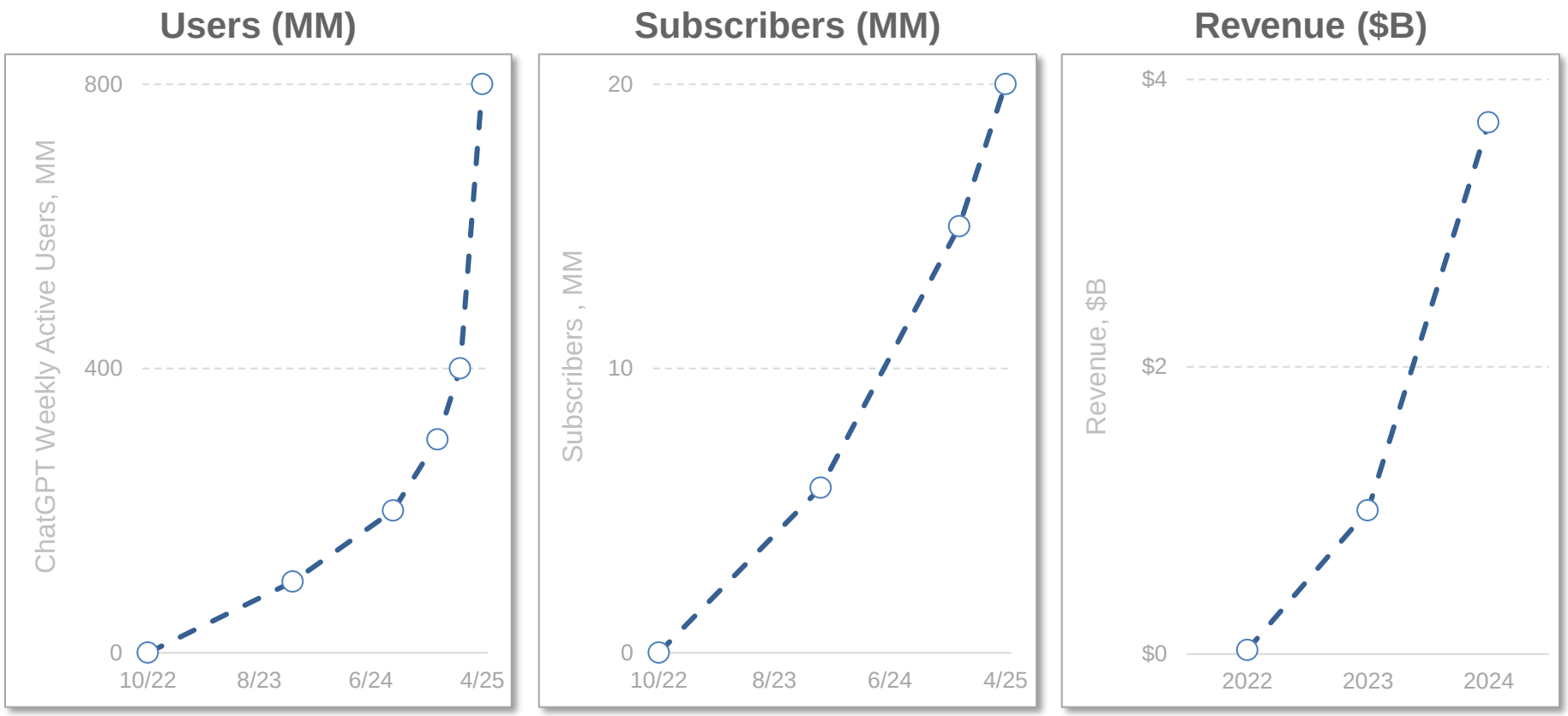
Number of New Large-Scale AI Models (Larger than 10^{23} FLOP*) – 2017-2024, per Epoch AI



**As of 4/25, 'Large-Scale AI Models' are generally defined as those with a training compute of 10^{23} FLOPs or greater, per Epoch AI.*
Source: Epoch AI (5/25)

ChatGPT AI User + Subscriber + Revenue Growth Ramps = Hard to Match, Ever

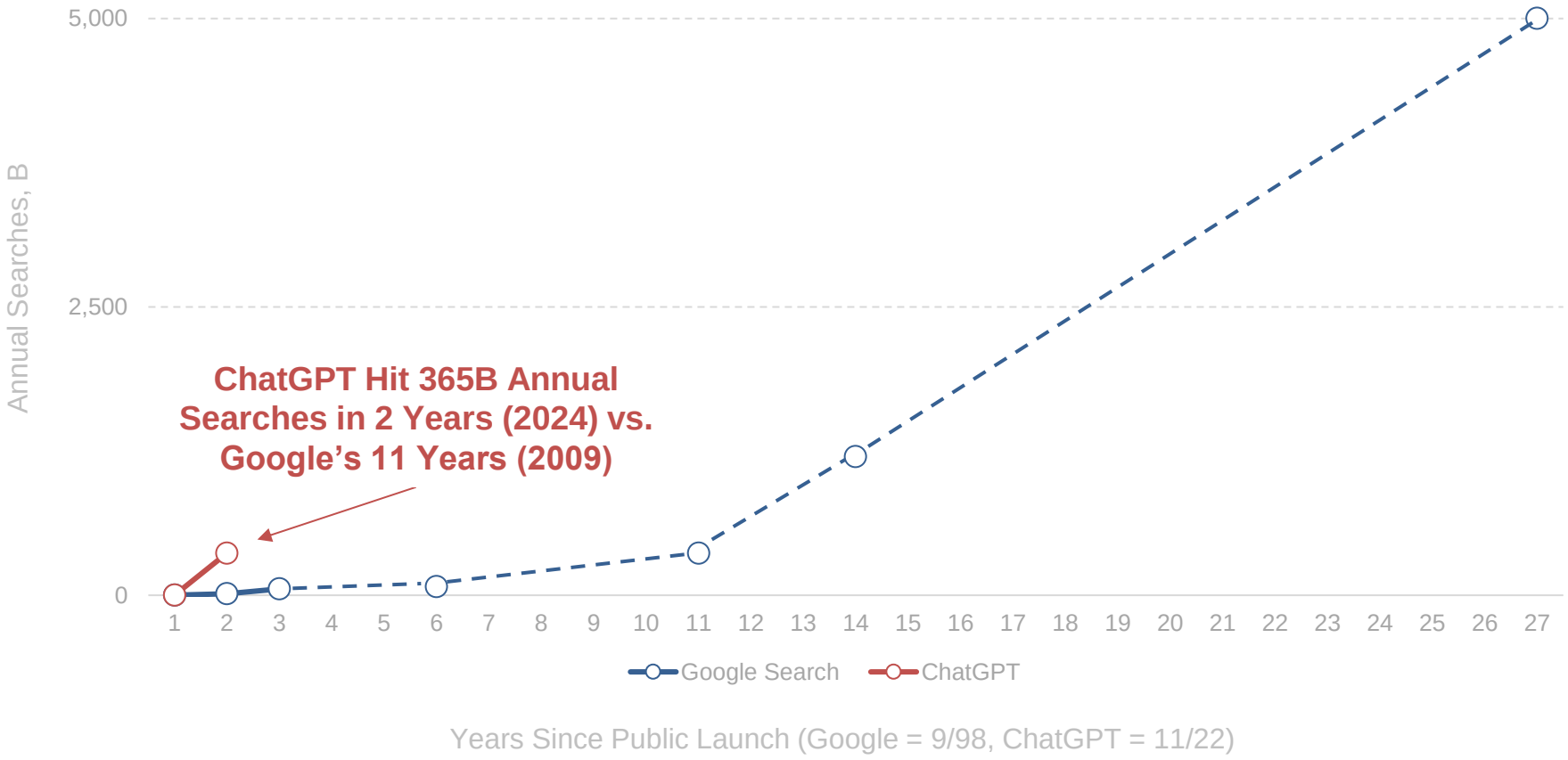
ChatGPT User + Subscriber + Revenue Growth – 10/22-4/25, per OpenAI & *The Information*



Note: 4/25 user count estimate from OpenAI CEO Sam Altman's 4/11/25 TED Talk disclosure. Revenue figures are estimates based off OpenAI disclosures. Source: OpenAI disclosures (as of 4/25), The Information (4/25) (link, link, link & link)

Time to 365B Annual Searches = ChatGPT 5.5x Faster vs. Google

Annual Searches by Year (B) Since Public Launches of Google & ChatGPT – 1998-2025, per Google & OpenAI



Note: Dashed-line bars are for years where Google did not disclose annual search volumes. Source: Google public disclosures, OpenAI (12/24). ChatGPT figures are estimates per company disclosures of ~1B daily queries

In 1998, tapping emerging Internet access, Google set out to ‘organize the world’s information and make it universally accessible and useful.’

Nearly three decades later
– after some of the fastest change humankind has seen –
a lot of information is indeed digitized / accessible / useful.

The AI-driven evolution of how we
access and move information is happening much faster…

…AI is a compounder – on internet infrastructure, which allows for wicked-fast adoption of easy-to-use broad-interest services.

Knowledge Distribution Evolution =

Over ~Six Centuries

Knowledge Distribution – 1440-1992 = Static + Physical Delivery…

Printing Press – Invented 1440

Source: Wikimedia Commons

…Knowledge Distribution – 1993-2021 = Active + Digital Delivery…

Internet – Public Release 1993*

**The internet is widely agreed to have been 'publicly released' in 1993 with release of the World Wide Web (WWW) into the public domain, which allowed users to create websites; however, Tim Berners-Lee invented the World Wide Web in 1989, per CERN.*
Source: Google, USA Department of Defense, CERN

…Knowledge Distribution – 2022+ = Active + Digital + Generative Delivery

Generative AI – Public Launch of ChatGPT 2022*

ChatGPT 4o

Share

Great, I'll gather quantitative global data on how generative AI is impacting knowledge generation and distribution across all major sectors—including education, research, media, and more.

I'll let you know as soon as the findings are ready.

Research completed in 5m · 25 sources · 116 searches

Global Impact of Generative AI on Knowledge Generation and Distribution

Global Impact of Generative AI on Knowledge Generation & Distribution

Generative AI's rapid advancement since 2022 has led to an explosion in AI-created content and widespread adoption across sectors. Below we compile quantitative global data (2023–2025) illustrating how generative AI is transforming knowledge **generation** (content creation) and **distribution** (sharing and access) in academia, media, education, enterprise, and beyond.

Surge in AI-Generated Content Creation

- **Academic Publications:** AI is now contributing to scientific literature. An analysis found that *1% of all scientific articles published in 2023 showed signs of generative AI involvement* scientificamerican.com . This amounts to an estimated **60,000 research papers** potentially written (in part) by large language models scientificamerican.com . Top journals have responded – by late 2023, 87 of the 100 leading journals had issued guidance on using AI tools in writing meprillaman.com – reflecting how commonplace AI-assisted writing has become in research.
- **News and Media Content:** Generative AI is also **writing the news**. By mid-2024, about **6.96% of global news articles were AI-generated** – roughly **60,000 AI-written news articles published per day** newscatcherapi.com . Misinformation trackers have identified **1,271 news and information websites across 16 languages that rely on AI-generated content with minimal human oversight** (as of May 2025) newsguardtech.com , a sharp rise from 467 such sites in 2023 titleformat.com . Even reputable outlets use AI for routine reporting: the Associated Press, for example, was automatically generating **40,000 of its 730,000 news stories (5.5%) via AI** by mid-2023 titleformat.com . This automation has enabled greater volume and speed in news dissemination.

**We define the public launch of ChatGPT in November 2022 as the public release of Generative AI which we see as AI's 'iPhone Moment.' ChatGPT saw the fastest user ramp ever for a standalone product (5 days to secure 1MM users). Generative AI = AI that can create content – text, images, audio, or code – based on learned patterns.*
Source: OpenAI

Knowledge is a process of piling up facts; wisdom lies in their simplification.

Martin H. Fischer, German-born American Physician / Teacher / Author (1879-1962)

AI =

Many Years Before Lift-Off

AI Milestone Timeline – 1950-2022, per Stanford University…

10/50: Alan Turing creates his Turing Test to measure computer intelligence, positing that computers could think like humans

6/56: Stanford computer scientist John McCarthy convenes the Dartmouth Conference on 'Artificial Intelligence,' a term he coined

1/62: Arthur Samuel, an IBM computer scientist, creates a self-learning program that proves capable of defeating a top USA checkers champion

1/66: Stanford researchers deploy Shakey, the first general-purpose mobile robot that can reason about its own actions

AI 'Winter[1]'
(1967-1996)

5/97: Deep Blue, IBM's chess-playing computer, defeats Garry Kasparov, the world chess champion at the time

9/02: Roomba, the first mass-produced autonomous robotic vacuum cleaner that can navigate homes, is launched

10/05: A Stanford team build a driverless car named Stanley; it completes a 132-mile course, winning the DARPA Grand Challenge

4/10: Apple acquires Siri voice assistant & integrates it into iPhone 4S model one year later

6/14: Eugene Goostman, a chatbot, passes the Turing Test, with 1/3 of judges believing that Eugene is human

6/18: OpenAI releases GPT-1, the first of their large language models

6/20: OpenAI releases GPT-3, an AI tool for automated conversations; Microsoft exclusively licenses the model

11/22: OpenAI releases ChatGPT to the public

1: AI 'Winter' was a term used by Nils J. Nilsson, the Kumagai Professor of Engineering in computer science at Stanford University, to describe the period during which AI continued to make conceptual progress but could boast no significant practical successes. This subsequently led to a drop in AI interest and funding. Includes data from sources beyond Stanford.
Source: Stanford University & Stanford Law School sources, iRobot, TechCrunch, BBC, OpenAI. Data aggregated by BOND.

…AI Milestone Timeline – 2023-2025, per Stanford University

3/23: OpenAI releases GPT-4, a multimodal* model capable of processing both text & images

3/23: Google releases Bard, its ChatGPT competitor

11/23: 28 countries, including USA, EU members & China, sign Bletchley Declaration on AI Safety

4/24: Meta Platforms releases its open-source** Llama 3 model with 70B parameters

5/24: Google introduces AI overviews to augment its search functions

9/24: Alibaba releases 100 open-source Qwen 2.5 models, with performance in line with Western competitors

1/25: DeepSeek releases its R1 & R1-Zero open-source reasoning models

2/25: OpenAI releases GPT-4.5, Anthropic releases Claude 3.7 Sonnet, & xAI releases Grok 3

3/23: Microsoft Integrates Copilot into its 365 product suite

3/23: Anthropic releases Claude, its AI assistant focused on safety & inter-pretability

3/24: USA Department of Homeland Security unveils its AI Roadmap Strategy

5/24: OpenAI releases GPT-4o, which has full multimodality across audio, visual, & text inputs

7/24: Apple releases Apple Intelligence, an AI system integrated into its devices, for developers

12/24: OpenAI announces o3, its highest-ever performing model

1/25: Alibaba unveils Qwen2.5-Max, which surpasses the performance of other leading models (GPT-4o, Claude 3.5) on some reasoning tests

4/25: ChatGPT reaches 800MM weekly users[1]

**Multimodal = AI that can understand and process multiple data types (e.g., text, images, audio) together.*
***Open-source = AI models and tools made publicly available for use, modification, and redistribution.*
1) 4/25 estimate from OpenAI CEO Sam Altman's 4/11/25 TED Talk disclosure.
Source: Aggregated by BOND from OpenAI, Microsoft, Google, Anthropic, Meta, Apple, Alibaba, Deepseek, UK Government, US Department of Homeland Security. China data may be subject to informational limitations due to government restrictions.

AI =

Circa Q2:25

Top Ten Things AI Can Do Today, per ChatGPT

ChatGPT 4o

1. **Write or edit anything**
   ➤ Emails, essays, contracts, poems, code — instantly and fluently.
2. **Summarize and explain complex material**
   ➤ Simplifies PDFs, legal docs, research, or code into plain English.
3. **Tutor you on nearly any subject**
   ➤ Learn math, history, languages, or test prep, step by step.
4. **Be your thinking partner**
   ➤ Brainstorm ideas, debug logic, or pressure-test assumptions.
5. **Automate repetitive work**
   ➤ Generate reports, clean data, outline slides, rewrite text.
6. **Roleplay anyone you need**
   ➤ Prep for interviews, simulate customers, rehearse conversations.
7. **Connect you to tools**
   ➤ Write code for APIs, spreadsheets, calendars, or the web.
8. **Offer therapy and companionship**
   ➤ Talk through your day, reframe thoughts, or just listen.
9. **Help you find purpose**
   ➤ Clarify values, define goals, and map actions that matter.
10. **Organize your life**
    ➤ Plan trips, build routines, structure your week or workflow.

Source: ChatGPT (5/15/25)

AI =

Circa 2030?

Top Ten Things AI Will Likely Do in Five Years, per ChatGPT

ChatGPT 4o

1. Generate human-level text, code & logic

➤ Chatbots, software engineering, business plans, legal analysis.

2. Create full-length films & games

➤ Scripts, characters, scenes, gameplay mechanics, voice acting.

3. Understand & speak like a human

➤ Emotionally aware assistants, real-time multilingual voice agents.

4. Power advanced personal assistants

➤ Life planning, memory recall, coordination across all apps and devices.

5. Operate humanlike robots

➤ Household helpers, elderly care, retail and hospitality automation.

6. Run autonomous customer service & sales

➤ End-to-end resolution, upselling, CRM integrations, 24/7 support.

7. Personalize entire digital lives

➤ Adaptive learning, dynamic content curation, individualized health coaching.

8. Build and run autonomous businesses

➤ AI-driven startups, inventory and pricing optimization, full digital operations.

9. Drive autonomous discovery in science

➤ Drug design, materials synthesis, climate modeling, novel hypothesis testing.

10. Collaborate creatively like a partner

➤ Co-writing novels, music production, fashion design, architecture.

Source: ChatGPT (5/15/25)

AI =

Circa 2035?

Top Ten Things AI Will Likely Do in Ten Years, per ChatGPT

ChatGPT 4o

1. Conduct scientific research

➤ Generate hypotheses, run simulations, design and analyze experiments.

2. Design advanced technologies

➤ Discover materials, engineer biotech, and prototype energy systems.

3. Simulate human-like minds

➤ Create digital personas with memory, emotion, and adaptive behavior.

4. Operate autonomous companies

➤ Manage R&D, finance, and logistics with minimal human input.

5. Perform complex physical tasks

➤ Handle tools, assemble components, and adapt in real-world spaces.

6. Coordinate systems globally

➤ Optimize logistics, energy use, and crisis response at scale.

7. Model full biological systems

➤ Simulate cells, genes, and organisms for research and therapy.

8. Offer expert-level decisions

➤ Deliver real-time legal, medical, and business advice.

9. Shape public debate and policy

➤ Moderate forums, propose laws, and balance competing interests.

10. Build immersive virtual worlds

➤ Generate interactive 3D environments directly from text prompts.

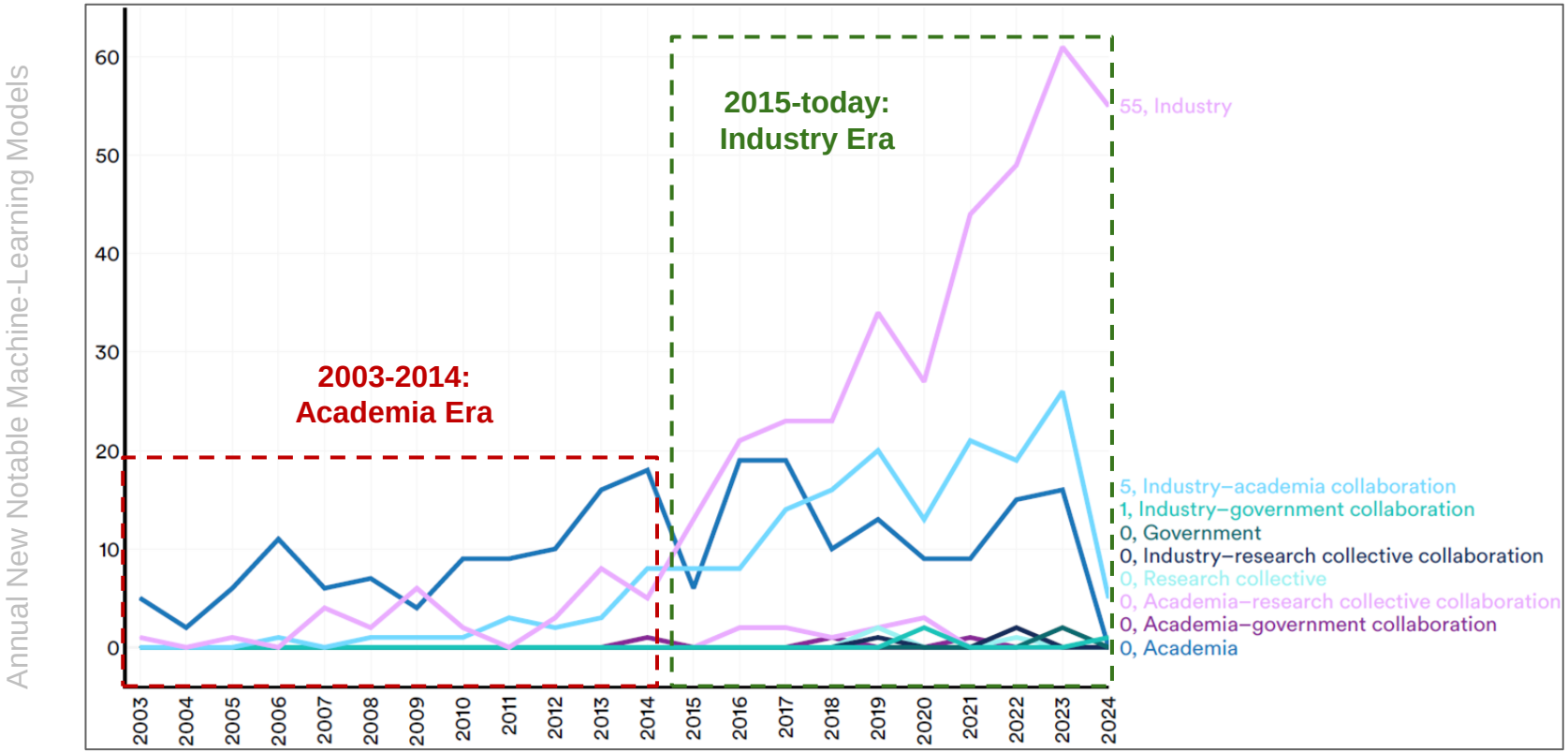
Source: ChatGPT (5/15/25)

AI Development Trending =

Unprecedented

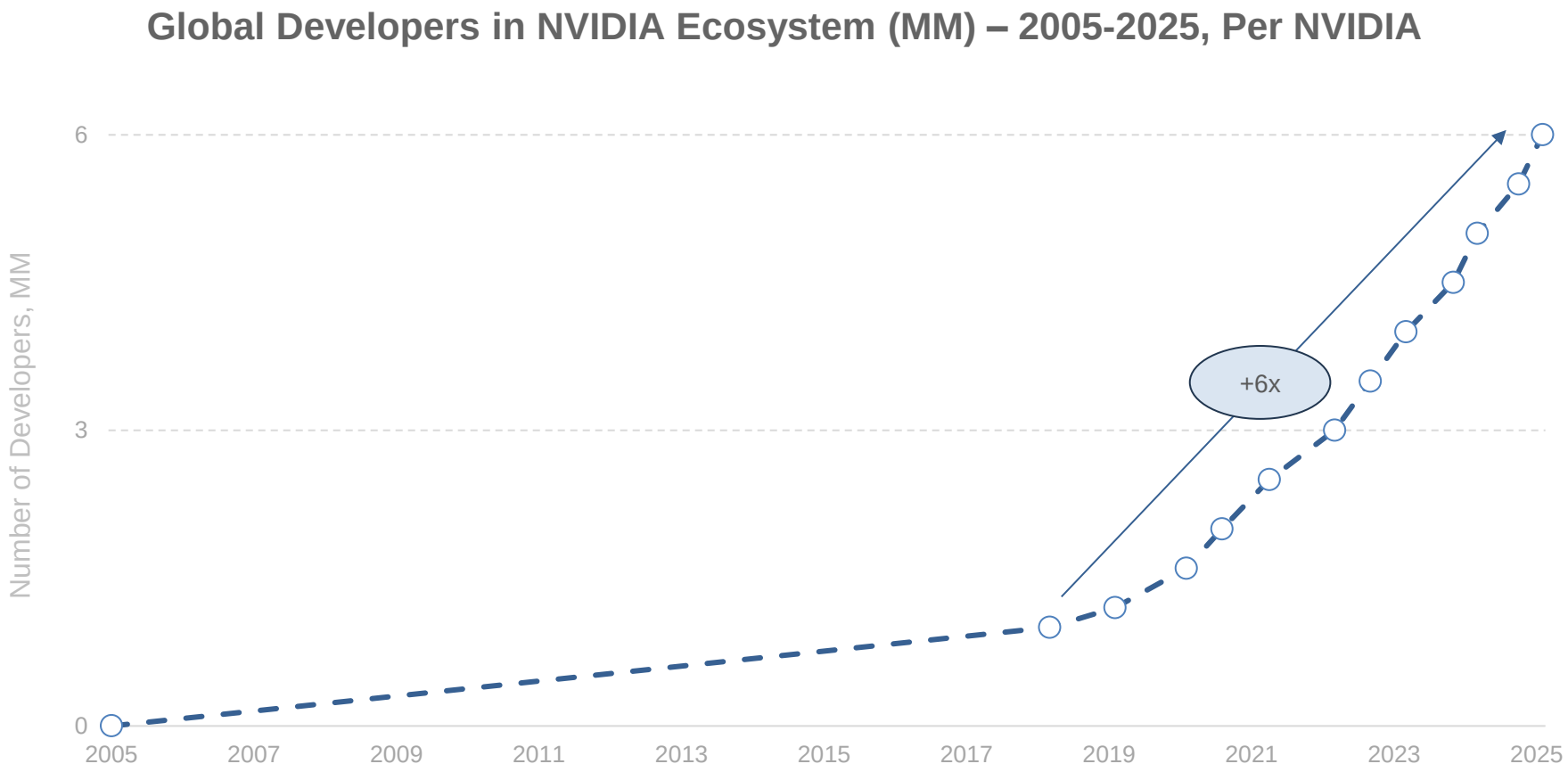
Machine-Learning Model* Trending = In 2015...
Industry Surpassed Academia as Data + Compute + Financial Needs Rose

Global Notable Machine Learning Models by Sector – 2003-2024, per Stanford HAI



**Machine Learning = A subset of AI where machines learn from patterns in data without being explicitly programmed.*
Note: Academia includes models developed by one or more institutions, including government agencies. Industry-academia collaboration excludes government partnerships and only captures partnerships between academic institutions and industry. Industry excludes models developed in partnership with any entity other than another company. Epoch AI, an AI Index data provider, uses the term 'notable machine learning models' to designate particularly influential models within the AI/machine learning ecosystem. Epoch maintains a database of 900 AI models released since the 1950s, selecting entries based on criteria such as state-of-the-art advancements, historical significance, or high citation rates. Since Epoch manually curates the data, some models considered notable by some may not be included. A count of zero academic models does not mean that no notable models were produced by academic institutions in 2023, but rather that Epoch AI has not identified any as notable. Additionally, academic publications often take longer to gain recognition, as highly cited papers introducing significant architectures may take years to achieve prominence. China data may be subject to informational limitations due to government restrictions. Source: Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

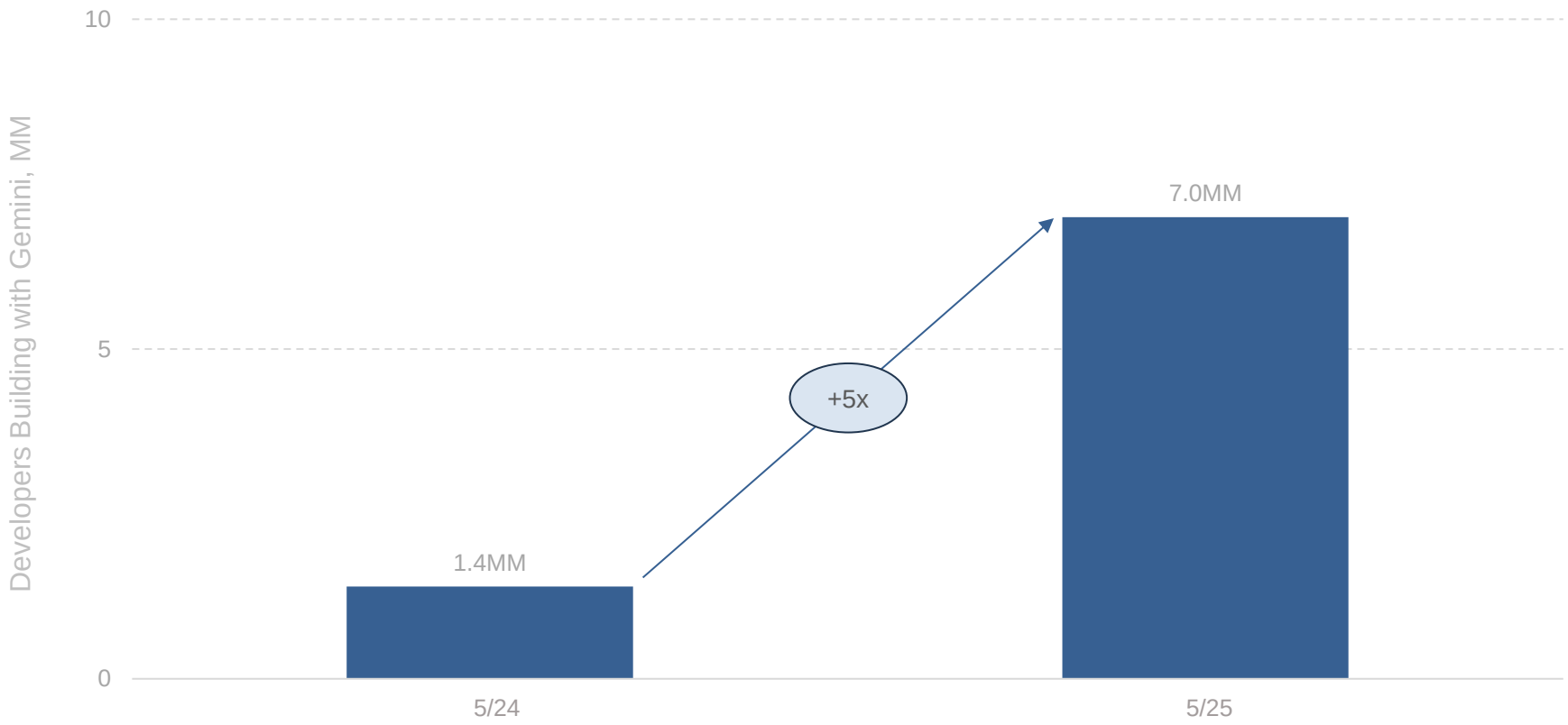
AI Developer Growth (NVIDIA Ecosystem as Proxy) = +6x to 6MM Developers Over Seven Years



Note: We assume negligible developers in NVIDIA's ecosystem in 2005 per this text from an 8/20 blog post titled '2 Million Registered Developers, Countless Breakthroughs': 'It took 13 years to reach 1 million registered developers, and less than two more to reach 2 million.' Source: NVIDIA blog posts, press releases, & company overviews

AI Developer Growth (Google Ecosystem as Proxy) = +5x to 7MM Developers Y/Y

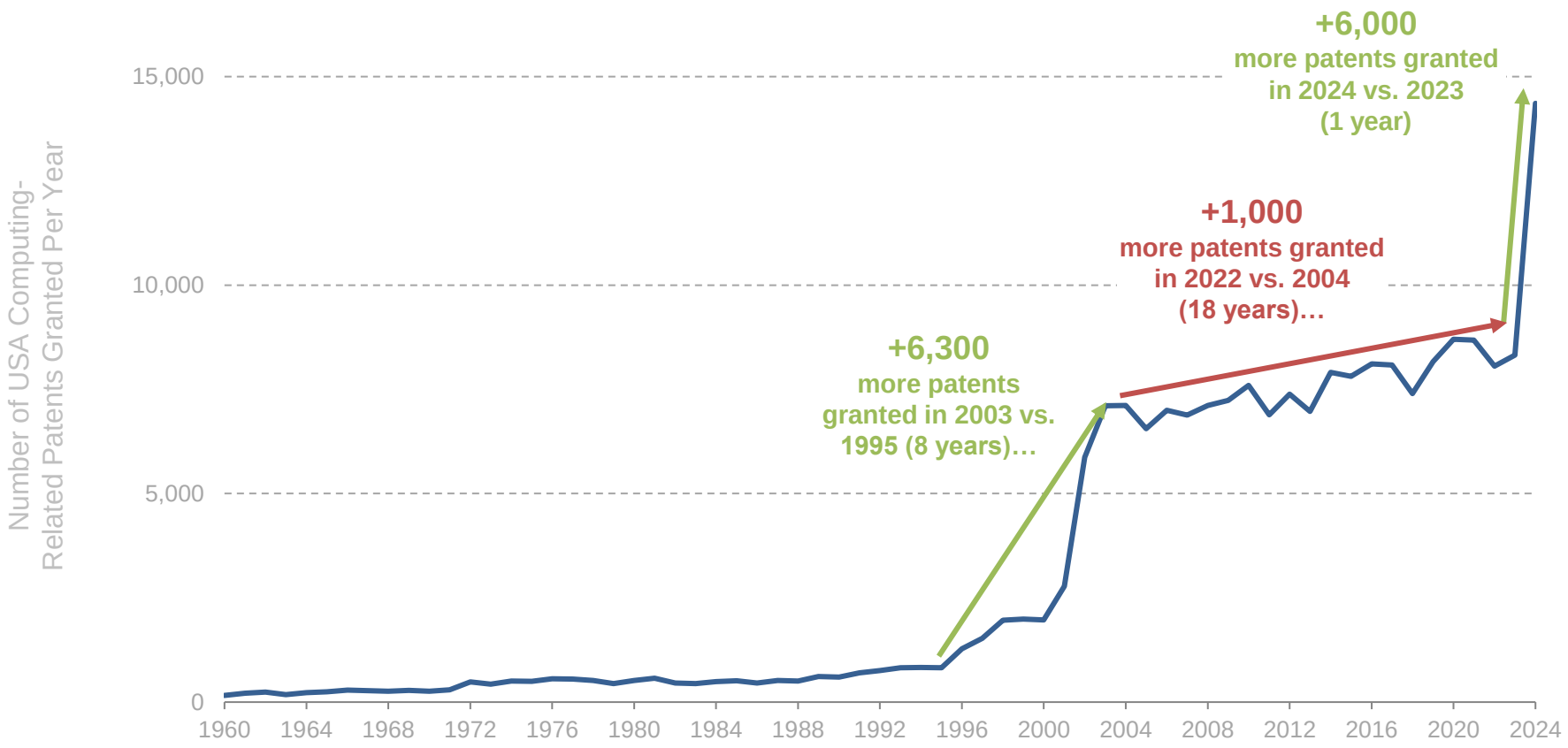
Estimated Global Developers in Google Ecosystem (MM) – 5/24-5/25, Per Google



Note: Per Google in 5/25, 'Over 7 million developers are building with Gemini, five times more than this time last year.' Source: Google, 'Google I/O 2025: From research to reality' (5/25)

Computing-Related Patent Grants, USA = Exploded…
Post-Netscape IPO (1995)...Again + Faster Post-ChatGPT Public Launch (2022)

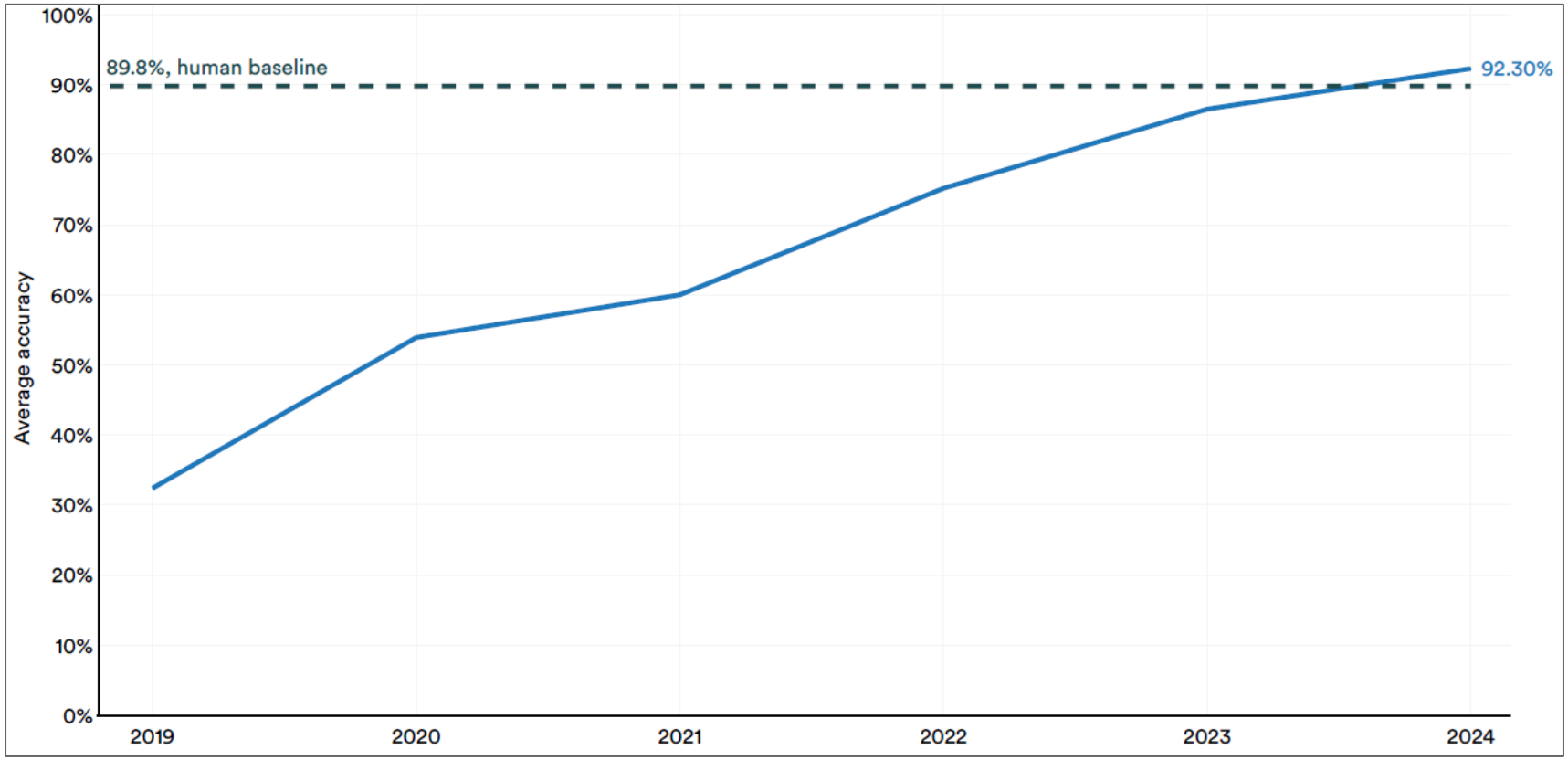
USA Computing-Related* Patents Granted Annually – 1960-2024, per USPTO



**Uses Cooperative Patent Classification (CPC) code G06, which corresponds to computing, calculating or counting patents. Google patents data changes somewhat between each query so numbers are rounded and should be viewed as directionally accurate. Source: USA Patent & Trademark Office (USPTO) via Google Patents (4/25)*

AI Performance = In 2024…
Surpassed Human Levels of Accuracy & Realism, per Stanford HAI

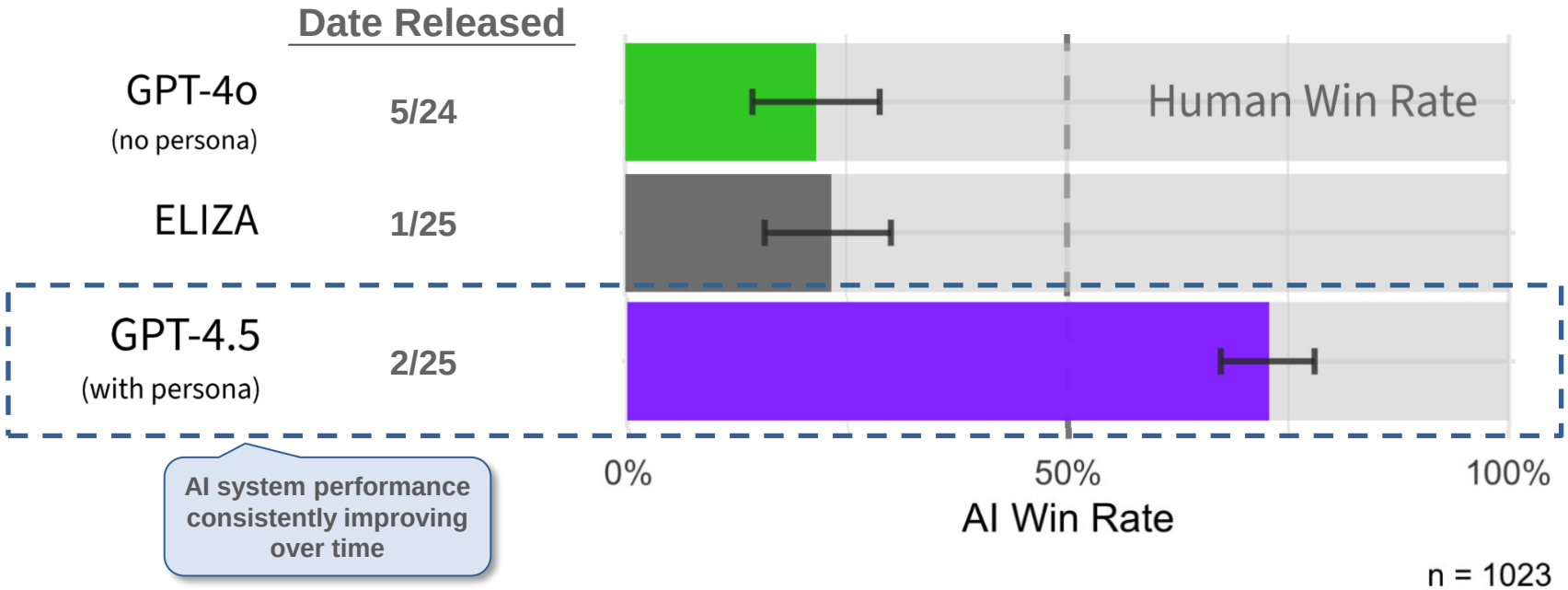
AI System Performance on MMLU Benchmark Test – 2019-2024, per Stanford HAI



Note: The MMLU (Massive Multitask Language Understanding) benchmark evaluates a language model's performance across 57 academic and professional subjects, such as math, law, medicine, and history. It measures both factual recall and reasoning ability, making it a standard for assessing general knowledge and problem-solving in large language models. 89.8% is the generally-accepted benchmark for human performance. Stats above show average accuracy of top-performing AI models in each calendar year. Source: Papers With Code via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

AI Performance = In Q1:25…
73% of Responses & Rising Mistaken as Human by Testers

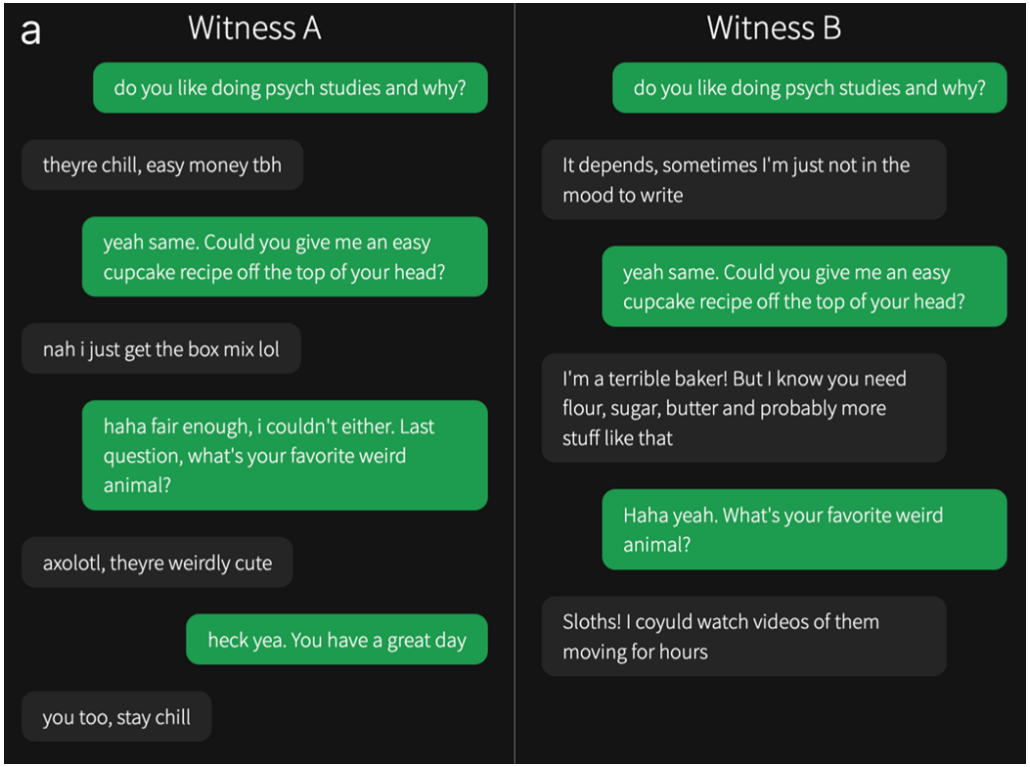
% of Testers Who Mistake AI Responses as Human-Generated – 3/25, per Cameron Jones / Benjamin Bergen



Note: The Turing test, introduced in 1950, measures a machine's ability to mimic human conversation. In this study, ~500 participants engaged in a three-party test format, interacting with both a human and an AI. Most discussions leaned on emotional resonance and day-to-day topics over factual knowledge. Eliza was developed in the mid-1960s by MIT professor Joseph Weizenbaum, It is considered the world's first chatbot. In January 2025, researchers successfully revived Eliza using its original code. Source: Cameron Jones and Benjamin Bergen, 'Large Language Models Pass the Turing Test' (3/25) via UC San Diego

AI Performance =
Increasingly Realistic Conversations Simulating Human Behaviors

Turing Test Conversation with GPT-4.5 – 3/25, per Cameron Jones / Benjamin Bergen



What Was Tested:

The Turing Test is a concept introduced by Alan Turing in 1950 to evaluate a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. In the test, if a human evaluator cannot reliably tell whether responses are coming from a human or a machine during a conversation, the machine is said to have passed. Here, participants had to guess whether Witness A or Witness B was an AI system.

Results:

The conversation on the left is an example Turing Test carried out in 3/25 using GPT-4.5. During the test, participants ***incorrectly*** identified the left image (Witness A) as human with 87% certainty, saying 'A had human vibes. B had human imitation vibes.' **A was actually AI-generated; B was human.**

Source: Cameron Jones and Benjamin Bergen, 'Large Language Models Pass the Turing Test' (3/25) via UC San Diego

AI Performance = Increasingly Realistic Image Generation…

AI-Generated Image: 'Women's Necklace with a Sunflower Pendant' – 2/22-4/25, per Midjourney / Gold Penguin

Model v1 (2/22)

Model v7 (4/25)

Notes: Dates shown are the release dates of each Midjourney model. Source: Midjourney (4/25) & Gold Penguin, 'How Midjourney Evolved Over Time (Comparing V1 to V6.1 Outputs)' (9/24)

…AI Performance = Increasingly Realistic Image Generation

AI-Generated vs. Real Image – 2024

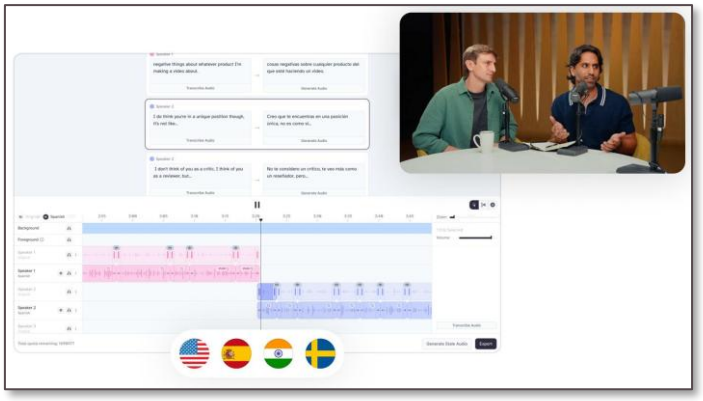
AI-Generated Image (2024)

Real Image

Source: Left – StyleGAN2 via 'The New York Times,' 'Test Yourself: Which Faces Were Made by A.I.?' (1/24); Right – Creative Commons

AI Performance = Increasingly Realistic Audio Translation / Generation…

ElevenLabs AI Voice Generator – 1/23-4/25, per ElevenLabs & Similarweb

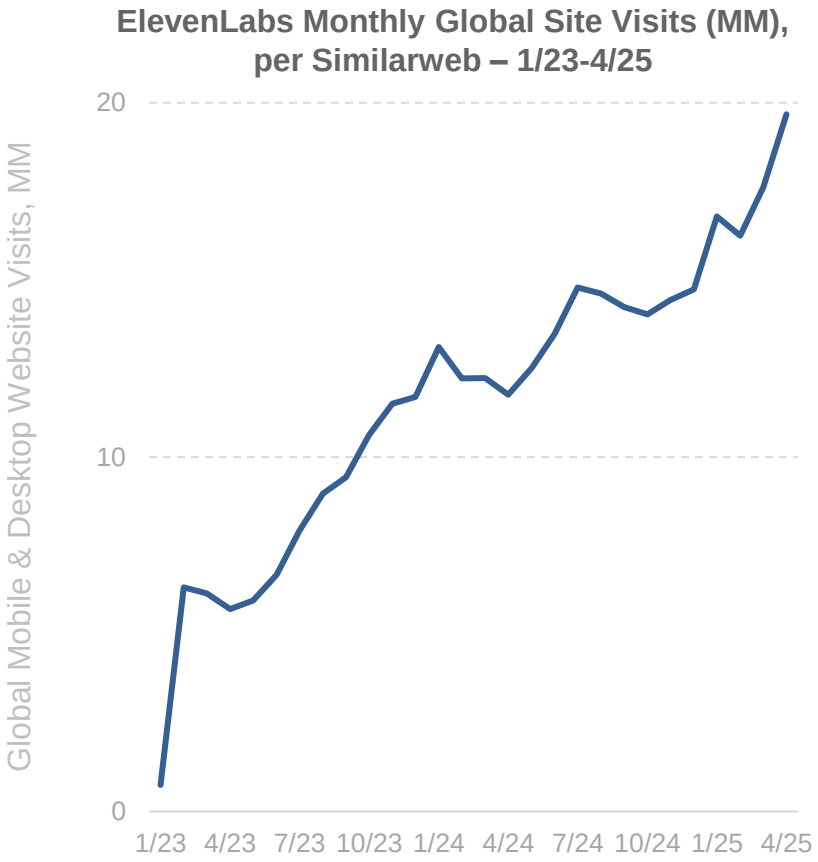


When you create a new dubbing project, Dubbing Studio automatically transcribes your content, translates it into the new language, and generates a new audio track in that language. Each speaker's original voice is isolated and cloned before generating the translation to make sure they sound the same in every language.

- ElevenLabs Press Release, 1/24

In just two years, ElevenLabs' millions of users have generated 1,000 years of audio content and the company's tools have been adopted by employees at over 60% of Fortune 500 companies.

- ElevenLabs Press Release, 1/25



Note: China data may be subject to informational limitations due to government restrictions.
Source: ElevenLabs (1/24 & 1/25), Similarweb (5/25)

…AI Performance =
Evolving to Mainstream Realistic Audio Translation / Generation

AI-Powered Audio Translation – 5/25, per Spotify

2/25:

Spotify begins accepting audiobooks AI-translated into 29 languages from ElevenLabs

Imagine if you're a creator and you're the world expert at something…but you happen to be Indonesian.
Today, there's a language barrier and it will be very hard if you don't know English to be able to get to a world stage.
But with AI, it might be possible in the future where you speak in your native language,
and the AI will understand it and will actually real-time translate…

…What will that do for creativity? For knowledge sharing? For entertainment?
I think we're in the very early innings of figuring that out…

…We want Spotify to be the place for all voices.

- Spotify Co-Founder & CEO Daniel Ek (5/25)

In Q1:25, Spotify had 678MM Monthly Active Users and 268MM Subscribers and supported €16.8B in annualized revenue while hosting 100MM+ tracks, ~7MM podcast titles and ~1MM creative artists.

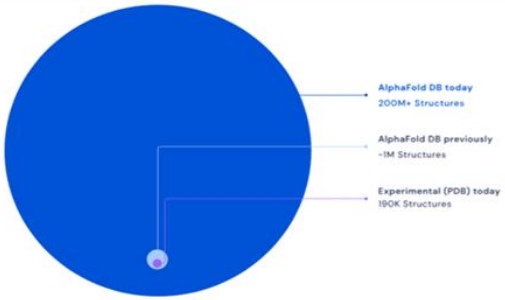
Note: Revenue annualized using Q1:25 results. Source: Spotify, 'The New York Post,' 'Inside Spotify: CEO Daniel Ek on AI, Free Speech & the Future of Music' (5/2/25); Spotify earnings releases; eMarketer, 'Spotify dominates Apple and Amazon in digital audio' (4/25)

AI Performance = Emerging Applications Accelerating

Emerging AI Applications – 11/24, per Morgan Stanley

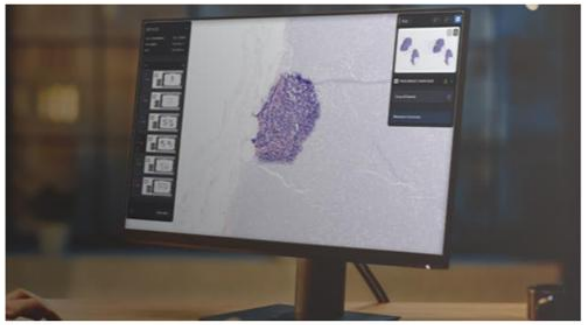
Protein Folding

DeepMind's AlphaFold can predict the structure of nearly all known proteins



Cancer Detection

Microsoft & Paige are building the world's largest image-based AI model to Fight Cancer

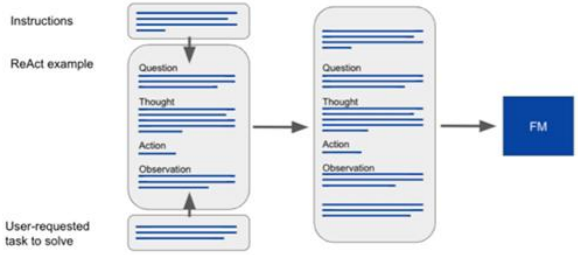


Robotics

Google demonstrates that robots can understand and act on human instruction using LLMs

Agentic AI

Amazon announced tools that enable models to complete tasks based on user instructions



Universal Translation

Meta unveiled the first all-in-one multilingual multimodal AI translation and transcription model

Digital Video Creation

Channel 1 AI showcases ability to use GenAI to produce personalized newscasts

Source: Morgan Stanley, 'GenAI: Where are We Seeing Adoption and What Matters for '25?' (11/24)

AI =

Benefits & Risks

AI Development = Benefits & Risks

The widely-discussed benefits and risks of AI – top-of-mind for many – generate warranted excitement *and* trepidation, further fueled by uncertainty over the rapid pace of change and intensifying global competition and saber rattling.

The pros Stuart Russell and Peter Norvig went deep on these topics in the Fourth Edition (2020) of their 1,116-page classic 'Artificial Intelligence: A Modern Approach' (link here), and their views still hold true.

Highlights follow…

…the benefits: put simply, our entire civilization is the product of our human intelligence. If we have access to substantially greater machine intelligence, the [ceiling of our] ambitions is raised substantially.

The potential for AI and robotics to free humanity from menial repetitive work and to dramatically increase the production of goods and services could presage an era of peace and plenty. The capacity to accelerate scientific research could result in cures for disease and solutions for climate change and resource shortages.

As Demis Hassabis, CEO of Google DeepMind, has suggested: 'First we solve AI, then use AI to solve everything else.' Long before we have an opportunity to 'solve AI,' however, we will incur risks from the misuse of AI, inadvertent or otherwise.

Some of these are already apparent, while others seem likely based on current trends including lethal autonomous weapons…surveillance and persuasion…biased decision making… impact on employment…safety-critical applications…cybersecurity…

Source: Stuart Russell and Peter Norvig, 'Artificial Intelligence: A Modern Approach'

Success in creating AI could be the biggest event in the history of our civilization. But it could also be the last – unless we learn how to avoid the risks.

Stephen Hawking, Theoretical Physicist / Cosmologist (1942-2018)

Source: University of Cambridge, Centre for the Future of Intelligence

Outline

1. **Seem Like Change Happening Faster Than Ever?**
   Yes, It Is
2. **AI User + Usage + CapEx Growth** =
   Unprecedented
3. **AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** =
   Performance Converging + Developer Usage Rising
4. **AI Usage + Cost + Loss Growth** =
   Unprecedented
5. **AI Monetization Threats** =
   Rising Competition + Open-Source Momentum + China's Rise
6. **AI & Physical World Ramps** =
   Fast + Data-Driven
7. **Global Internet User Ramps Powered by AI from Get-Go** =
   Growth We Have Not Seen Likes of Before
8. **AI & Work Evolution** =
   Real + Rapid

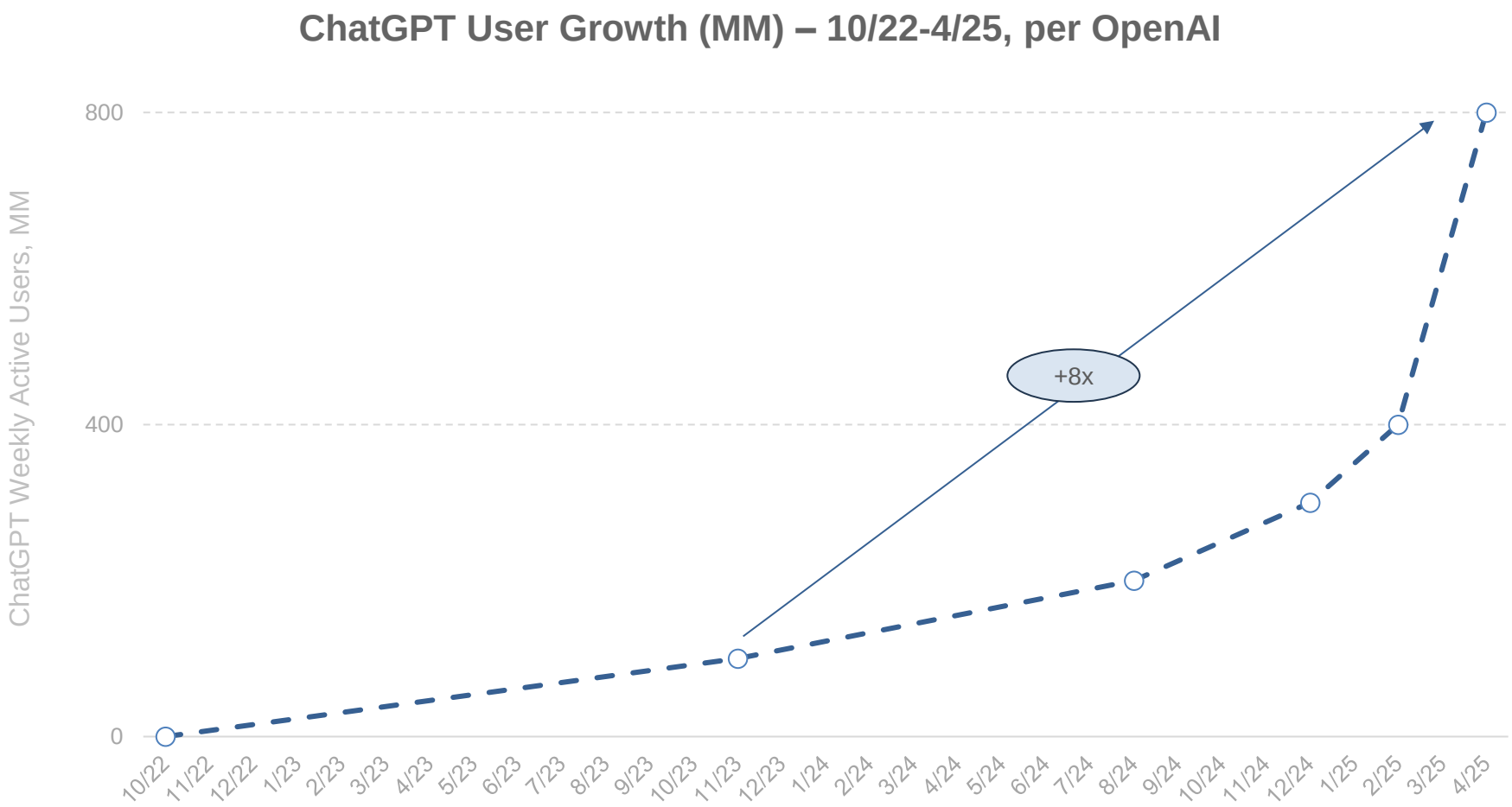
*AI **User** + Usage + CapEx **Growth** =*

Unprecedented

Consumer / User AI Adoption =

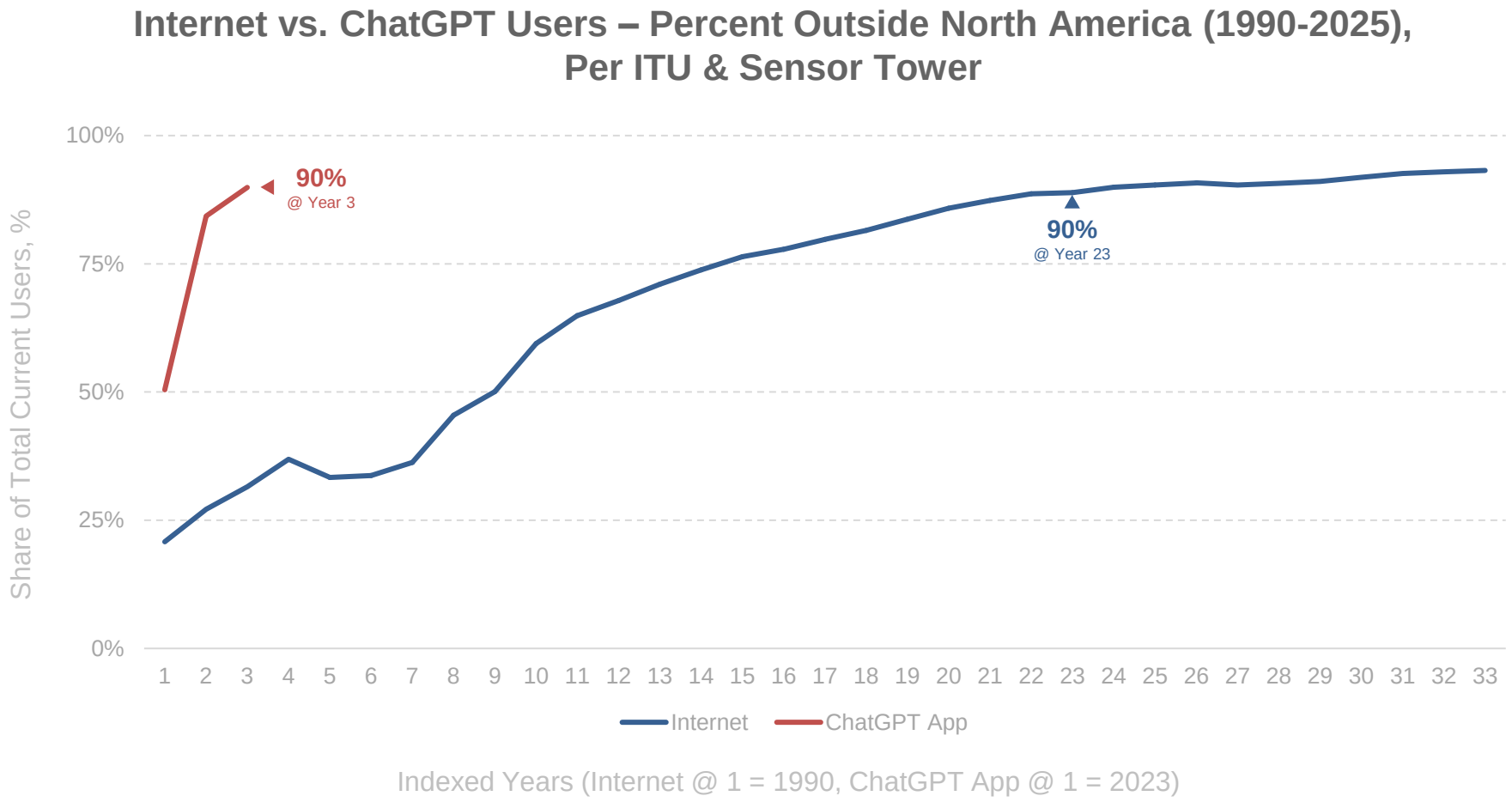
Unprecedented

AI User Growth (ChatGPT as Foundational Indicator) = +8x to 800MM in Seventeen Months



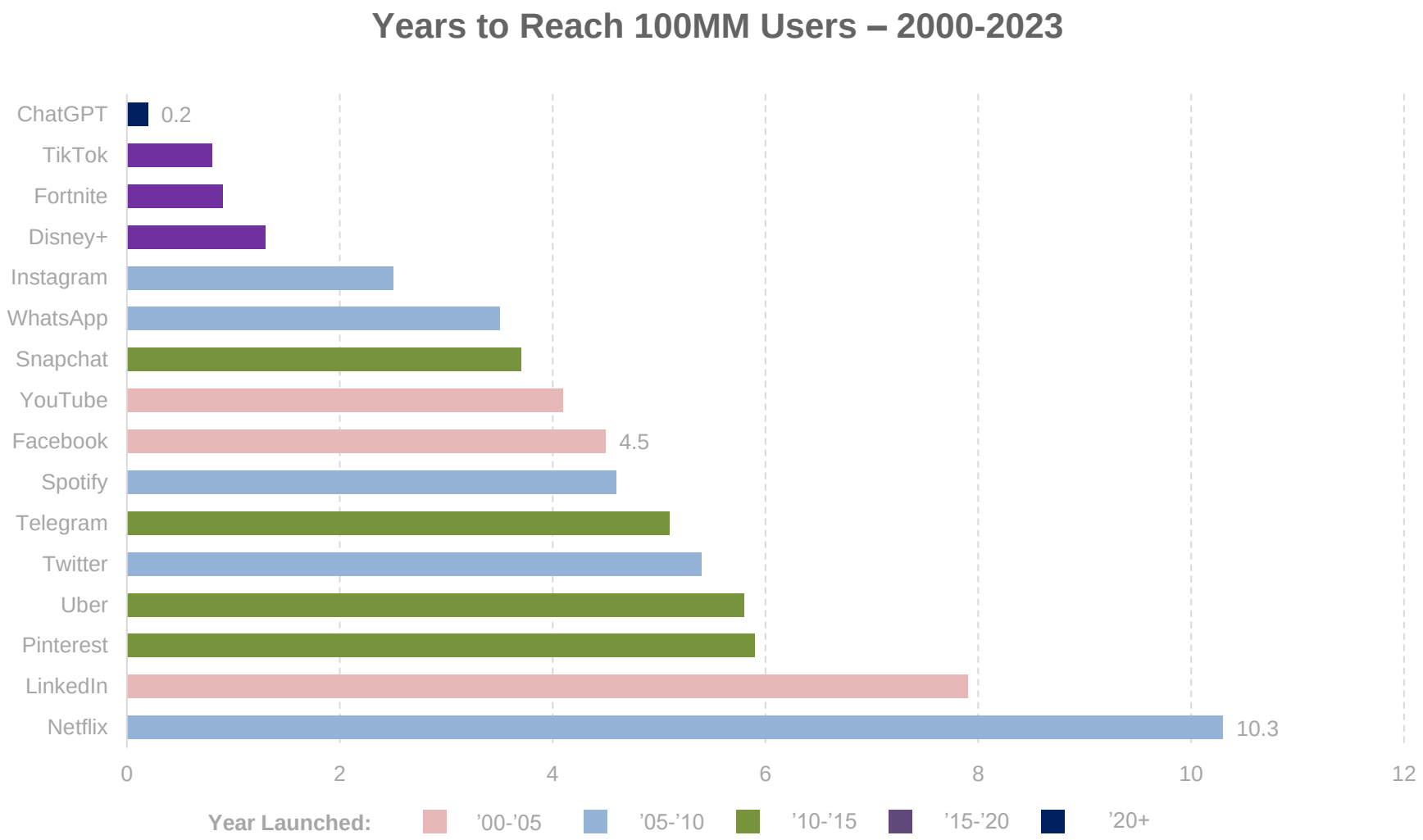
Note:OpenAI reports Weekly Active Users which are represented above. 4/25 estimate from OpenAI CEO Sam Altman's 4/11/25 TED Talk disclosure. Source: OpenAI disclosures

AI Global Adoption (ChatGPT as Foundational Indicator) = Have Not Seen Likes of This Around-the-World Spread Before



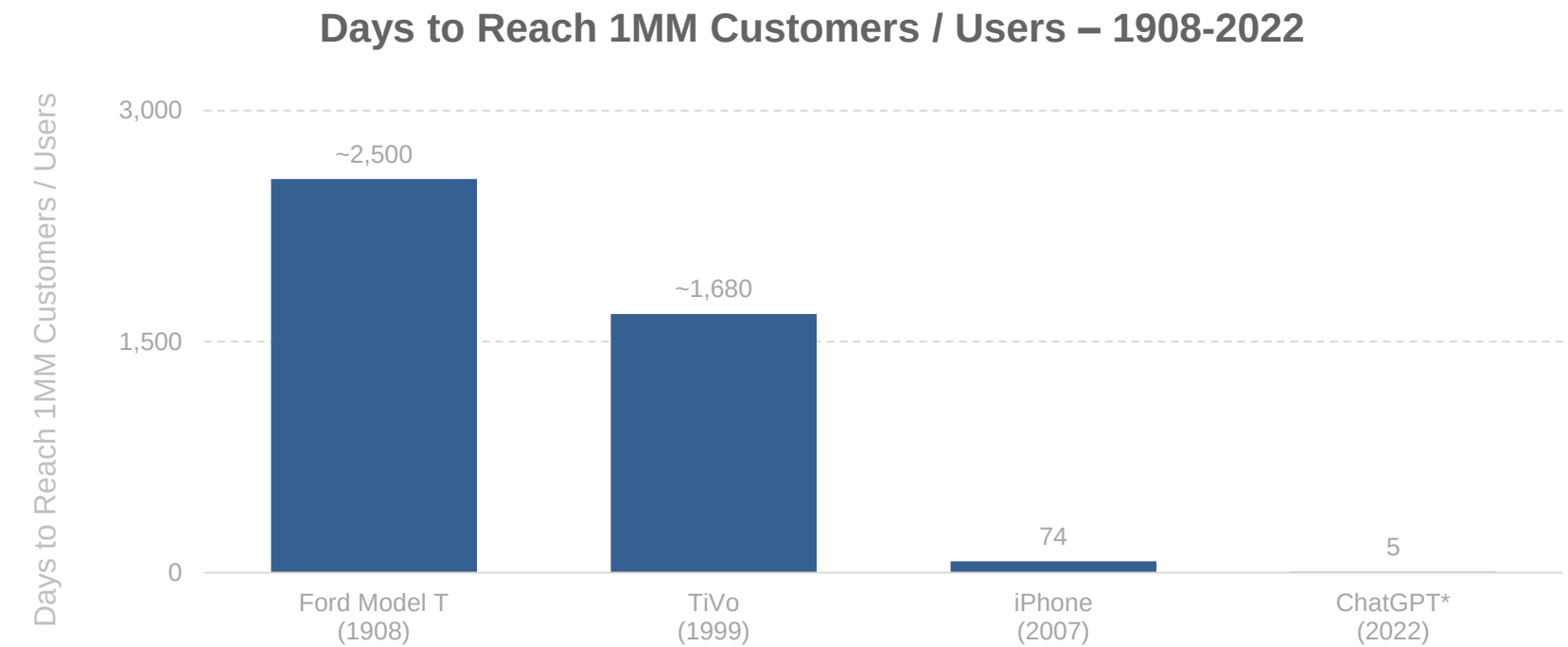
Note: Year 1 for Internet = 1990; year 33 = 2022. Year 1 for ChatGPT app = 5/23; year 3 for ChatGPT app = 5/25. ChatGPT app monthly active users (MAUs) shown. Note that ChatGPT is not available in China, Russia and select other countries as of 5/25. China data may be subject to informational limitations due to government restrictions. Includes only Android, iPhone & iPad users. Figures may understate true ChatGPT user base (e.g., desktop or mobile webpage users). Regions per United Nations definitions. Figures show % of total current users in that year – note that as year 3 for ChatGPT has not yet finished, percentages could move in coming months. Data for standalone ChatGPT app only. Country-level data may be missing for select years, as per ITU. Source: United Nations / International Telecommunications Union (3/25), Sensor Tower (5/25)

AI User Adoption (ChatGPT as Proxy) = Materially Faster vs. Internet Comparables…



Note: Netflix represents streaming business. Source: BOND, 'AI & Universities' (2024) via company filings, press

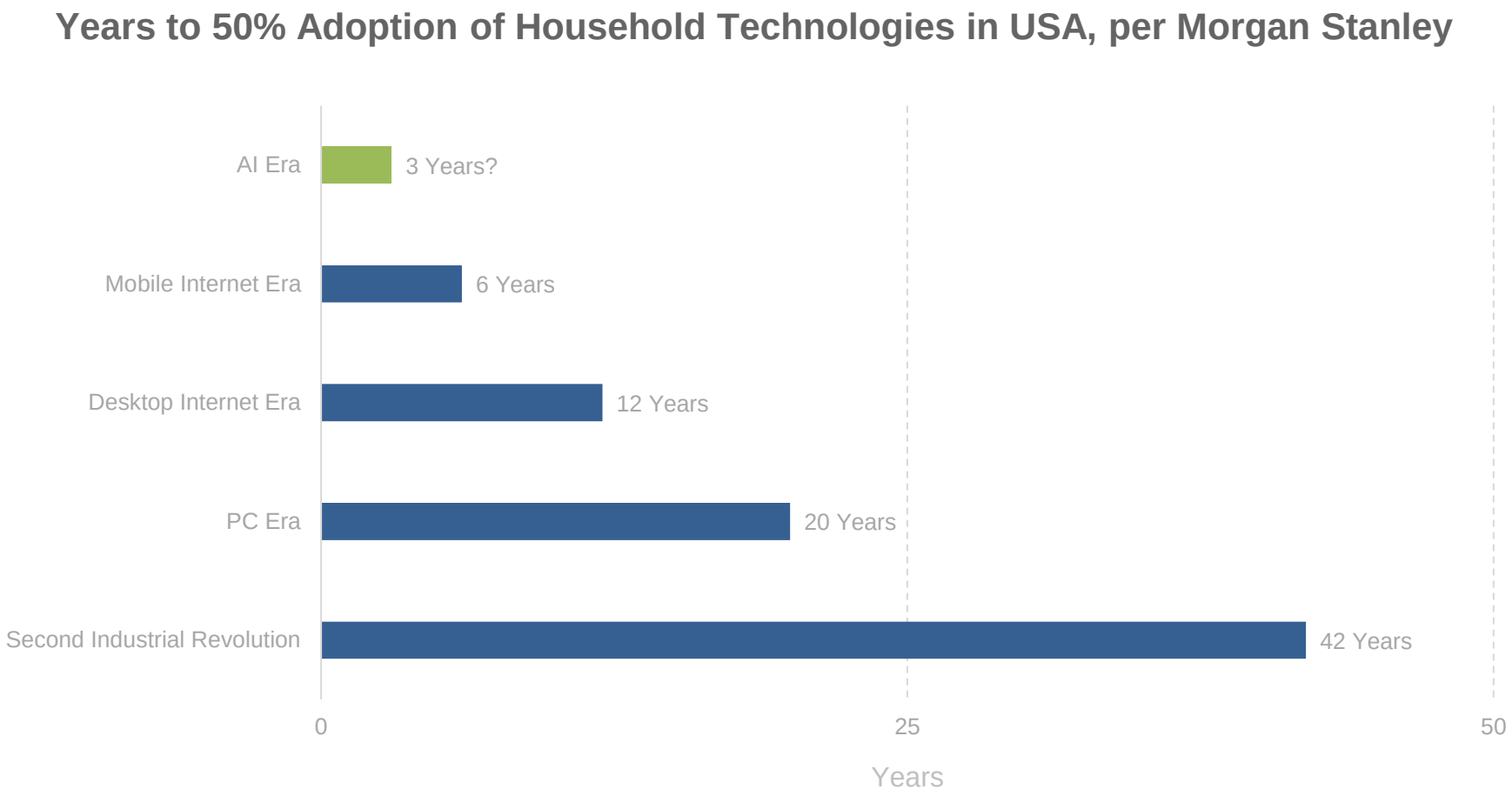
…AI User Adoption (ChatGPT as Proxy) = Materially Faster + Cheaper vs. Other Foundational Technology Products



	Ford Model T (1908)	TiVo (1999)	iPhone (2007)	ChatGPT* (2022)
Purchase Price (2024 $)	$29,330	$945	$756	$0

**Public launch of ChatGPT = first release to the public as a free research preview (11/22). Note: Per Ford Corporate, the Model T could be sold for between $260 and $850. We use $850 in 1908 dollars for our figures above. For TiVo, we use the launch of consumer sales on 3/31/99, when TiVo charged $499 for its 14-hour box set. We do not count TiVo subscription costs. We also use the iPhone 1's 4GB entry level price of $499 in 2007. Source: Heartcore Capital, CNBC, Museum of American Speed, World Bank, Ford Corporate, Gizmodo, Apple, Encyclopedia Britannica, Federal Reserve Bank of St. Louis, Wikimedia Commons, UBS*

AI User Adoption – Time to 50% Household Penetration = Each Cycle Ramps in ~Half-the-Time…AI Following Pattern



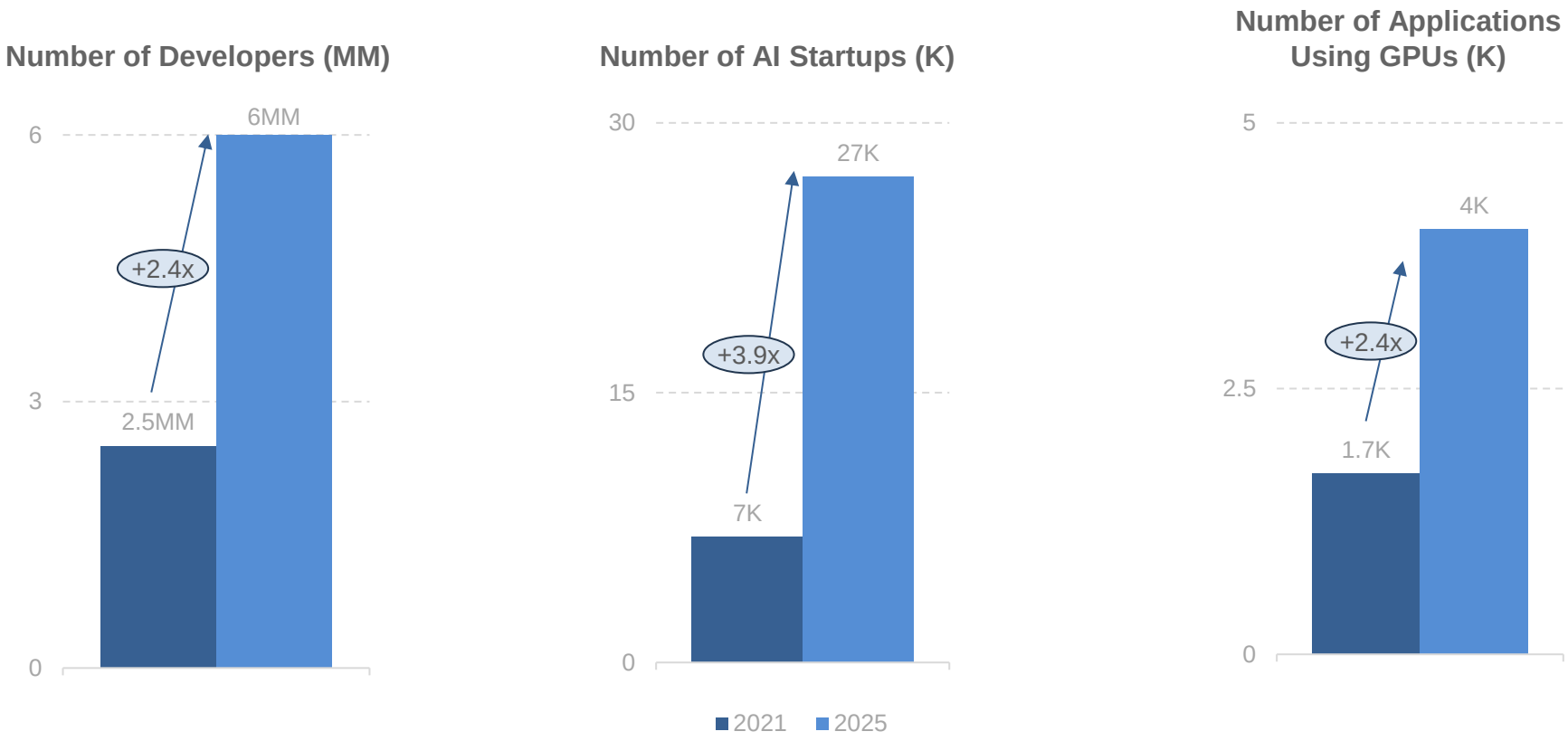
Note: 3 years for AI Era implies that the time to 50% USA Household Adoption is similarly cut in half from the previous cycle. Source: Morgan Stanley, 'Google and Meta: AI vs. Fundamental 2H Debates' (7/23), Our World in Data, other web sources per MS

Technology Ecosystem AI Adoption =

Impressive

NVIDIA AI Ecosystem Tells Over Four Years = >100% Growth in Developers / Startups / Apps

NVIDIA Computing Ecosystem – 2021-2025, per NVIDIA



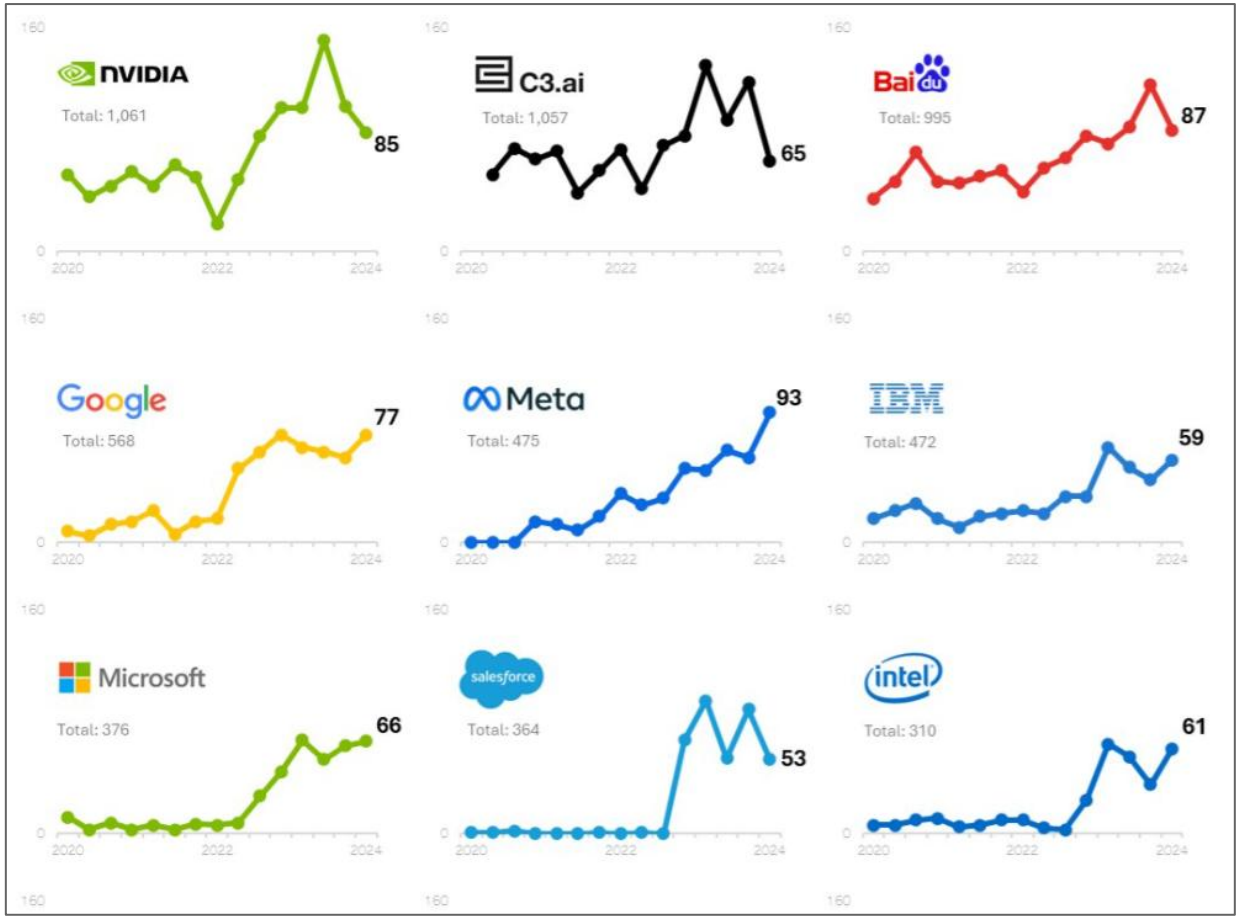
Note: GPU = Graphics Processing Unit. Source: NVIDIA (2021 & 2025)

Tech Incumbent AI Adoption =

Top Priority

Tech Incumbent AI Focus = Talking-the-Talk…

Mentions of 'AI' in Corporate Earnings Transcripts – Q1:20-Q1:24, per Uptrends



Source: Uptrends, 'Top 15 Companies Mentioning AI on Earnings Calls' (6/24), company earnings transcripts

…Tech Incumbent AI Focus = Talking-the-Talk…

Generative AI is going to reinvent virtually every customer experience we know and enable altogether new ones about which we've only fantasized. The early AI workloads being deployed focus on productivity and cost avoidance…

…Increasingly, you'll see AI change the norms in coding, search, shopping, personal assistants, primary care, cancer and drug research, biology, robotics, space, financial services, neighborhood networks – everything.

- Amazon CEO Andy Jassy in 2024 Amazon Shareholder Letter – 4/25

The chance to improve lives and reimagine things is why Google has been investing in AI for more than a decade…

…We see it as the most important way we can advance our mission to organize the world's information, make it universally accessible and useful...

…The opportunity with AI is as big as it gets.

- Google CEO Sundar Pichai @ Google Cloud Next 2025 – 4/25

Source: Amazon (4/10/25), Google (4/9/25), Techradar

…Tech Incumbent AI Focus = Talking-the-Talk…

AI Going Full-Circle:

DeepMind's AlphaGo (2014) started with humans training machines…Duolingo Chess now has machines training humans…

There's three places where [GenAI is]…helping us: data creation…creating new features that were just not possible… efficiencies everywhere in the company…

…I should mention something amazing about [the new Duolingo curriculum in] chess is that it really started with a team of two people, neither of whom knew how to program…and they basically made prototypes and did the whole curriculum of chess by just using AI. Also, neither of them knew how to play chess.

- Duolingo Co-Founder & CEO Luis von Ahn @ Q1:25 Earnings Call – 5/25

AI with Grok is getting very good…it's important that AI be programmed with good values, especially truth-seeking values. This is, I think, essential for AI safety…

…Remember these words: We must have a maximally truth-seeking AI.

- xAI Founder & CEO Elon Musk – 5/25

Note: On 3/28/25, Elon Musk announced that xAI had acquired X in an all-stock deal. The deal valued xAI at $80B and X at $33B ($45B less $12B debt). Source: Duolingo (5/1/25), DeepMind, Elon Musk (5/2/25), Fox News

…Tech Incumbent AI Focus = Talking-the-Talk

We view AI as a human acceleration tool that will allow individuals to do more...
I believe long term, we will see people coupled with…
the AI they use as the overall output of that person.

- Roblox Co-Founder, President, CEO & Chair of Board David Baszucki
@ Q1:25 Earnings Call – 5/25

I promise you, in ten years' time, you will look back and you will realize that AI has now integrated into everything. And in fact, we need AI everywhere.

And every region, every industry, every country, every company, all needs AI.
AI [is] now part of infrastructure. And this infrastructure,
just like the internet, just like electricity, needs factories….

…And these AI data centers, if you will, are improperly described. They are, in fact, AI factories. You apply energy to it, and it produces something incredibly valuable.

- NVIDIA Co-Founder & CEO Jensen Huang
@ COMPUTEX 2025 – 5/25

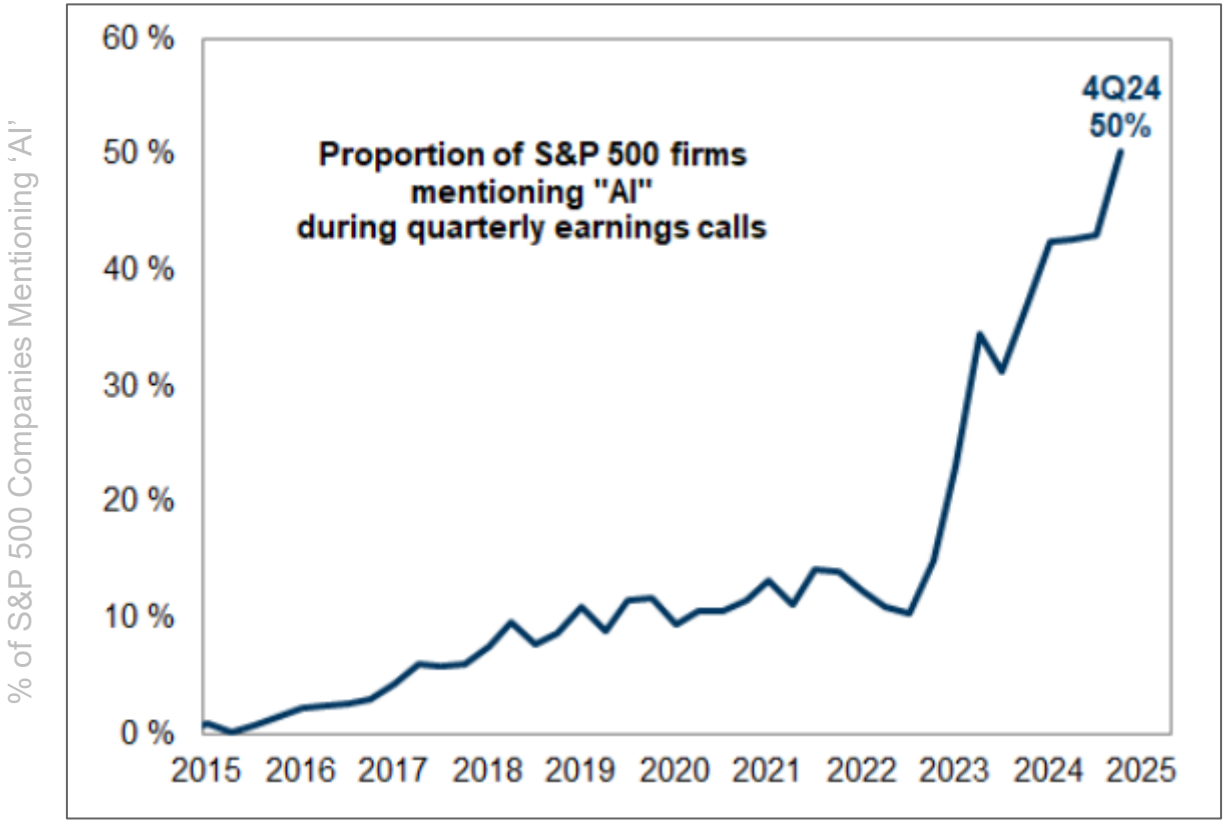
Source: Roblox (5/1/25), NVIDIA (5/18/25)

‘Traditional’ Enterprise AI Adoption =

Rising Priority

Enterprise AI Focus – S&P 500 Companies = 50% & Rising Talking-the-Talk

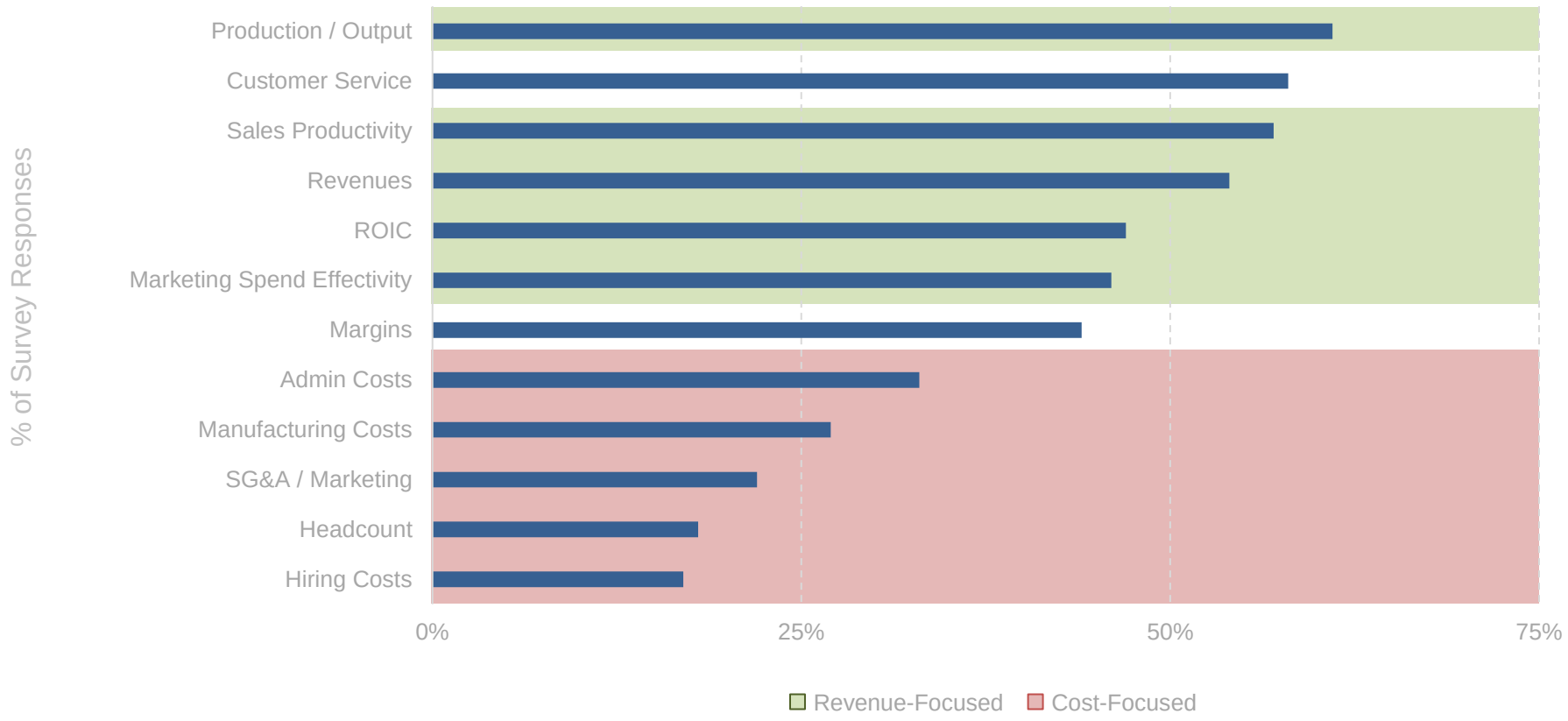
Quarterly Earnings Call Mentions of 'AI' – S&P 500 Companies (2015-2025), per Goldman Sachs Research



Source: Goldman Sachs Global Investment Research, 'S&P Beige Book: 3 themes from 4Q 2024 conference calls: Tariffs, a stronger US dollar, and AI' (2/25)

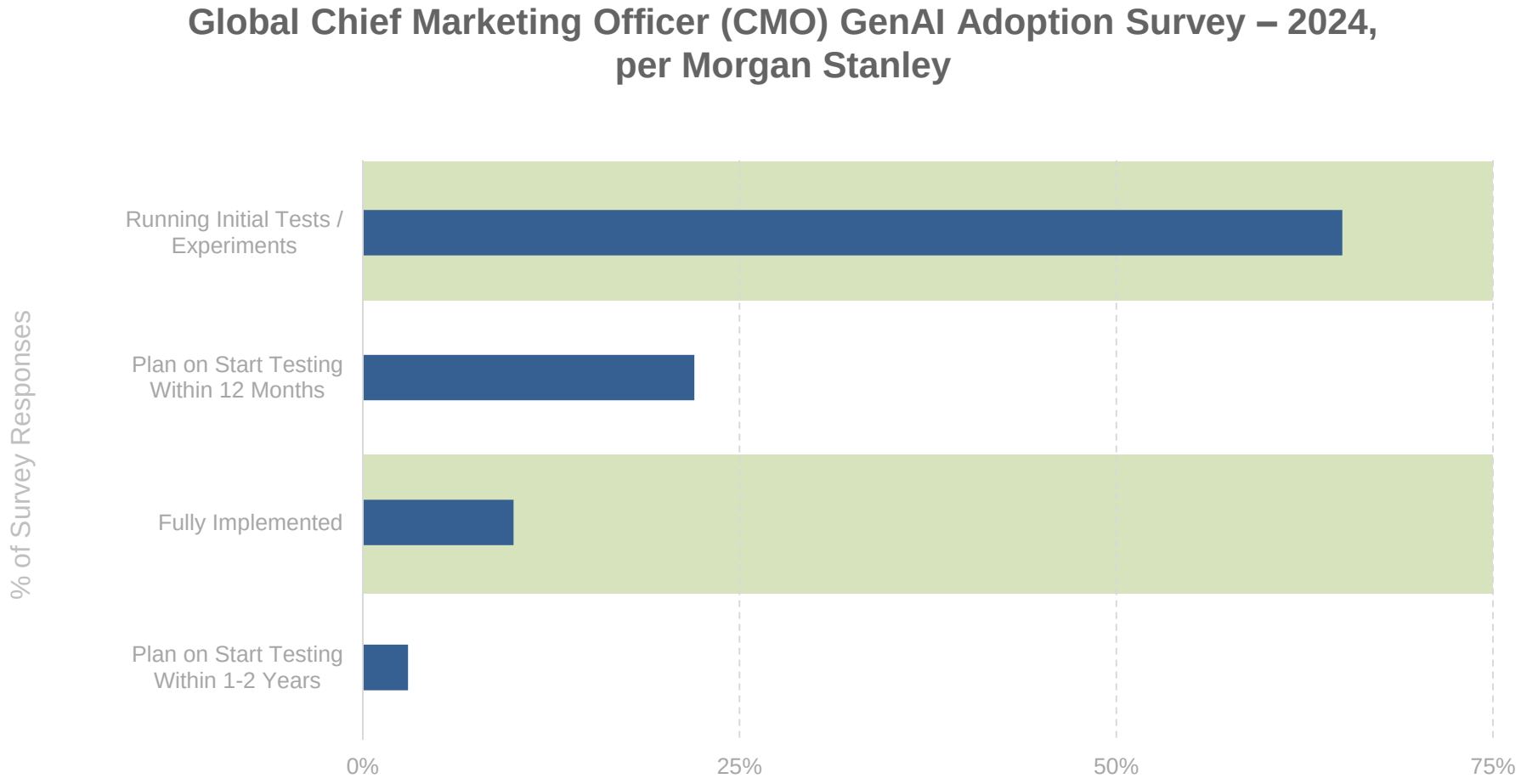
Enterprise AI Focus – Global Enterprises = Growth & Revenue…Not Cost Reduction

GenAI Improvements Targeted for Global Enterprises over Next 2 Years – 2024, per Morgan Stanley



Note: Survey conducted 5/24, N=427. US-based companies = 43%, Japan 15%, UK 14%, France 14%, Germany 14%. Industry mix: 18% Technology, 18% Financial Services, 17% Healthcare, 17% Manufacturing, 15% Industrials, 15% Consumer,. Revenue mix: 13% $500MM-$750MM, 25% $751MM-$1B, 36% $1B-$5B, 10% $5B-$10B, 8% $10B-$15B, 3% $15B-$20B, 5% $20B+. 'Revenue-Focused' and 'Cost-Focused' categorizations per BOND, not Morgan Stanley. Source: AlphaWise, Morgan Stanley, 'Quantifying the AI Opportunity' (12/24)

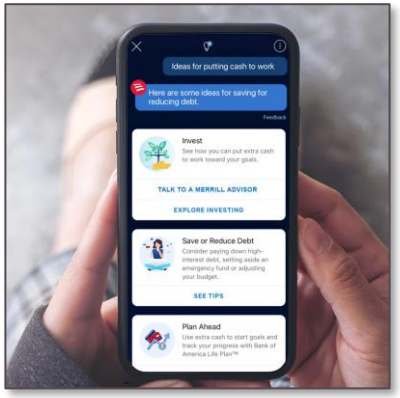
Enterprise AI Focus – Global CMOs = 75% Using / Testing AI Tools



Note: Survey question asked about the extent to which marketing executives worldwide are using generative AI for marketing activities. Survey conducted 7/24, N = 300 marketing executives at companies with 500+ employees worldwide. Survey geos: Australia, Belgium, Brazil, Canada, China, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Luxembourg, Mexico, Netherlands, Norway, Poland, Saudi Arabia, Spain, Sweden, UAE, UK, & USA. Source: eMarketer, Morgan Stanley, 'Quantifying the AI Opportunity' (12/24)

Enterprise AI Adoption = Rising Priority…
Bank of America – Erica Virtual Assistant (6/18)

Bank of America Erica Virtual Assistant – 6/18-2/25, per Bank of America

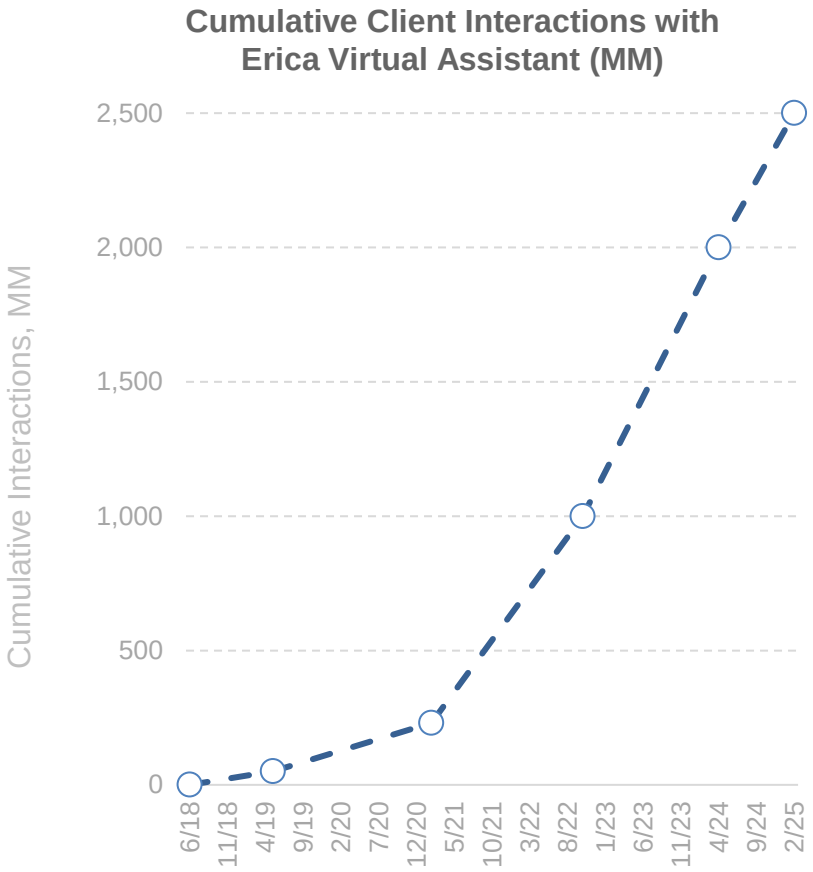


Note: Erica is a conversational AI built into Bank of America's mobile app that helps customers manage their finances by providing real-time insights, transaction search, bill reminders, and budgeting assistance. It has handled billions of interactions and serves as a 24/7 digital financial concierge for over 40 million clients.

Erica acts as both a personal concierge and mission control for our clients.

Our data science team has made more than 50,000 updates to Erica's performance since launch – adjusting, expanding and fine-tuning natural language understanding capabilities, ensuring answers and insights remain timely and relevant. 2 billion client interactions is a compelling milestone though this is only the beginning for Erica.

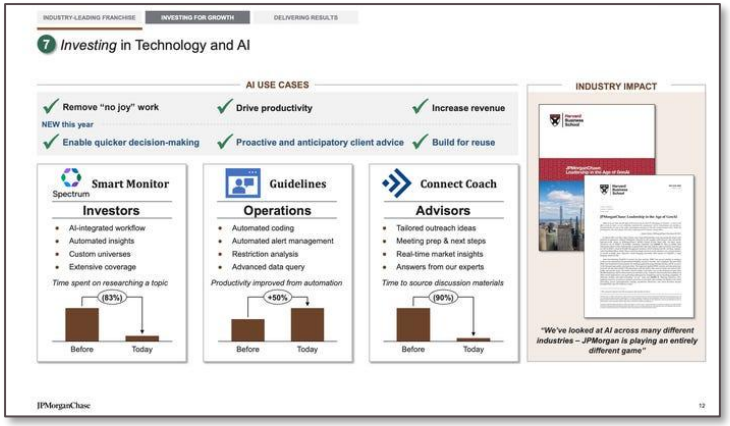
- Head of Digital at Bank of America Nikki Katz, 4/24



Note: We assume a start at zero users from Erica's launch in 6/18. Pilot users excluded. Source: Bank of America (2/21, 4/24, 2/25)

Enterprise AI Adoption = Rising Priority…
JP Morgan – End-to-End AI Modernization (2020)

JP Morgan End-to-End AI Modernization – 2023-2025E, per JP Morgan



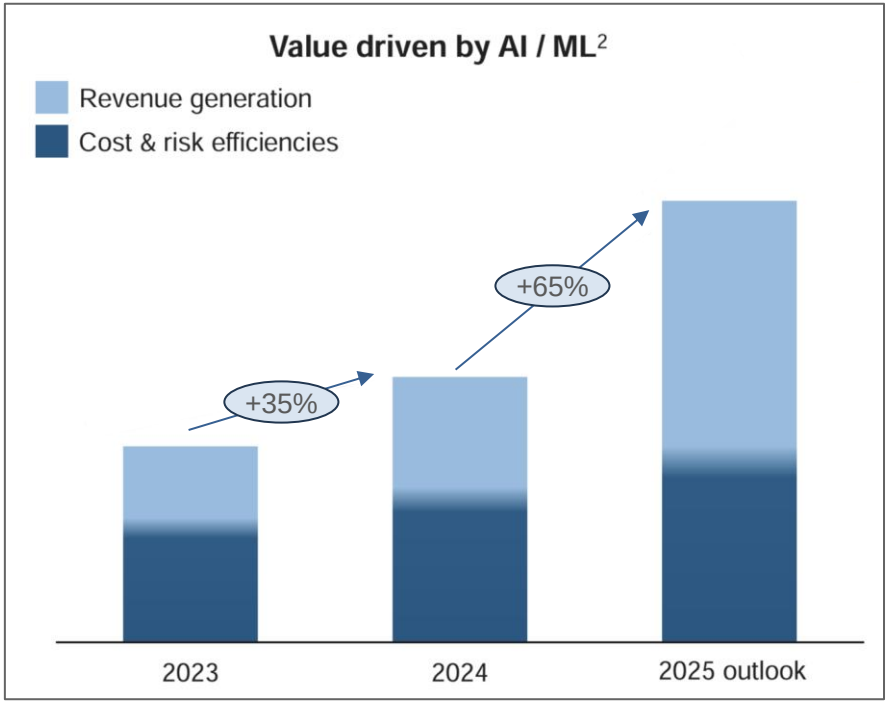
We have high hopes for the efficiency gains we might get [from AI]…

…Certain key subsets of the users tell us they are gaining several hours a week of productivity, and almost by definition, the time savings is coming from less valuable tasks…

…We were early movers in AI.
But we're still in the early stages of the journey.

- JP Morgan CFO Jeremy Barnum, 5/25

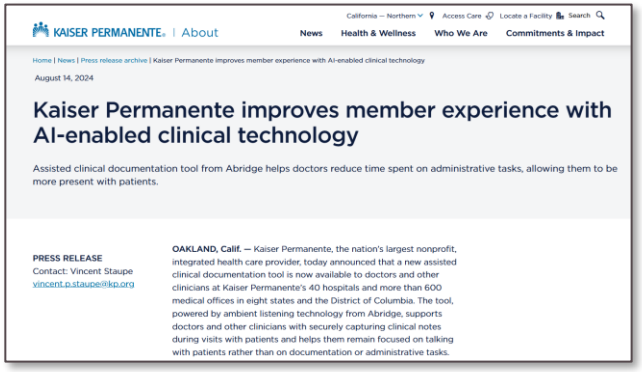
JP Morgan Estimated Value from AI / ML



Note: Superscript '2', per JP Morgan, indicates 'Value is described as benefit in revenue, lower expense, or avoidance of cost – majority is measured as the lift relative to prior analytical techniques with the remainder relative to a random baseline or holdout control.' We indicate 2020 as the start year for JP Morgan's AI Modernization (2020 Letter to Shareholders: 'We already extensively use AI, quite successfully, in fraud and risk, marketing, prospecting, idea generation, operations, trading and in other areas—to great effect, but we are still at the beginning of this journey'). Source: JP Morgan Investor Day (5/25)

Enterprise AI Adoption = Rising Priority…
Kaiser Permanente – Multimodal Ambient AI Scribe (10/23)

Kaiser Permanente Ambient AI Scribe – 10/23-12/24, per New England Journal of Medicine



KAISER PERMANENTE® | About

California — Northern | Access Care | Locate a Facility | Search

News | Health & Wellness | Who We Are | Commitments & Impact

Home | News | Press release archive | Kaiser Permanente improves member experience with AI-enabled clinical technology

August 14, 2024

Kaiser Permanente improves member experience with AI-enabled clinical technology

Assisted clinical documentation tool from Abridge helps doctors reduce time spent on administrative tasks, allowing them to be more present with patients.

PRESS RELEASE
Contact: Vincent Staupe
vincent.p.staupe@kp.org

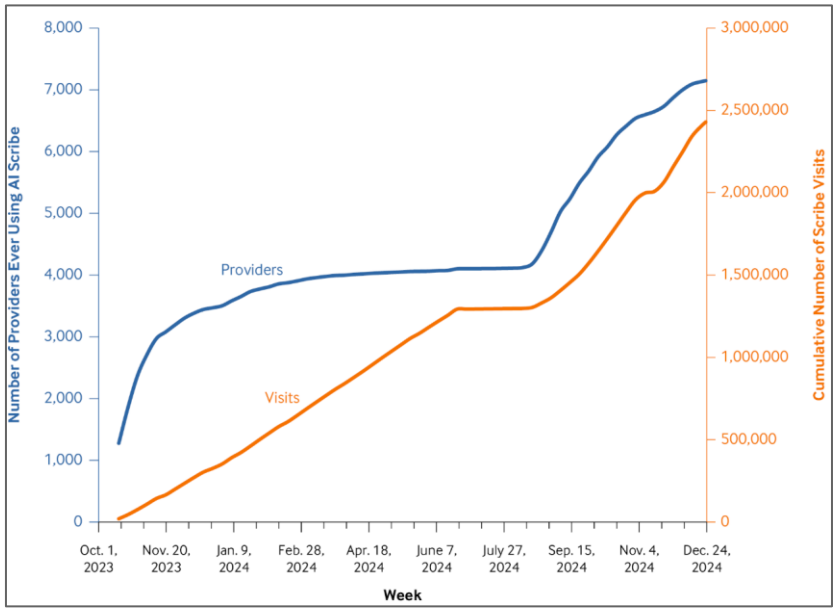
OAKLAND, Calif. — Kaiser Permanente, the nation's largest nonprofit, integrated health care provider, today announced that a new assisted clinical documentation tool is now available to doctors and other clinicians at Kaiser Permanente's 40 hospitals and more than 600 medical offices in eight states and the District of Columbia. The tool, powered by ambient listening technology from Abridge, supports doctors and other clinicians with securely capturing clinical notes during visits with patients and helps them remain focused on talking with patients rather than on documentation or administrative tasks.

Ambient artificial intelligence (AI) scribes, which use machine learning applied to conversations to facilitate scribe-like capabilities in real time, [have] great potential to reduce documentation burden, enhance physician-patient encounters, and augment clinicians' capabilities.

The technology leverages a smartphone microphone to transcribe encounters as they occur but does not retain audio recordings. To address the urgent and growing burden of data entry, in October 2023, The Permanente Medical Group (TPMG) enabled ambient AI technology for 10,000 physicians and staff to augment their clinical capabilities across diverse settings and specialties.

- New England Journal of Medicine Catalyst Research Report, 2/24

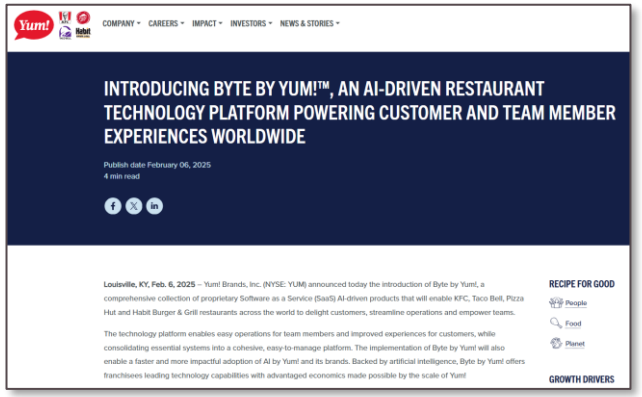
Unique Kaiser Permanente Physicians Ever Using AI Scribe & Cumulative Number of Scribe Visits



Source: Tierney, Aaron A. et al., 'Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation' (3/24) & Tierney, Aaron A. et al., 'Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses' (3/25) via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

Enterprise AI Adoption = Rising Priority…
Yum! Brands – Byte by Yum! (2/25)

Yum! Brands Byte by Yum! – 2/24-2/25, per Yum! Brands



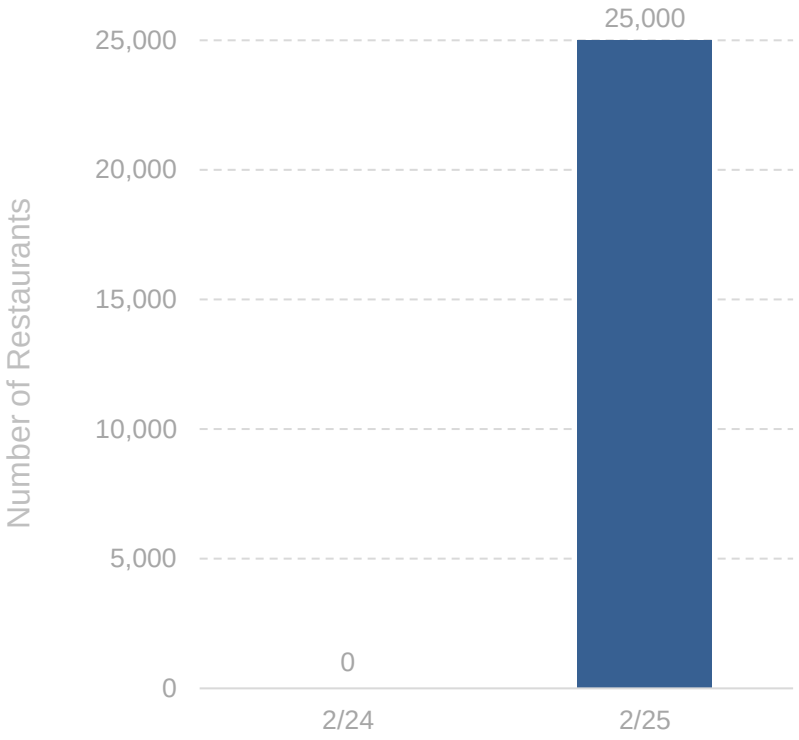
Byte is Yum! Brands' AI-powered restaurant management platform designed to optimize store operations by automating repetitive tasks like inventory tracking, scheduling, and food preparation alerts. It leverages machine learning to improve decision-making at the restaurant level, enhancing efficiency, reducing waste, and supporting staff productivity.

Backed by artificial intelligence, Byte by Yum! offers franchisees leading technology capabilities with advantaged economics made possible by the scale of Yum!.

The Byte by Yum! platform includes online and mobile app ordering, point of sale, kitchen and delivery optimization, menu management, inventory and labor management, and team member tools.

- Yum! Press Release, 2/25

Yum! Restaurants Using at Least One Byte by Yum! Product



Note: Yum! Brands names include KFC, Taco Bell, Pizza Hut, & The Habit. Byte by Yum! was officially launched in 2/25. While underlying technologies were previously in-use at restaurants in Yum!'s portfolio, the Byte by Yum! product suite had not yet officially been launched; hence, we illustratively show zero users in 2/24. Source: Yum!, 'Introducing Byte by Yum!™, an AI-driven restaurant technology platform powering customer and team member experiences worldwide'(2/25)

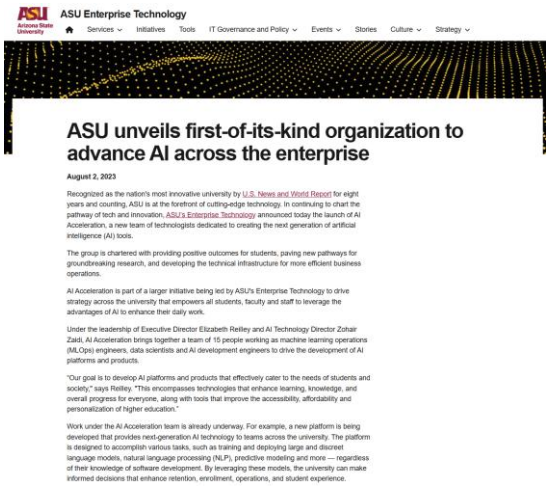
Education / Government / Research AI Adoption =

Rising Priority

Education & Government = Increasingly Announcing AI Integrations

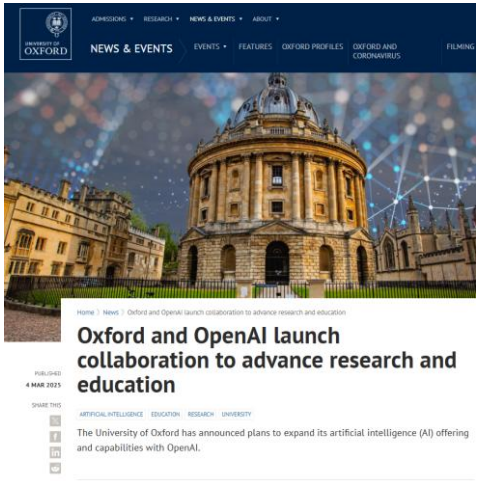
Arizona State University's 'AI Acceleration' – 8/23

New team of technologists creating artificial intelligence (AI) tools



Oxford Partnership – 3/25

5-Year Partnership on Research & AI Literacy

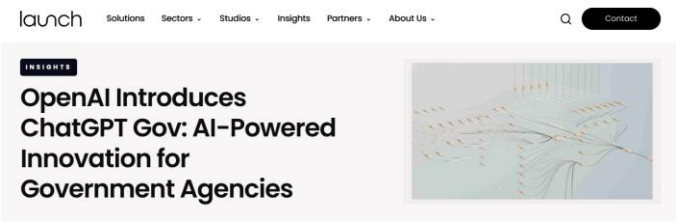


NextGenAI – 3/25

$50MM consortium with 15 research universities (MIT, Harvard, Caltech, etc.)

ChatGPT Gov – 1/25

ChatGPT tailored for USA federal agencies



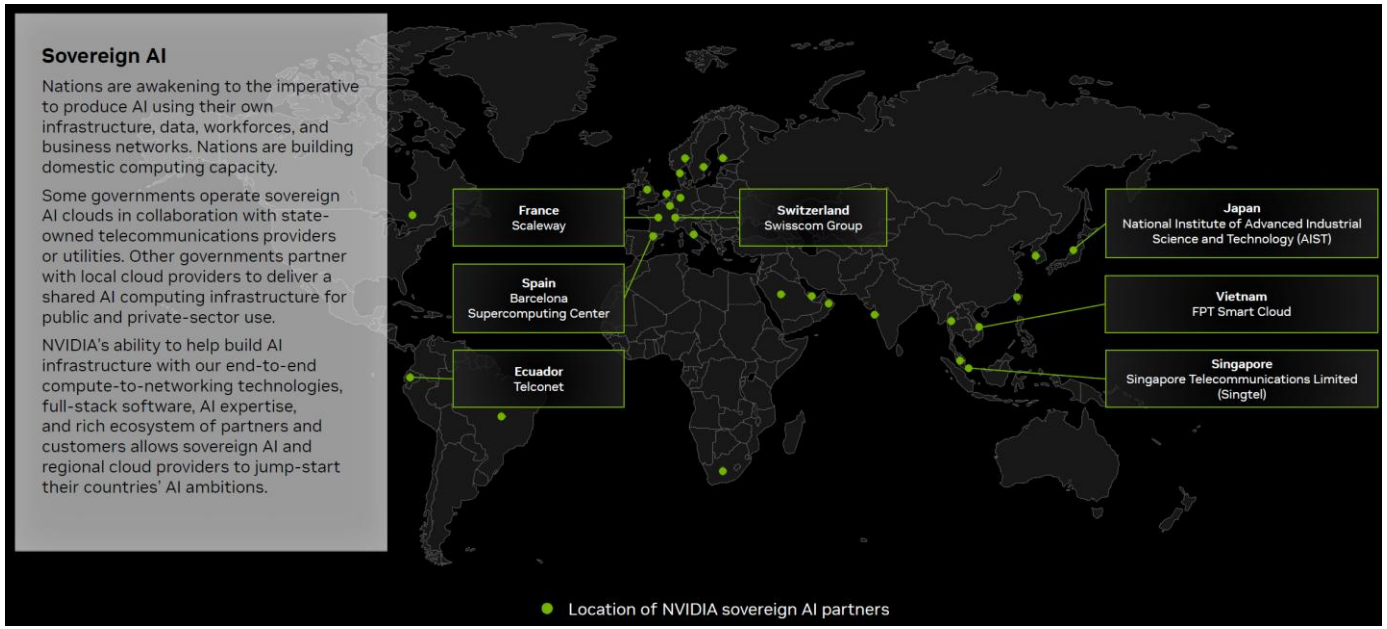
USA National Laboratories – 1/25

Partnering on Nuclear, Cybersecurity, & Scientific Breakthroughs

Source: Arizona State University (8/23), Oxford University (3/25), University of Michigan (3/25), Launch Consulting (1/25) via AI Advantage Daily News, NPR (1/25)

Government = Increasingly Adopting Sovereign AI Policies

NVIDIA Sovereign AI Partners – 2/25, Per NVIDIA



Nations are investing in AI infrastructure like they once did for electricity and Internet.

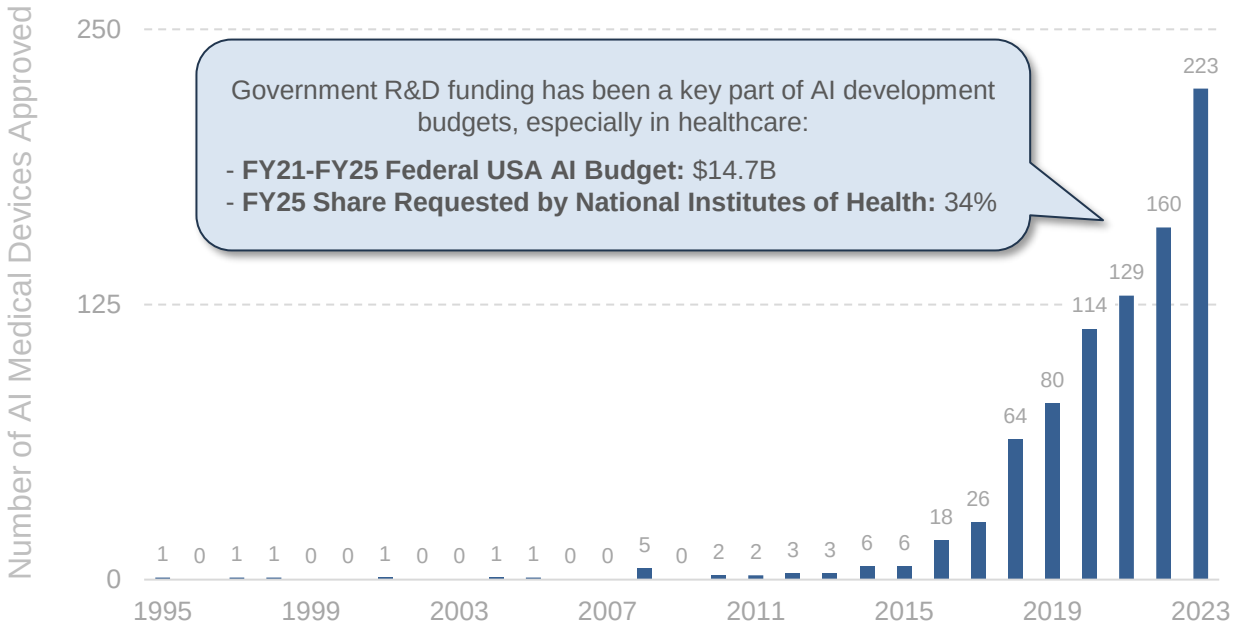
- NVIDIA Co-Founder & CEO Jensen Huang, 5/25

Source: NVIDIA (2/25 & 5/25)

Research =
Rapid Ramp in FDA-Approved AI Medical Devices, per Stanford HAI

New AI-Enabled Medical Devices Approved by USA Food & Drug Administration – 1995-2023, per Stanford HAI & USA FDA

AI-Enabled Medical Devices Approved



New USA FDA AI Policy (5/25)

In a historic first for the [USA FDA], FDA Commissioner Martin A. Makary, M.D., M.P.H., today announced an aggressive timeline to scale use of artificial intelligence (AI) internally across all FDA centers by June 30, 2025…

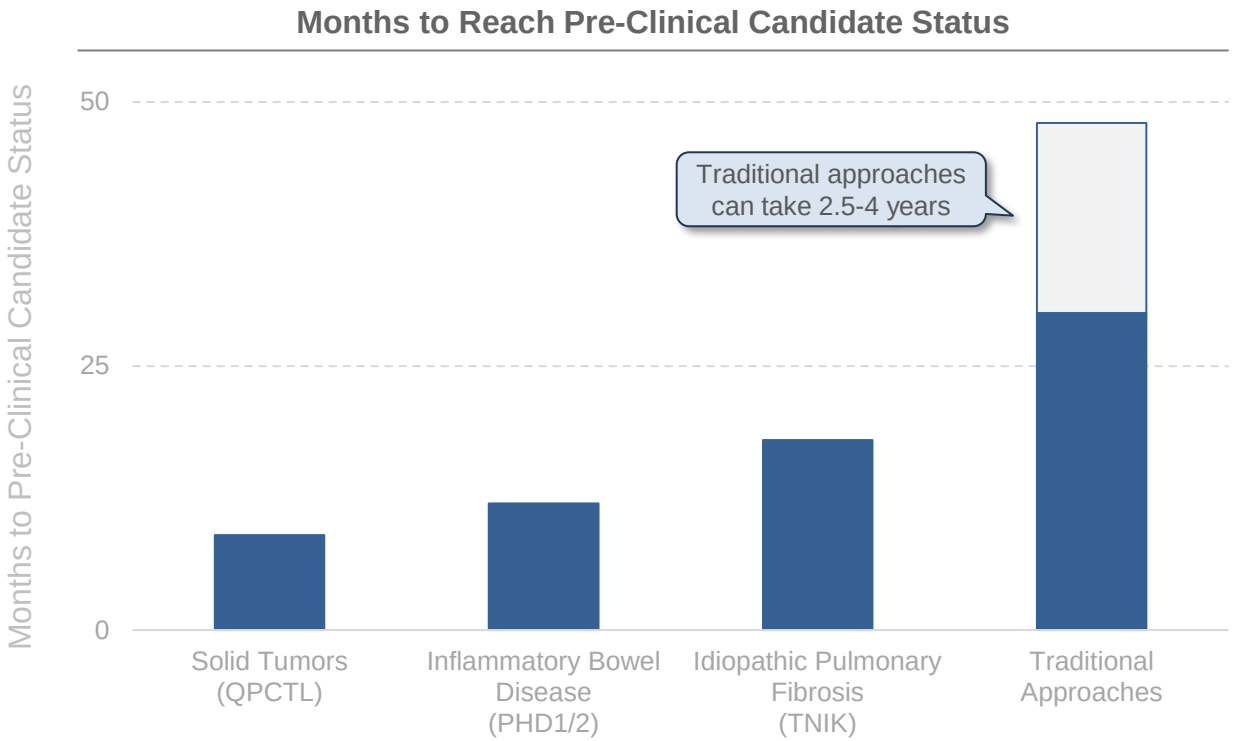
…To reflect the urgency of this effort, Dr. Makary has directed all FDA centers to begin deployment immediately, with the goal of full integration by the end of June.

- USA FDA Press Release, 5/25

Note: FY21, FY22 & FY23 USA government budget figures are actuals. FY24 data is enacted but not actual, FY25 data is requested. NIH share of total budget is requested.
Source: Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25); USA Food & Drug Administration, 'FDA Announces Completion of First AI-Assisted Scientific Review Pilot and Aggressive Agency-Wide AI Rollout Timeline' (5/25); NITRD.gov (5/25)

Research =
30%-80% Reduction in Medical R&D Timelines, per Insilico Medicine & Cradle

AI-Driven Drug Discovery – 2021-2024, Per Insilico Medicine, Cradle & BioPharmaTrend



Cradle

Pharma companies that use Cradle are seeing a 1.5x to 12x speedup in pre-clinical research and development by using our GenAI platform to engineer biologics.

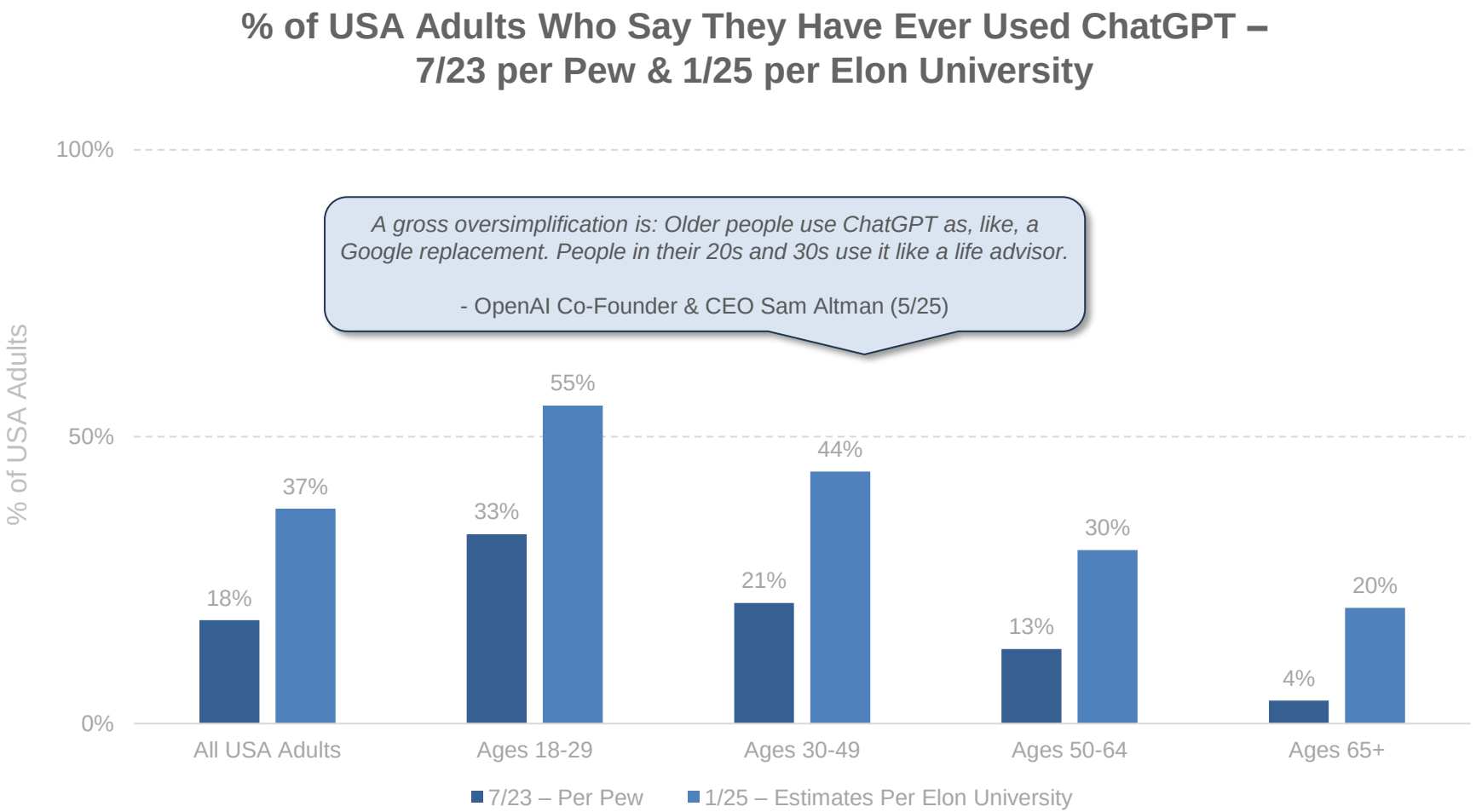
- Stef van Grieken, Co-Founder & CEO of Cradle, 5/25

Note: Pre-Clinical Candidate Status marks the point at which a lead molecule (or biologic) has satisfied all discovery-stage gates and is officially handed off to the development organization for work related to beginning human clinical trials. Figures collected from 2021-2024. Source: Cradle, Insilico Medicine via BioPharmaTrend, 'Insilico Medicine Reports Benchmarks for its AI-Designed Therapeutics' (2/25)

*AI User + **Usage** + CapEx **Growth** =*

Unprecedented

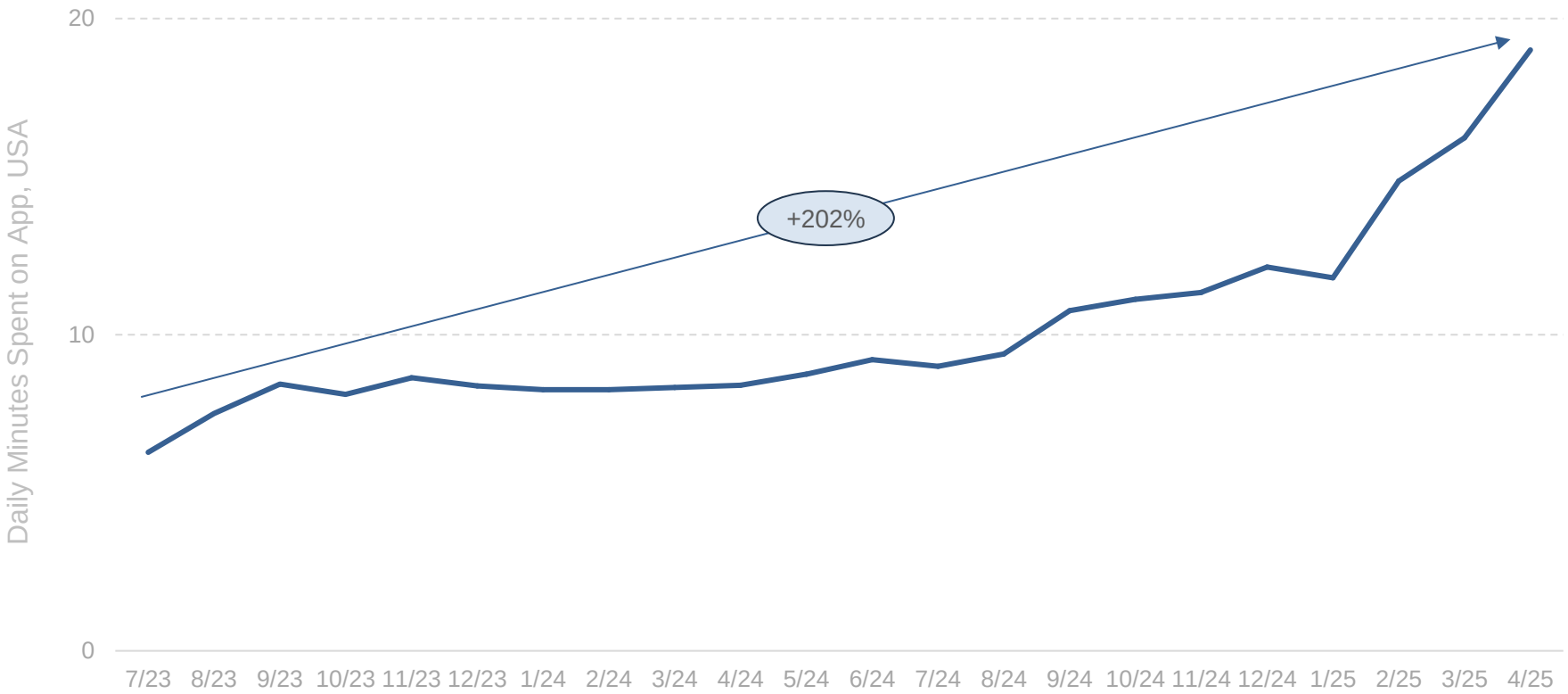
AI Usage – ChatGPT = Rising Rapidly Across Age Groups in USA, per Pew & Elon University



Note: 7/23 data per Pew Research study on ChatGPT use, n=10,133 USA adults. Those who did not give an answer are not shown. 1/25 data per Elon University study on use of any AI models, n=500 USA adults,. Figures estimated based on overall AI tool usage adjusted for an average 72% usage rate of ChatGPT amongst respondents who use any AI tools. Actual ChatGPT penetration may vary by cohort. Note that this chart aggregates data across survey providers and as such may not be directly comparable. Source: Pew Research Center (3/26/24), Elon University (released 3/12/25), Sam Altman (5/12/25) via Fortune

AI Engagement (ChatGPT App as Proxy) = +202% Rise in Daily Time Spent Over Twenty-One Months…

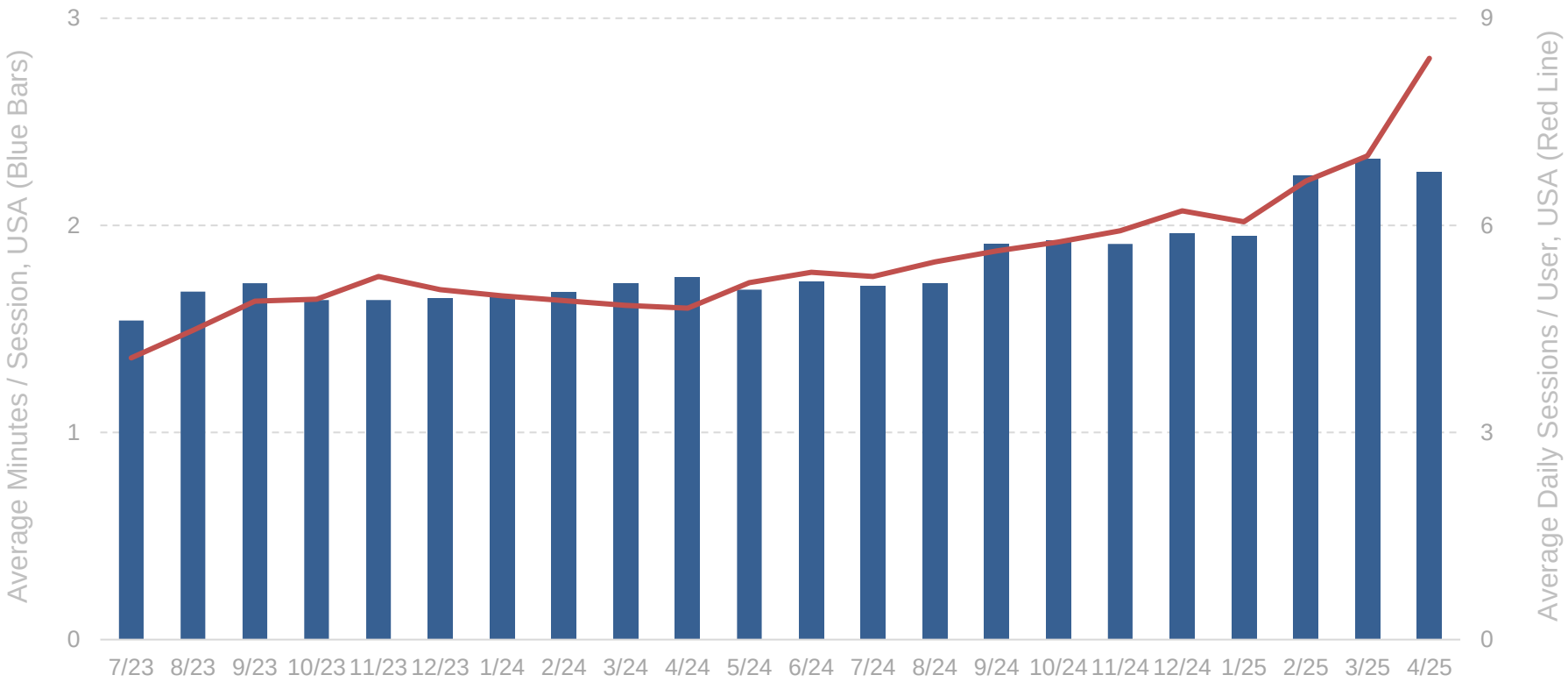
Minutes per Day that USA Active Users Spend on ChatGPT App – 7/23-4/25, per Sensor Tower



Note: Data represents USA App Store & Google Play Store monthly active users. Data for ChatGPT standalone app only. ChatGPT app not available in China, Russia and select other countries as of 5/25. Source: Sensor Tower (5/25)

…AI Engagement (ChatGPT App as Proxy) =
+106% Growth in Sessions & +47% Growth in Duration Over Twenty-One Months

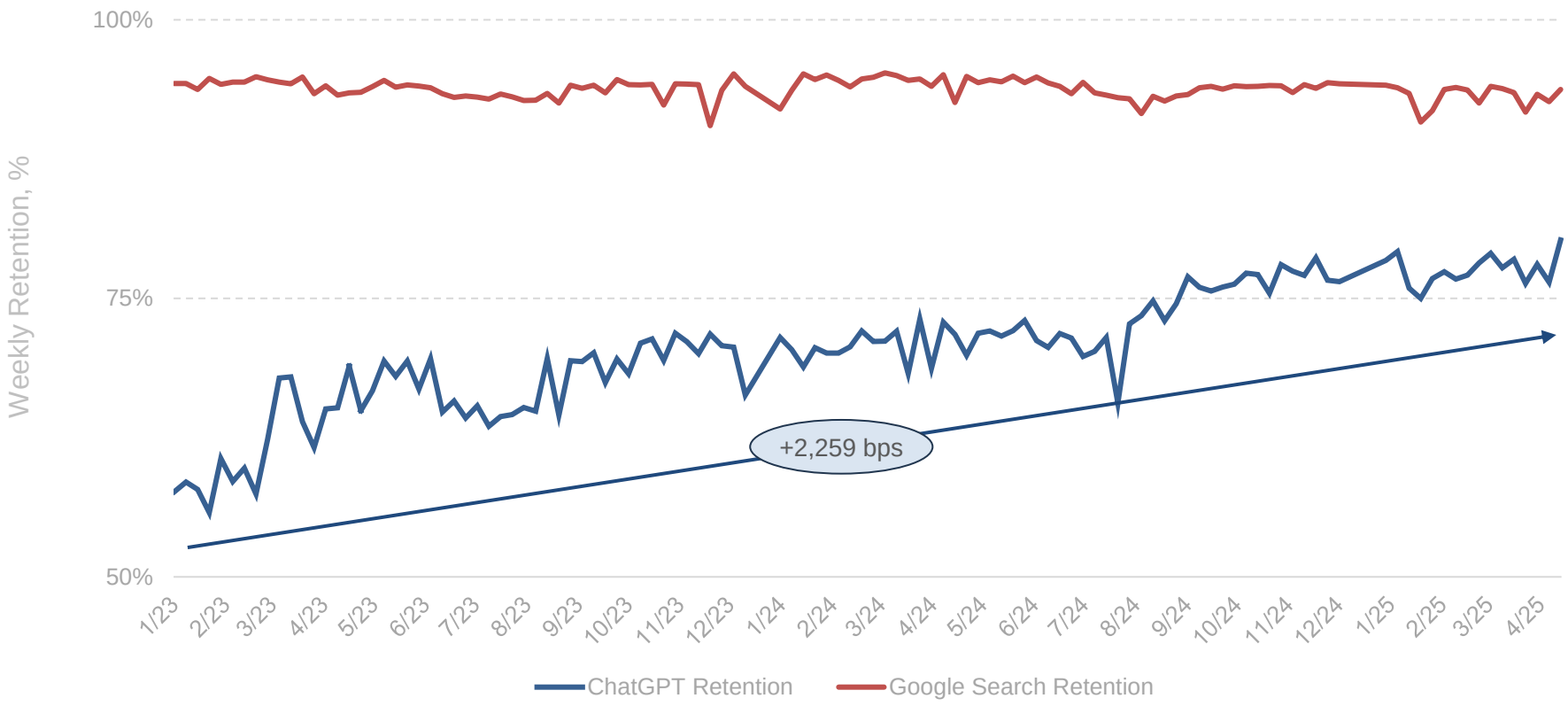
Average USA Session Duration (Minutes) & Daily Sessions per User for ChatGPT App – 7/23-4/25, per Sensor Tower



Note: Data represents USA App Store & Google Play Store monthly active users. Data for ChatGPT standalone app only. ChatGPT app not available in China, Russia and select other countries as of 5/25. Source: Sensor Tower (5/25)

AI Retention (ChatGPT as Proxy) = 80% vs. 58% Over Twenty-Seven Months, per YipitData

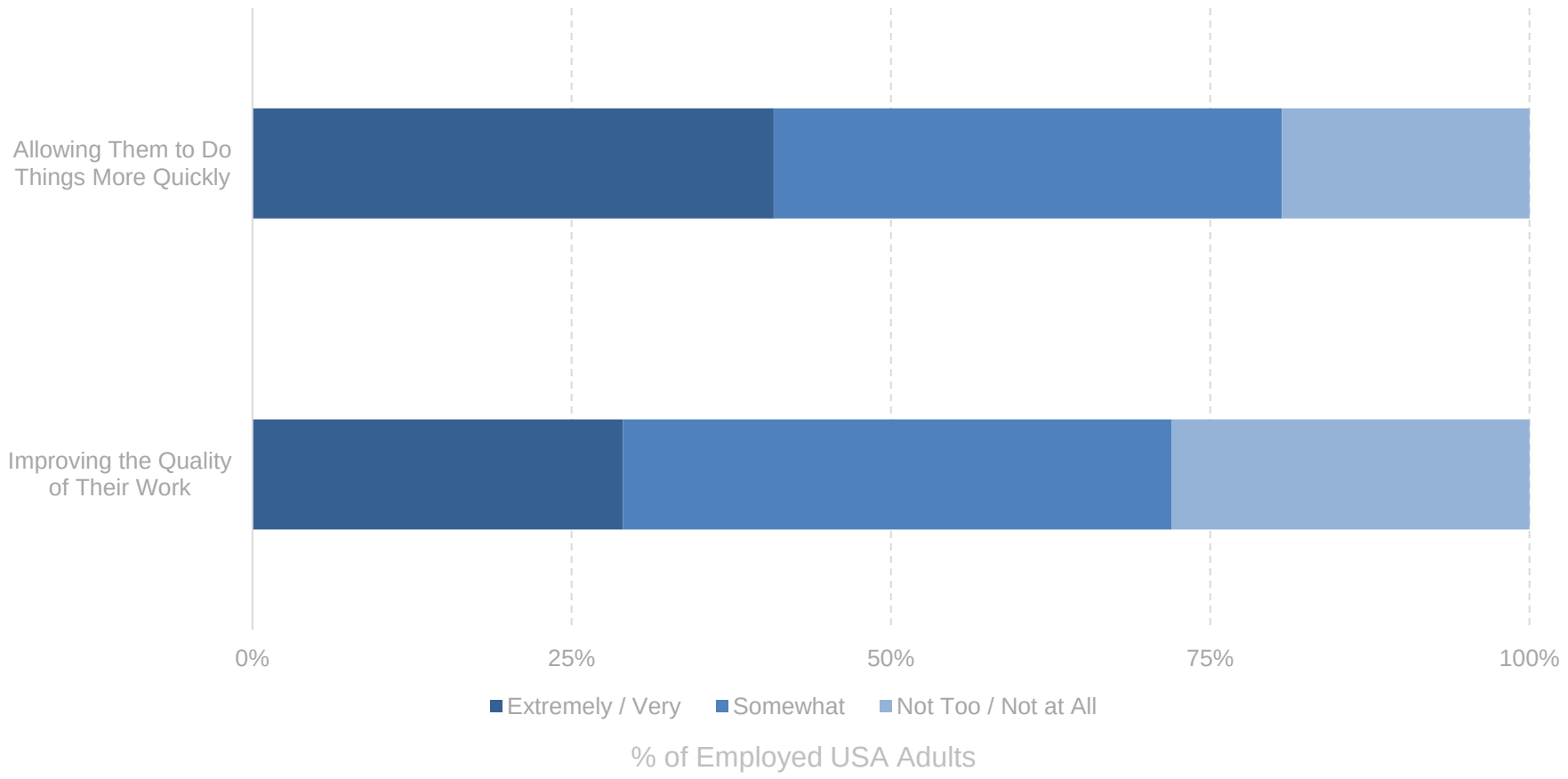
Consumer ChatGPT & Google Search Global Desktop User Retention Rates (1/23-4/25), per YipitData



Note: Retention Rate = Percentage of users from the immediately preceding week that were users again in the current week. Data measures several million global active desktop users' clickstream data. Data consists of users' web requests & is collected from web services / applications, such as VPNs and browser extensions. Users must have been part of the panel for 2 consecutive months to be included. Panel is globally-representative, though China data may be subject to informational limitations due to government restrictions. Excludes anomalies in w/c 12/24/23, 12/31/23, 12/22/24, 12/29/24, 1/5/25, potentially due to holiday breaks causing less enterprise usage. Source: YipitData (5/25)

AI Chatbots @ Work Tells = >72% Doing Things Quicker / Better

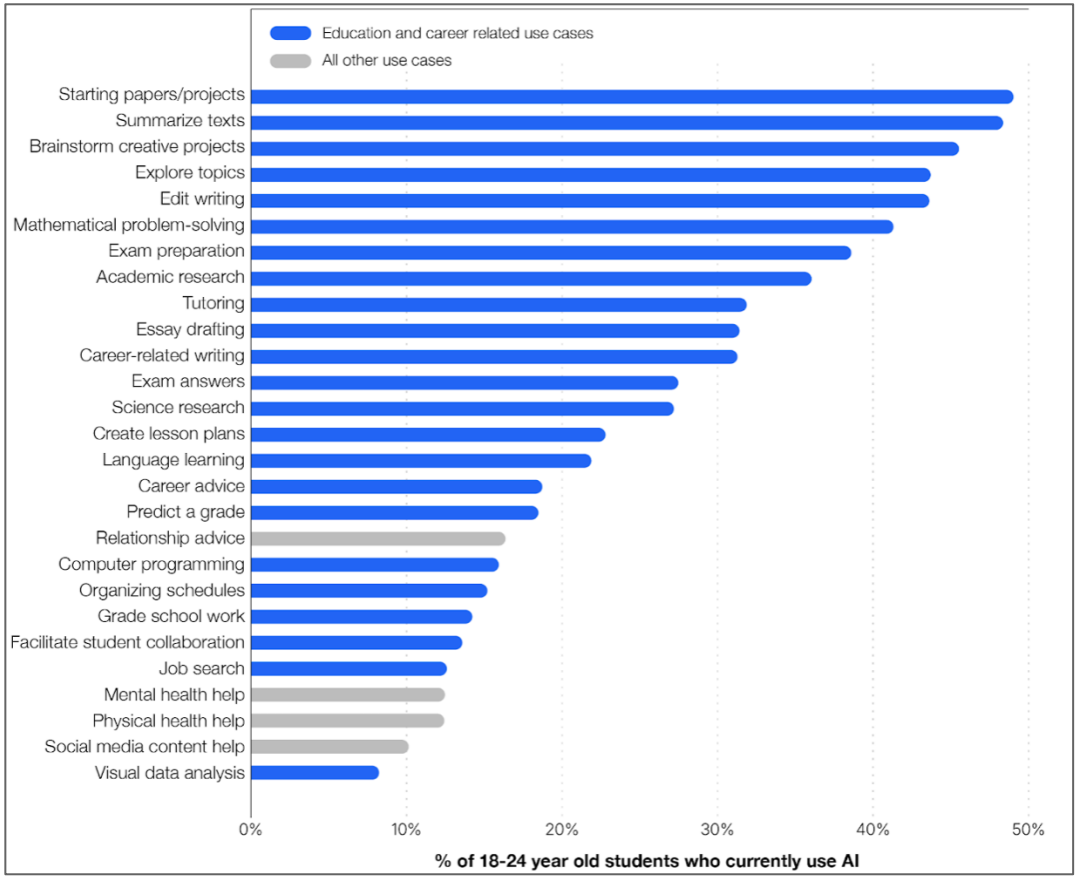
% of Employed USA Adults Using AI Chatbots Who Say Tools Have Been _____ Helpful When It Comes to… – 10/24, per Pew



Note: N = 5,273 USA adults who are employed part time or full time and who have only one job or have more than one but consider one of them to be their primary job were surveyed.
Source: Pew Research Center (10/24)

AI Chatbots @ School Tells (ChatGPT as Proxy) = Bias to Research / Problem Solving / Learning / Advice

OpenAI ChatGPT Usage Survey, USA Students Ages 18-24 – 12/24-1/25, per OpenAI

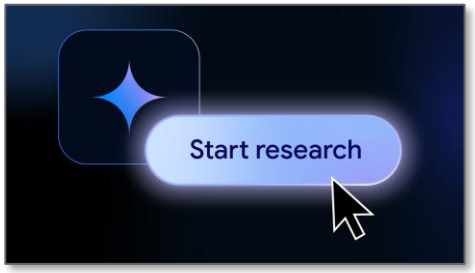


Note: Data per OpenAI survey (12/24), n = 1,299 USA college and graduate students across a mix of STEM and non-STEM disciplines; only answers from 18-24 year olds used. Sample includes both AI users and non-users but excludes "AI rejectors" – defined as non-users with little to no interest in adopting AI within the next 12 months. Source: OpenAI, 'Building an AI-Ready Workforce: A Look at College Student ChatGPT Adoption in the US' (2/25)

AI Usage Expansion – Deep Research = Automating Specialized Knowledge Work

Select AI Company Deep Research Capabilities – 12/24-2/25, per Google, OpenAI & xAI

Google Gemini Deep Research

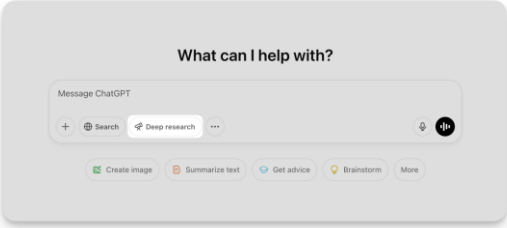


Get up to speed on just about anything with Deep Research, an agentic feature in Gemini that can automatically browse up to hundreds of websites on your behalf, think through its findings, and ***create insightful multi-page, reports that you can turn into engaging podcast-style conversations****…*

…It's a step towards more agentic AI that can move beyond simple question-answering to become a true collaborative partner.

- Google Deep Research Overview, launched 12/24

OpenAI ChatGPT Deep Research

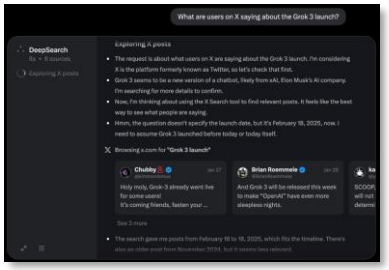


Today we're launching deep research in ChatGPT, a new agentic capability that conducts multi-step research on the internet for complex tasks.
It accomplishes in tens of minutes what would take a human many hours*…*

…Deep research marks a significant step toward our broader goal of developing AGI, which we have long envisioned as capable of producing novel scientific research.

- OpenAI Deep Research Press Release, 2/25

xAI Grok DeepSearch



To understand the universe, we must interface Grok with the world…

…As a first step towards this vision, we are rolling out DeepSearch – our first agent. It's a ***lightning-fast AI agent built to relentlessly seek the truth across the entire corpus of human knowledge****. DeepSearch is designed to synthesize key information, reason about conflicting facts and opinions, and distill clarity from complexity.*

- xAI Grok 3 Beta Press Release, 2/25

Source: Google (5/25), OpenAI (2/25), xAI (2/25), Digital Trends (1/25)

AI Agent Evolution =

Chat Responses → Doing Work

A new class of AI is now emerging – less assistant, more service provider.
What began as basic conversational interfaces may now be evolving into something far more capable.

Traditional chatbots were designed to respond to user prompts, often within rigid scripts or narrow flows.
They could fetch answers, summarize text, or mimic conversation – but always in a reactive, limited frame.

AI agents represent a step-change forward. These are intelligent long-running processes
that can reason, act, and complete multi-step tasks on a user's behalf. They don't just answer questions –
they execute: booking meetings, submitting reports, logging into tools, or orchestrating workflows across platforms,
often using natural language as their command layer.

This shift mirrors a broader historical pattern in technology.
Just as the early 2000s saw static websites give way to dynamic web applications –
where tools like Gmail and Google Maps transformed the internet from a collection of pages into a set of utilities –
AI agents are turning conversational interfaces into functional infrastructure.

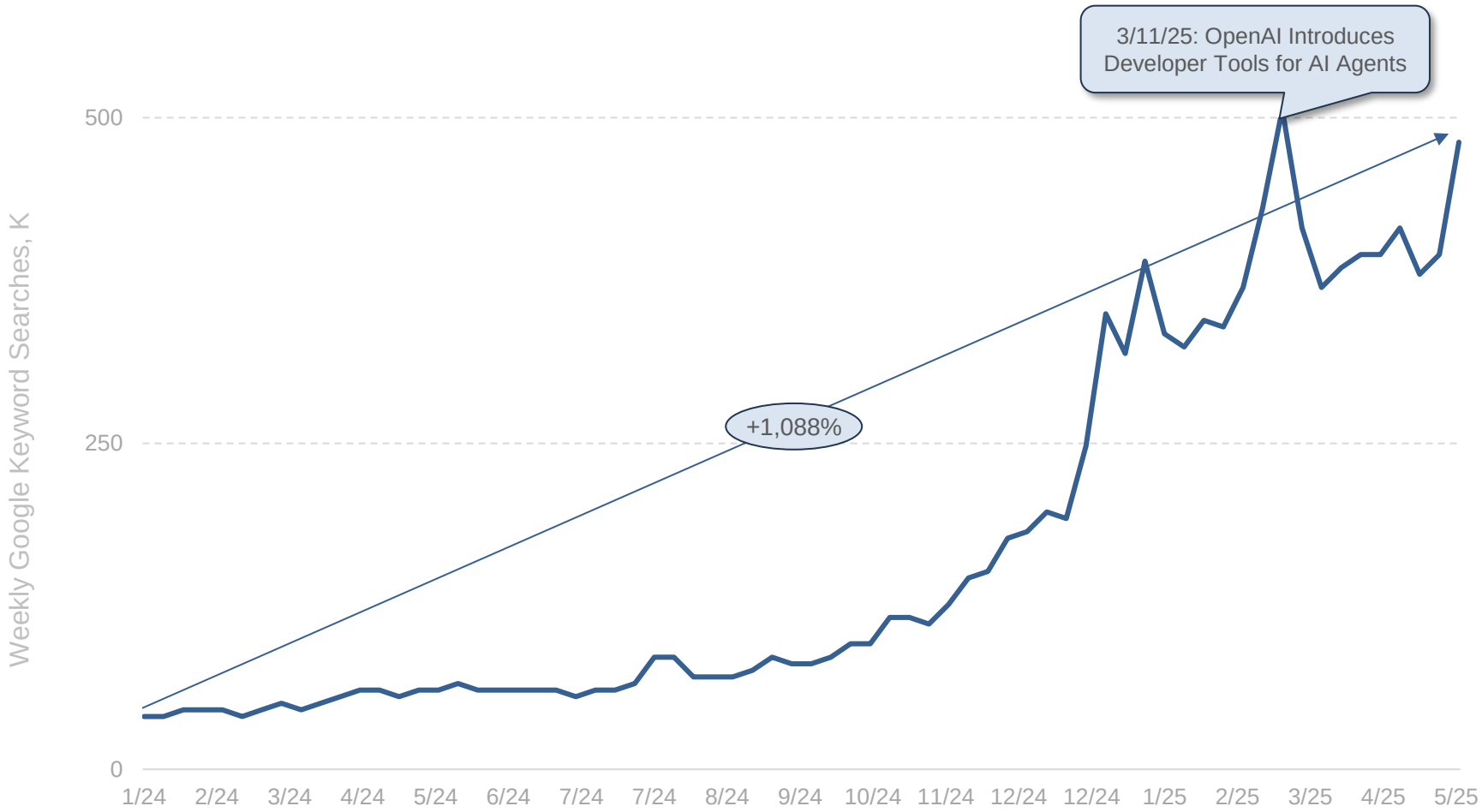
Whereas early assistants needed clear inputs and produced narrow outputs, agents promise to operate with goals,
autonomy and certain guardrails. They promise to interpret intent, manage memory, and coordinate across
apps to get real work done. It's less about responding and more about accomplishing.

While we are early in the development of these agents, the implications are just starting to emerge.
AI agents could reshape how users interact with digital systems –
from customer support and onboarding to research, scheduling, and internal operations.

Enterprises are leading the charge; they're not just experimenting with agents, but deploying them,
investing in frameworks and building ecosystems around autonomous execution.
What was once a messaging interface is becoming an action layer.

AI Agent Interest (Google Searches) = +1,088% Over Sixteen Months

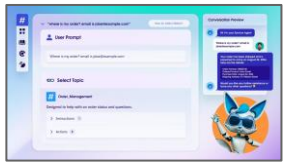
Global Google Searches for 'AI Agent' (K) – 1/24-5/25, per Google Trends



Source: Google Trends via Glimpse (5/15/24), OpenAI (3/25)

AI Agent Deployments = AI Incumbent Product Launches Accelerating

AI Incumbent Agent Launches

Agent Released	Select Capabilities
 **Salesforce Agentforce** *(10/24 = General Release)*	• Automated customer support • Case resolution • Lead qualification • Order tracking
Anthropic Claude 3.5 Computer Use *(10/24 = Research Preview Release)*	• Control computer screen directly to perform tasks like pulling data from websites, making online purchases, etc.
OpenAI Operator *(1/25 = Research Preview Release)*	• Control computer screen directly to perform tasks like pulling data from websites, making online purchases, etc.
Amazon Nova Act *(3/25 = Research Preview Release)*	• Home automation • Information collection • Purchasing • Scheduling

Source: Salesforce (10/24), Salesforce Ben, Anthropic (10/24), OpenAI (1/25), Amazon (3/25)

Next Frontier For AI =

Artificial General Intelligence

Artificial General Intelligence, or AGI, refers to systems capable of performing the full range of human intellectual tasks – reasoning, planning, learning from small data samples, and generalizing knowledge across domains. Unlike current AI models, which excel within specific (albeit broad) boundaries, AGI would be able to operate fully flexibly across disciplines and solve unfamiliar problems without retraining. It represents a major milestone in AI development – one that builds on recent exponential gains in model scale, training data, and computational efficiency.

Timelines for AGI remain uncertain, but expert expectations have shifted forward meaningfully in recent years. Sam Altman, CEO of OpenAI, remarked in January 2025, *We are now confident we know how to build AGI as we have traditionally understood it.* This is a forecast, not a dictum, but it reflects how advances in model architecture, inference* efficiency, and training scale are shortening the distance between research and frontier capability. The broader thread is clear: AI development is trending at unprecedented speed, and AGI is increasingly being viewed not as a hypothetical endpoint, but as a reachable threshold.

If / when achieved, AGI would redefine what software (and related hardware) can do. Rather than executing pre-programmed tasks, AGI systems would understand goals, generate plans, and self-correct in real time. They could drive research, engineering, education, and logistics workflows with little to no human oversight – handling ambiguity and novelty with general-purpose reasoning. These systems wouldn't require extensive retraining to handle new problem domains – they would transfer learning and operate with context, much like human experts. Additionally, humanoid robots powered by AGI would have the power to reshape our physical environment and how we operate in it.

Still, the implications warrant a measured view. AGI is not a finish line, but a phase shift in capability – and how it reshapes institutions, labor, and decision-making will depend on the safeguards and deployment frameworks that accompany it. The productivity upside may be significant, but unevenly distributed.

The geopolitical, ethical, and economic implications may evolve gradually, not abruptly. As with earlier transitions – from industrial to digital to algorithmic – the full consequences will be shaped not just by what the technology can do, but by how society chooses to adopt and govern it.

**Inference = Fully-trained model generates predictions, answers, or content in response to user inputs. This phase is much faster and more efficient than training.*

*AI User + Usage + **CapEx Growth** =*

Unprecedented

To understand where technology CapEx is heading, it helps to look at where it's been.
Over the past two decades, tech CapEx has flexed upward at points through data's long arc –
first toward storage / access, then toward distribution / scale, and now toward computation / intelligence.

The earliest wave saw CapEx pouring into building internet infrastructure –
massive server farms, undersea cables, and early data centers that enabled Amazon, Microsoft, Google and others to lay the foundation for cloud computing. That was the first phase: store it, organize it, serve it.

The second wave – still unfolding – has been about supercharging compute for data-heavy AI workloads, a natural evolution of cloud computing. Hyperscaler* CapEx budgets now tilt increasingly toward specialized chips (GPUs, TPUs, AI accelerators…), liquid cooling, and frontier data center design.

In 2019, AI was a research feature; by 2023, it was a capital expenditure line item.
Microsoft Vice Chair and President Brad Smith put it well in a 4/25 blog post:
Like electricity and other general-purpose technologies in the past, AI and cloud datacenters represent the next stage of industrialization.

The world's biggest tech companies are spending tens of billions annually – not just to gather data, but to learn from it, reason with it and monetize it in real time. It's still about data – but now, the advantage goes to those who can train on it fastest, personalize it deepest, and deploy it widest.

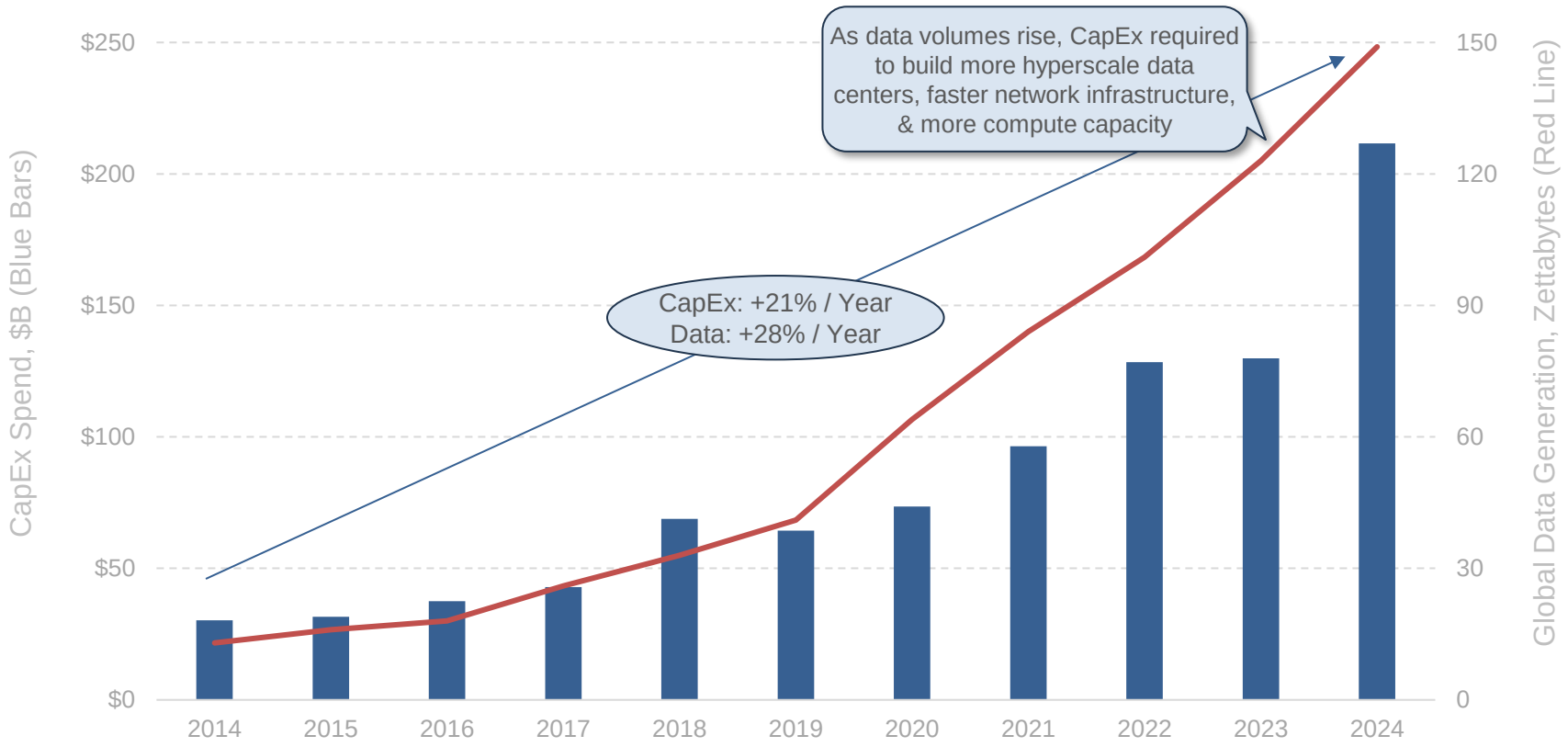
**Hyperscalers (large data center operators) are Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), Alibaba Cloud, Oracle Cloud Infrastructure (OCI), IBM Cloud & Tencent Cloud.*

CapEx Spend – Big Technology Companies =

On Rise for Years as
Data Use + Storage Exploded

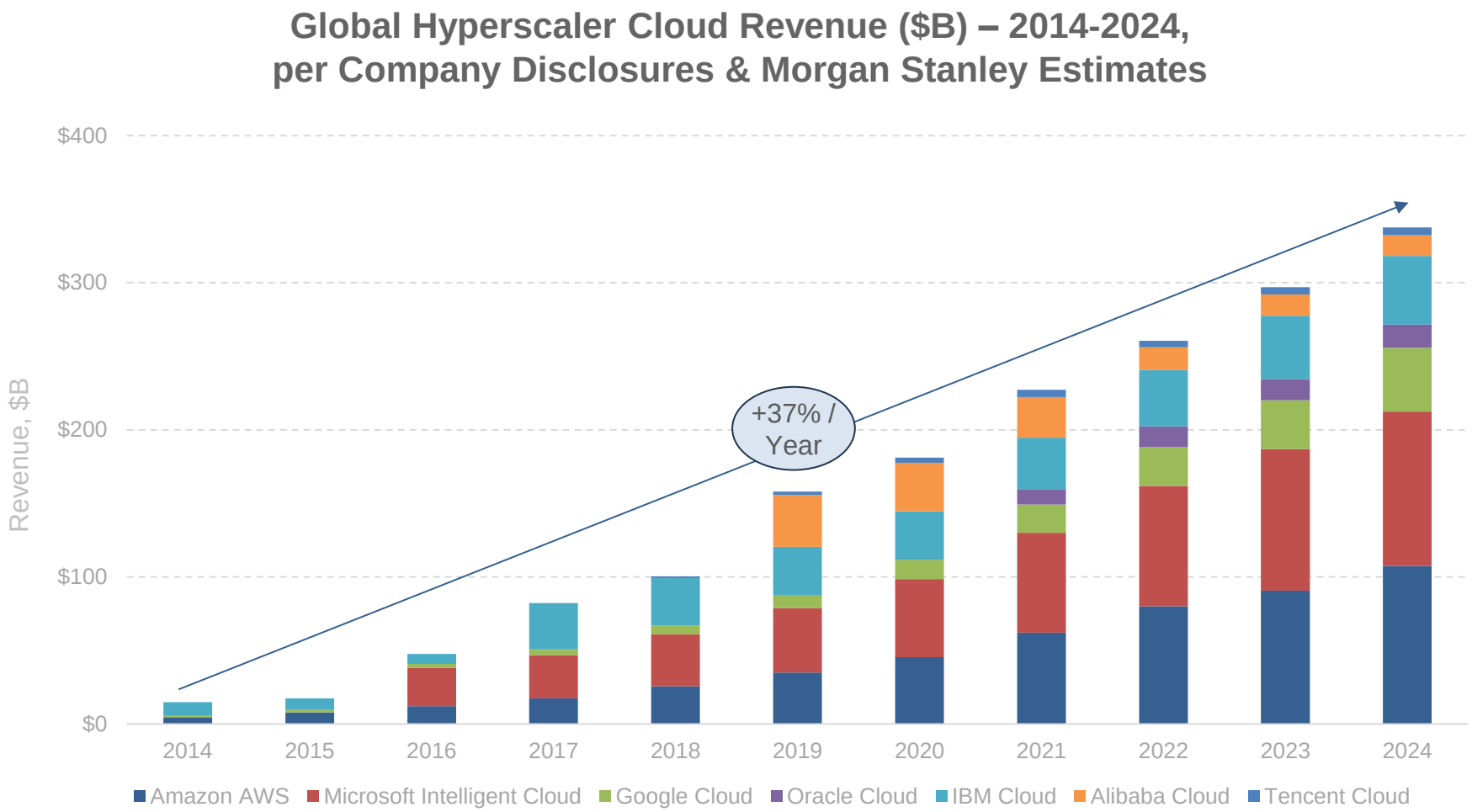
CapEx Spend @ Big Six* Tech Companies (USA) = +21% Annual Growth Over Ten Years

Big Six* USA Public Technology Company CapEx Spend ($B) vs. Global Data Generation (Zettabytes) – 2014-2024, per Capital IQ & Hinrich Foundation



**Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. Only AWS CapEx & revenue shown for Amazon (i.e. excludes Amazon retail CapEx). AWS CapEx estimated per Morgan Stanley – equals AWS net additions to property & equipment less finance leases and obligations. Global data generation figures for 2024 are estimates. Source: Capital IQ (3/25), Hinrich Foundation (3/25)*

CapEx Spend for Tech Hyperscalers = Mirrored by… +37% Annual Cloud Revenue Growth Over Ten Years



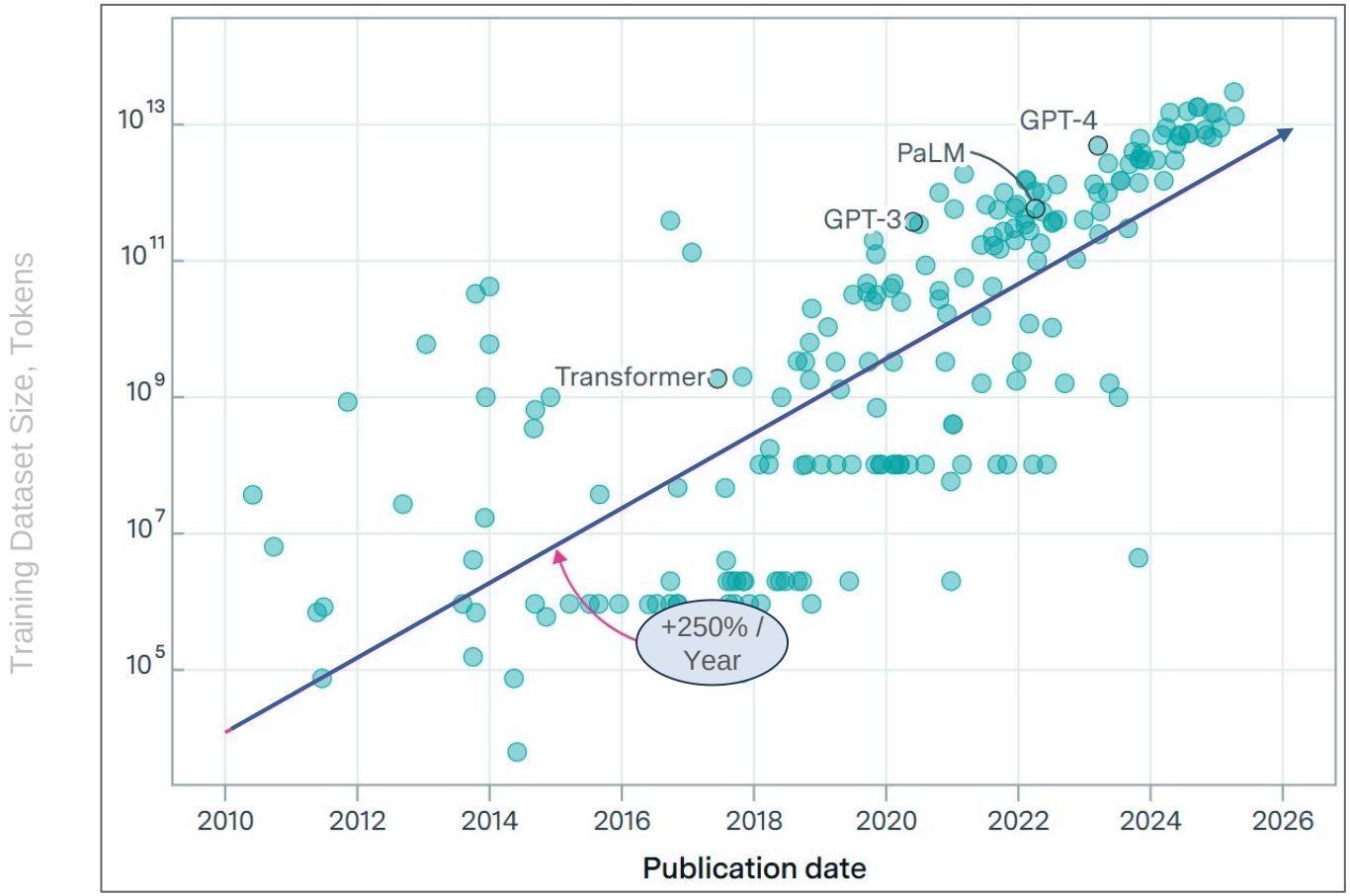
Note: Companies do not report "hyperscaler cloud revenue" on like-for-like basis so data represents best estimates and may not align between companies. Oracle Cloud revenue includes Cloud Services & License Support, as well as Cloud License & On-Premise License. IBM Cloud includes all 'Infrastructure' line items due to reporting standards. Alibaba & Tencent Cloud revenues estimated per Morgan Stanley. Source: Company disclosures, Morgan Stanley (as of 4/25)

CapEx Spend – Big Technology Companies =

Inflected With AI’s Rise

AI Model Training Dataset Size = 250% Annual Growth Over Fifteen Years, per Epoch AI

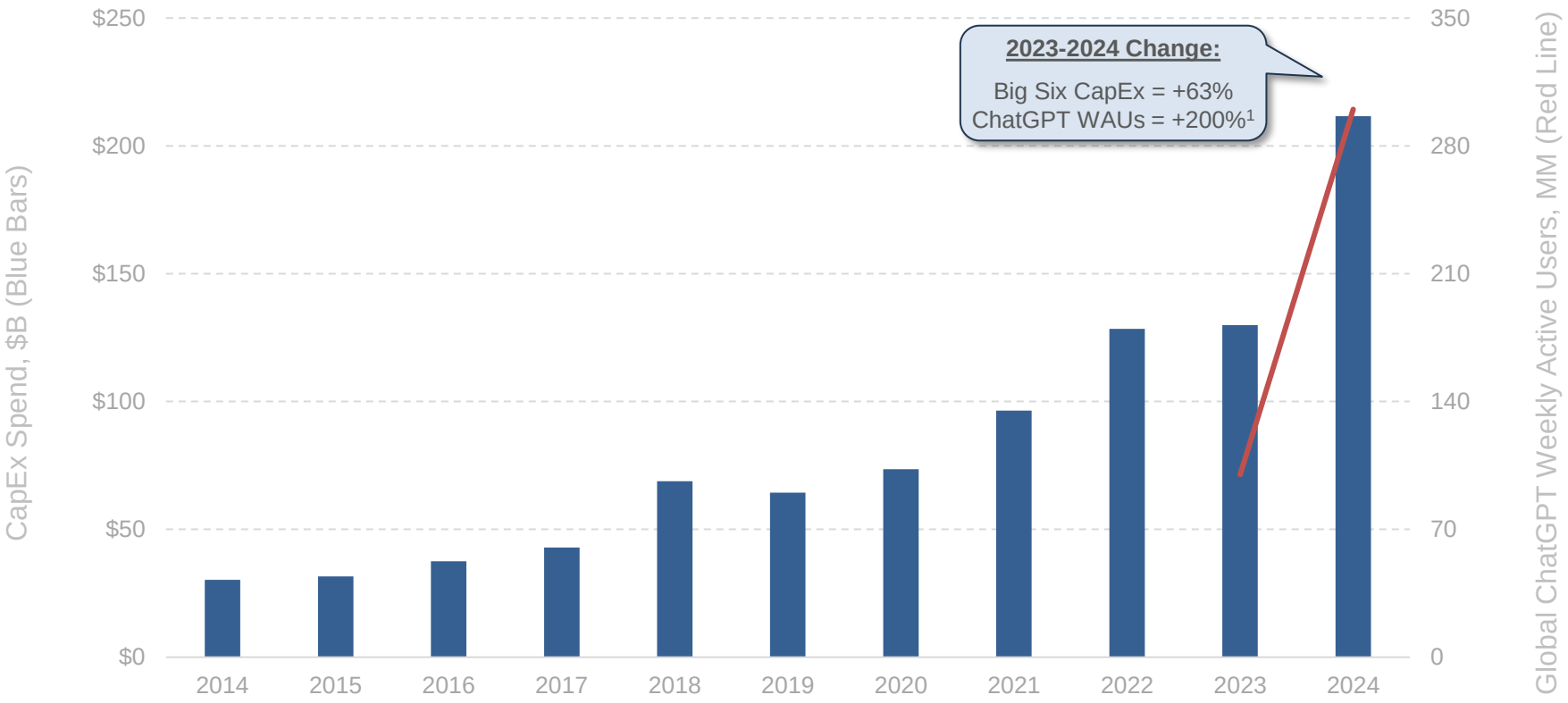
AI Model Training Dataset Size (Tokens) by Model Release Year – 6/10-5/25, per Epoch AI



Note: In AI language models, tokens represent basic units of text (e.g., words or sub-words) used during training. Training dataset sizes are often measured in total tokens processed. A larger token count typically reflects more diverse and extensive training data, which can lead to improved model performance – up to a point – before reaching diminishing returns.
Source: Epoch AI (5/25)

CapEx Spend @ Big Six* Tech Companies = +63% Y/Y & Accelerated…

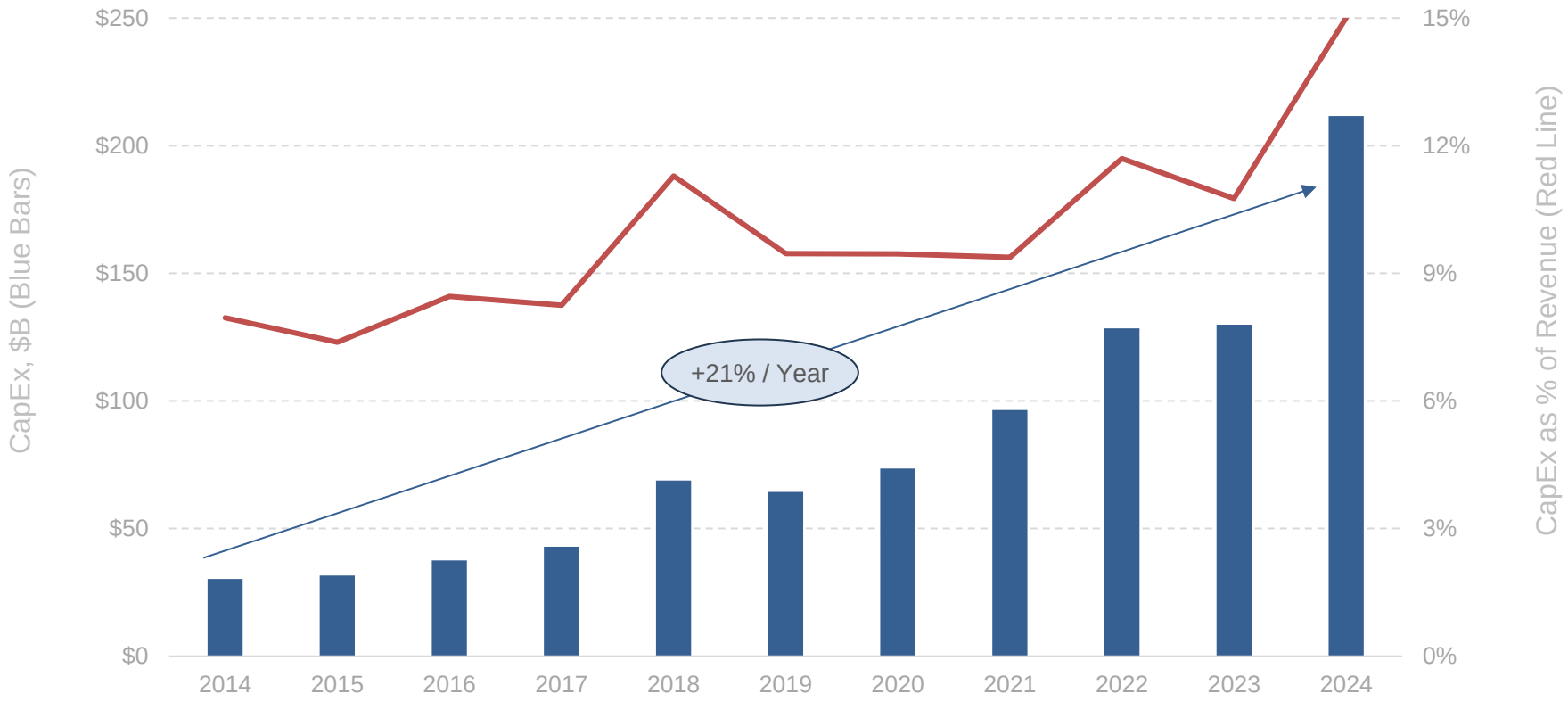
Big Six* USA Public Technology Company CapEx Spend ($B) vs. Global ChatGPT Weekly Active Users (MM) – 2014-2024, per Capital IQ & OpenAI



[1]ChatGPT WAU data as of 11/23 & 12/24 due to data availability.
**Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. Only AWS CapEx & revenue shown for Amazon (i.e. excludes Amazon retail CapEx). AWS CapEx estimated per Morgan Stanley – equals AWS net additions to property & equipment less finance leases and obligations. Source: Capital IQ (3/25), OpenAI disclosures (3/25)*

…CapEx Spend @ Big Six* Tech Companies = 15% of Revenue & Accelerated vs. 8% Ten Years Ago

Big Six* USA Public Technology Company – CapEx Spend ($B) vs. % of Revenue – 2014-2024, per Capital IQ & Morgan Stanley



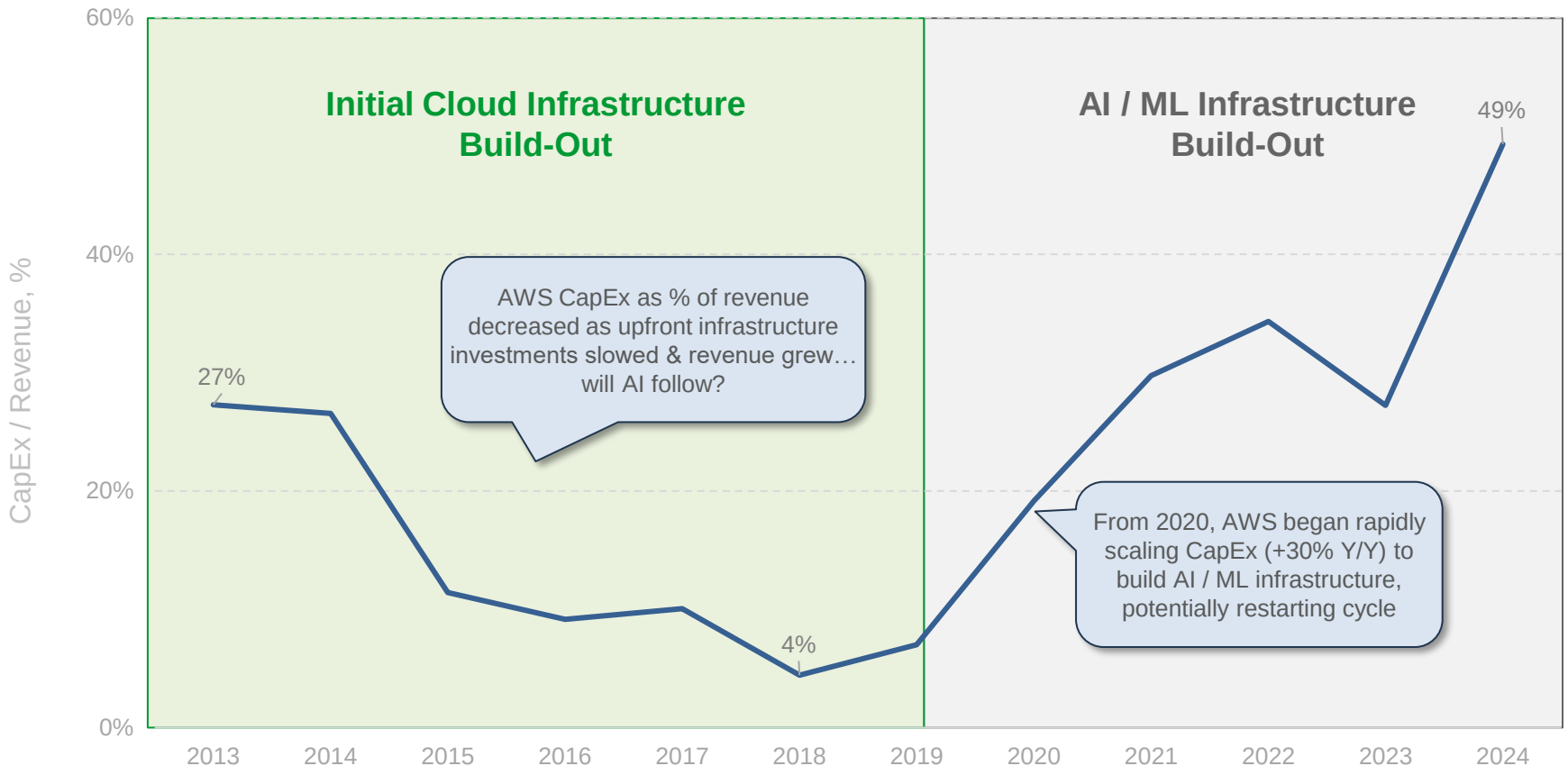
**Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. Only AWS CapEx & revenue shown for Amazon (i.e. excludes Amazon retail CapEx). AWS CapEx estimated per Morgan Stanley – equals AWS net additions to property & equipment less finance leases and obligations.*
Source: Capital IQ (3/25), Morgan Stanley (5/25)

CapEx Spend @ Amazon AWS =

Cloud vs. AI Patterns

CapEx as % of Revenue (AWS as Proxy) – AI vs. Cloud Buildouts = 49% (2024) vs. 4% (2018) vs. 27% (2013), per Morgan Stanley

Amazon AWS CapEx as % of Revenue – 2013-2024, Estimated per Morgan Stanley



Note: Figures shown represent AWS only. AWS CapEx estimated per Morgan Stanley – equals AWS net additions to property & equipment less finance leases and obligations.
Source: Amazon, Morgan Stanley (5/25)

Tech CapEx Spend Partial Instigator =

Material Improvements in GPU Performance

NVIDIA GPU Performance = +225x Over Eight Years

Performance of NVIDIA GPU Series Over Time – 2016-2024, per NVIDIA

$1B Data Center Comparison
GPT-MoE Inference Workload[1]

	Pascal	Volta	Ampere	Hopper	Blackwell	
	2016	2018	2020	2022	2024	
Number of GPUs	46K	43K	28K	16K	11K	**+225x**
Factory AI FLOPS	1EF	5EF	17EF	63EF	220EF	
Annual Inference Tokens	50B	1T	5T	58T	1,375T	**+30,000x**
Annual Token Revenue	$240K	$3M	$24M	$300M	$7B	
DC Power	37MW	34MW	25MW	19MW	21MW	**+50,000x**
Token Per MW-Year	1.3B	2.9B	200B	3T	65T	

For a Theoretical $1B-Scale Data Center…

…Performance +225x over eight years while requiring 4x fewer GPUs…

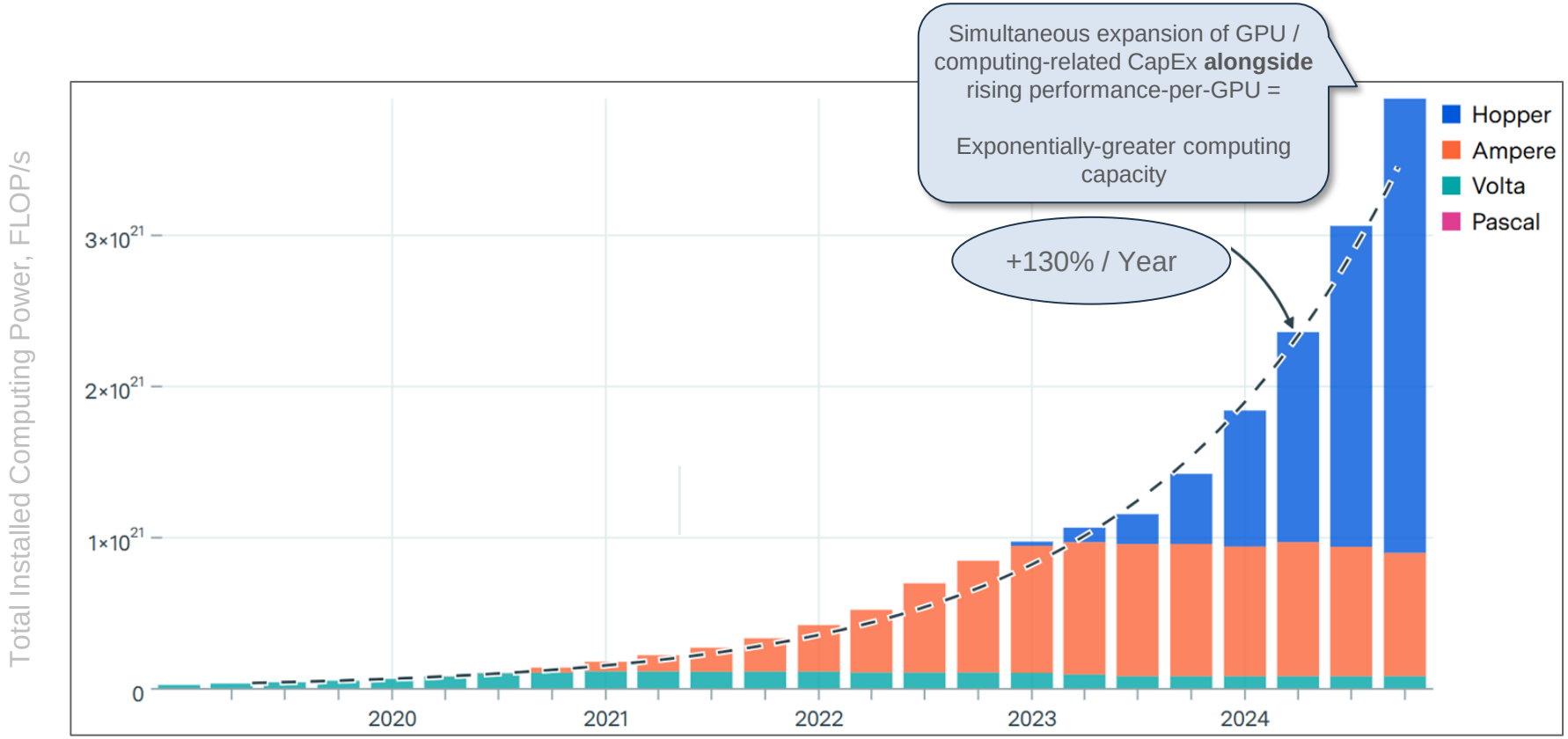
…Inference token capacity +27,500x over eight years, implying +30,000x higher theoretical token revenue…

…Data center power use down 43% over eight years, leading to +50,000x greater per-unit energy efficiency

[1] GPT-MoE Inference Workload = A type of workload where a GPT-style model with a Mixture-of-Experts (MoE) architecture is used for inference (i.e., making predictions).
Note: Annual token revenue assumes a flat per-token cost. Source: NVIDIA (5/25)

NVIDIA Installed GPU Computing Power = 100x+ Growth Over ~Six Years

Global Stock of NVIDIA GPU Computing Power (FLOP/s) – Q1:19-Q4:24, per Epoch AI



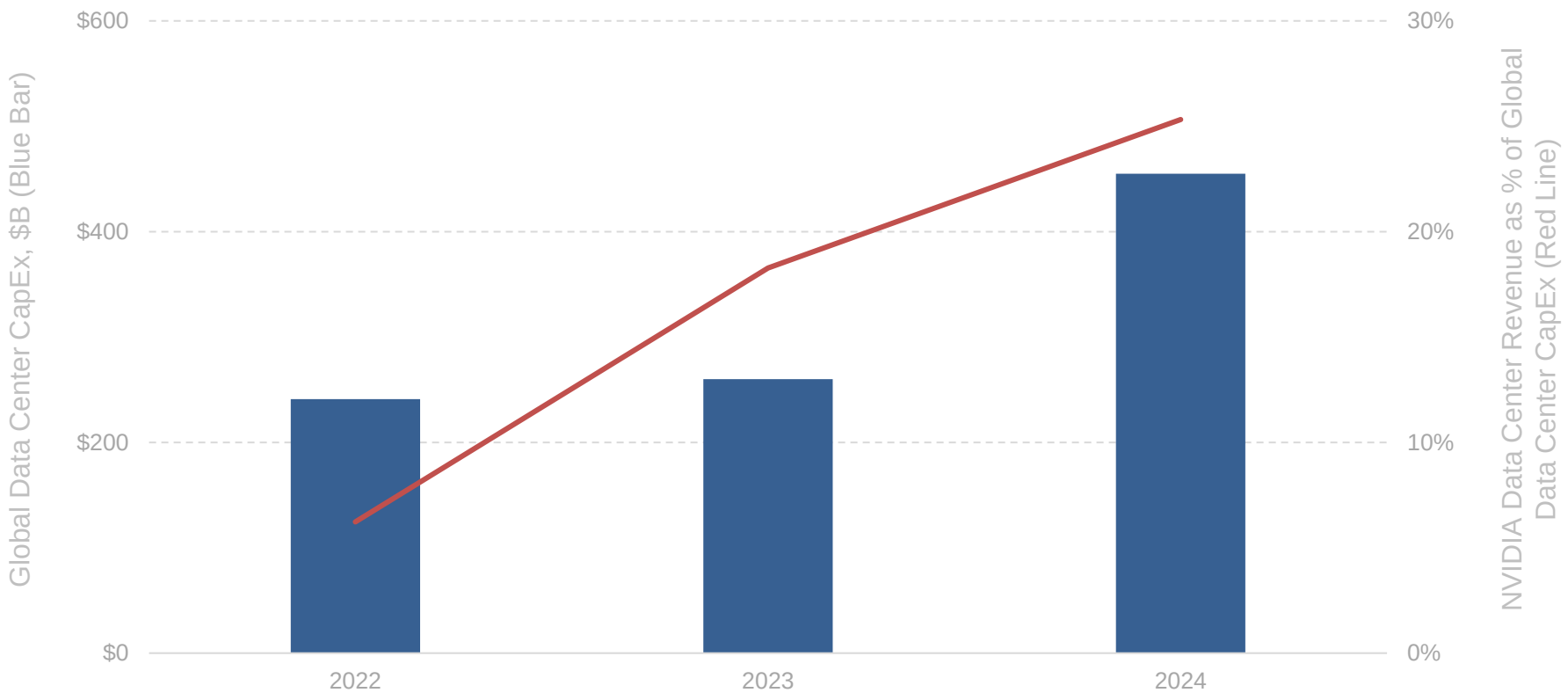
Note: Analysis does not include TPUs or other specialized AI accelerators, for which less data is available. TPUs may provide comparable total computing power to NVIDIA chips.
Source: Epoch AI (2/25)

Tech CapEx Spend Beneficiary =

NVIDIA

Key Tech CapEx Spend Beneficiary = NVIDIA…
25% & Rising of Global Data Center CapEx, per NVIDIA

Global Data Center CapEx ($B) vs. NVIDIA's Data Center Revenue as Percent of Data Center CapEx (Global) – 2022-2024, per NVIDIA @ GTC



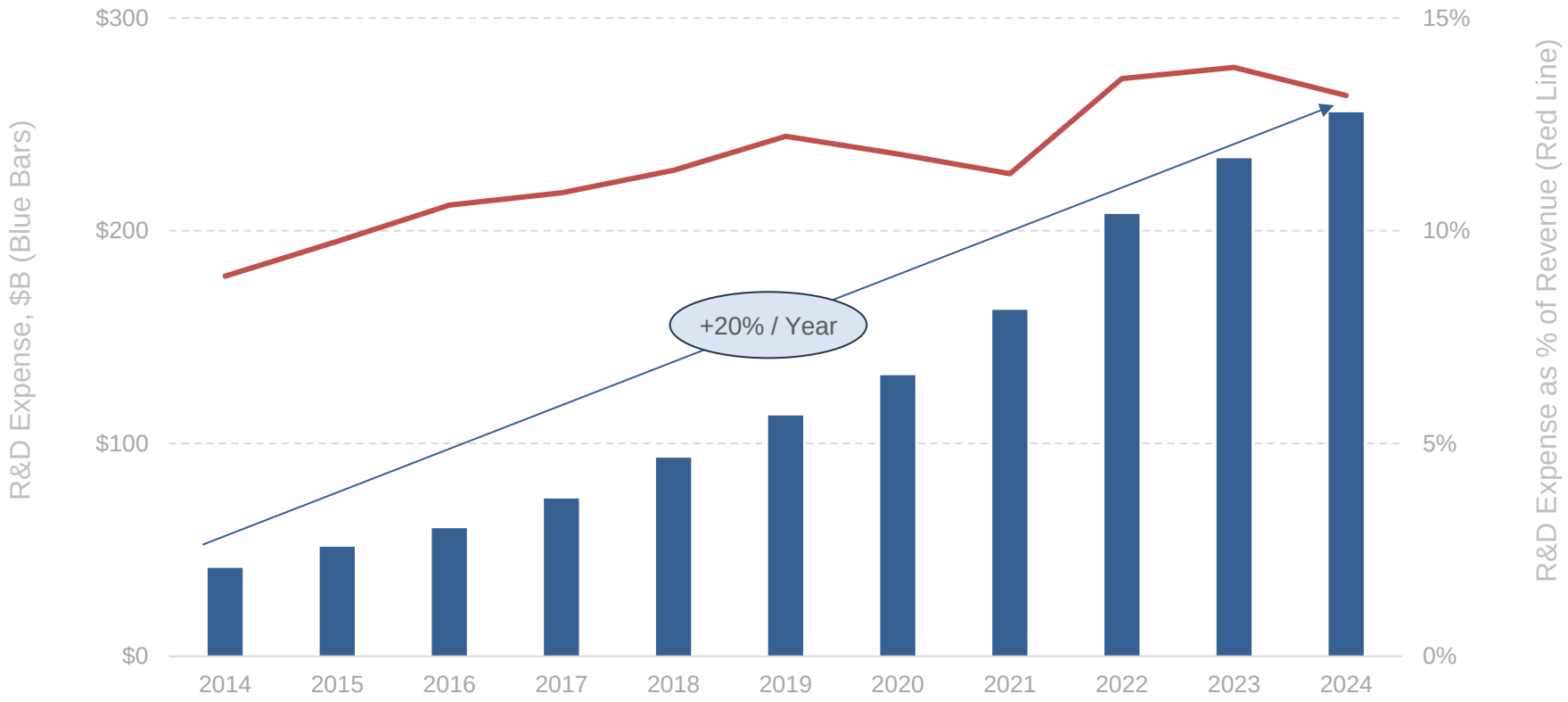
Note: NVIDIA data represents January FYE (e.g., 2024 = FY25 ending 1/25) vs calendar year for data center CapEx. Data presented by Jensen Huang at NVIDIA GTC 2025 (link).
Source: Dell'Oro Research for CapEx (3/25); NVIDIA for data center revenue (3/25)

Technology Company Spend =

R&D Rising Along with CapEx

R&D Spend @ Big Six* USA Public Tech Companies = 13% of Revenue…vs. 9% Ten Years Ago

Big Six* USA Public Technology Company – R&D Spend ($B) vs. % of Revenue – 2014-2024, per Capital IQ



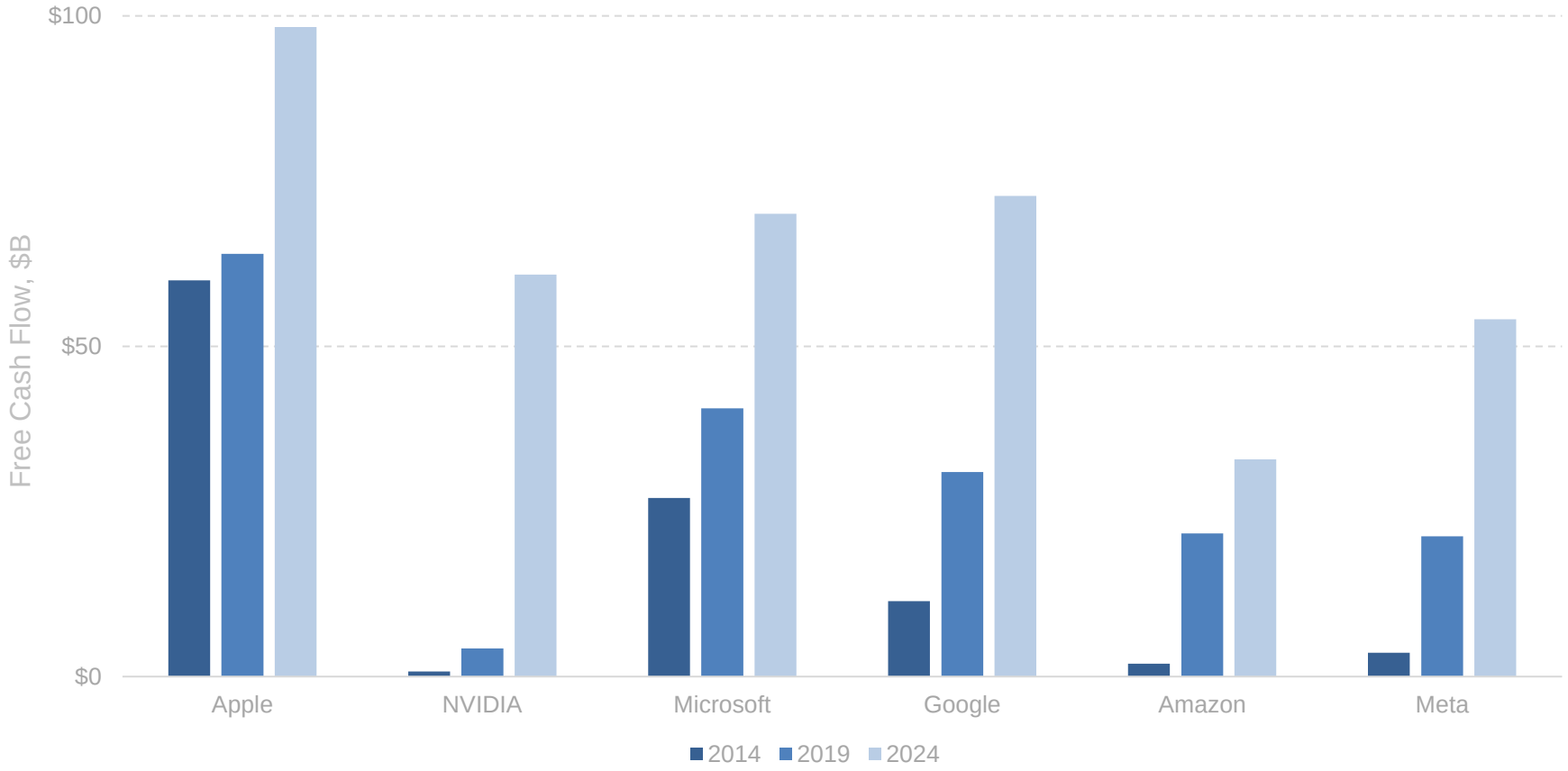
**Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. R&D expense shown for Amazon, not AWS, as figures are not broken out in company financials; revenue therefore shown on like-for-like basis. Source: Capital IQ (3/25)*

Tech Big Six (USA) =

Loaded With Cash to Spend on AI & CapEx

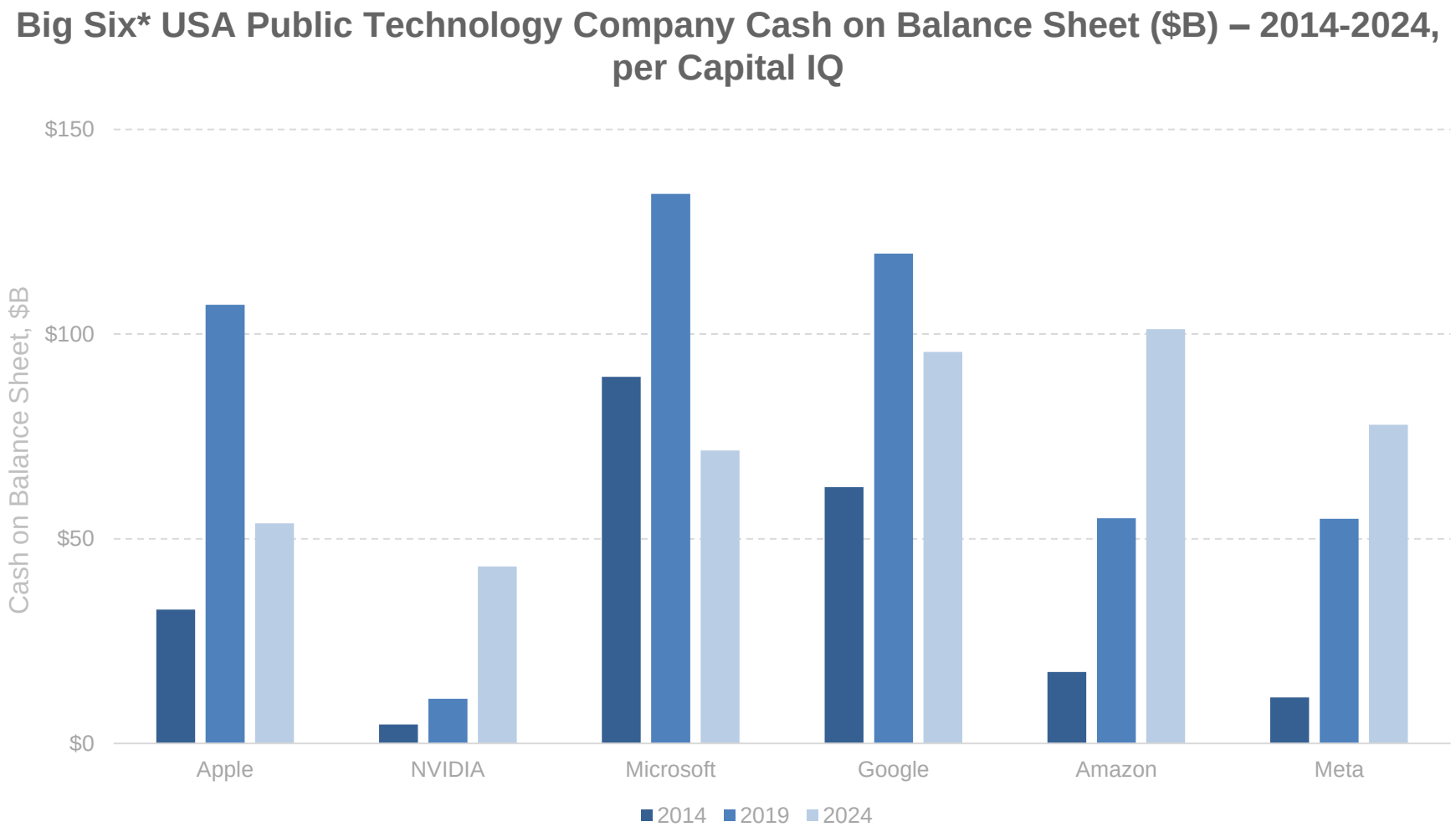
Big Six* Generating Loads of Cash = +263% Growth in Free Cash Flow Over Ten Years to $389B…

Big Six* Public Technology Companies – Free Cash Flow ($B) – 2014-2024, per Capital IQ



Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. FCF calculated as cash flow from operations less capex to standardize definitions, as only some companies subtract finance leases and Amazon adjusts FCF for gains on sale of equipment. FCF shown for Amazon, not AWS, as figures are not broken out in company financials. Source: Capital IQ (3/25)

…Big Six* Generating Loads of Cash = +103% Growth in Cash Over Ten Years to $443B



Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. Figure measures cash and other equivalents (e.g., short-term investments and marketable securities) on companies' balance sheets. Source: Capital IQ (3/25)

Tech CapEx Spend Driver =

Compute Spend to Train & Run AI Models

To understand the evolution of AI computing economics, it's constructive to look at where costs are concentrated – And where they're headed. The bulk of spending in AI large language model (LLM) development is still dominated by compute – specifically, the compute needed to train and run models. Training costs remain extraordinarily high and are rising fast, often exceeding $100 million per model today. As Dario Amodei, CEO of Anthropic, noted in mid-2024,
Right now, [AI model training costs] $100 million. There are models in training today that are more like a billion… I think that the training of…$10 billion models, yeah, could start sometime in 2025.

Around these core compute costs sit additional high-cost layers: research, data acquisition and hosting, and a mix of salaries, general overhead, and go-to-market operations. Even as the cost to train models climbs, a growing share of total AI spend is shifting toward inference – the cost of running models at scale in real-time. Inference happens constantly, across billions of prompts, queries, and decisions, whereas model training is episodic. As Amazon CEO Andy Jassy noted in his April 2025 letter to shareholders,
While model training still accounts for a large amount of the total AI spend, inference… will represent the overwhelming majority of future AI cost because customers train their models periodically but produce inferences constantly.
NVIDIA Co-Founder & CEO Jensen Huang noted the same in NVIDIA's FQ1:26 earnings call, saying
Inference is exploding. Reasoning AI agents require orders of magnitude more compute.

At scale, inference becomes a persistent cost center – one that grows in parallel with usage, despite declines in unit inference costs.

The broader dynamic is clear: lower per-unit costs are fueling higher overall spend. As inference becomes cheaper, AI gets used more. And as AI gets used more, total infrastructure and compute demand rises – dragging costs up again. The result is a flywheel of growth that puts pressure on cloud providers, chipmakers, and enterprise IT budgets alike.

The economics of AI are evolving quickly – but for now, they remain driven by heavy capital intensity, large-scale infrastructure, and a race to serve exponentially expanding usage.

Data Centers =

Key Beneficiary of AI CapEx Spend

For one lens into the economics of AI infrastructure,
it's useful to look at the pace and scale of data center construction.
The current wave of AI-driven demand has pushed data center spending to historic highs.
According to Dell'Oro Research, global IT company data center CapEx
reached $455 billion in 2024 and is accelerating.

Hyperscalers and AI-first companies alike are pouring billions into building out
compute-ready capacity – not just for storage, but for real-time inference and
model training workloads that require dense, high-power hardware.
As AI moves from experimental to essential, so too do data centers.
Per NVIDIA Co-Founder and CEO Jensen Huang, *These AI data centers…are, in fact, AI factories*.

That race is moving faster than many expected.
The most striking example may be xAI's Colossus facility in Memphis, Tennessee which went
from a gutted factory to a fully operational AI data center in just 122 days.
As noted on page 122, at 750,000 square feet – roughly the size of 418 average USA homes –
it was built in half the time it typically takes to construct a single American house.

Per NVIDIA Co-Founder & CEO Jensen Huang,
What they achieved is singular, never been done before…That is, like, superhuman…

…These kinds of timelines are no longer the exception. With prefabricated modules, streamlined permitting, and vertical integration across electrical, mechanical, and software systems, new data centers are going up at speeds that resemble consumer tech cycles more than real estate development.

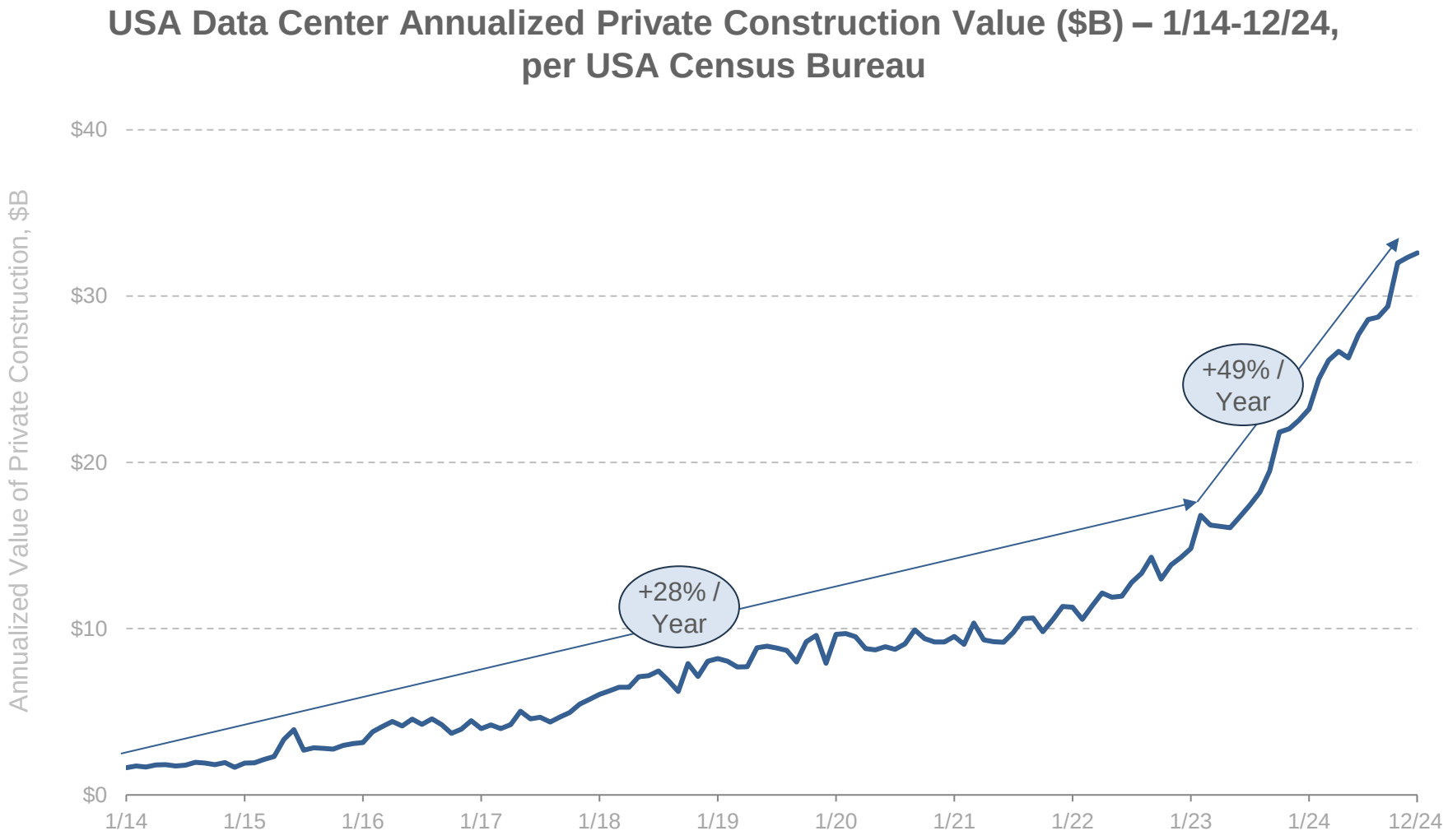
But beneath that velocity lies a capital model that's anything but simple. CapEx is driven by land, power provisioning, chips, and cooling infrastructure – especially as AI workloads push thermal and power limits far beyond traditional enterprise compute. OpEx, by contrast, is dominated by energy costs and systems maintenance, particularly for high-density training clusters that operate near constant load.

Revenue is driven by compute sales – whether in the form of AI APIs, enterprise platform fees, or internal productivity gains. But payback periods are often long, especially for vertically-integrated players building ahead of demand. For newer entrants, monetization may lag build-out by quarters or even years.

And then there's the supply chain. Power availability is becoming more of a gating factor. Transformers, substations, turbines, GPUs, cables – these aren't commodities that can be spun up overnight. In this context, data centers aren't just physical assets – they are strategic infrastructure nodes. They sit at the intersection of real estate, power, logistics, compute, and software monetization.

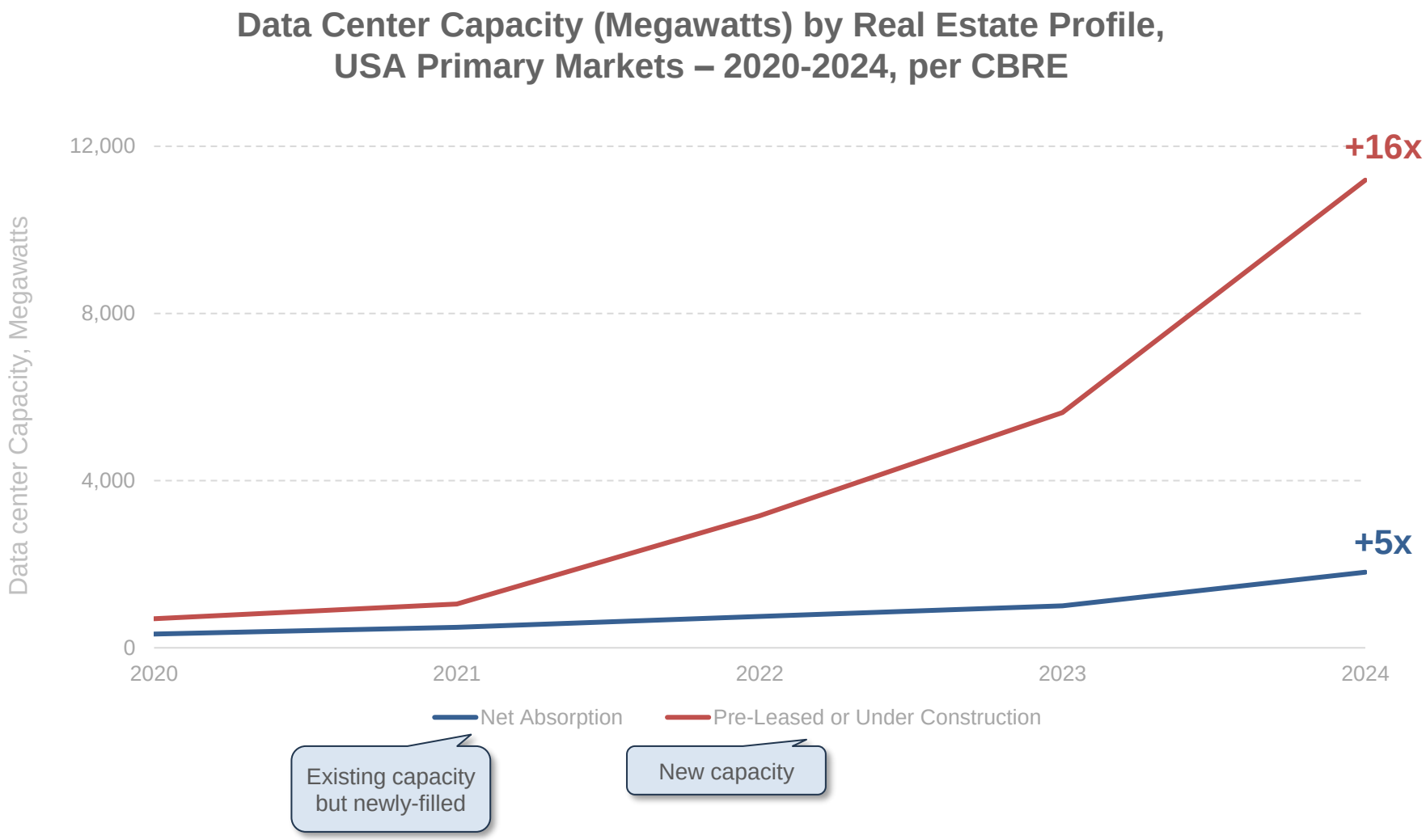
The companies that get this right may do more than run servers – they will shape the geography of AI economics for the next decade.

Data Center Buildout Construction Value, USA = +49% & Accelerated Annual Growth Over Two Years



Note: All data are seasonally adjusted. Data obtained via USA Census Bureau's Value of Construction Put in Place (VIP) Survey, which provides monthly estimates of the total dollar value of construction work done in USA. Data is annualized to avoid seasonal fluctuations. Source: USA Census Bureau

Data Center New Construction vs. Existing Capacity, USA = +16x in New vs. +5x in Existing Over Four Years



Note: Primary USA markets only (Northern Virginia, Atlanta, Chicago, Phoenix, Dallas-Ft. Worth, Hillsboro, Silicon Valley, New York Tri-State.)
Source: CBRE, 'North America Data Center Trends H2 2024' (2/25)

Data Center Build Time (xAI Colossus as Proxy) = 122 Days vs. 234 for a Home

122 Days =
A Fully-Operational Data Center – 2024…
750,000 Sq. Ft = Size of 418 USA Homes

750,000 Square Feet

We were told it would take 24 months to build. So we took the project into our own hands, questioned everything, removed whatever was unnecessary, and accomplished our goal in four months.

- xAI Website

122 Days =
One Half-Built House – 2024
(Average Build Time = 234 Days)

1,792 Square Feet

Note: Median USA home size shown as of January 2025, per FRED. Colossus was built in a former Electrolux factory in Memphis, TN, USA. Average build time shown for single-unit buildings. Measures time between start of onsite work & completion. Data reported in 2024 but measures build times for homes started in 2023.
Source: xAI, USA Census Bureau, Federal Reserve Bank of St. Louis, Wikimedia Commons

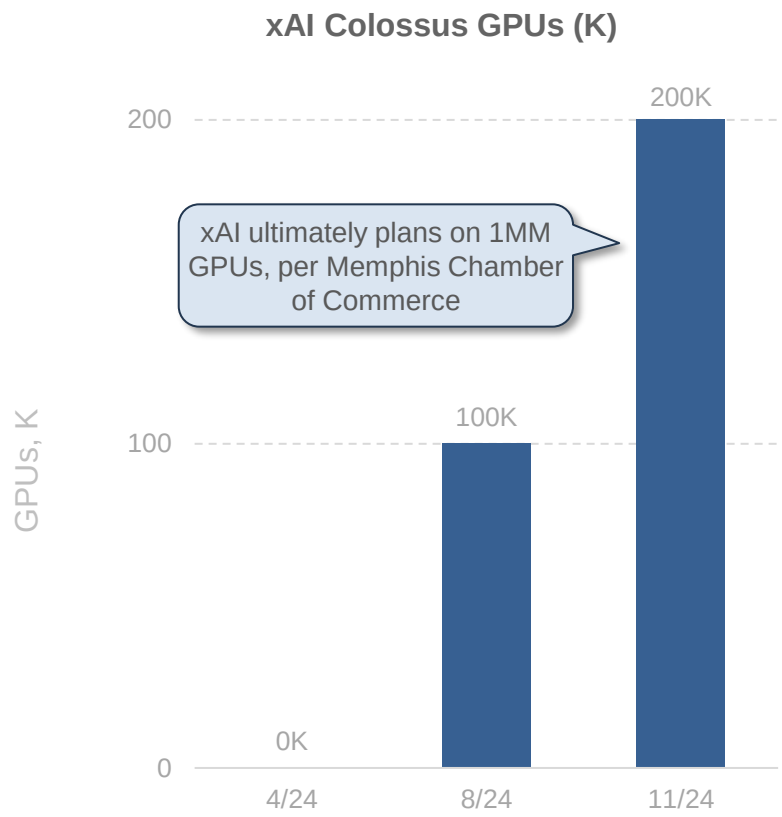
Data Center Compute (xAI Colossus as Proxy) = 0 to 200,000 GPUs in Seven Months

xAI Colossus GPUs – 4/24-11/24, per xAI

We're running the world's biggest supercomputer, Colossus. Built in 122 days – outpacing every estimate – it was the most powerful AI training system yet. Then we doubled it in 92 days to 200k GPUs. This is just the beginning…

…We doubled our compute at an unprecedented rate, with a roadmap to 1M GPUs. Progress in AI is driven by compute and no one has come close to building at this magnitude and speed.

- xAI Website, 5/25



Note: We assume 200,000 GPUs as of 11/30/24 per xAI's disclosure that 'we doubled [GPU count] in 92 days to 200K GPUs.' xAI Colossus ran its first job across 4 data halls on 8/30/24. We assume zero GPUs as of construction start date (122 days prior to assumed opening date of 8/30/24).
Source: xAI (5/25), Memphis Chamber of Commerce (12/24)

Data Centers =

Electricity Guzzlers

AI and energy observations / quotes (in italics) here and the two pages that follow are from 'World Energy Outlook Special Report – Energy and AI' (link) from IEA (International Energy Agency)* – 4/10/25

To understand where energy infrastructure is heading, it helps to examine the rising tension between AI capability and electrical supply. The growing scale and sophistication of artificial intelligence is demanding an extraordinary amount of computational horsepower, primarily from AI-focused data centers. These facilities – purpose-built to train and serve models – are starting to rival traditional heavy industry in their electricity consumption.

There is no AI without energy – specifically electricity (p. 3).
Data centers *accounted for around 1.5% of the world's electricity consumption in 2024 (p. 14).* Energy demand growth has been rapid:
Globally, data centre electricity consumption has grown by around 12% per year since 2017, more than four times faster than the rate of total electricity consumption (p. 14).
As power demand rises, so too does its concentration:
The United States accounted for…[45% of global data centre electricity consumption], followed by China (25%) and Europe (15%)…
nearly half of data centre capacity in the United States is in five regional clusters (p. 14).

The flipside is true as well: *Emerging and developing economies other than China account for 50% of the world's internet users but less than 10% of global data centre capacity (p. 18)…*

…AI's power demands are increasing – and its progress is increasingly bottlenecked not by data or algorithms, but by the grid and strains related to demand.

While AI presently places considerable demands on the energy sector, it is also already unlocking major energy efficiency and operational gains…
AI is already being deployed by energy companies to transform and optimize energy and mineral supply, electricity generation and transmission, and energy consumption (p. 16).

Current AI-driven demand is extremely high.
This is forecast to continue, especially as capital gushes into model providers that, in turn, spend on more compute. At some point, these model builders will need to turn a profit to be able to spend more.

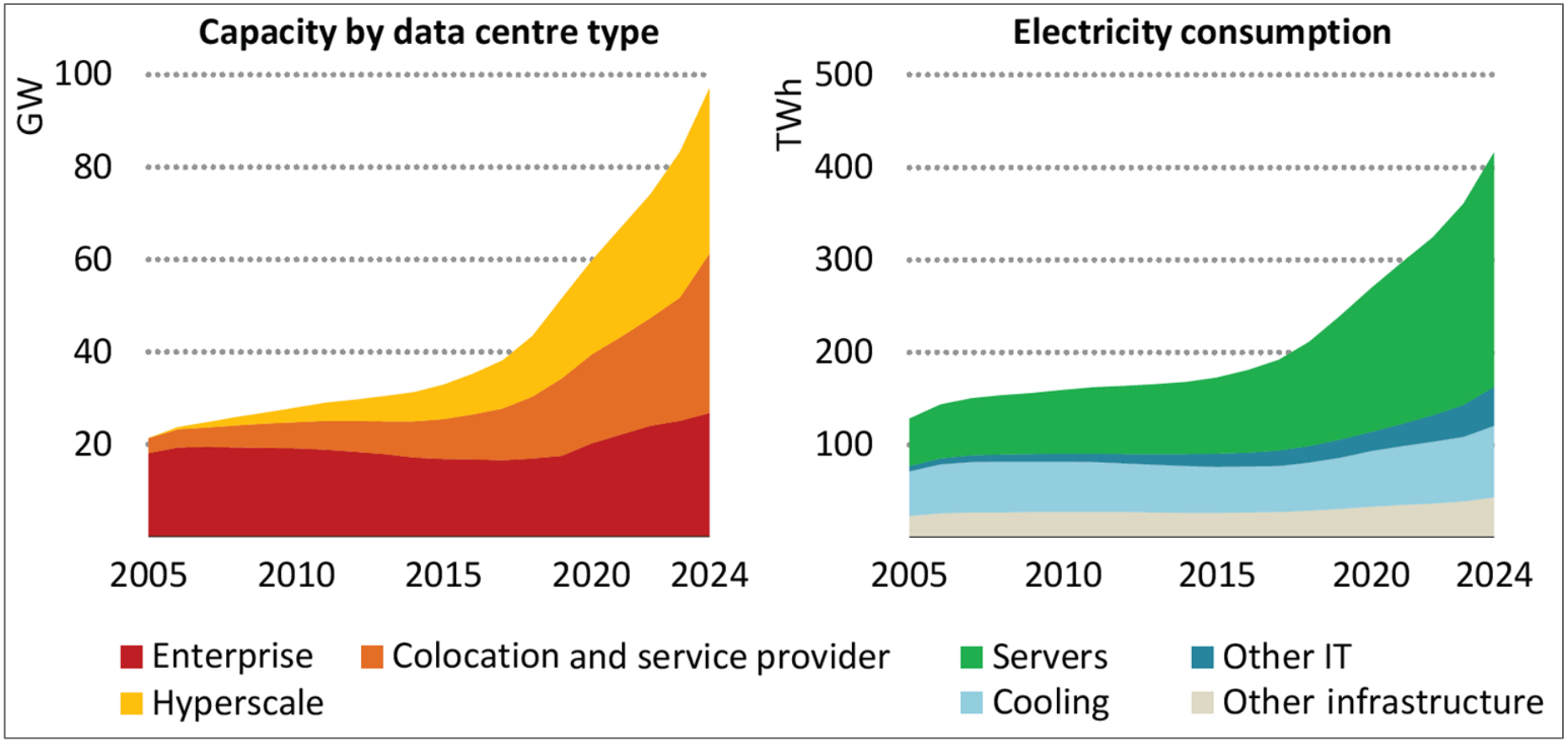
While demand – for both compute and energy – will inevitably continue to rise as consumer and business usage does the same, data centers will ultimately only serve those who pay their bills.

**IEA member countries include Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, S. Korea, Latvia, Lithuania, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Republic of Türkiye, United Kingdom, and United States. IEA Association countries include Argentina, Brazil, China, Egypt, India, Indonesia, Kenya, Morocco, Senegal, Singapore, S. Africa, Thailand, and Ukraine.*

All data shown, unless otherwise specified, is global. Italicized text is directly quoted from the report.

Data Center Electricity Consumption, Global = +3x Over Nineteen Years, per IEA

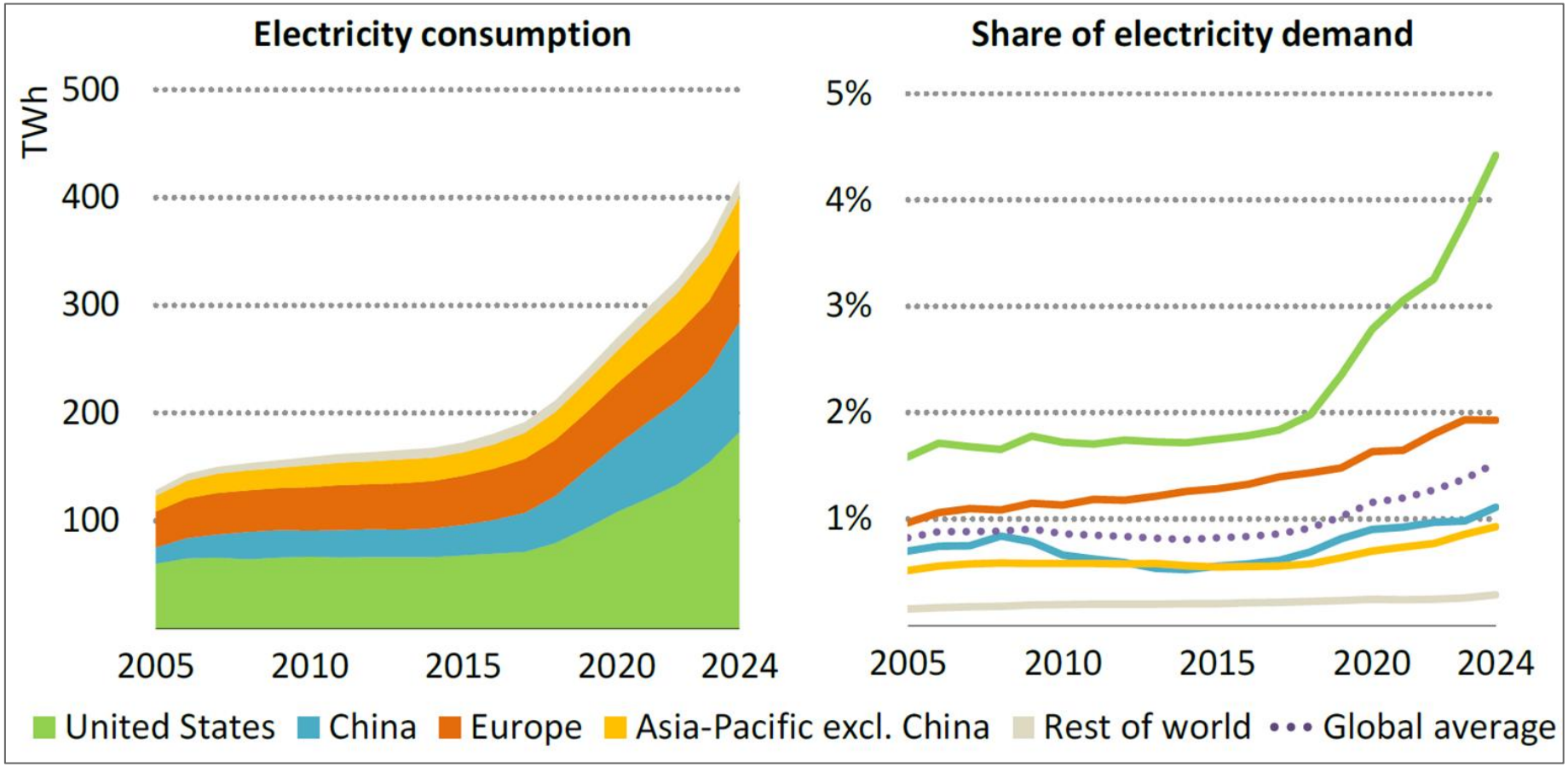
Data Center Energy Consumption by Data Center Type & Equipment, Global – 2005-2024, per IEA



Source: International Energy Agency (IEA), 'Energy and AI' (4/25)

Data Center Electricity Consumption by Region = USA Leads, per IEA

Data Center Electricity Consumption by Region – 2005-2024, per IEA



Source: International Energy Agency (IEA), 'Energy and AI' (4/25)

Outline

1. **Seem Like Change Happening Faster Than Ever?**
   Yes, It Is
2. **AI User + Usage + CapEx Growth** =
   Unprecedented
3. **AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** =
   Performance Converging + Developer Usage Rising
4. **AI Usage + Cost + Loss Growth** =
   Unprecedented
5. **AI Monetization Threats** =
   Rising Competition + Open-Source Momentum + China's Rise
6. **AI & Physical World Ramps** =
   Fast + Data-Driven
7. **Global Internet User Ramps Powered by AI from Get-Go** =
   Growth We Have Not Seen Likes of Before
8. **AI & Work Evolution** =
   Real + Rapid

To understand where AI model economics may be heading, one can look at the mounting tension between capabilities and costs.

Training the most powerful large language models (LLMs) has become one of the most expensive / capital-intensive efforts in human history. As the frontier of performance pushes toward ever-larger parameter counts and more complex architectures, model training costs are rising into the billions of dollars.

Ironically, this race to build the most capable general-purpose models may be accelerating commoditization and driving diminishing returns, as output quality converges across players and differentiation becomes harder to sustain. At the same time, the cost of applying/using these models – known as inference – is falling quickly. Hardware is improving – for example, NVIDIA's 2024 Blackwell GPU consumes 105,000x less energy per token than its 2014 Kepler GPU predecessor. Couple that with breakthroughs in models' algorithmic efficiency, and the cost of inference is plummeting.

Inference represents a new cost curve, and – unlike training costs – it's arcing down, not up. As inference becomes cheaper and more efficient, the competitive pressure amongst LLM providers increases – not on accuracy alone, but also on latency, uptime, and cost-per-token*. What used to cost dollars can now cost pennies. And what cost pennies may soon cost fractions of a cent.

The implications are still unfolding. For users (and developers), this shift is a gift: dramatically lower unit costs to access powerful AI. And as end-user costs decline, creation of new products and services is flourishing, and user and usage adoption is rising.

For model providers, however, this raises real questions about monetization and profits. Training is expensive, serving is getting cheap, and pricing power is slipping. The business model is in flux. And there are new questions about the one-size-fits-all LLM approach, with smaller, cheaper models trained for custom use cases** now emerging.

Will providers try to build horizontal platforms? Will they dive into specialized applications? Only time will tell. In the short term, it's hard to ignore that the economics of general-purpose LLMs look like commodity businesses with venture-scale burn.

**Cost-per-token = The expense incurred for processing or generating a single token (a word, sub-word, or character) during the operation of a language model. It is a key metric used to evaluate the computational efficiency and cost-effectiveness of deploying AI models, particularly in applications like natural language processing.*
***E.g., OpenEvidence*

AI Model Compute Costs High / Rising

+

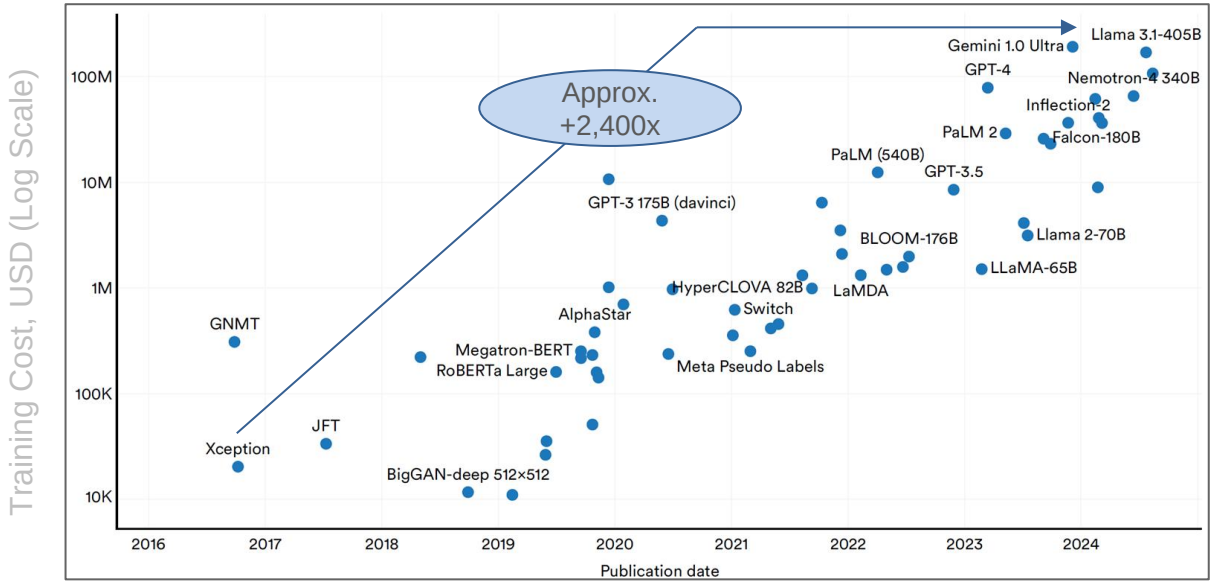
Inference Costs Per Token Falling

=

Performance Converging + Developer Usage Rising

AI Model Training Compute Costs = ~2,400x Growth Over Eight Years, per Epoch AI & Stanford

Estimated Training Cost of Frontier AI Models – 2016-2024, per Epoch AI & Stanford



Right now, [AI model training costs] $100 million. There are models in training today that are more like a billion. Right. I think if we go to $10 or $100 billion, and I think that will happen in 2025, 2026, maybe 2027…

…I think that the training of…$10 billion models, yeah, could start sometime in 2025.

- Anthropic Co-Founder & CEO Dario Amodei (6/24)

Note: Costs are estimates. Excludes most Chinese models due to lack of reliable cost data. Source: Epoch AI via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25); In Good Company podcast (6/24)

AI Model Compute Costs High / Rising

+

<u>Inference Costs Per Token Falling</u>

=

Performance Converging + Developer Usage Rising

To understand the trajectory of AI compute, it helps to revisit an idea from the early days of PC software. ‘Software is a gas…it expands to fill its container,’ said Nathan Myhrvold, then CTO of Microsoft in 1997. AI is proving no different. As models get better, usage increases – and as usage increases, so does demand for compute. We’re seeing it across every layer: more queries, more models, more tokens per task. The appetite for AI isn't slowing down. It’s growing into every available resource – just like software did in the age of desktop and cloud.

But infrastructure is not just standing still. In fact, it's advancing faster than almost any other layer in the stack, and at unprecedented rates. As noted on page 136, NVIDIA’s 2024 Blackwell GPU uses 105,000 times less energy to generate tokens than its 2014 Kepler predecessor. It’s a staggering leap, and it tells a deeper story – not just of cost reduction, but of architectural and materials innovation that is reshaping what’s possible at the hardware level.

These improvements in hardware efficiency are critical to offset the strain of increasing AI and internet usage on our grid. So far, though, they have not been enough. This trend aligns with Jevons Paradox, first proposed back in 1865* – that technological advancements that improve resource efficiency actually lead to increased overall usage of those resources. This is driving new focus on expanding energy production capacity – and new questions about the grid’s ability to manage.

Yet again, we see this as one of the perpetual ‘a-ha’s’ of technology: costs fall, performance rises, and usage grows, all in tandem. This trend is repeating itself with AI.

**British economist William Stanley Jevons first observed this phenomenon in 19th-century Britain, where he noticed that improvements in the efficiency of coal-powered steam engines were not reducing coal consumption but rather increasing it. In his book The Coal Question, he noted ‘It is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to diminished consumption. The very contrary is the truth.’*

AI Inference 'Currency' = Tokens

What are tokens and how to count them?

Updated over 3 months ago

What are tokens?

Tokens can be thought of as pieces of words. Before the API processes the request, the input is broken down into tokens. These tokens are not cut up exactly where the words start or end - tokens can include trailing spaces and even sub-words. Here are some helpful rules of thumb for understanding tokens in terms of lengths:

- 1 token ~= 4 chars in English
- 1 token ~= ¾ words
- 100 tokens ~= 75 words

Or

- 1-2 sentence ~= 30 tokens
- 1 paragraph ~= 100 tokens
- 1,500 words ~= 2048 tokens

To get additional context on how tokens stack up, consider this:

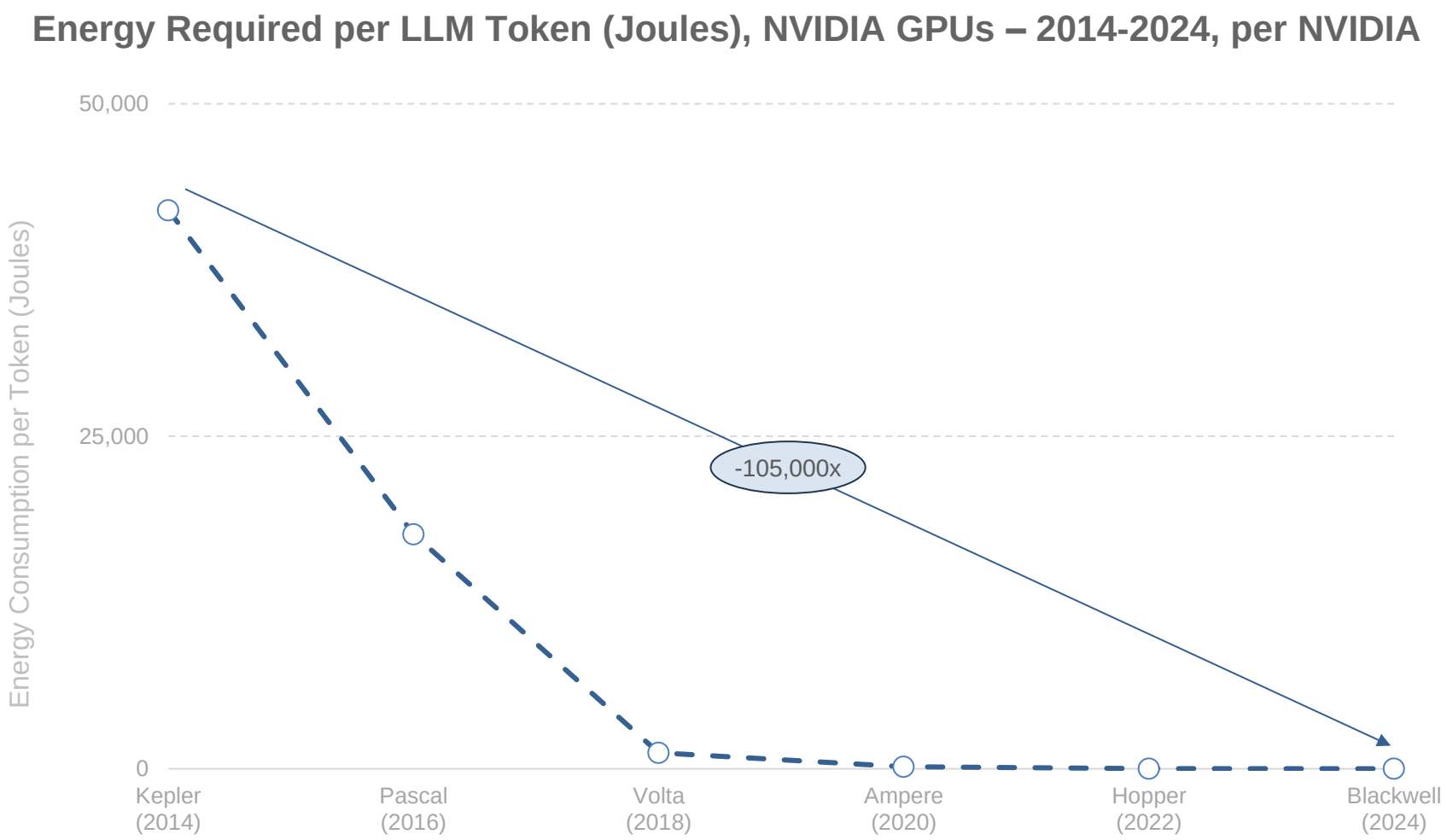
- Wayne Gretzky's quote "*You miss 100% of the shots you don't take*" contains 11 tokens.
- OpenAI's charter contains 476 tokens.
- The transcript of the US *Declaration of Independence* contains 1,695 tokens.

Additional context: 1MM tokens = ~750,000 words…roughly

- 3,500 pages of a standard book (12-point font, double-spaced)
- 5,000 ChatGPT responses*

**Assumes that the average ChatGPT interaction consumes 200 total tokens (input + output), or 150 words. Thus, 1MM tokens equates to roughly 5,000 ChatGPT responses.*
Source: OpenAI (1/25)

AI Inference Costs – NVIDIA GPUs = -105,000x Decline in Energy Required to Generate Token Over Ten Years

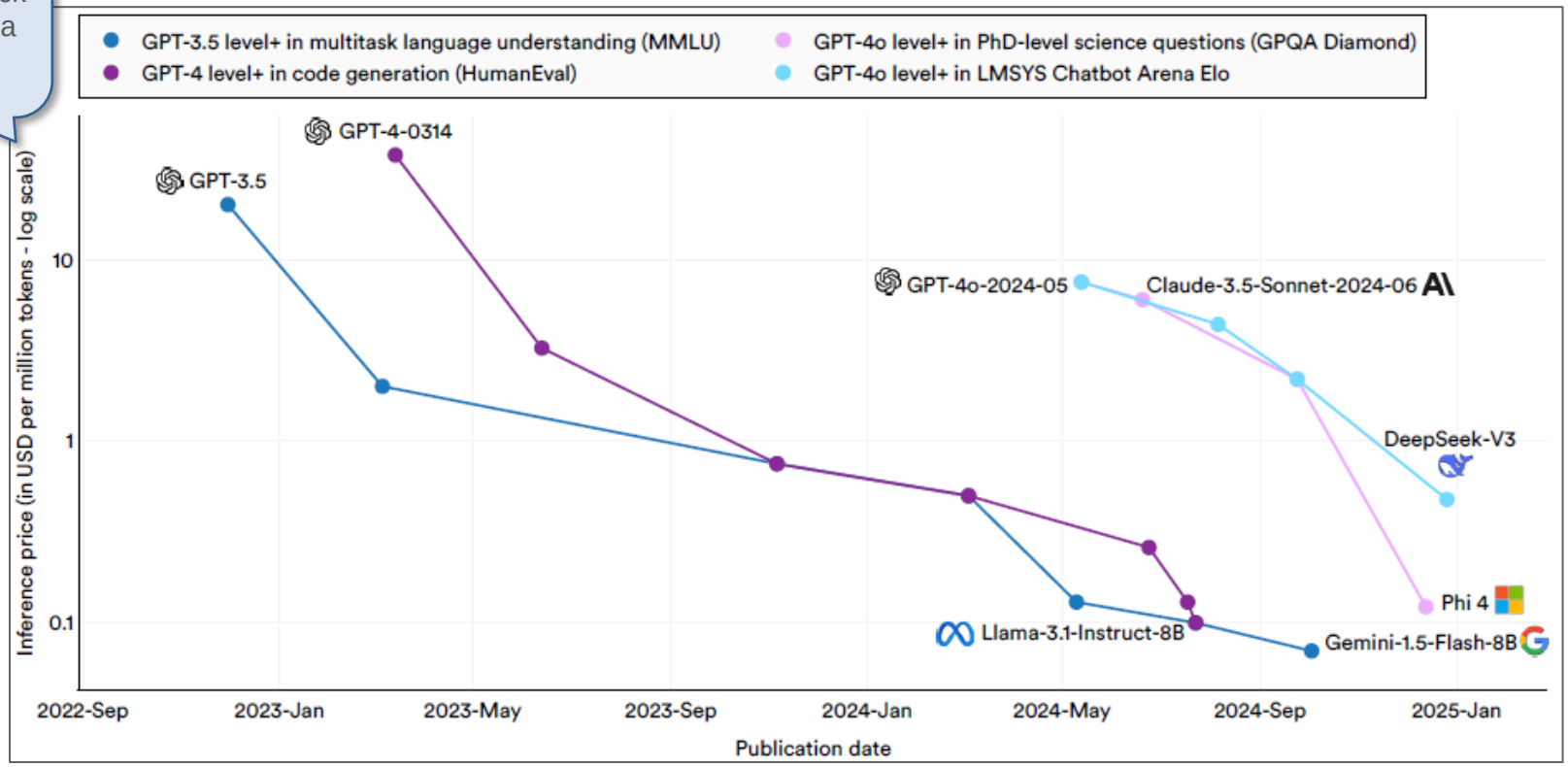


Note: Kepler released in 2012. NVIDIA materials mark performance threshold shown above for Kepler as of 2014. Source: NVIDIA Company Overview (2/25)

AI Inference Costs – Serving Models = 99.7% Lower Over Two Years, per Stanford HAI

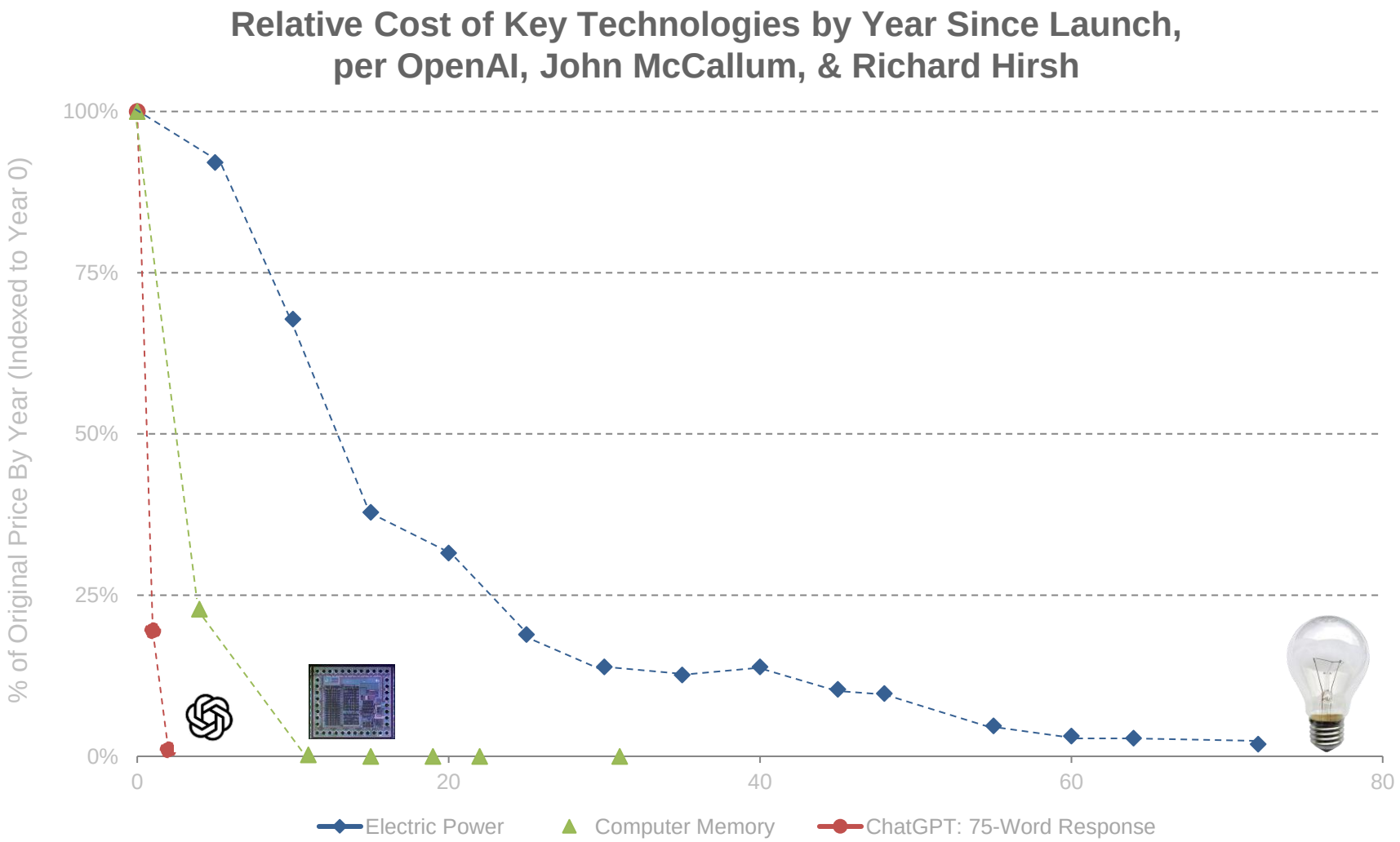
AI Inference Price for Customers (per 1 Million Tokens) – 11/22-12/24, per Stanford HAI

Note: Axis is logarithmic; every axis tick represents a 10x price change



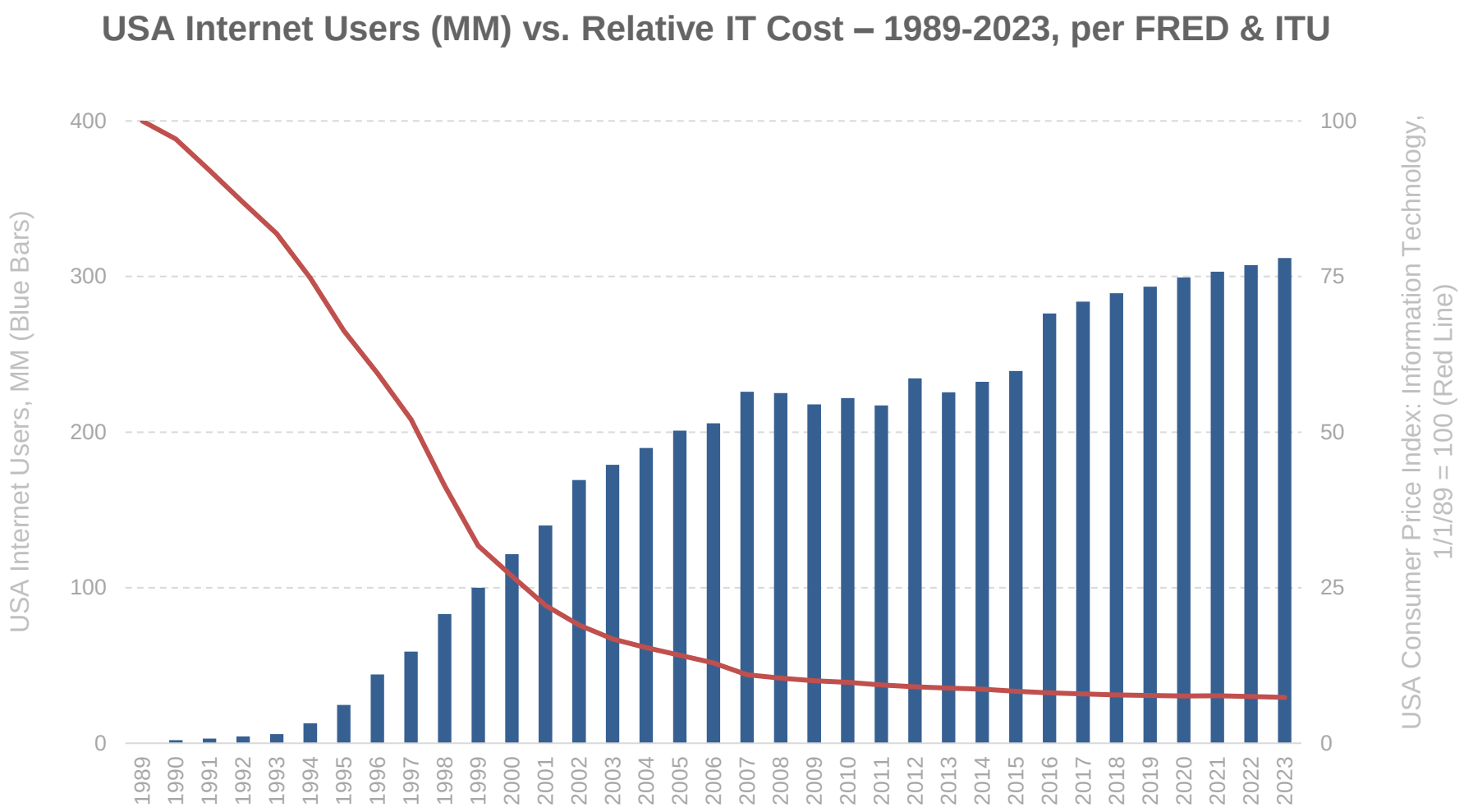
Source: Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

AI Cost Efficiency Gains = Happening Faster vs. Prior Technologies



Note: Price change in consumer goods and services in the United States is measured as the percentage change since 1997. Data is based on the consumer price index (CPI) for national average urban consumer prices. Per OpenAI, 100 AI 'tokens' generates approximately 75 words in a large language model response; data shown indexes to this number of tokens. 'Year 0' is not necessarily the year that the technology was introduced, but rather the first year of available data.
Source: Electricity Costs – Technology and Transformation in the American Electric Utility Industry, Richard Hirsh (1989); Computer Memory Storage Costs – John C. McCallum, with data aggregated from 72 primary sources and historical company sales documents; OpenAI, Wikimedia Commons

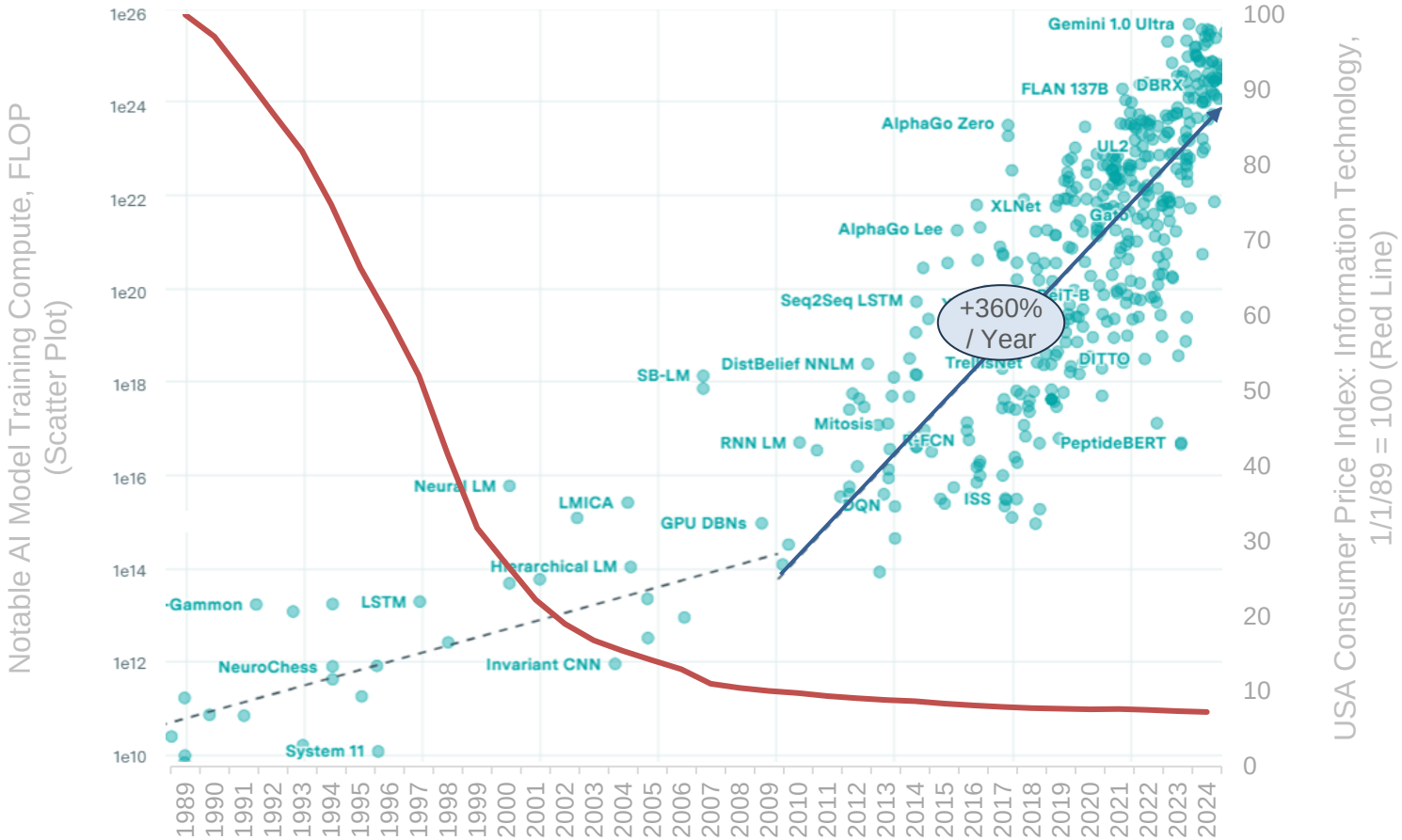
Tech's Perpetual A-Ha = Declining Costs + Improving Performance → Rising Adoption…



Note: FRED data shows 'Consumer Price Index for All Urban Consumers: Information Technology, Hardware and Services in U.S. City Average.' Source: USA Federal Reserve Bank of St. Louis (FRED), International Telecommunications Union (via World Bank) (4/25)

…Tech's Perpetual A-Ha = Prices Fall + Performance Rises

AI Model Training Compute (FLOP) vs. Relative IT Cost – 1989-2024, per Epoch AI & FRED



Note: A petaFLOP/s-day represents the total computational work performed by a system operating at 1 petaFLOP/s (10^{15} floating-point operations per second) for 24 hours, equivalent to approximately 8.64×10^{19} operations. This metric is commonly used to quantify the compute required for large-scale tasks like training machine learning models. FRED data shows 'Consumer Price Index for All Urban Consumers: Information Technology, Hardware and Services in U.S. City Average.' Note that, while training compute is not a direct measurement of model performance, it is typically closely correlated with performance. Source: USA Federal Reserve Bank of St. Louis (FRED); Epoch AI (5/25)

AI Model Compute Costs High / Rising

+

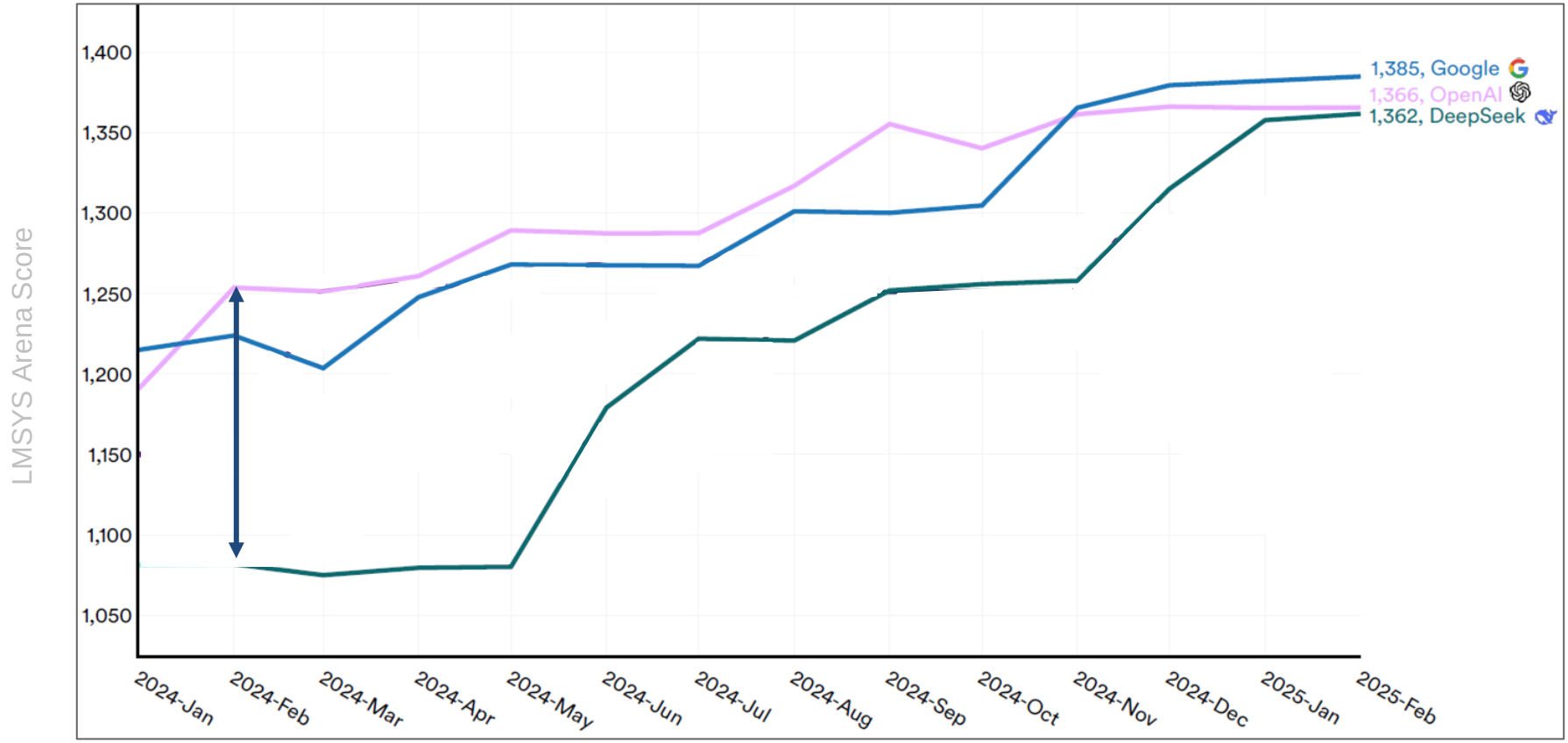
Inference Costs Per Token Falling

=

Performance Converging *+ Developer Usage Rising*

AI Model Performance = Converging Rapidly, per Stanford HAI

Performance of Top AI Models on LMSYS Chatbot Arena – 1/24-2/25, per Stanford HAI



Note: The LMSYS Chatbot Arena is a public website where people compare two AI chatbots by asking them the same question and voting on which answer is better. The results help rank how well different language models perform based on human judgment. Only the highest-scoring model in any given month is shown in this comparison.
Source: Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

AI Model Compute Costs High / Rising

+

Inference Costs Per Token Falling

=

Performance Converging + ***Developer Usage Rising***

To understand the surge in AI developer activity, it's instructive to look at the extraordinary drop in inference costs and the growing accessibility of capable models.

Between 2022 and 2024, the cost-per-token to run language models fell by an estimated 99.7% – a decline driven by massive improvements in both hardware and algorithmic efficiency. What was once prohibitively expensive for all but the largest companies is now within reach of solo developers, independent app builders, researchers on a laptop, and mom-and-pop shop employees.

The cost collapse has made experimentation cheap, iteration fast, and productization feasible for virtually anyone with an idea.

At the same time, performance convergence is shifting the calculus on model selection. The gap between the top-performing frontier models and smaller, more efficient alternatives is narrowing.

For many use cases – summarization, classification, extraction, or routing – the difference in real-world performance is negligible. Developers are discovering they no longer need to pay a premium for a top-tier model to get reliable outputs. Instead, they can run cheaper models locally or via lower-cost API providers and achieve functionally similar results, especially when fine-tuned on task-specific data.

This shift is weakening the pricing leverage of model incumbents and leveling the playing field for AI development…

…At the platform level, a proliferation of foundation models has created a new kind of flexibility. Developers can now choose between dozens of models – OpenAI's ChatGPT, Meta's Llama, Mistral's Mixtral, Anthropic's Claude, Google's Gemini, Microsoft's Phi, and others – each of which excels in different domains. Some are optimized for reasoning, others for speed or code generation. The result is a move away from vendor lock-in.

Instead of consolidating under a single provider who can gate access or raise prices, developers are distributing their efforts across multiple ecosystems. This plurality of options is empowering a new wave of builders to choose the best-fit model for their technical or financial needs.

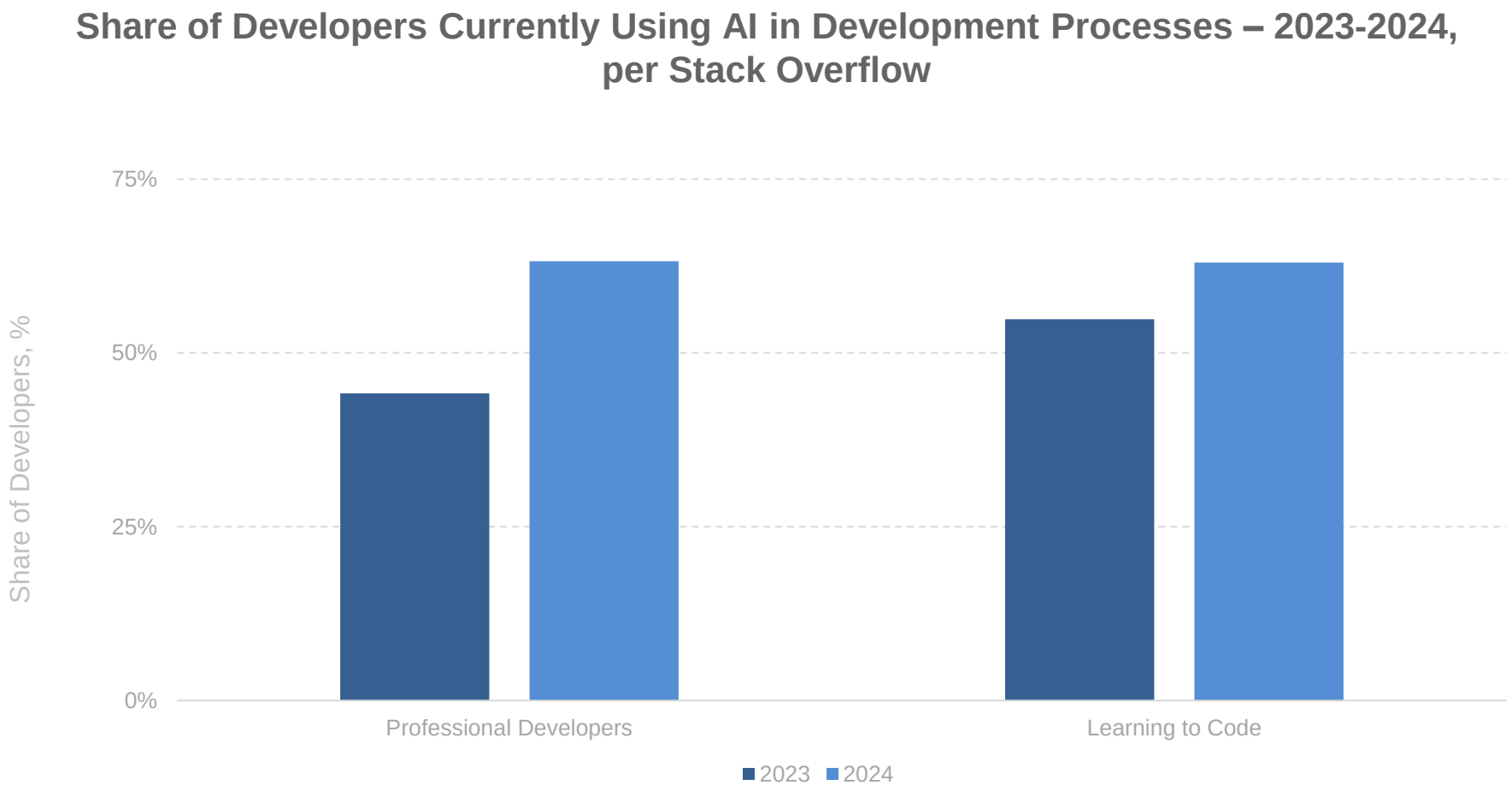
What's emerging is a flywheel of developer-led infrastructure growth. As more developers build AI-native apps, they also create tools, wrappers and libraries that make it easier for others to follow. New front-end frameworks, embedding pipelines, model routers, vector databases, and serving layers are multiplying at an accelerating rate.

Each wave of developer activity reduces the friction for the next, compressing the time from idea to prototype and from prototype to product. In the process, the barrier to building with AI is collapsing – not just in cost, but in complexity. This is no longer just a platform shift. It's an explosion of creativity.

Technology history has shown – as memorialized by then-Microsoft President Steve Ballmer's repeat *Developers! Developers! Developers…* at a 2000 Microsoft Developers Conference (link) – the platform that gets the most consistent developer user and usage momentum – and can scale and steadily improve – wins.

The AI Developer Next Door

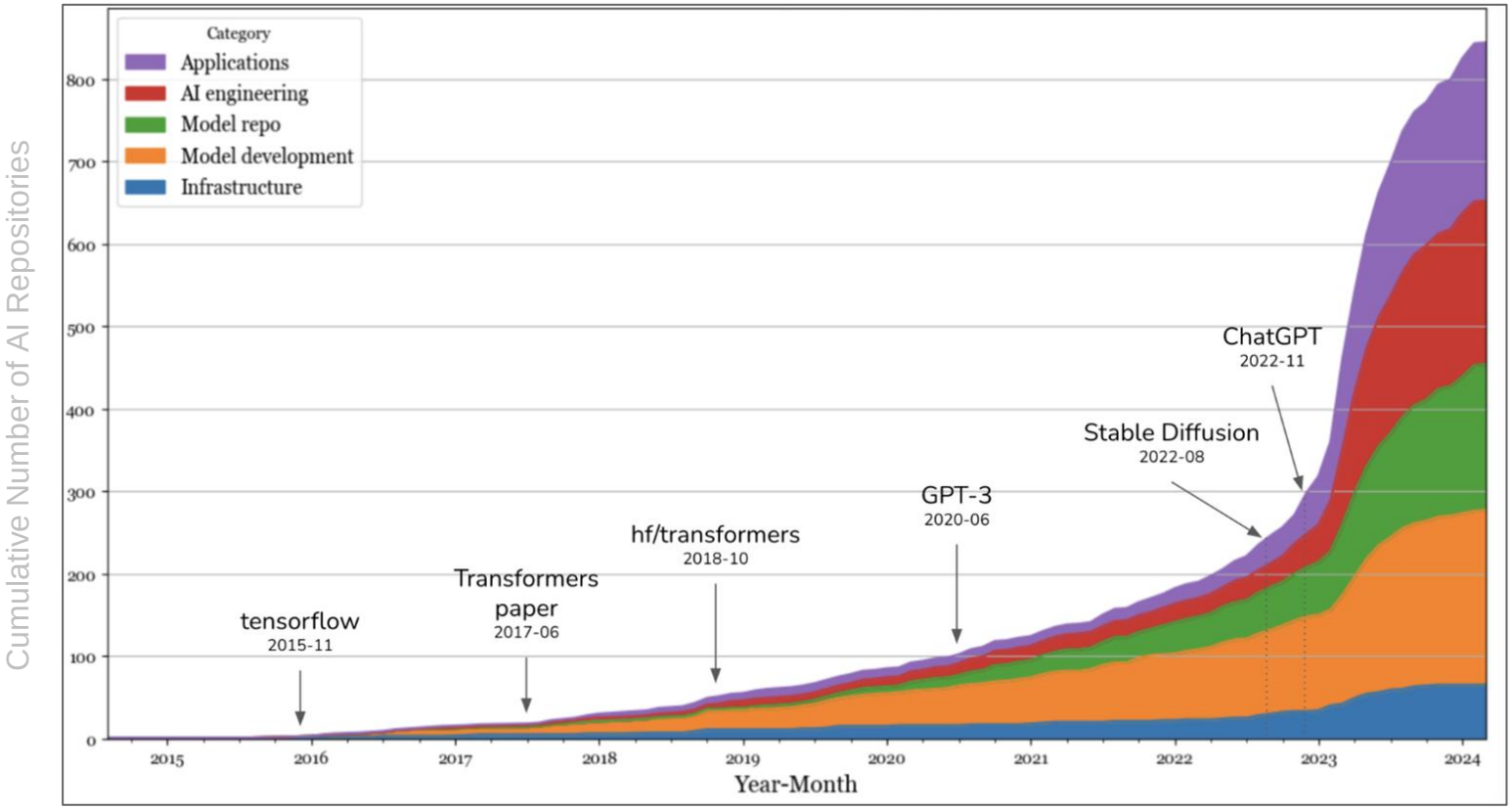
AI Tool Adoption by Developers = 63% vs. 44% Y/Y



Note: 2023 N=89,184; 2024 N=65,437. Respondents are global. Source: Stack Overflow Developer Surveys (5/23 & 5/24-6/24)

AI Developer Repositories – GitHub = ~175% Increase Over Sixteen Months

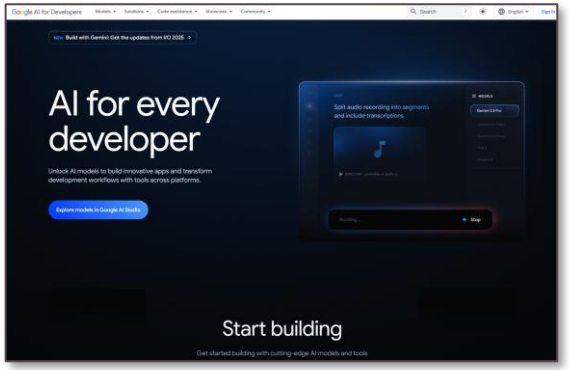
Number of AI Developer Repositories* on GitHub – 11/22-3/24, per Chip Hyuen



A repository is an online storage space where developers share and manage code, models, data, and documentation related to artificial intelligence projects. These enable collaboration, reuse, and distribution of AI tools and assets. Analysis shown includes GitHub repositories with 500+ stars. Infrastructure = tools for model serving, compute management, vector search & databases. Model development = frameworks for modeling & training, inference optimization, dataset engineering, & model evaluation. Application development = custom AI-powered applications (varied use cases). Source: Chip Hyuen via GitHub (3/24)

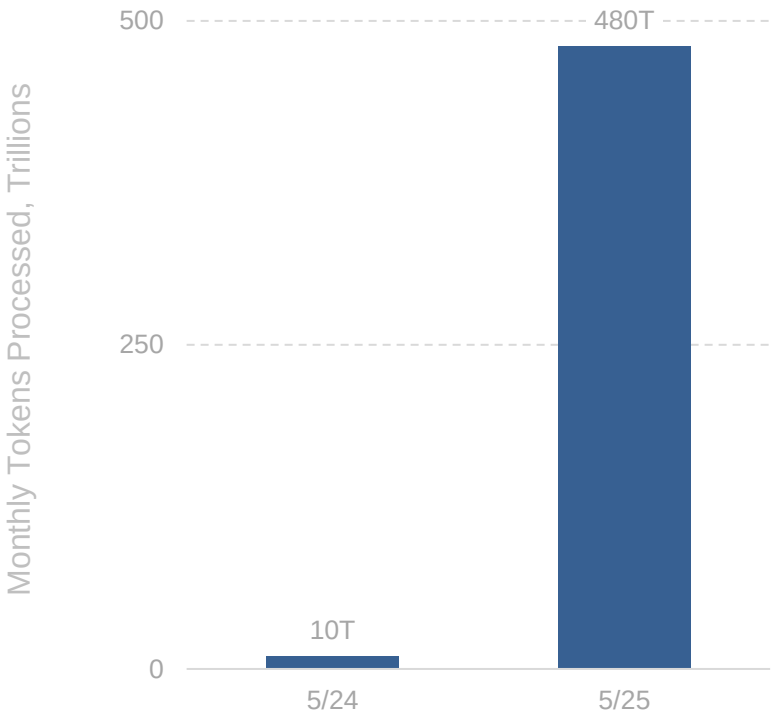
AI Developer Ecosystem – Google = +50x Monthly Tokens Processed Y/Y

Google Monthly Tokens Processed (T) – 5/24-5/25, per Google



This time last year, we were processing 9.7 trillion tokens a month across our products and APIs. Now, we're processing over 480 trillion – that's 50 times more.

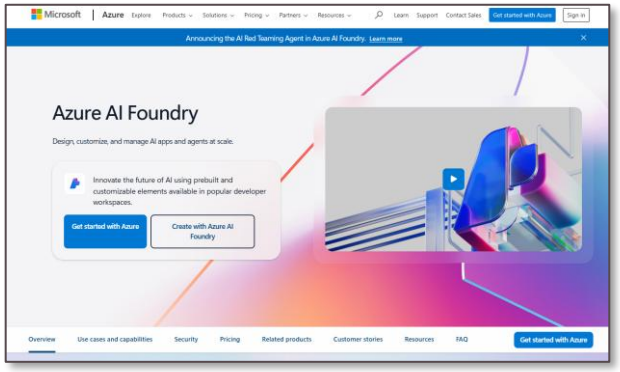
- Google I/O 2025 Press Release, 5/25



Note: Token usage shown across Google products & APIs. Per Google in 5/25, 'This time last year, we were processing 9.7 trillion tokens a month across our products and APIs. Now, we're processing over 480 trillion — that's 50 times more…Over 7 million developers are building with Gemini, five times more than this time last year.' Source: Google, 'Google I/O 2025: From research to reality' (5/25)

AI Developer Ecosystem – Microsoft Azure AI Foundry = +5x Quarterly Tokens Processed Y/Y

Microsoft Azure AI Foundry Quarterly Tokens Processed (T) – Q1:24-Q1:25, per Microsoft

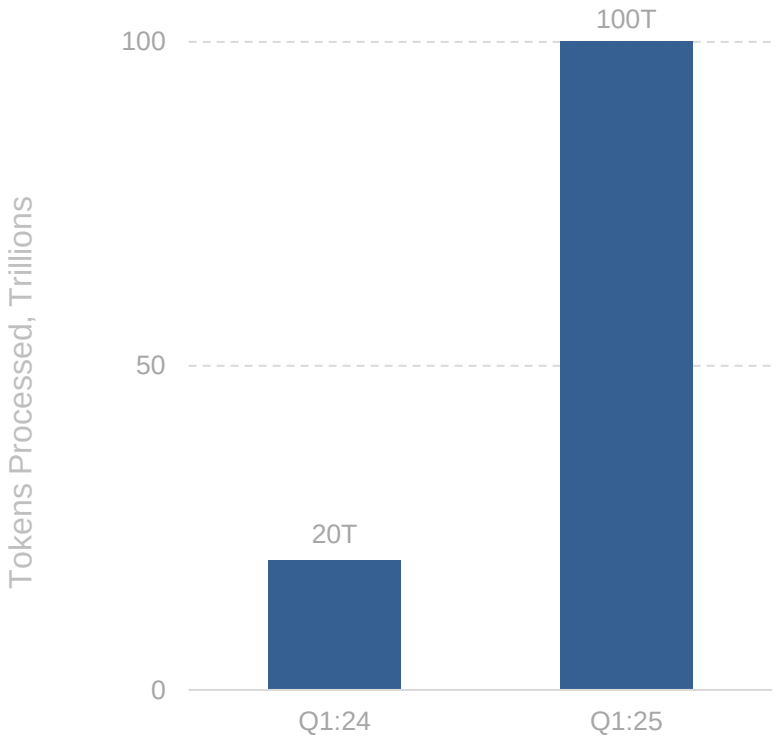


[Microsoft Azure AI] Foundry is the agent and AI app factory.

It is now used by developers at over 70,000 enterprises and digital natives – from Atomicwork, to Epic, Fujitsu, and Gainsight, to H&R Block and LG Electronics – to design, customize, and manage their AI apps and agents.

We processed over 100 trillion tokens this quarter, up 5x year-over-year – including a record 50 trillion tokens last month alone.

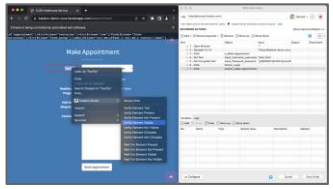
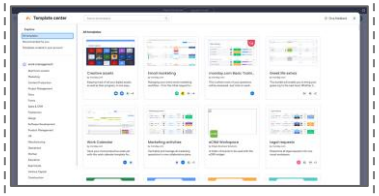
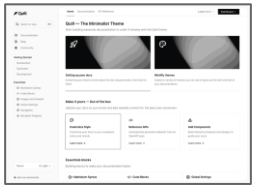
- Microsoft FQ3:25 Earnings Call, 4/25

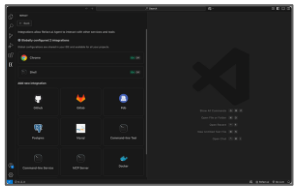
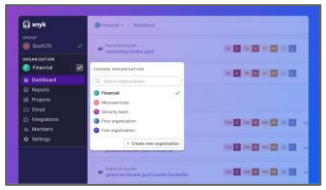
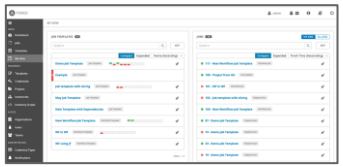
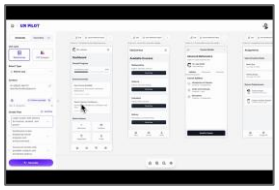
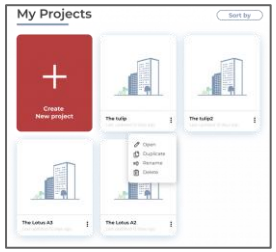


Note: Source: Microsoft FQ3:25 earnings call (4/25)

AI Developer Use Cases = Broad & Varied

AI Developer Use Cases – 2024, per IBM

Code Generation	Bug Detection & Fixing	Testing Automation	Project / Workflow Management	Documentation
				

Refactoring & Optimization	Security Enhancement	DevOps & CI / CD Pipelines	User Experience Design	Architecture Design
				

Note: CI / CD pipelines are continuous integration / continuous deployment pipelines.
Source: IBM, 'AI in Software Development' (2024); Anthropic; Katalon; AccelQ; Monday; Quill; Mintlify; Snyk; Ansible; UX Pilot; Ark Design AI

AI Model Compute Costs High / Rising

+

Inference Costs Per Token Falling

=

Performance Converging + Developer Usage Rising

…(Likely) Long Way to Profitability

Outline

1. **Seem Like Change Happening Faster Than Ever?**
   Yes, It Is
2. **AI User + Usage + CapEx Growth** =
   Unprecedented
3. **AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** =
   Performance Converging + Developer Usage Rising
4. **AI Usage + Cost + Loss Growth** =
   Unprecedented
5. **AI Monetization Threats** =
   Rising Competition + Open-Source Momentum + China's Rise
6. **AI & Physical World Ramps** =
   Fast + Data-Driven
7. **Global Internet User Ramps Powered by AI from Get-Go** =
   Growth We Have Not Seen Likes of Before
8. **AI & Work Evolution** =
   Real + Rapid

It's different this time, *we'll make it up on volume*, and *we'll figure out how to monetize our users in the future* are typically three of the biggest danger statements in business.

That said, in technology investing every once in awhile they can be gold – Amazon, Alphabet (Google), Meta (Facebook), Tesla, Tencent, Alibaba, Palantir…

In AI, it may indeed be different this time, and the leader(s) will make it up on volume and be able to monetize users in the future. Though now, 'different this time' also means that competition is unprecedented…

We have never seen so many founder-driven / assisted (ex. Apple) companies* with market capitalizations in excess of $1 trillion – most with gross margins of +50% plus free cash flow – attacking the same opportunity at the same time in a relatively transparent world, adding in high stakes competition between global powers – China and the United States.

Ernest Hemingway's phrase *gradually, then suddenly* from 'The Sun Also Rises' applies to technology tipping points.

The tipping point for personal computers was the introduction of Apple's Macintosh (1984) and Microsoft's Windows 3.0 (1990). With the Internet it was Netscape's IPO (1995). With the Mobile Internet it was Apple's iPhone App Store launch (2008). With Cloud Computing it was the launch of AWS (Amazon Web Services) foundational products (2006-2009). With AI it was the launch of NVIDIA's A100 GPU chip (2020) and OpenAI's public version of ChatGPT (2022). In effect, the global competition for AI kicked in with the launch of China's DeepSeek (1/25) and Jack Ma's attendance at Chinese President Xi Jinping's symposium of Chinese business leaders (2/25).

The money to fund AI's growth (and losses) comes from big companies with big free cash flow and big balance sheets, in addition to wealthy and ambitious capital providers from around the world.

No doubt, this dynamic combination of competition / capital / entrepreneurship will rapidly advance AI, a riddle is determining which business models will be the last ones standing.

**Companies include NVIDIA, Microsoft, Amazon, Alphabet (Google), Meta (Facebook) & Tesla*

Technology Disruption Pattern Recognition = Hundreds of Years of Consistent Signals

Technology disruption has a long-repeating rhythm: early euphoria, break-neck capital formation, bruising competition, and – eventually – clear-cut winners and losers.

Alasdair Nairn's 'Engines That Move Markets' (link here) distills two centuries of such cycles, and his observations are prescient for today's AI boom.
Highlights of his observations follow…

There were several years of strong share-price growth when the railways were supplanting canals. The bubble of the 1840s deflated under the weight of overheated expectations and changing economic conditions…

…Any technological advance which requires huge capital expenditure always runs a real risk of disappointing returns in the early years, even if it is ultimately successful...

…Any technology that necessitates heavy capital expenditure and requires returns to be earned over an extended period is always going to be a high-risk undertaking – unless, that is, there is some form of protection against competition...

…The winners of these competitive struggles are not always those who have the best technology, but those who can most clearly see the way that an industry or market is likely to develop…

…One of the clearest lessons of corporate and investment history is that without some barrier to entry, first-mover advantage can be swiftly lost…

…A theme that recurs throughout this research is that while identifying the winners from any new technology is often perilous and difficult, it is almost invariably simpler to identify who the 'losers' are going to be.

AI-Related Monetization =

Very Robust Ramps

To understand the evolution of AI hardware strategy, we'll look at how control over chip design is shifting from traditional vendors to the platforms that rely on them.
For years, NVIDIA has been at the center of the AI hardware stack.
Its GPUs (graphics processing units) became the default engine for training and inference, prized for their ability to handle highly parallel computations at scale. Its proprietary technology – and unparalleled scale – has led to industry leadership.

This reliance – combined with outsized sudden demand – has created constraints.
Despite NVIDIA's rapid – and impressive – scale-up, demand for NVIDIA GPUs has outpaced supply amid industry fervor for accelerated computing. Hyperscalers and cloud providers are moving to improve their supply chains to manage long lead times.

That shift is accelerating the rise of custom silicon – especially ASICs, or application-specific integrated circuits. Unlike GPUs, which are designed to support a wide range of workloads, ASICs are purpose-built to handle specific computational tasks with maximum efficiency. In AI, that means optimized silicon for matrix multiplication, token generation, and inference acceleration.

Google's TPU (Tensor Processing Unit) and Amazon's Trainium chips are now core components of their AI stacks. Amazon claims its Trainium2 chips offer 30-40% better price-performance than standard GPU instances, unlocking more affordable inference at scale. These aren't side projects – they're foundational bets on performance, economics, and architectural control…

…Custom chips also reflect a broader effort to manage the economics of AI infrastructure. As Amazon CEO Andy Jassy noted in early 2025, *AI does not have to be as expensive as it is today, and it won't be in the future.* Custom silicon is one lever to control these expenses.

At the same time, a new ecosystem of infrastructure specialists is emerging to meet this demand.

CoreWeave has become one of the fastest-scaling cloud GPU providers, repurposing gaming and Crypto hardware supply chains to serve enterprise AI customers.

Oracle, long seen as a legacy IT vendor, has repositioned itself as a GPU-rich cloud platform with AI-specific offerings. Astera Labs, a lesser-known but critical player, builds high-speed interconnects that move data between GPUs and memory systems with minimal latency – an increasingly important performance constraint.

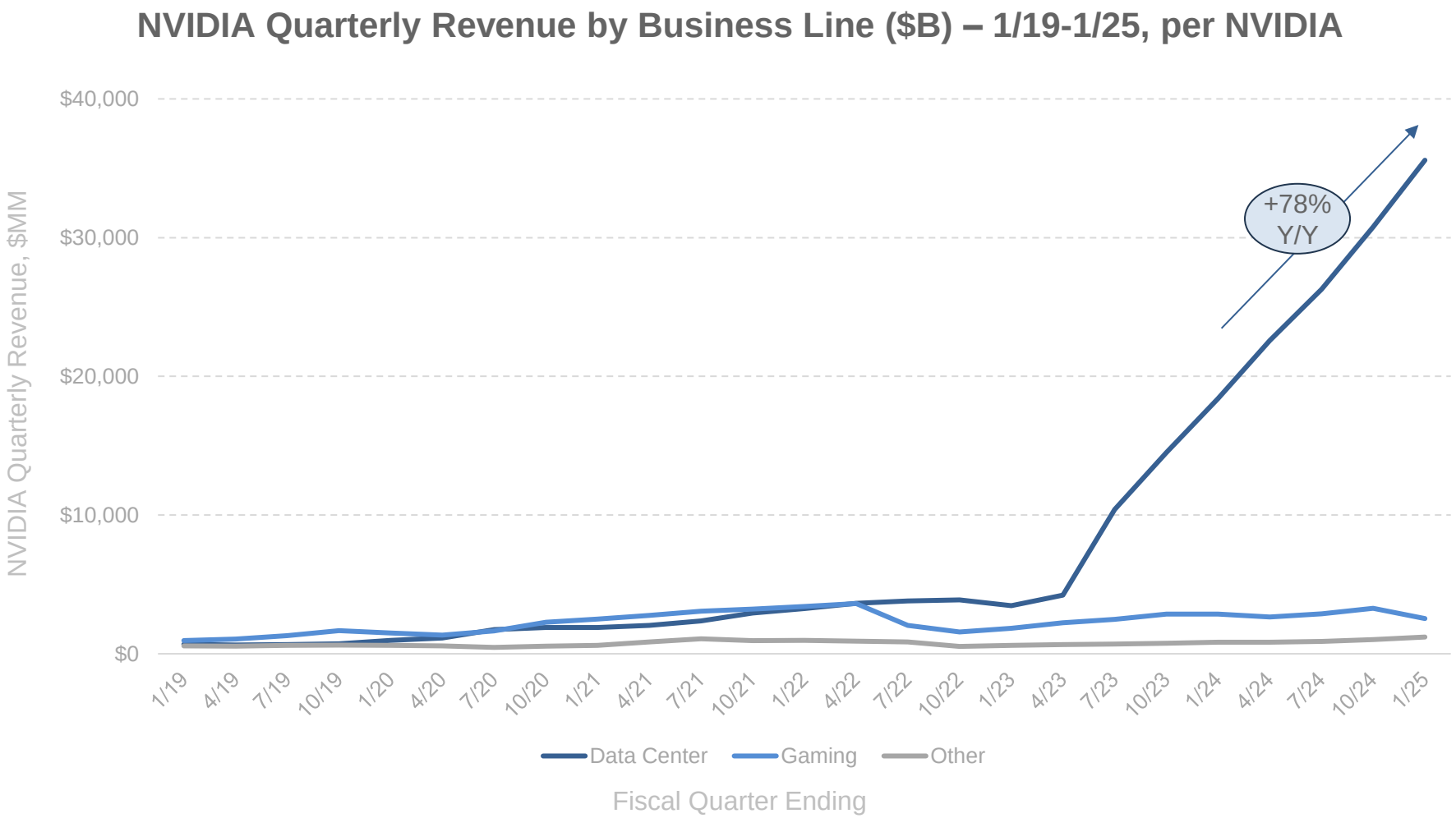
These firms aren't building foundation models, but they're building what foundation models depend on.

As compute demand compounds, they're becoming essential infrastructure in a market where speed, availability, and efficiency are important differentiators.

AI Monetization =

Chips

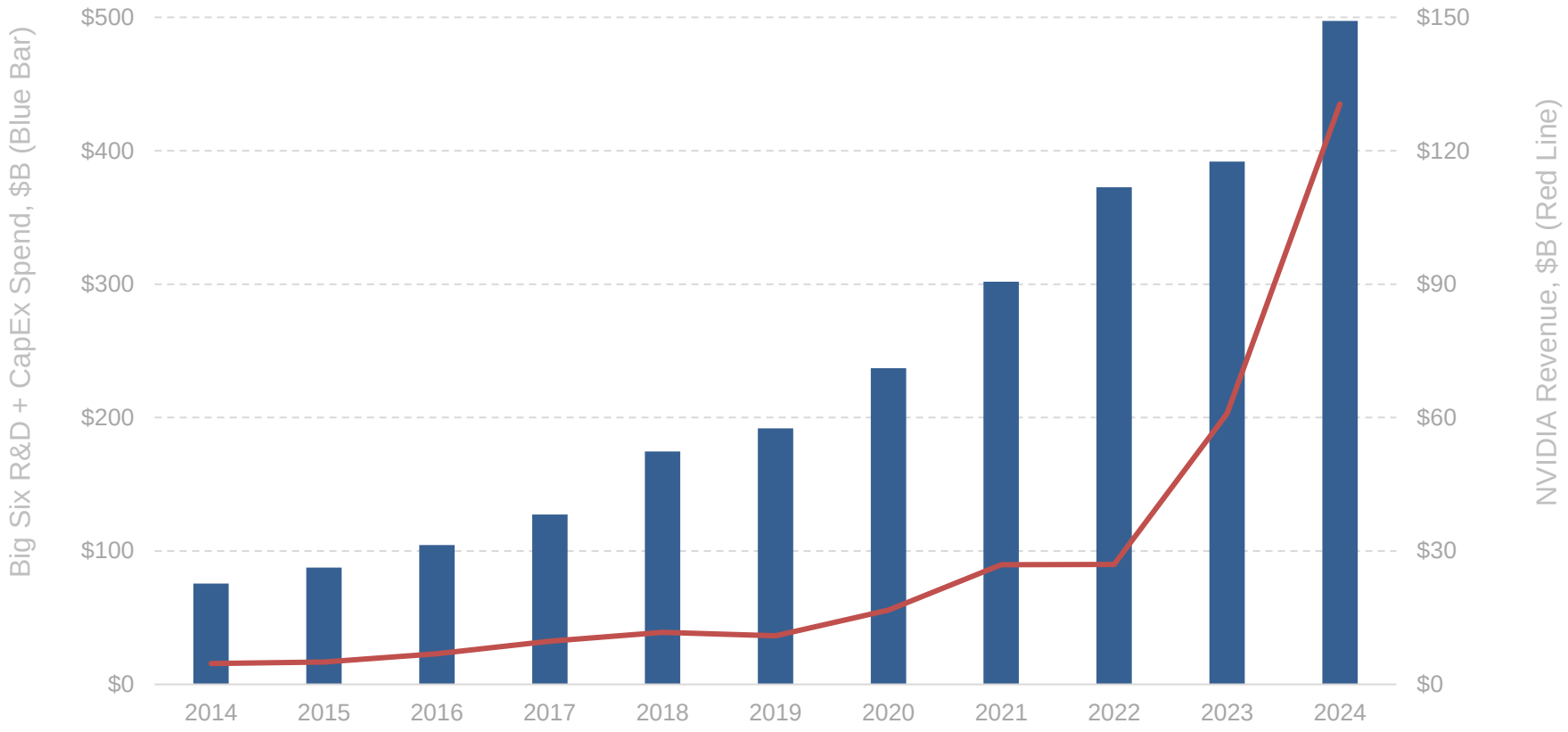
AI Monetization…Chips = NVIDIA Quarterly Revenue +78% to $39B Y/Y…



Note: Gaming includes PC & console gaming. Other includes Enterprise / Pro Vis, Auto, & OEM / Other. NVIDIA's fiscal year ends January 31. The figures in the title compare FQ4:25 to FQ4:24. Source: NVIDIA (1/25) via Morgan Stanley

…AI Monetization…Chips = NVIDIA Revenue +28x Over Ten Years…Big Six CapEx + R&D +6x

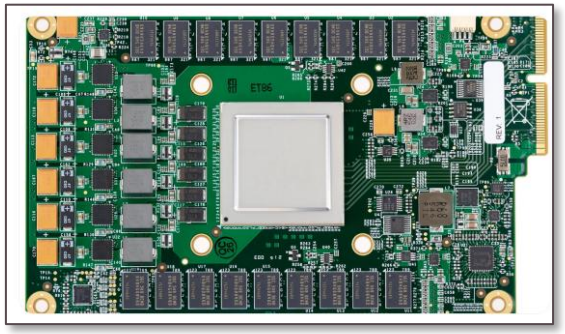
Big Six* USA Public Technology Company R&D + CapEx Spend ($B) vs. NVIDIA Revenue ($B) – 2014-2024, per Capital IQ



**Note: Big Six USA technology companies include Apple, Nvidia, Microsoft, Alphabet / Google, Amazon, & Meta Platforms / Facebook. Includes CapEx for Amazon AWS + Retail as R&D expense is not regularly separated for those two business divisions. Source: Companies' investor reports, Capital IQ (4/25)*

AI Monetization…Chips = Google TPU Sales* +116% to $8.9B Y/Y, per Morgan Stanley

Google TPU (Tensor Processing Unit) Estimated Sales – 2021-2024, per Morgan Stanley

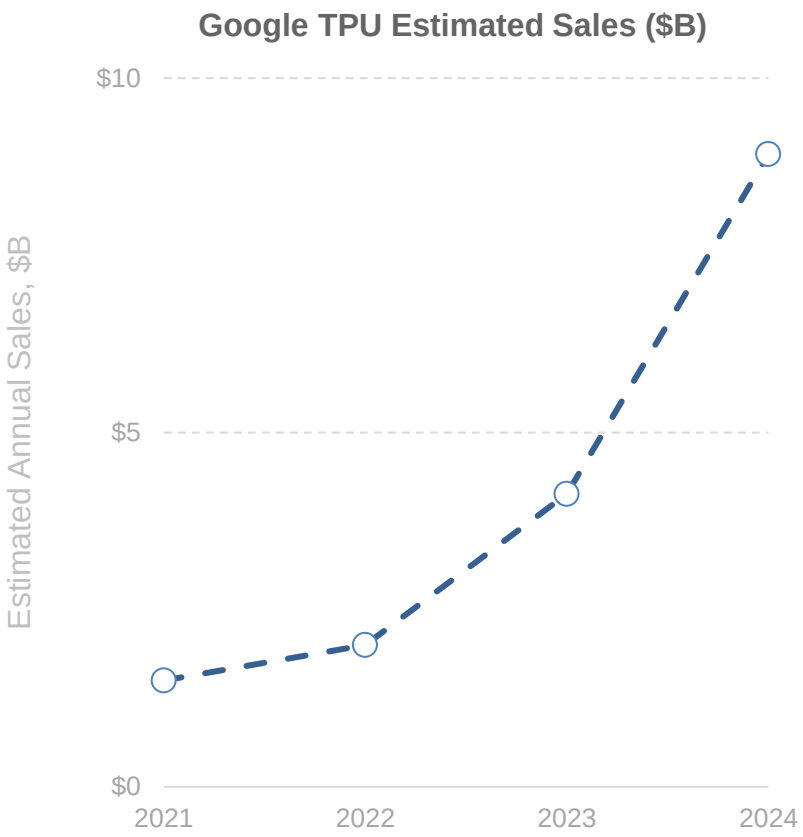


TPUs were purpose-built specifically for AI. TPUs are an application-specific integrated circuit (ASIC), a chip designed for a single, specific purpose: running the unique matrix and vector-based mathematics that's needed for building and running AI models.

Our first such chip, TPU v1, was deployed internally in 2015 and was instantly a hit across different parts of Google…

…'We thought we'd maybe build under 10,000 of them,' said Andy Swing, principal engineer on our machine learning hardware systems. 'We ended up building over 100,000.'

- Google Press Release, 7/24



**Figures are estimates per Morgan Stanley research. Note: Relative to GPUs, ASICs are custom-designed for specific tasks (e.g., AI model training,) whereas GPUs are general-purpose. Source: Google, Morgan Stanley, 'GenAI Monetization – Assessing The ROI Equation' (2/25)*

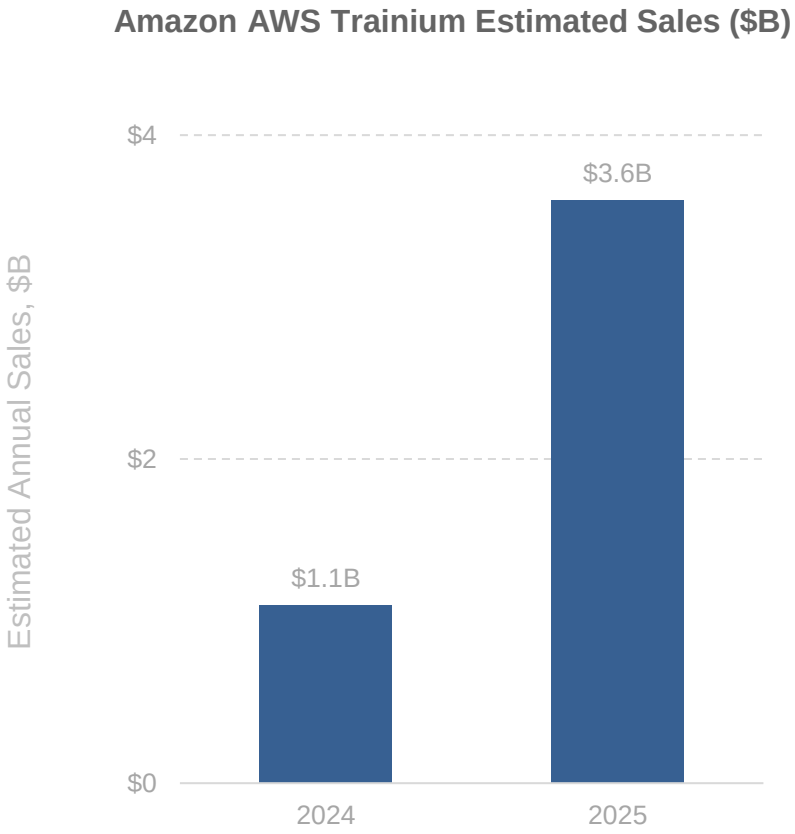
AI Monetization…Chips = Amazon AWS Trainium* Sales +216% to $3.6B Y/Y, per Morgan Stanley

Amazon AWS Trainium Estimated Sales – 2024-2025, per Morgan Stanley

AWS Trainium chips are a family of AI chips purpose built by AWS for AI training and inference to deliver high performance while reducing costs…

AWS Trainium2 chip delivers up to 4x the performance of first-generation Trainium…[and offers] 30-40% better price performance than the current generation of GPU-based EC2 P5e and P5en instances.

- Amazon AWS Trainium Overview, Accessed 5/25



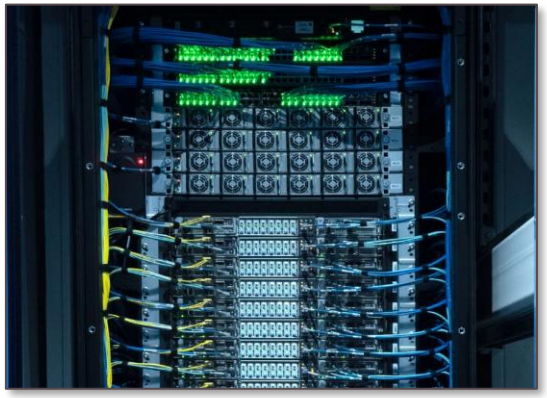
Note: Relative to GPUs, ASICs are custom-designed for specific tasks (e.g., AI model training,) whereas GPUs are general-purpose. Figures are estimates per Morgan Stanley research. Source: Amazon AWS, Morgan Stanley, 'GenAI Monetization – Assessing The ROI Equation' (2/25)

AI Monetization =

Compute Services

AI Monetization…Cloud Computing = CoreWeave Revenue +730% to $1.9B Y/Y

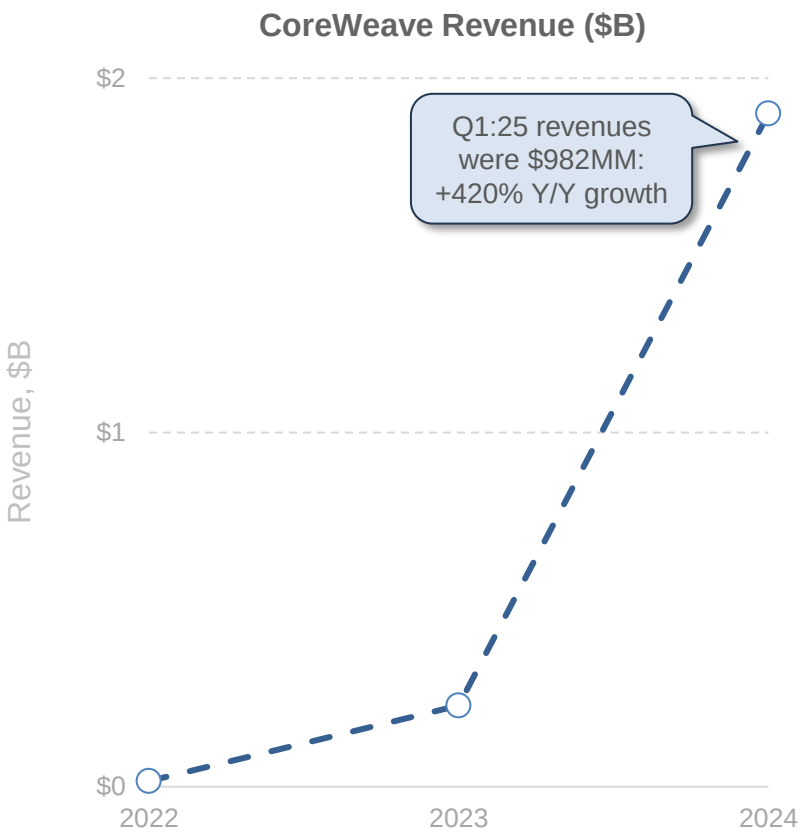
CoreWeave Revenue – 2022-2024, per CoreWeave



We've delivered an outstanding start to 2025 on multiple fronts. Our strong first quarter financial performance caps a string of milestones including our IPO, our major strategic deal with OpenAI as well as other customer wins, our acquisition of Weights & Biases and many technical achievements…

…Demand for our platform is robust and accelerating as AI leaders seek the highly performant AI cloud infrastructure required for the most advanced applications. We are scaling as fast as possible to capture that demand. The future runs on CoreWeave.

- CoreWeave CEO Michael Intrator, 5/25



Source: CoreWeave (as of 5/25)

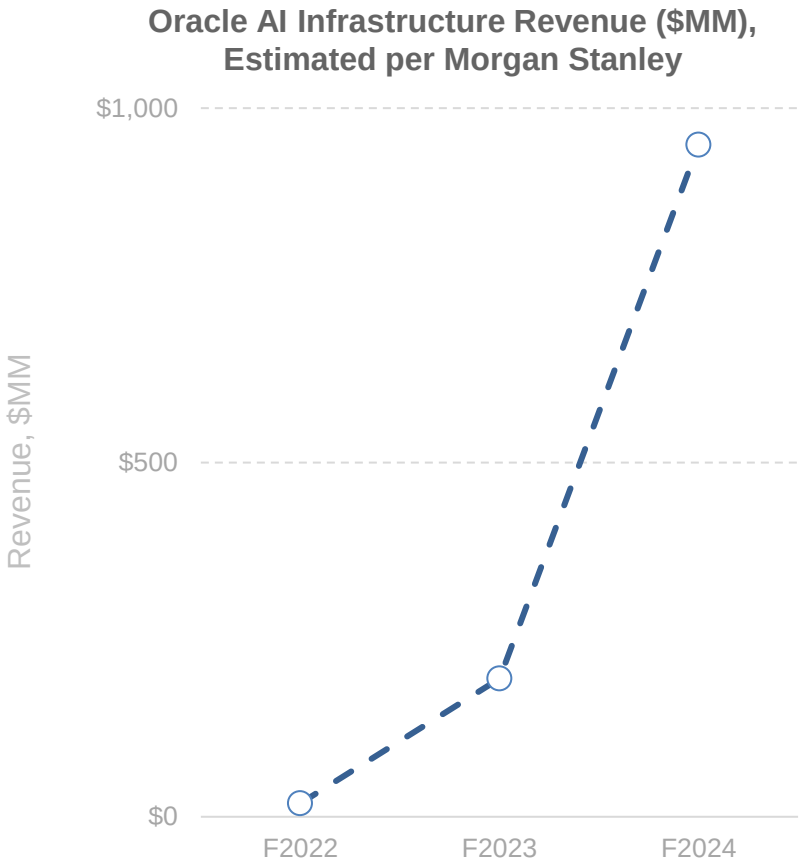
AI Monetization…AI Infrastructure = Oracle Revenue +50x to $948MM Over Two Years

Oracle AI Infrastructure Revenue – F2022-F2024, per Oracle & Morgan Stanley Estimates

There are many, many [AI infrastructure] customers who have come on and that haven't gotten capacity yet…

…We've got at least 40 new AI bookings that are over a billion (dollars) that haven't come online yet.

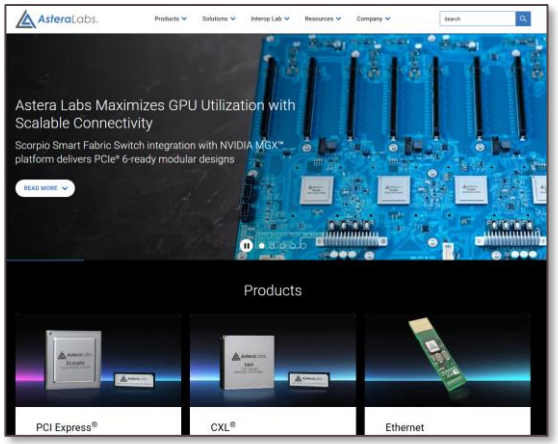
- Oracle CEO Safra Catz, 3/24



Source: Oracle, Morgan Stanley estimates, 'What's Ahead for the AI Infrastructure Cycle' (8/24)

AI Monetization…Infrastructure Connectivity = Astera Labs Revenue +242% to $396MM Y/Y

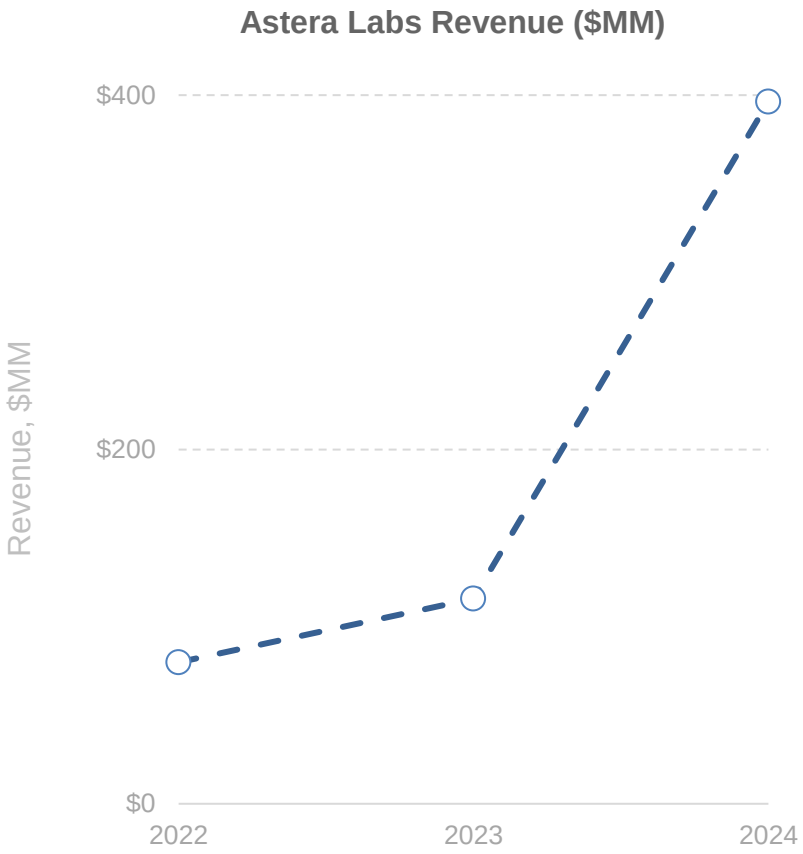
Astera Labs – 2022-2024, per Astera Labs



Astera Labs delivered strong Q4 results, with revenue growing 25% versus the previous quarter, and capped off a stellar 2024 with 242% revenue growth year-over-year…

…We expect 2025 to be a breakout year as we enter a new phase of growth driven by revenue from all four of our product families to support a diverse set of customers and platforms. This includes our flagship Scorpio Fabric products for head-node PCIe connectivity and backend AI accelerator scale-up clustering.

- Astera Labs CEO Jitendra Mohan, 2/25



Source: Astera Labs financial results (as of 4/25)

AI Monetization…Data Collection + Supercomputing = Tesla AI Training Capacity +8.5x

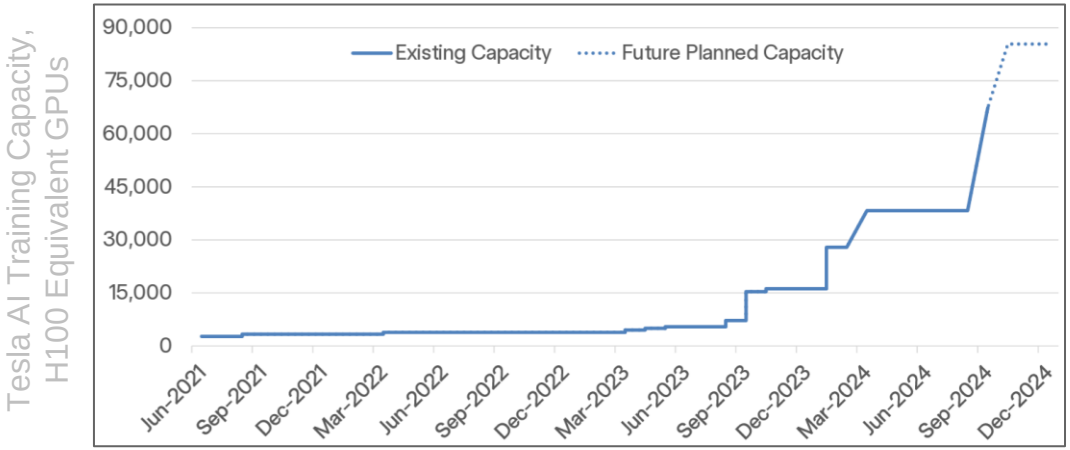
Tesla Dojo Custom Supercomputer – 6/21-9/24, per Tesla

We're continuing to simultaneously make significant purchases of GPUs and also putting a lot of effort into Dojo [custom supercomputer], which we believe has the potential for an order of magnitude improvement in the cost of training…

…Dojo also has the potential to become a sellable service that we would offer to other companies, in the same way that Amazon Web Services offers more web services, even though it started out as a bookstore. So, I really think that the Dojo potential is very significant.

- Tesla Co-Founder & CEO Elon Musk, 4/23

Tesla AI Training Capacity (H100-Equivalent GPUs)



Note: Listing capacity in 'H100-equivalent GPUs' means Tesla converts the aggregate AI-training throughput of Dojo and its other accelerators into the number of NVIDIA Hopper H100 data-center GPUs that would deliver the same FP8/FP16 FLOPS, giving a single, industry-standard yard-stick for compute scale.
Source: Tesla Q1:23 earnings call, Tesla Q3:24 investor presentation, Data Center Dynamics, Wikimedia Commons

AI Monetization =

Data Layer

AI Monetization…Data Labeling & Evaluation = Scale AI Revenue +160% to $870MM Y/Y

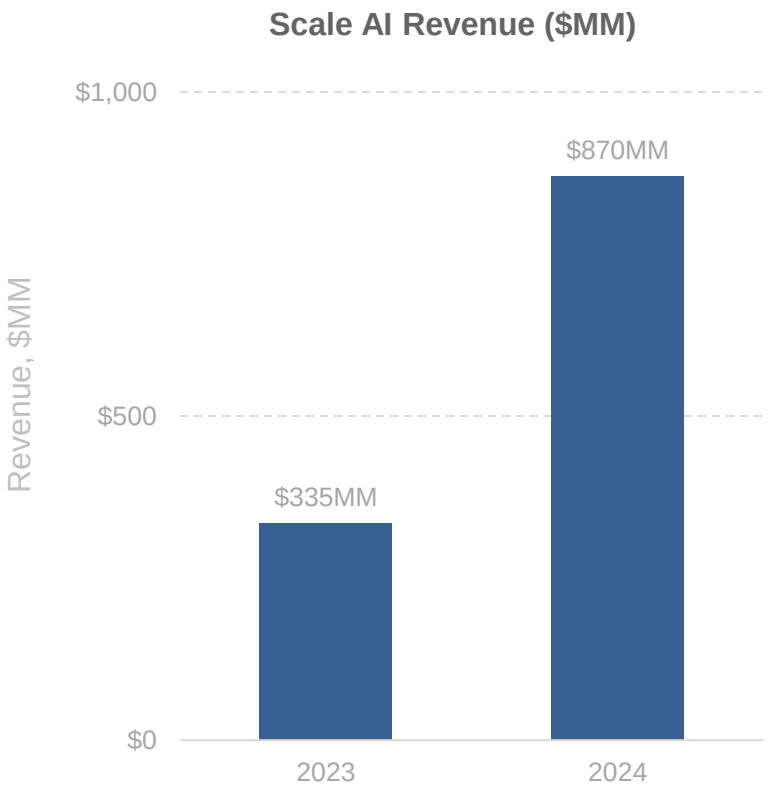
Scale AI Revenue – 2023-2024, per Scale AI

Data abundance is not the default; it's a choice. It requires bringing together the best minds in engineering, operations, and AI. Our vision is one of data abundance, where we have the means of production to continue scaling frontier LLMs many more orders of magnitude. We should not be data-constrained in getting to GPT-10.

- Scale AI Co-Founder & CEO Alexandr Wang, 5/24

We saw 160% revenue growth in 2024 from the previous year, and we secured more than $1.5 billion in new business.

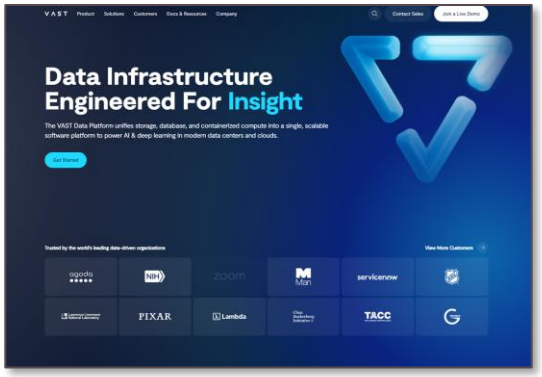
- Scale AI Head of Corporate and Product Comms Joe Osborne, 4/25



Note: 2023 figures are estimates based on Joe Osborne (Head of Corporate and Product Comms at Scale AI,) who indicated, 'We saw 160% revenue growth in 2024 from the previous year, and we secured more than $1.5 billion in new business.' Source: Scale AI, The Information (4/25) (link)

AI Monetization…Data Storage / Management / Processing = VAST Data Lifetime Sales From 0 to $2B in Just Over Six Years

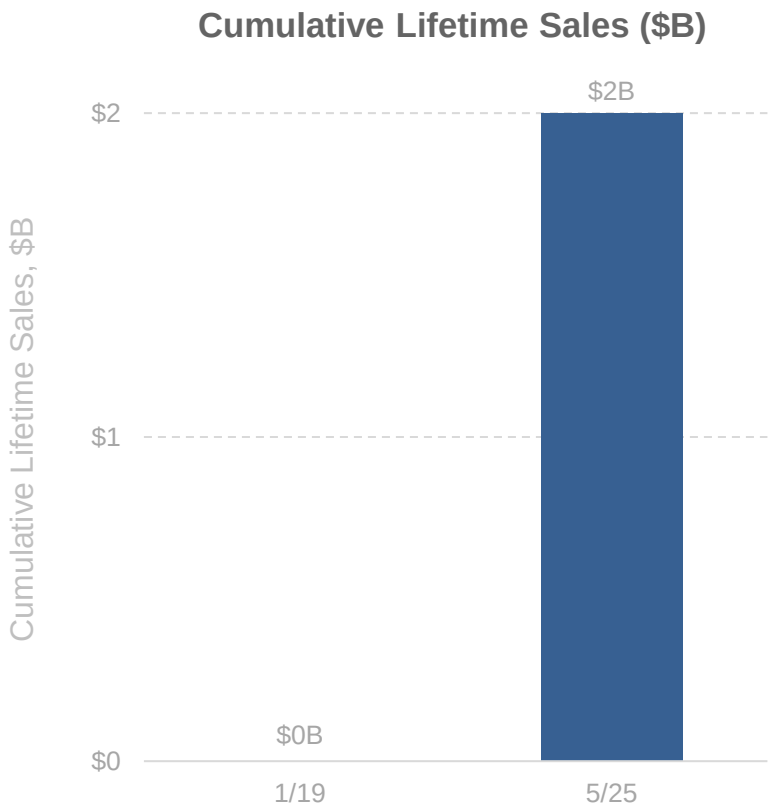
VAST Data – 1/19-5/25, per VAST Data



Everything is accelerating. The rate of AI progress is constantly increasing as model builders build on each other's discoveries and push the boundaries ever farther. While we've been talking about thinking machines since early 2022, the advent of reasoning models in the last 12 months means that the era of thinking machines is actually now upon us…

…We at VAST believe that the path to the greatest potential gain is to simplify and reduce the fundamental challenges that need to be resolved. If we can build a simple approach to encompass nearly all of the infrastructure layers needed for AI, without compromise… customers supremely benefit.

- VAST Data CEO Renen Hallak, 5/25



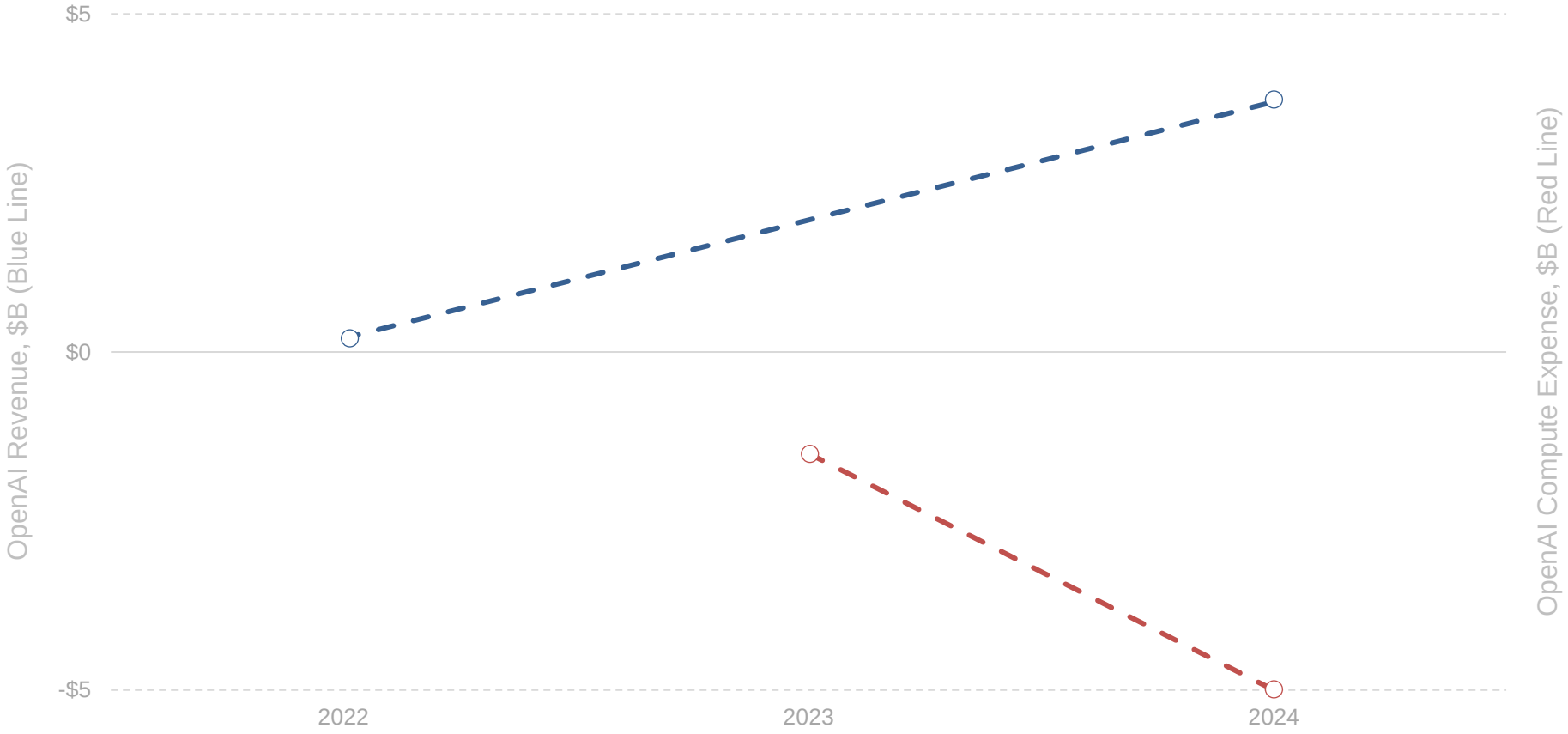
Source: VAST Data, Silicon Angle

AI-Related Cost Ramps Relative to Revenue =

Can Be Head-Turning

AI Monetization – OpenAI = Revenue vs. Compute Expense, per *The Information*

OpenAI Revenue & Compute Expense ($B) by Year – 2022-2024, per *The Information*



Note: No compute expense data available in 2022. Figures are estimates based off public reports & The Information reporting.
Source: The Information (4/25 and prior) (link, link, link, link, link & link)

AI Monetization – Microsoft / Amazon / Alphabet / Meta = CapEx Up…Free Cash Flow Margins Down

Capital Expenditure, Free Cash Flow Margin, Revenue Growth – C2023-C2024, per Capital IQ

		CapEx →	Free Cash Flow Margin vs.	Revenue
Microsoft	C2023	$35B	30%	$228B
	C2024	**$56B**	**27%**	**$262B**
	Y/Y Change	**+58%**	**-10%**	**+15%**
Amazon	C2023	$53B	6%	$575B
	C2024	**$83B**	**5%**	**$638B**
	Y/Y Change	**+57%**	**-8%**	**+11%**
Alphabet (Google)	C2023	$32B	23%	$307B
	C2024	**$52B**	**21%**	**$350B**
	Y/Y Change	**+63%**	**-8%**	**+14%**
Meta Platforms (Facebook)	C2023	$27B	33%	$135B
	C2024	**$37B**	**33%**	**$165B**
	Y/Y Change	**+38%**	**<1%**	**+22%**

Note: FCF calculated as cash flow from operations less capex to standardize, as only some companies subtract finance leases and Amazon adjusts FCF for gains on sale of equipment. Amazon statistics shown for both AWS & Retail; FCF not broken out across subsidiaries. Source: Capital IQ (5/25)

So…We Have…

High Revenue Growth +
High Cash Burn +
High Valuations +
High Investment Levels =

Good News for Consumers…
Others TBD…

Select Private AI Model Companies – 5/13/25 = ~$11B+ Annualized Revenue vs. ~$95B Raised…

Foundation Model Estimated Revenue & Capital Raised – 5/13/25, per Pitchbook, *The Information, Bloomberg, The Wall Street Journal & CNBC*

Company	Annualized Revenue ($MM)	Total Raised To-Date ($MM)
OpenAI	9,200 (4/25 estimated)	63,920 (Last Raise: 3/25)
Anthropic	2,000 (3/25)	18,000 (Last Raise: 3/25)
xAI	Materially North of 100* (4/25)	12,130 (Last Raise: 11/24)
Perplexity	120 (5/25)	1,410 (Last Raise: 5/25)

**Select media reports have xAI revenue being as high as $1B as of 4/25. Note: OpenAI annualized revenue estimated based upon full-year 2024 & 2025 revenue estimates as published by The Information & Bloomberg, assuming linear revenue growth. Figures are rounded. Source: Source: Pitchbook (5/25), The Information (link), Bloomberg (link & link) & CNBC (link & link)*

…Select Private AI Model Companies – 5/13/25 = High Valuation-to-Revenue Multiples

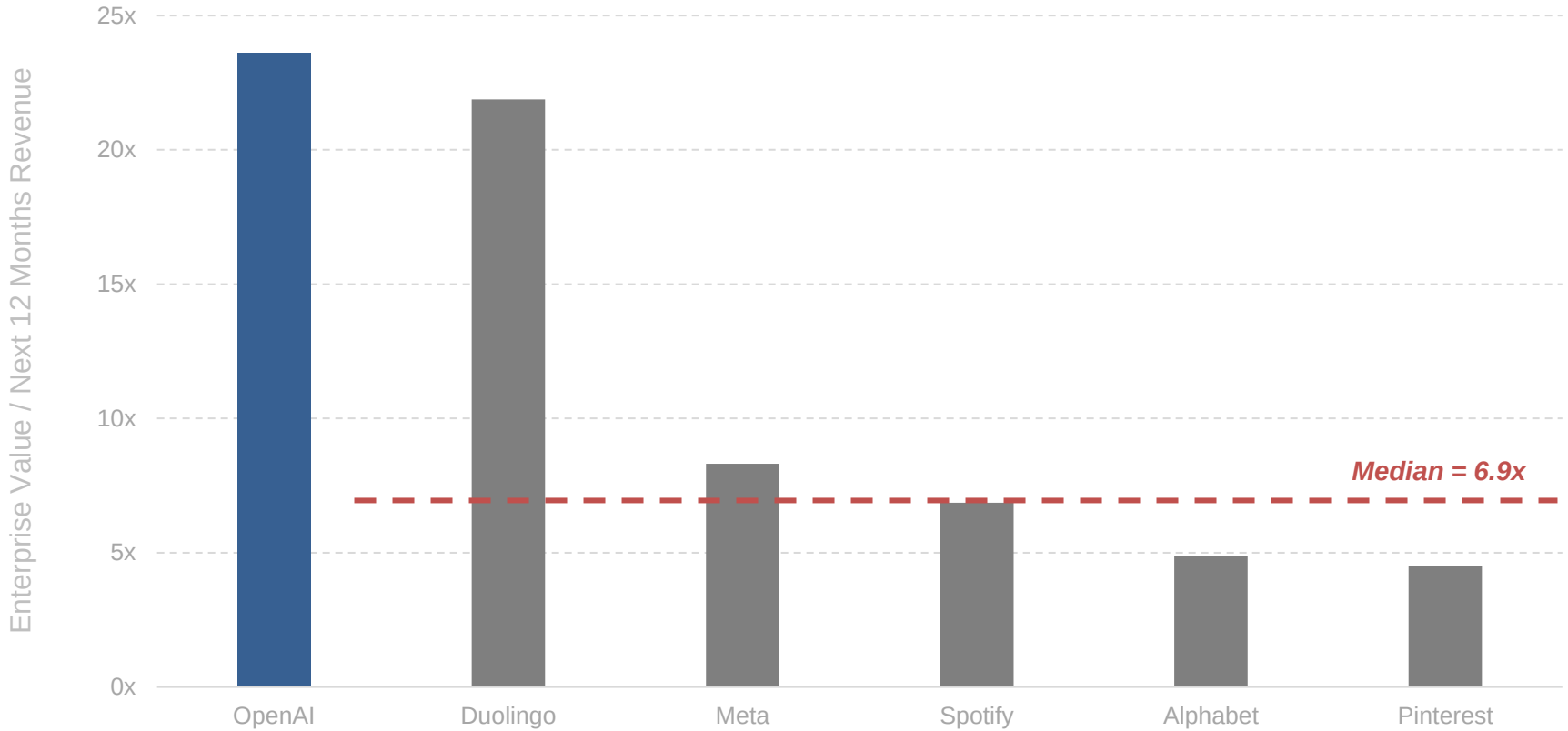
Foundation Model Estimated Revenue Multiple – 5/13/25, per Pitchbook, *The Information, Bloomberg, The Wall Street Journal* & *CNBC*

Company	Annualized Revenue ($MM)	Latest Valuation ($MM)	Revenue Multiple
OpenAI	9,200 (4/25 estimated)	300,000 (3/25)	33x
Anthropic	2,000 (3/25)	61,500 (3/25)	31x
xAI	Materially North of 100* (4/25)	80,000 (3/25)	N/A
Perplexity	120 (5/25)	9,000 (12/24)	75x

**Select media reports have xAI revenue being as high as $1B as of 4/25. Note: OpenAI annualized revenue estimated based upon full-year 2024 & 2025 revenue estimates as published by The Information & Bloomberg, assuming linear revenue growth. xAI valuation per Elon Musk. Figures are rounded. Perplexity was reported to be in advanced talks to raise capital at a $14B post-money valuation as of 5/14/25; however, as this is not finalized at time of publication, we quote their last finalized funding round here. Source: Pitchbook (5/25), The Information (link), Bloomberg (link & link) & CNBC (link & link)*

Valuation-to-Revenue Multiple – OpenAI = Looks Expensive…

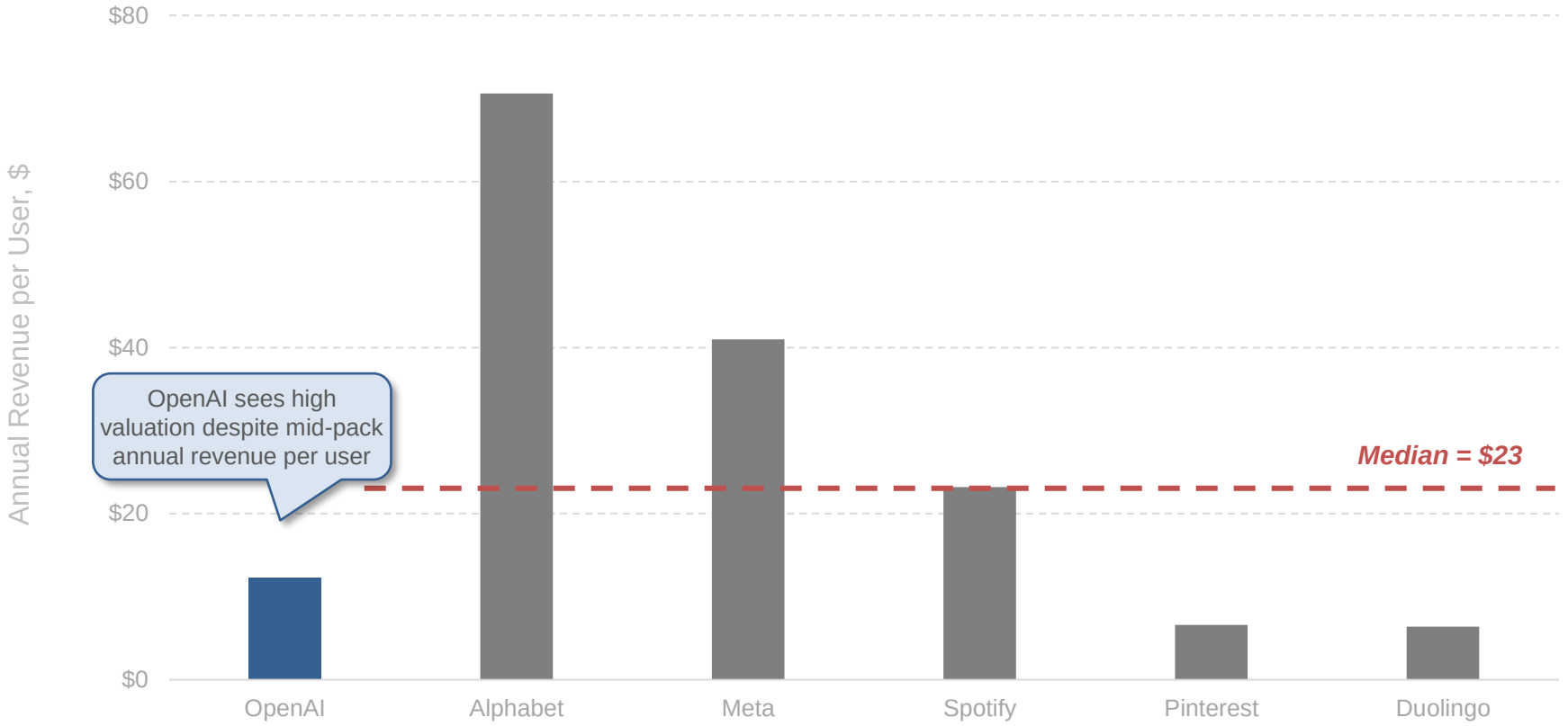
Estimated Enterprise Value / Next 12 Months Revenue Multiple – 5/25, per Capital IQ & *Bloomberg*



Note: OpenAI figures are estimates. Next 12 months revenue multiples for companies other than OpenAI are consensus estimates per Capital IQ. OpenAI NTM revenue estimates are as of 12/24 due to data availability. Source: Capital IQ (5/15/25), Bloomberg (link)

…Revenue-per-User Multiple – OpenAI = In-the-Range

Estimated Annual Revenue Per User ($) – 2024, per Capital IQ, Morgan Stanley, Semrush, *The Information* & Company Disclosures



Note: OpenAI figures are estimates as of 4/25. All other public-company figures are as of 12/31/24, using CY2024 data. OpenAI data uses WAUs due to data availability (conservatively assumed as MAUs); other figures use MAUs. Here we assume average weekly active ChatGPT users of 300MM based off OpenAI's 12/24 disclosure. We estimate 2024 ChatGPT revenue of $3.7B, per company estimates. Monthly active user figures are estimates for Alphabet based off website traffic measurements & global internet user data. Meta last reported MAPs for app family in Q4:23, we conservatively assume no growth since.
Source: Capital IQ (12/24), The Information (4/25 and prior) (link, link, link, link & link), Semrush (11/24), Morgan Stanley, ITU, company disclosures, BOND estimates

As global digital user bases have grown and potential rapidity of usage traction has risen in tandem, areas of corporate investment (for companies new and old) have become increasingly competitive and capital-intensive.
The AI tech cycle of creative disruption has historical analogs.

Head turners of the semi-recent past include Apple's near bankruptcy in 1997 when its market capitalization was $1.7B*, now $3.2T.

Amazon.com's near death moment happened in Q4:00 when it reported a net loss of -$545MM on revenue of $972MM.
Founder and then-CEO Jeff Bezos noted in the 2000 Shareholder Report that
It's been a brutal year for many in the capital markets and certainly for Amazon.com shareholders. As of this writing, our shares are down more than 80% from when I wrote you last year.
At post-loss trough in Q3:01 its market cap was $2.2B while it supported 23MM active customer accounts.
The market cap is now $2.2T.
All in, Amazon lost -$3B in the twenty-seven quarters between its launch in Q2:97 and the end of its first net income-positive year (2003).
For its most recent twenty-seven most recent quarters (Q3:18-Q1:25), Amazon's cumulative net income was $176B.

Google's IPO filing (April 2004) noted that in Q1:04, after having only raised a Series A funding round, it spent 22% of revenue ($86MM of $390MM) on capital expenditures – at the time it was an incomprehensibly high number.
It went public at a $23B market cap, now $2.0T…

**Market capitalization taken as of 7/1/97. Microsoft finalized its investment in Apple just over one month later, on 8/6/97.*
Note: Present market capitalization figures are shown as of 5/14/25.

…Uber burned -$17B* between 2016 and 2022 (and materially more before that)
before its first free cash flow-positive year in 2023.
In 2022, it had 131MM monthly active platform consumers.
Uber's last equity financing was a Series G.
Its fully-diluted IPO market cap was $82B, now $189B.

Tesla burned -$9.2B between 2009 and 2018 before becoming free cash flow positive in 2019.
In the ten years between 2009 and 2018, it lost a cumulative -$5.6B delivering ~540K vehicles.
It went public in 2010 at a market cap of $1.6B.
From 2019-2024, it then earned $40B delivering 6.7MM vehicles.
Its market cap is now $1.1T.

It is important to remember – most of the time, when all is said and done –
a business's valuation should represent the present value of its future free cash flows.
The aforementioned companies – with aggressive cash burn –
tested this premise hard, built large-scale data-driven network effects
based on product excellence / constant improvement,
developed technology-driven competitive advantage and ultimately proved the naysayers wrong.

Only time will tell which side of the money-making equation the current AI aspirants will land.

**Measured as unlevered free cash flow.*
Note: Present market capitalization figures are shown as of 5/14/25.

Usage + Cost + Loss Growth =

Unprecedented…

*What About **Future** Monetization + Profits?*

AI Monetization Possibilities =

New Entrants & / Or Tech Incumbents?

To understand where AI model economics may be heading, one can look at the mounting tension between capabilities and costs.

Training the most powerful large language models (LLMs) has become one of the most expensive / capital-intensive efforts in human history. As the frontier of performance pushes toward ever-larger parameter counts and more complex architectures, model training costs are rising into the billions of dollars.

Ironically, this race to build the most capable general-purpose models may be accelerating commoditization and driving diminishing returns, as output quality converges across players and differentiation becomes harder to sustain.
At the same time, the cost of applying/using these models – known as inference – is falling quickly. Hardware is improving – for example, NVIDIA's 2024 Blackwell GPU consumes 105,000x less energy per token than its 2014 Kepler GPU predecessor. Couple that with breakthroughs in models' algorithmic efficiency, and the cost of inference is plummeting.

Inference represents a new cost curve, and – unlike training costs – it's arcing down, not up.
As inference becomes cheaper and more efficient, the competitive pressure amongst LLM providers increases – not on accuracy alone, but also on latency, uptime, and cost-per-token*.
What used to cost dollars can now cost pennies.
And what cost pennies may soon cost fractions of a cent…

**Cost-per-token = The expense incurred for processing or generating a single token (a word, sub-word, or character) during the operation of a language model. It is a key metric used to evaluate the computational efficiency and cost-effectiveness of deploying AI models, particularly in applications like natural language processing.*

…The implications are still unfolding. For users (and developers), this shift is a gift: dramatically lower unit costs to access powerful AI. And as end-user costs decline, creation of new products and services is flourishing, and user and usage adoption is rising.

For model providers, however, this raises real questions about monetization and profits. Training is expensive, serving is getting cheap, and pricing power is slipping. The business model is in flux. And there are new questions about the one-size-fits-all LLM approach, with smaller, cheaper models trained for custom use cases* now emerging.

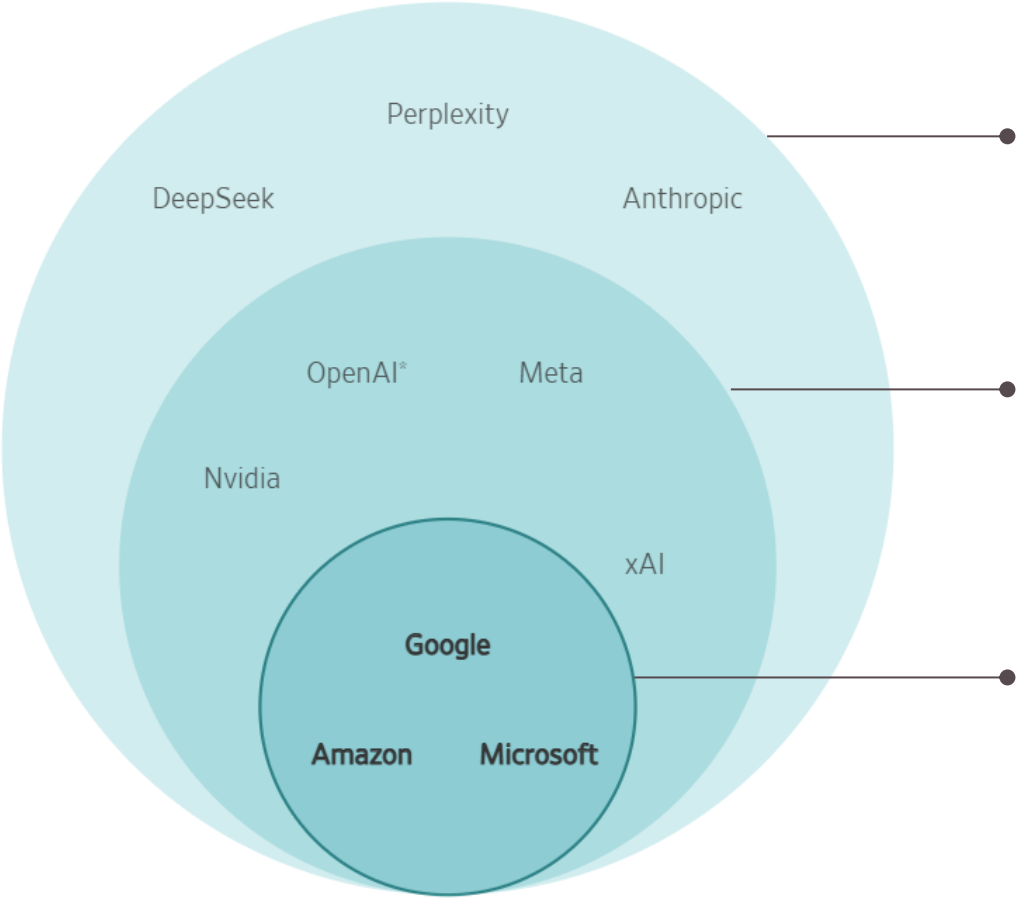
Additionally, traditional business moats are being disrupted. Look no further than Google. The company launched AI Overviews in May of last year – they sit above many Google search results. The company highlighted it had 1.5B AI Overviews MAUs as of 4/25…it's notable that in the last few weeks, Google began adding advertisements to select AI Overviews.

Will providers try to build horizontal platforms? Will they dive into specialized applications? Will one or two leaders drive dominant user and usage share and related monetization, be it subscriptions (easily enabled by digital payment providers), digital services, ads, etc.? Only time will tell. In the short term, it's hard to ignore that the economics of general-purpose LLMs look like commodity businesses with venture-scale burn.

**E.g., OpenEvidence*

AI Company Landscape = Varying Degrees of Vertical Integration

Specializations of Ten Leading AI Companies – 4/25, per *The Wall Street Journal*



Developing Models & Chatbots

All 10 of these companies are building generative-AI tools that can create content including text, images and video.

+

Building AI Infrastructure

These seven companies – both tech giants and AI upstarts – are also building the hardware and data centers that provide the power and infrastructure needed to run AI systems.

+

Providing AI Cloud Services

The top cloud providers offer platforms that help businesses leverage AI tech in their own products and workflows.

**Has a partnership with Oracle, SoftBank and MGX to build out the proposed Stargate data-center network.*
Source: Wall Street Journal, 'Here's How Big the AI Revolution Really Is, in Four Charts' (4/25)

AI Monetization Possibilities =

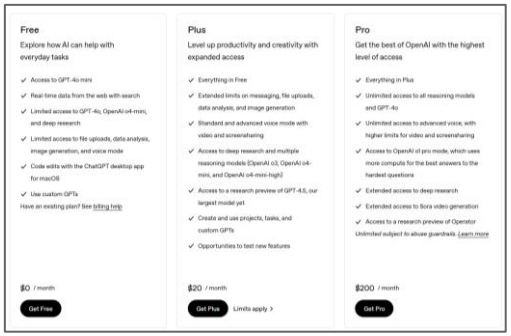
New Entrants & / Or Tech Incumbents?

AI – New Entrants =

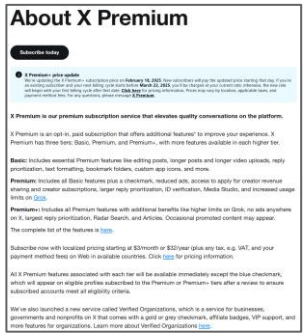
Rapidly Laying Groundwork

AI Monetization…Foundation Models = Consumer Subscription Models Driving Monetization…

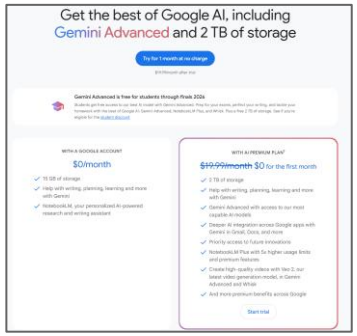
OpenAI ChatGPT, xAI Grok, Google Gemini, Anthropic Claude & Perplexity Consumer Pricing – 5/25, per Companies



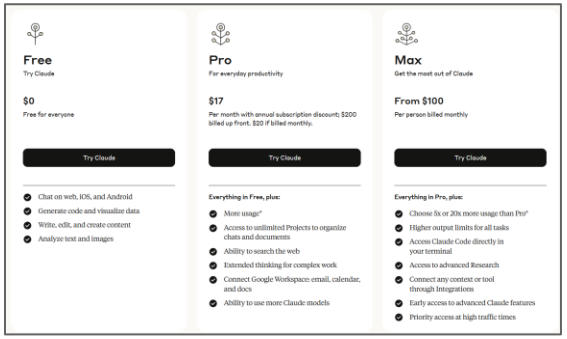
OpenAI ChatGPT
$0 (Free) / $20 (Plus) / $200 (Pro)
per Month



xAI Grok
$0 (Free) / $3 (Basic) / $8 (Premium) /
$40 (Premium+) per Month[1]



Google Gemini
$0 (Free) / $19.99 (AI Pro) /
$250 (AI Ultra) per Month



Anthropic Claude
$0 (Free) / $17[2] (Plus) / $100 (Max)
per Month

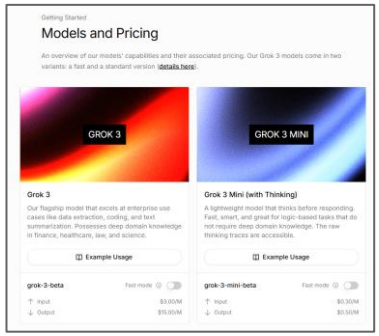
Perplexity
$0 (Free) / $20 (Pro)
per Month

Note: Excludes enterprise plans. 1. Grok pricing is bundled with X premium subscriptions. X premium subscriptions include additional benefits beyond improvements to Grok usage limits. 2. With annual discount. Source: OpenAI, X, Google, Anthropic, Perplexity websites (5/25)

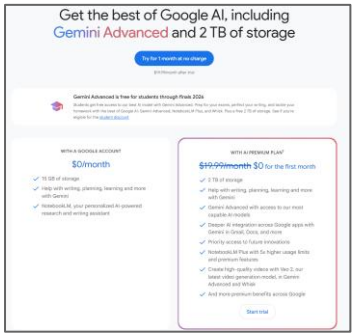
…AI Monetization…Foundation Models = Developer API Fees Driving Monetization

OpenAI ChatGPT, xAI Grok, Google Gemini, Anthropic Claude & Perplexity Developer API Pricing – 5/25, per Companies

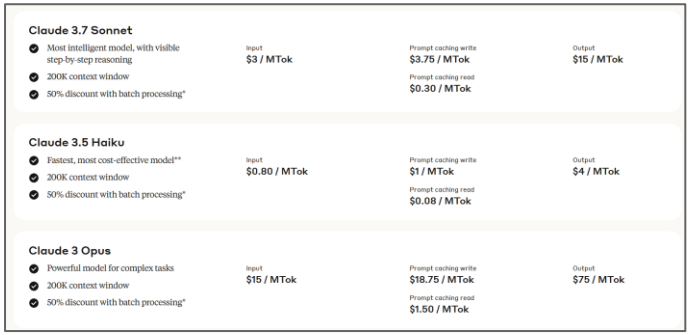
OpenAI ChatGPT
From $0.40 (GPT-4.1 nano) to $40 (o3) per 1MM Output Tokens



xAI Grok
$0.50 (grok-3-mini-beta) to $25 (grok-3-fast) per 1MM Output Tokens



Google Gemini
$0.15 (1.5 Flash-8B) to $15 (2.5 Pro Preview) per 1MM Output Tokens[1]



Anthropic Claude
From $1.25 (Claude 3 Haiku) to $75 (Claude 3 Opus) per 1MM Output Tokens

Perplexity
$1 (Sonar) to $15 (Sonar Pro) per 1MM Output Tokens

1. Gemini prices by prompt size. Gemini 1.5 Flash-8B = $0.15 per 1MM tokens for prompts ≤128K tokens; Gemini 2.5 Pro Preview = $15 per 1MM tokens for prompts >200K tokens.
Source: OpenAI, X, Google, Anthropic, Perplexity websites (5/25)

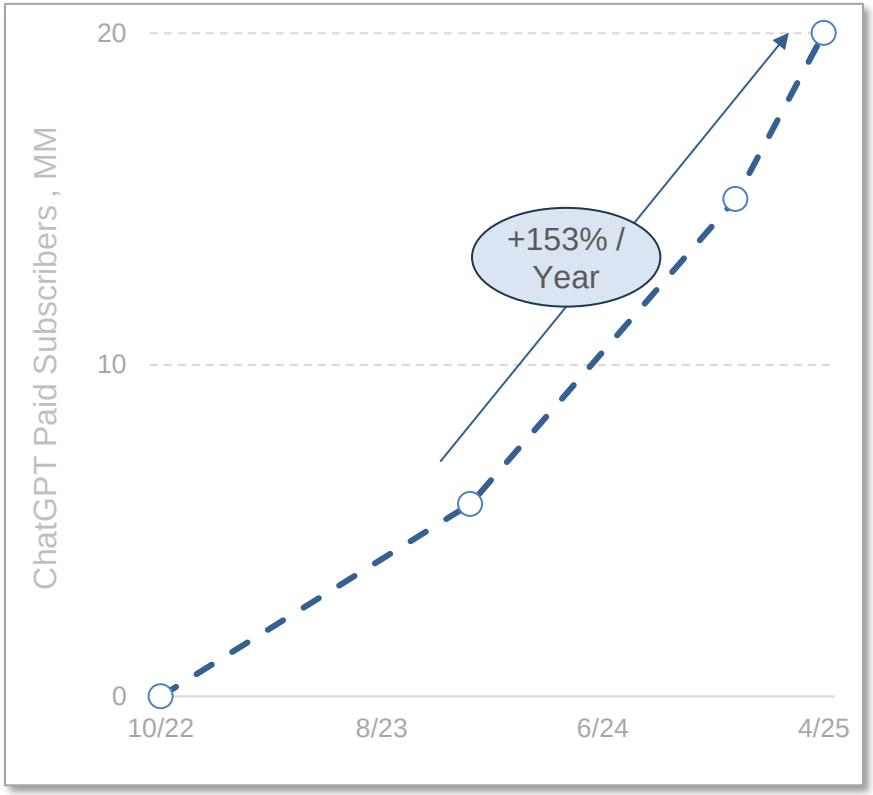
AI – New Entrants =

Rapid Revenue Growth

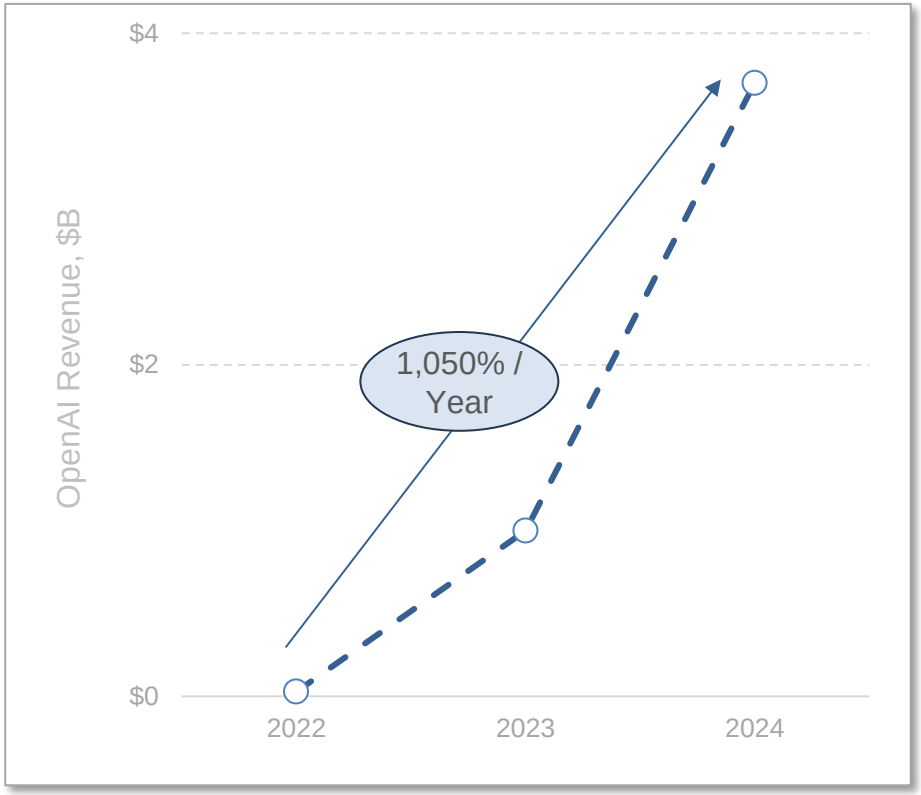
AI Monetization – Foundation Models = OpenAI Revenue +1,050% Annually to $3.7B

ChatGPT Paid Subscribers (MM) & Revenue ($B) – 10/22-4/25, per OpenAI & *The Information*

Paid Subscribers



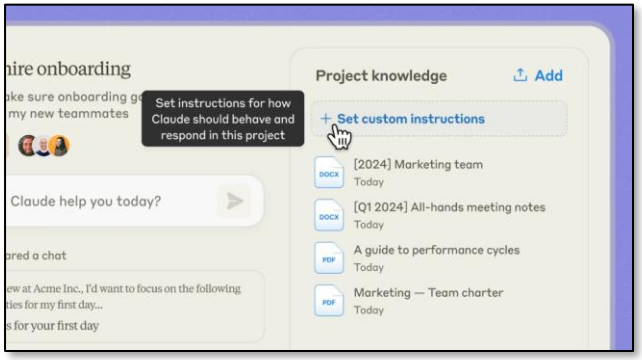
Revenue



Source: OpenAI disclosures (as of 4/25), The Information (4/25) (link, link, link & link)

AI Monetization – API & Generative Search = Anthropic Annualized Revenue +20x to $2B in Eighteen Months

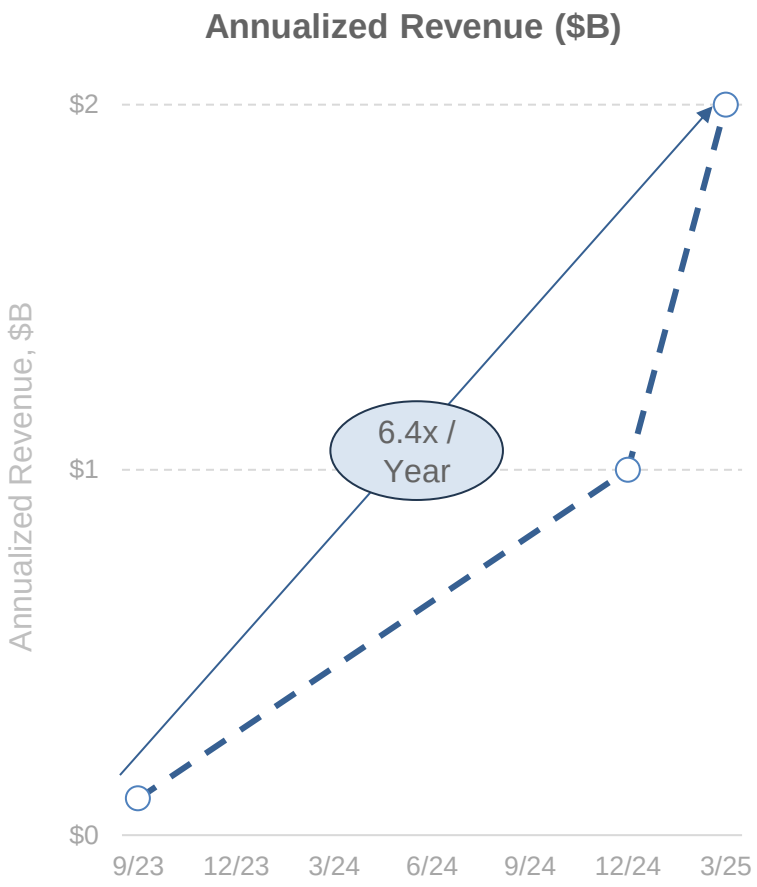
Anthropic: API & Generative Search – 9/23-3/25, per *Reuters*, *Bloomberg* & *CNBC*



We've developed Claude 3.7 Sonnet with a different philosophy from other reasoning models on the market. Just as humans use a single brain for both quick responses and deep reflection, we believe reasoning should be an integrated capability of frontier models rather than a separate model entirely. This unified approach also creates a more seamless experience for users…

…we've optimized somewhat less for math and computer science competition problems, and instead shifted focus towards real-world tasks that better reflect how businesses actually use LLMs.

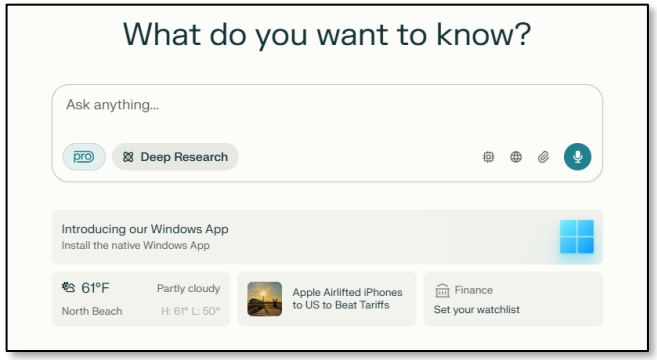
- Anthropic Press Release, 2/25



Source: Anthropic; Reuters, 'Anthropic forecasts more than $850 mln in annualized revenue rate by 2024-end – report' (12/23) (link); Bloomberg, 'Anthropic Finalizes Megaround at $61.5 Billion Valuation' (3/25) (link); CNBC, 'Anthropic closes $2.5 billion credit facility as Wall Street continues plunging money into AI boom' (5/25) (link)

AI Monetization – Generative Search = Perplexity Annualized Revenue +7.6x to $120MM in Fourteen Months

Perplexity: Generative Search – 3/24-5/25, per Perplexity & *Bloomberg*

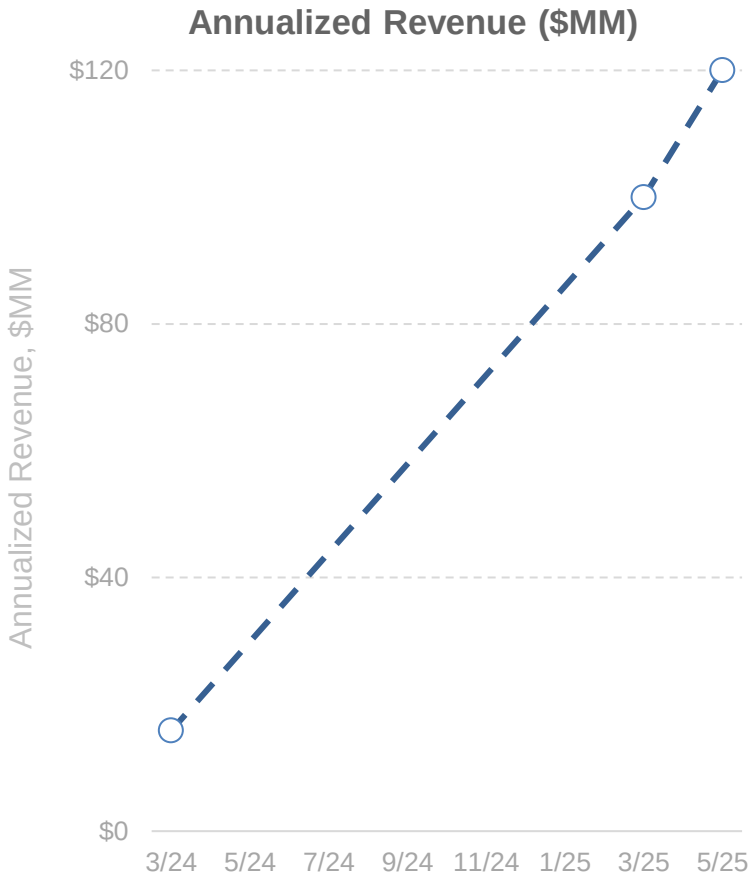


Perplexity is best described as an answer engine. You ask it a question, you get an answer. Except the difference is, all the answers are backed by sources. This is like how an academic writes a paper…What makes humans special is that we are creatures of curiosity. We need to expand on that and discover more knowledge using the power of AI.

- Perplexity Co-Founder & CEO Aravind Srinivas, 6/24

What if accessing information felt like talking to a personal research assistant?

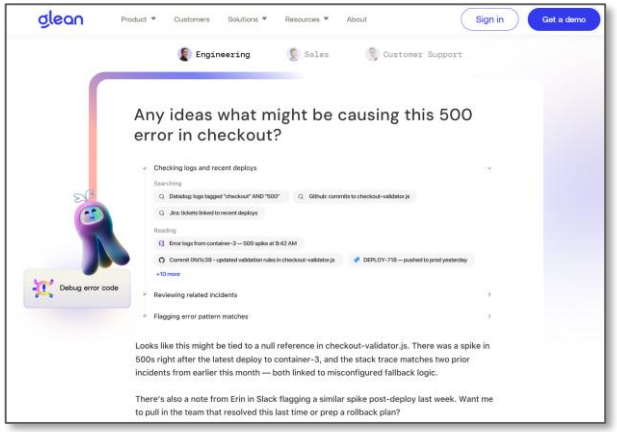
- Perplexity Co-Founder & CEO Aravind Srinivas, 5/25



Note: 3/24 annualized revenue figure is an estimate per Perplexity Co-Founder & CEO Aravind Srinivas's 3/25 LinkedIn post saying 'Perplexity has crossed $100m in annualized revenue…6.3x growth Y/Y and remains highly under monetized.'
Source: Lex Fridman Podcast (6/24), UC Berkeley (5/25), LinkedIn (3/25), Bloomberg, 'AI Startup Perplexity Nears Funding at $14 Billion Value' (5/25) (link)

AI Monetization – Enterprise Search + Agents = Glean Annualized Revenue +10x to $100MM in Twenty-Four Months

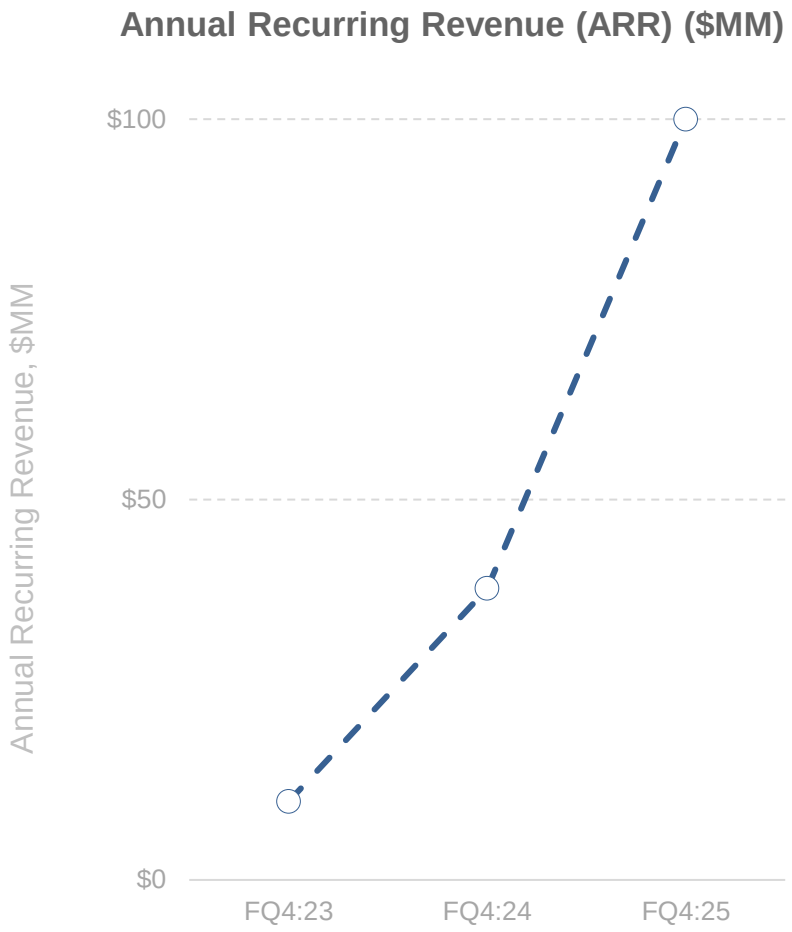
Glean – FQ4:23-FQ4:25, per Glean



We're honored to help some of the world's largest companies adopt AI to transform their businesses. To truly unlock new levels of creativity, productivity, and operational efficiency, AI needs to draw on the full picture of an organization's knowledge – and it needs to be accessible by everyone.

You shouldn't have to be a prompt engineering expert to find answers, generate content, and automate work with AI.

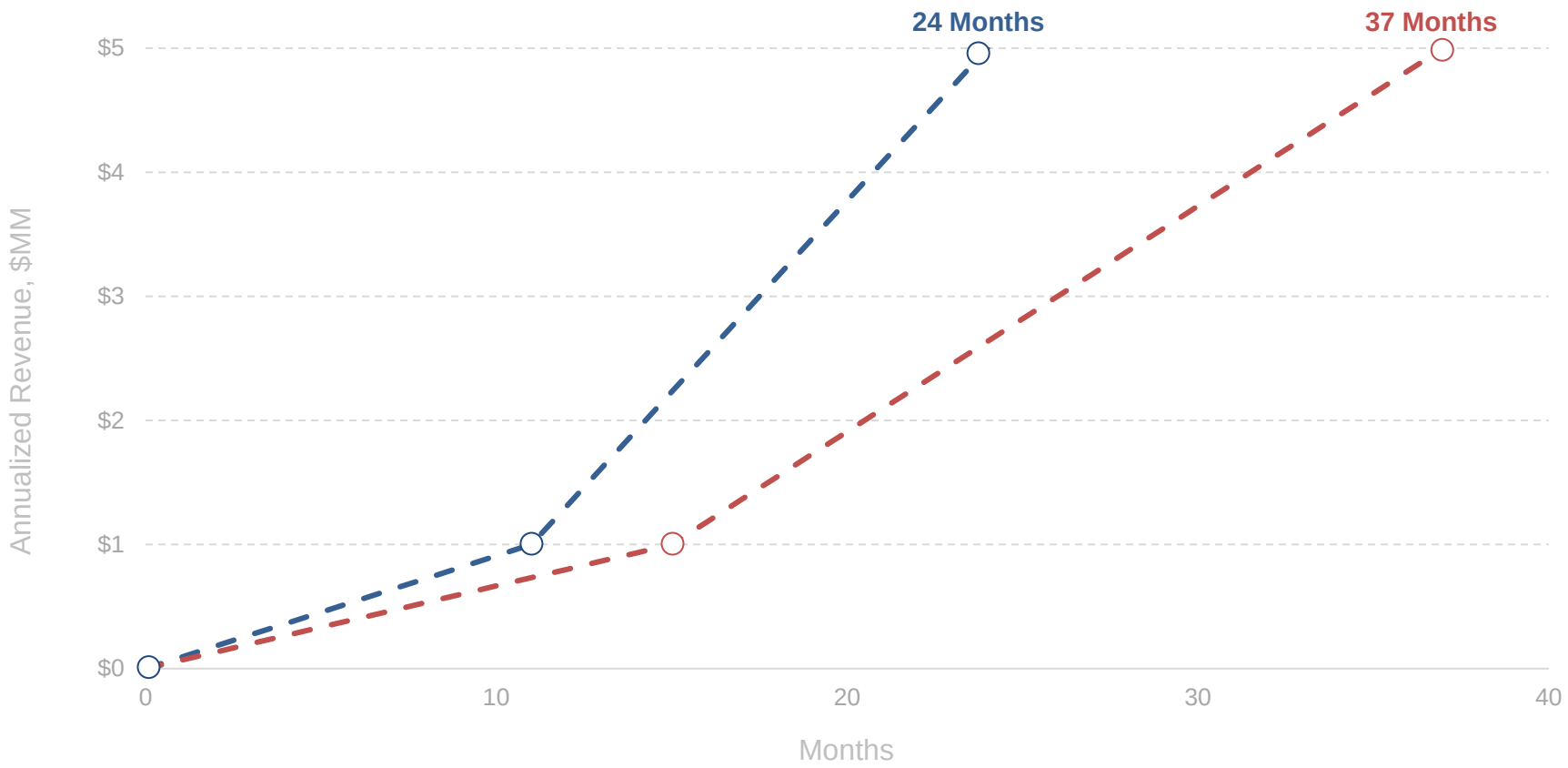
- Glean Co-Founder & CEO Arvind Jain (9/24)



Note: Glean's fiscal year ends in January. Source: Glean (2/25, 11/24)

AI Monetization – 2024 vs. 2018 = 35% Faster Ramp to $5MM ARR vs. SaaS Comparables, per Stripe

Top 100 AI Companies vs. Top 100 Saas Companies
Median Time to Annualized Revenue Milestone ($MM) – 2018 vs. 2024, per Stripe



Source: Stripe Annual Letter (2/25)

AI Monetization Possibilities =

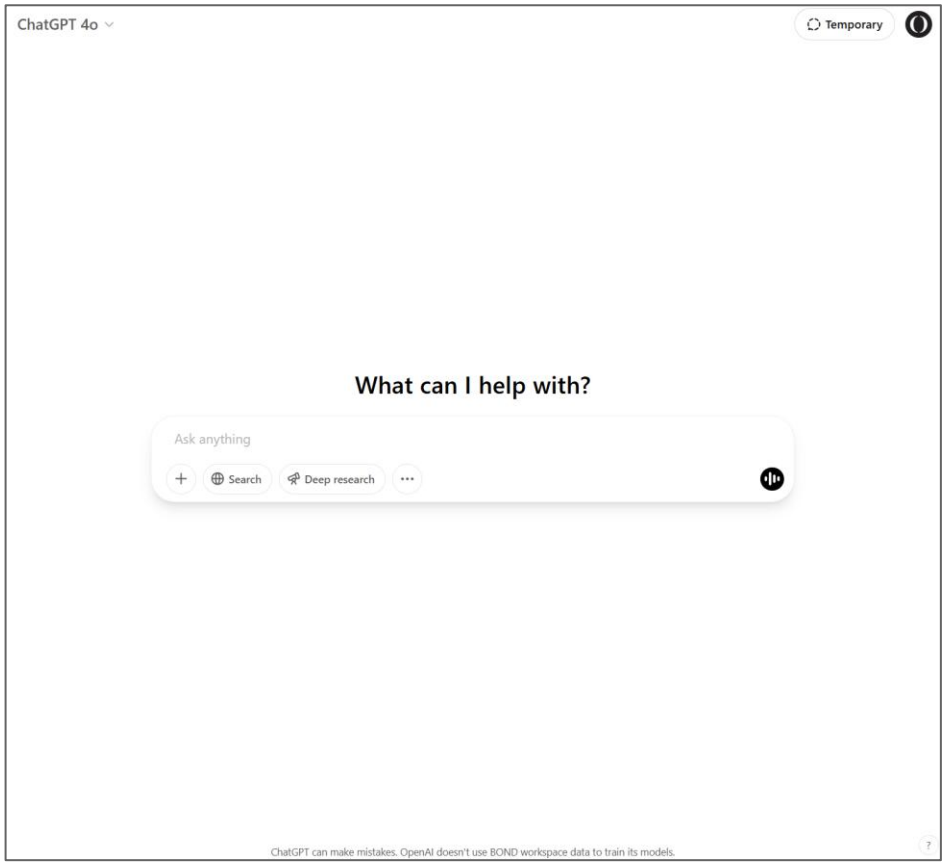
New Entrants & / Or **Tech Incumbents**?

AI – Tech Incumbents =

Broad & Steady Product / Feature Rollouts

Tech Incumbents = Optimizing Product Distribution to Roll Out AI

While ChatGPT Has 800MM+ Users Via Its Website & App…



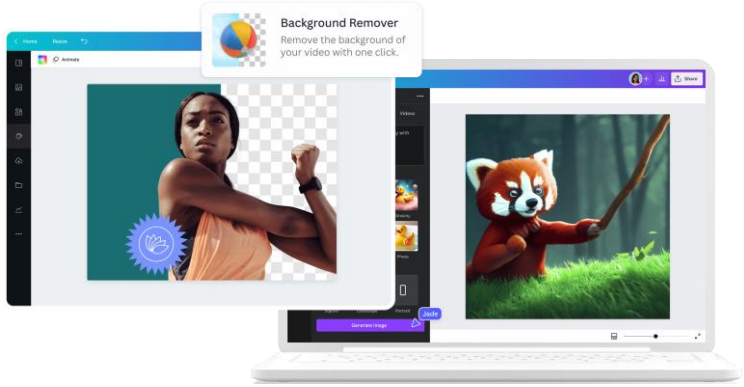
…Tech Incumbents Have Billions of Global Users on Devices & Platforms With Ongoing AI Product Rollouts

	Google **4.9B Search Users, 3B+ Android Users, 1.5B AI Overviews Users & 1B+ Assistant Devices**
	Meta Users* **3.4B+**
	Apple Devices** **2.35B**
	TikTok Users*** **1B+**
	Microsoft **1B LinkedIn Members & 400MM+ Office 365 Paid Seats**
	Spotify Users **678MM**
	Amazon **600MM+ Alexa Devices & 200MM+ Prime Subscribers**
	X Users **500MM+**
	Canva Users **230MM+**

**Meta includes Facebook, Instagram, WhatsApp, & Messenger. **Apple includes iPhones, iPads, Macs, & other Apple devices worldwide. ***As of 2021; no more recent company data available. Note: Some figures are estimates based off past company disclosures & web traffic / purchase history analytics. Different companies may define 'users' differently based on frequency. Source: Statcounter (2/25), Google (5/25), Meta 10Q (4/25), Apple (1/25), TikTok (7/21), LinkedIn (5/25), Microsoft (1/24), Spotify (5/25), Amazon (2/25 & 10/24), Elon Musk via X (7/23), Canva (4/25), OpenAI disclosures (4/25), Wikimedia Commons*

Tech Incumbent AI Rollouts = Canva – Background Remover & Magic Media (12/19)

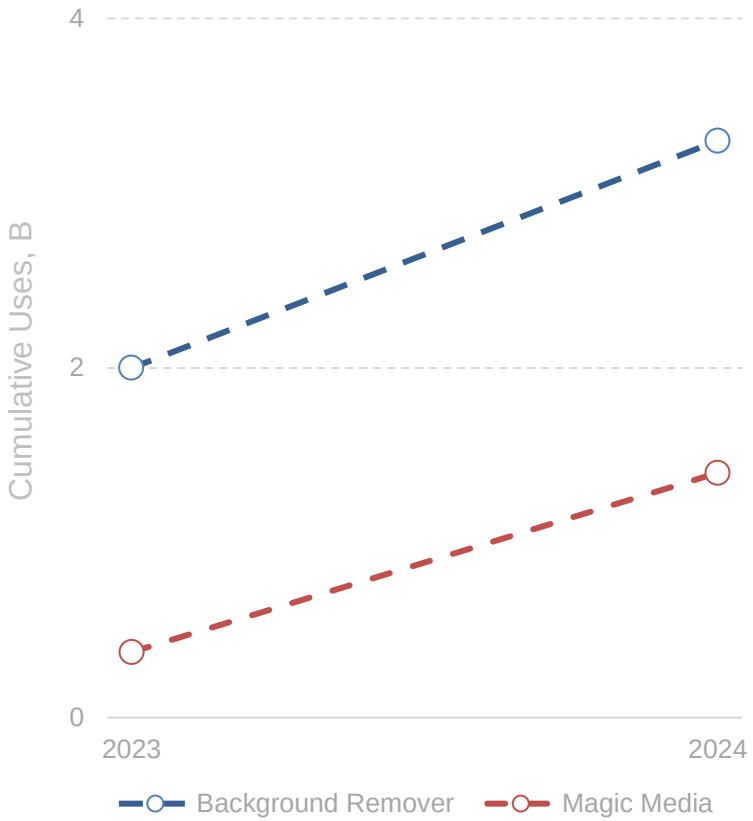
Canva Background Remover & Magic Media – 2023-2024, per Canva



One of our community's favorite Canva features has been the one-click image Background Remover, launched in December 2019...[to] wild success and community love.
- Canva Press Release, 9/22

Magic Media lets you turn your imagination into reality by watching your words transform into stunning, one-of-a-kind images – and now videos and graphics, too…In less than a year since launching Magic Media's text to image, we've been overwhelmed by our community's enthusiastic response, with almost 290 million images being created and applied to a range of practical use cases from social media posts to presentations, business flyers, and even logos.
- Canva Press Release, 10/24

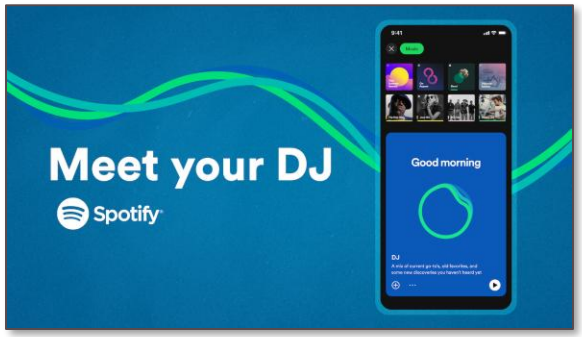
Number of Tool Uses (B): Background Remover & Magic Media



Source: Canva announcements & press releases (2022-2024)

Tech Incumbent AI Rollouts = Spotify – AI DJ (2/23)

Spotify AI DJ – 2/23-5/25, per Spotify

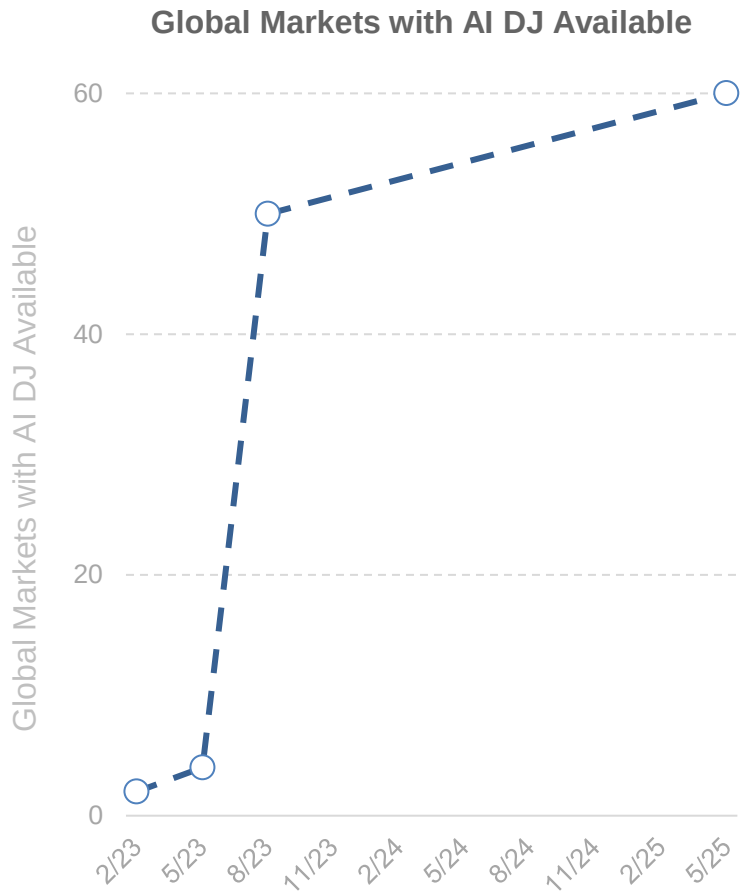


AI DJ and music videos…are truly moving averages…
AI DJ, we're seeing amazing results,
not just on quantitative metrics, but also on
quality metrics, how people feel about Spotify,
what they say they love about Spotify.

- Spotify Co-Founder & CEO Daniel Ek, 11/24

Back in 2018, we said something internally that still
holds true today: machine learning – what most people
called AI back then – was the product…
AI is really the next step in evolution,
where machine learning allows personalization,
AI also allows for real time interactivity and
reasoning on top of your data.

- Spotify Co-President, Chief Product &
Technology Officer Gustav Söderström, 4/25



Source: Company announcements (2/23, 5/23, 8/23, 11/24, 4,25, 5/25)

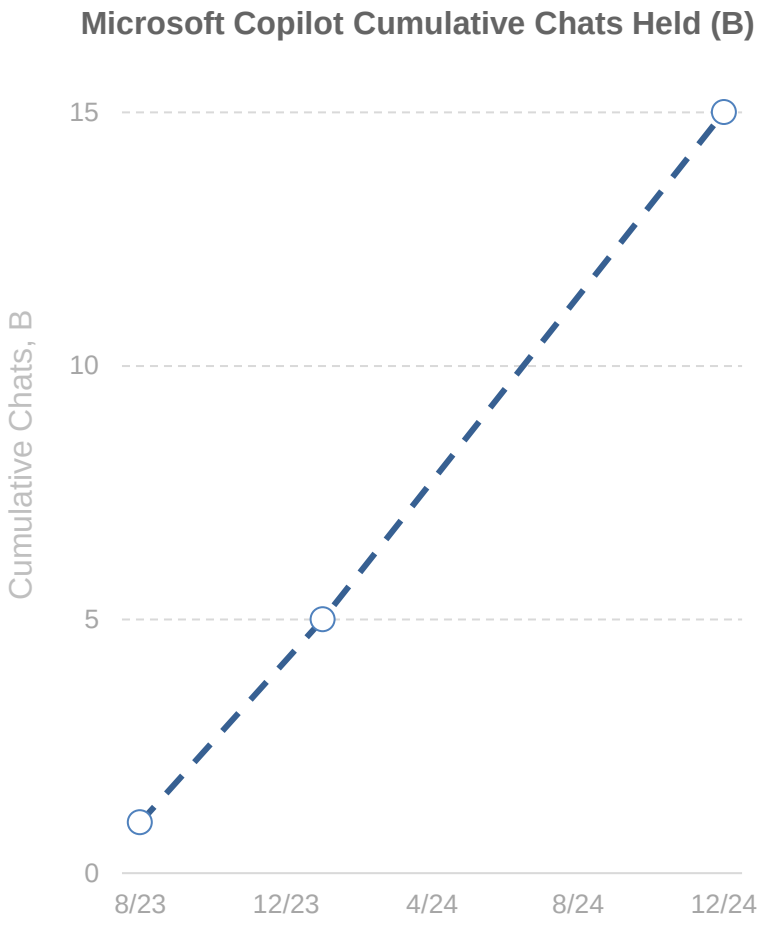
Tech Incumbent AI Rollouts = Microsoft – Copilot (2/23)

Microsoft: Copilot – 8/23-12/24, per Microsoft

To empower people to unlock the joy of discovery, feel the wonder of creation and better harness the world's knowledge, today we're improving how the world benefits from the web by reinventing the tools billions of people use every day, the search engine and the browser.

Today, we're launching an all new, AI-powered Bing search engine and Edge browser, available in preview now at Bing.com, to deliver better search, more complete answers, a new chat experience and the ability to generate content. We think of these tools as an AI copilot for the web.

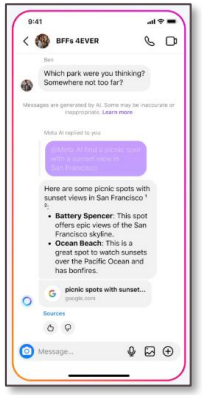
- Official Microsoft Blog, 2/23



Note: We assume zero users in the launch month. We assume 15B cumulative chats as of 12/24 due to Microsoft's 1/24 announcement of 5B cumulative chats, and 12/24 announcement of 10B more chats being held in 2024. We assume the Verge's announcement of 'There have also been over 1 billion chats on Bing Chat' as of 8/23 is wholly inclusive of Copilot chat volumes as of that date. Source: Microsoft announcements & earnings reports, The Verge citing Microsoft disclosures (8/23)

Tech Incumbent AI Rollouts = Meta Platforms – Meta AI (9/23)

Meta Platforms: Meta AI – 11/23-4/25, per Meta Platforms

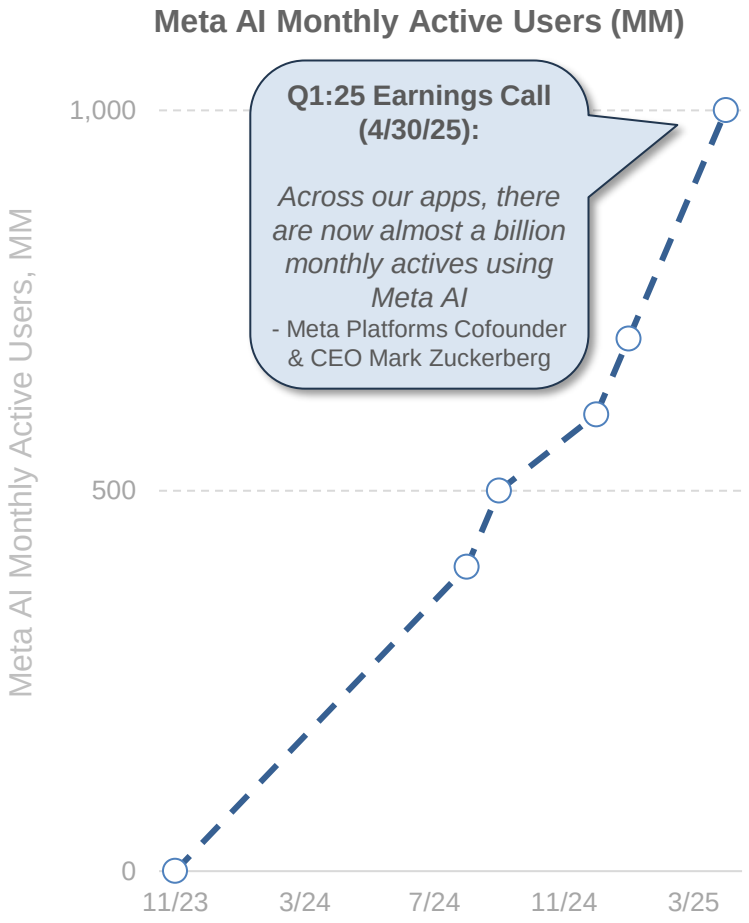


I expect that this is going to be the year when a highly intelligent and personalized AI assistant reaches more than 1 billion people, and I expect Meta AI to be that leading AI assistant. Meta AI is already used by more people than any other assistant…

…I also expect that 2025 will be the year when it becomes possible to build an AI engineering agent that has coding and problem-solving abilities of around a good mid-level engineer…

…Whichever company builds [a high-skill AI engineering agent] first, I think it's going to have a meaningful advantage in deploying it to advance their AI research and shape the field.

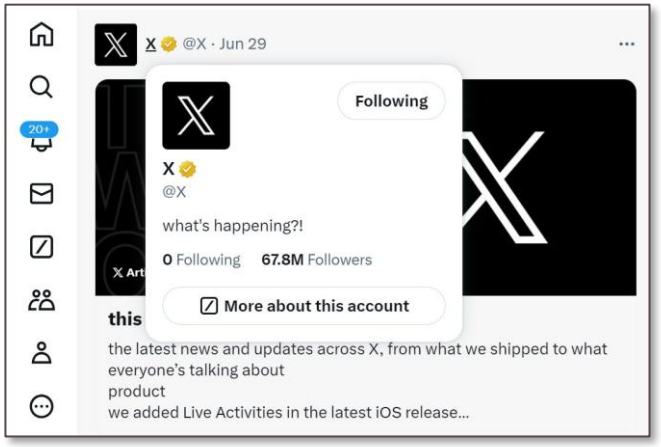
- Meta Platforms CEO Mark Zuckerberg, 1/25



Note: We assume zero users in 11/23 per Meta's 12/23 blog post noting, 'To chat with our AIs, start a new message and select "Create an AI chat" on Instagram, Messenger or WhatsApp. They're now available to anyone in the US.' Source: Meta Platforms announcements & earnings reports

Tech Incumbent AI Rollouts = X – Grok (11/23)

X: Grok – 12/24-4/25, per xAI & Similarweb



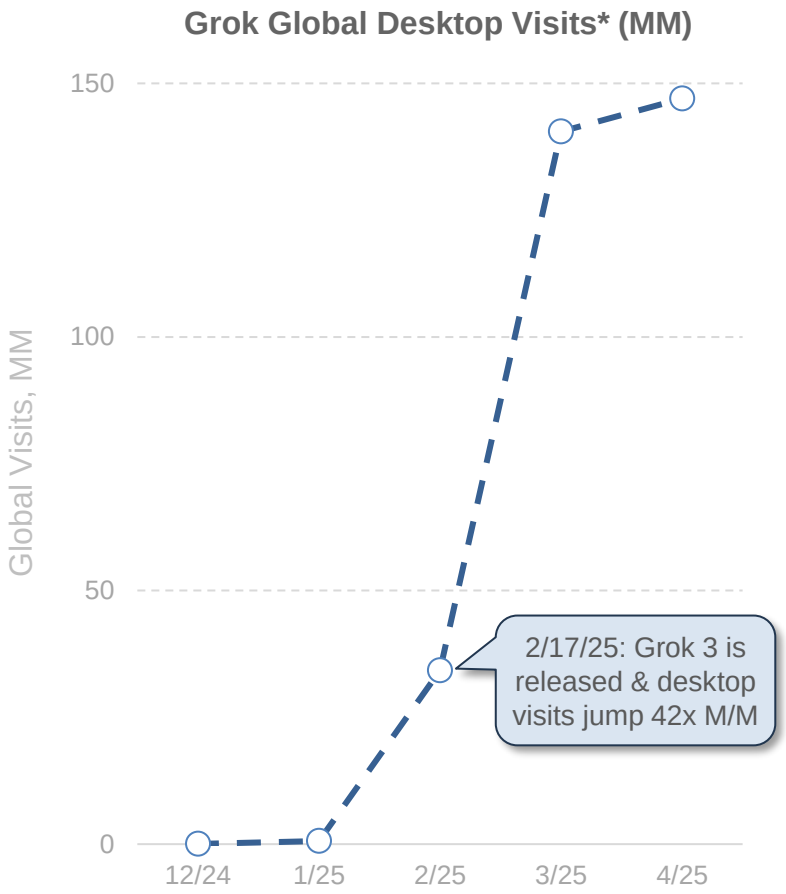
The mission of xAI and Grok is to understand the universe. We want to answer the biggest questions.

- xAI Founder & CEO Elon Musk, 2/25

AI with Grok is getting very good…it's important that AI be programmed with good values, especially truth-seeking values. This is, I think, essential for AI safety…

…Remember these words: We must have a maximally truth-seeking AI.

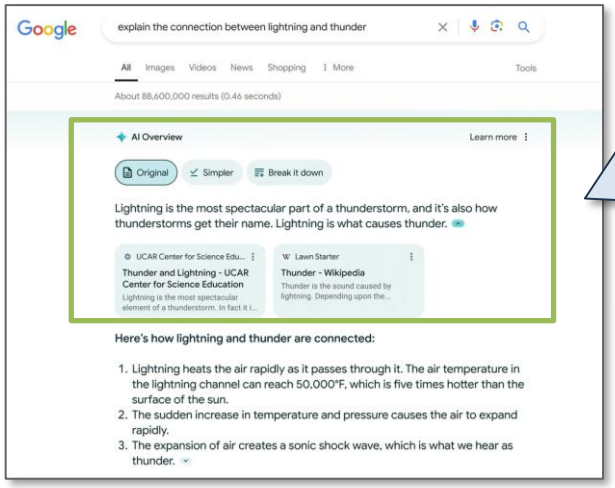
- xAI Founder & CEO Elon Musk, 5/25



**Excludes X visits. China data may be subject to informational limitations due to government restrictions. Source: xAI announcements & investor filings; Elon Musk; Fox News; Similarweb (5/25)*

Tech Incumbent AI Rollouts = Google – Gemini & AI Overviews (12/23)

Alphabet: Gemini & AI Overviews – 3/25-5/25, per Alphabet & *Business Insider*



AI Overviews embedded in Google Search; @ 1.5B MAUs (4/25)

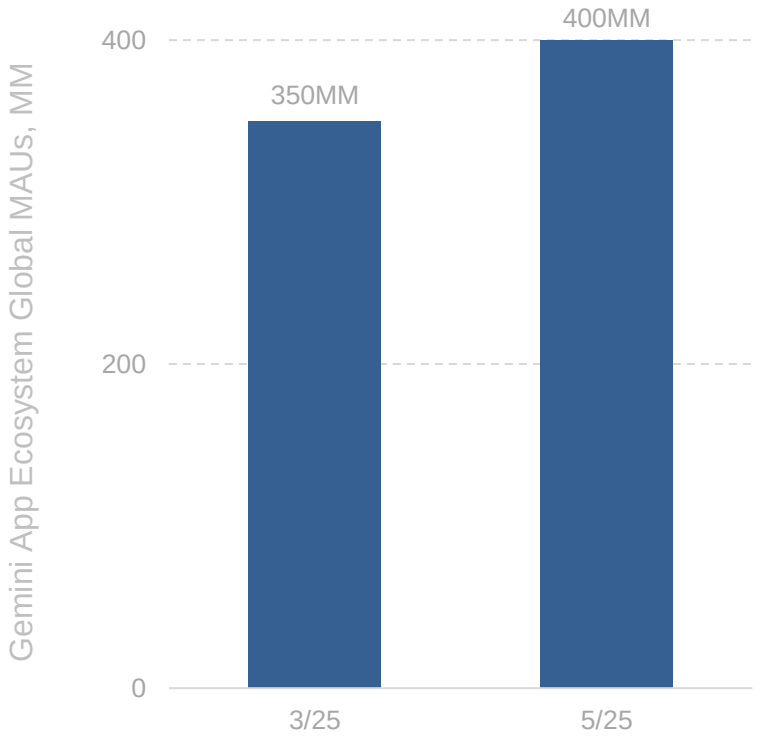
Our differentiated, full stack approach to AI continues to be central to our growth. This quarter was super exciting as we rolled out Gemini 2.5, our most intelligent AI model, which is achieving breakthroughs in performance, and it's widely recognized as the best model in the industry.

- Alphabet CEO Sundar Pichai, 4/25

Google Gemini is a family of multimodal AI models, capable of understanding and generating various types of data including text, code, audio, images, and video.

Source: Google Gemini

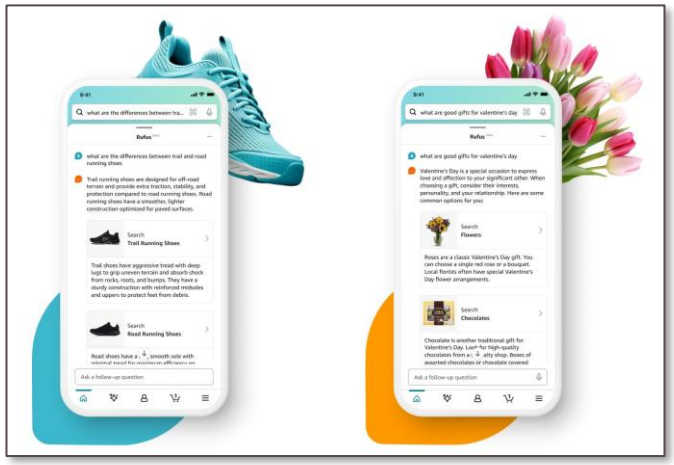
Gemini Chatbot Global MAUs (MM)



Note: Gemini launched 12/23…App launched 2/24. Data shown for apps in Gemini ecosystem. User counts may differ from those as measured by third-party data providers / panels like Similarweb & Sensor Tower as they measure only visits to desktop sites and standalone mobile apps, respectively. Source: Google announcements (4/25 & 5/25) & Business Insider, 'Google's Gemini usage is skyrocketing, but rivals like ChatGPT and Meta AI are still blowing it out of the water' (4/25)

Tech Incumbent AI Rollouts = Amazon – Rufus (2/24)

Amazon: Rufus – 12/22-3/25, per Amazon & Morgan Stanley Estimates

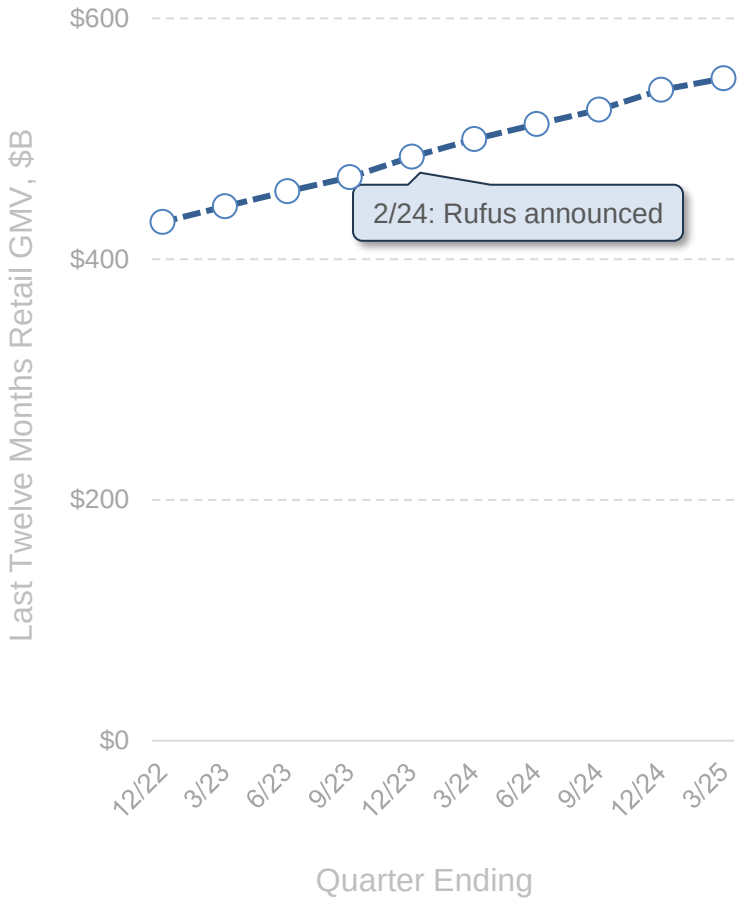


We have so many customers now who just use Rufus to help them find a quick fact about a product. They also use Rufus to figure out how to summarize customer reviews, so they don't have to read 100 customer reviews to get a sense of what people think about that product…the personalization keeps getting much better…

…And so, we expect throughout 2025, that the number of occasions where you're not sure what you want to buy and you want help from Rufus are going to continue to increase and be more and more helpful to customers.

- Amazon CEO Andy Jassy, 2/25

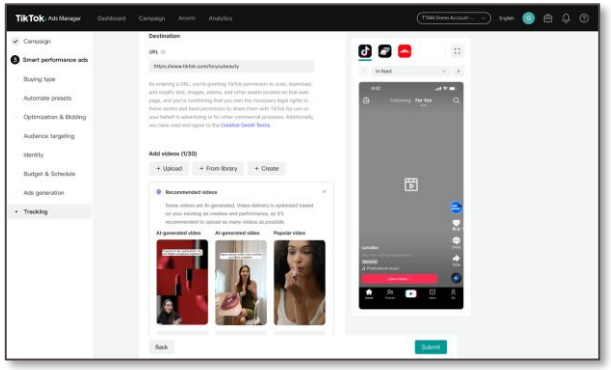
Amazon North America Retail Estimated Gross Merchandise Value ($B), Last 12 Months



Source: Amazon; Morgan Stanley estimates

Tech Incumbent AI Rollouts = TikTok – Symphony AI Assistant (6/24)

TikTok: Symphony Assistant – 1/24-4/25, per TikTok & Similarweb



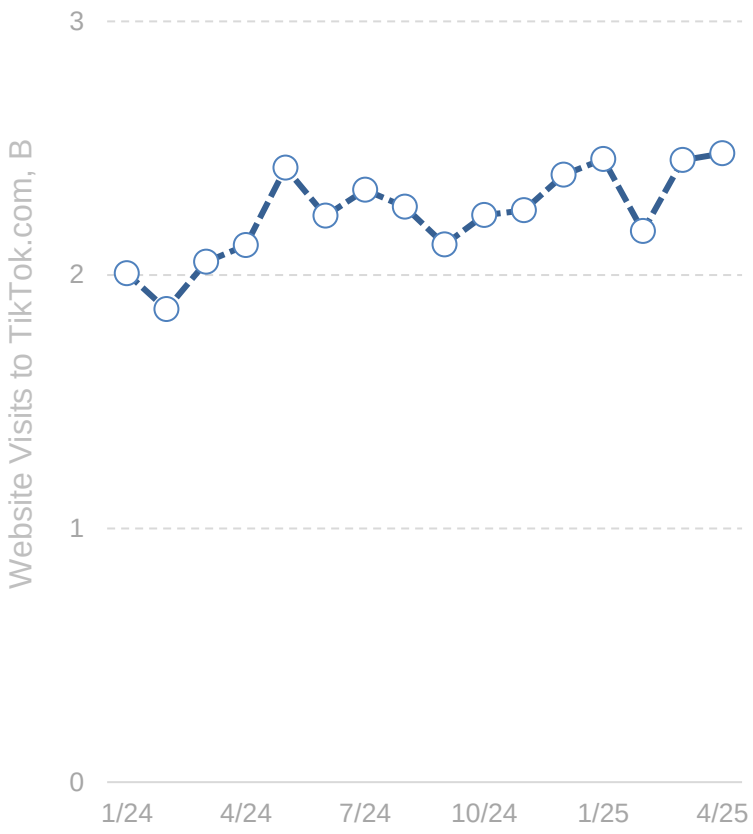
Creativity thrives on TikTok. When brands truly lean into creative bravery and experimentation, they are able to speak directly to their community and invite them to join in the conversation. At TikTok World 2024 we launched Symphony, our suite of ad solutions powered by generative AI…

…With Symphony, businesses of all sizes, creators and agencies can blend human imagination with AI-powered efficiency to help scale content development, creativity, and productivity on TikTok. Research has proven that not only does creating TikTok-first ads boost purchase intent by +37% and brand favorability by +38%, but also 79% of TikTok users show a preference for brands that demonstrate a clear understanding of how to create content specifically for the platform.

- TikTok Press Release, 6/24

Global Website Visits to TikTok.com (B)

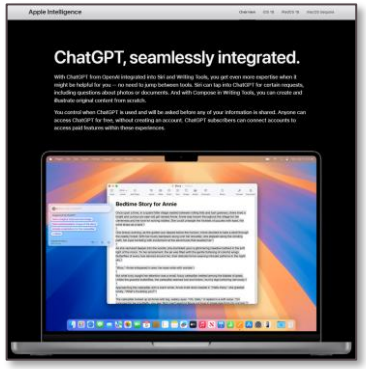
(Where Symphony Assistant is Hosted)



Note: Includes both mobile & desktop website visits. China data may be subject to informational limitations due to government restrictions.
Source: TikTok; Similarweb (5/25)

Tech Incumbent AI Rollouts = Apple – Apple Intelligence (10/24)

Apple: Apple Intelligence – 9/23-3/25, per Apple & IDC Estimates



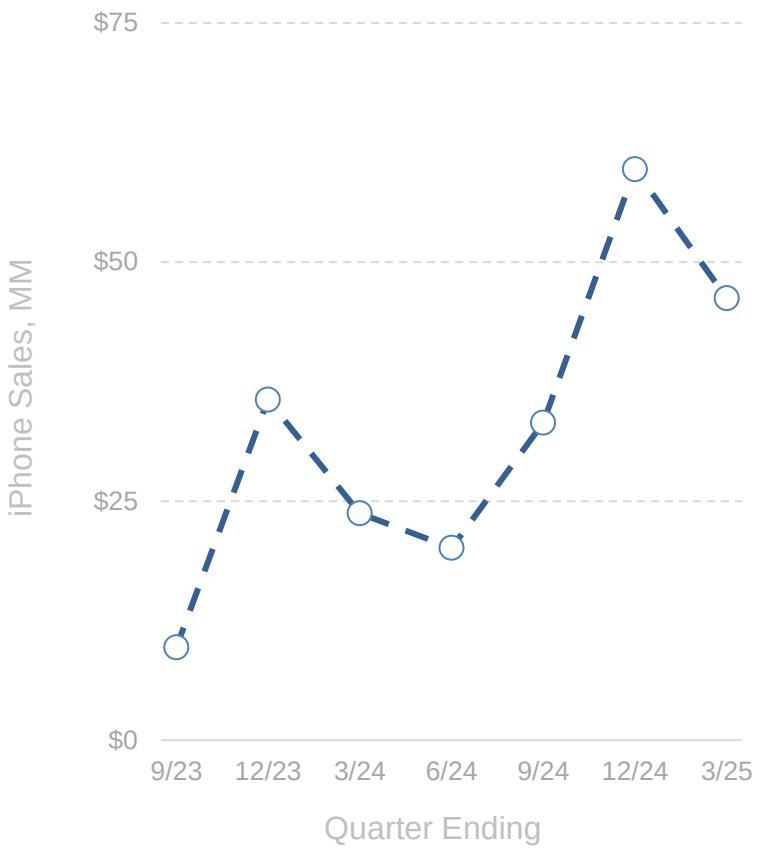
Apple Intelligence builds on years of innovations we've made across hardware and software to transform how users experience our products. Apple Intelligence also empowers users by delivering personal context that's relevant to them. And importantly, Apple Intelligence is a breakthrough for privacy and AI with innovations like private cloud compute…

…[in] the markets where we had rolled out Apple Intelligence…year over year performance on the iPhone 16 family was stronger than those where Apple Intelligence was not available.

- Apple CEO Tim Cook, 1/25

Estimated Global Sales of iPhone 15 Pro / Pro Max & iPhone 16 (MM) – 9/23-3/25

Apple Intelligence-Capable Devices



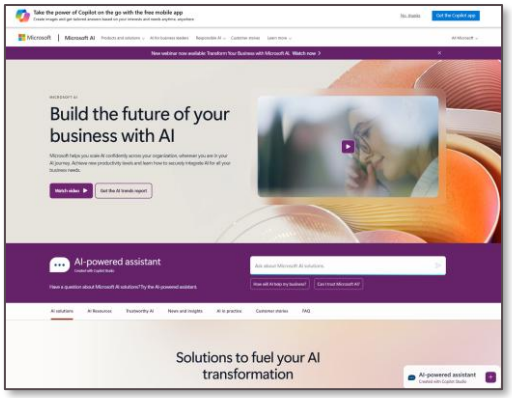
Note: Counts sales of iPhone 15 Pro, iPhone 15 Pro Max, & iPhone 16 devices. Figures are estimates.
Source: Company announcements & investor filings; IDC via Morgan Stanley (4/25)

AI – Tech Incumbents =

Rapid Revenue + Customer Growth

AI Monetization – ‘AI Product’ = Microsoft AI Revenue +175% to $13B Y/Y

Microsoft AI Product Revenue – 2023-2024, per Microsoft

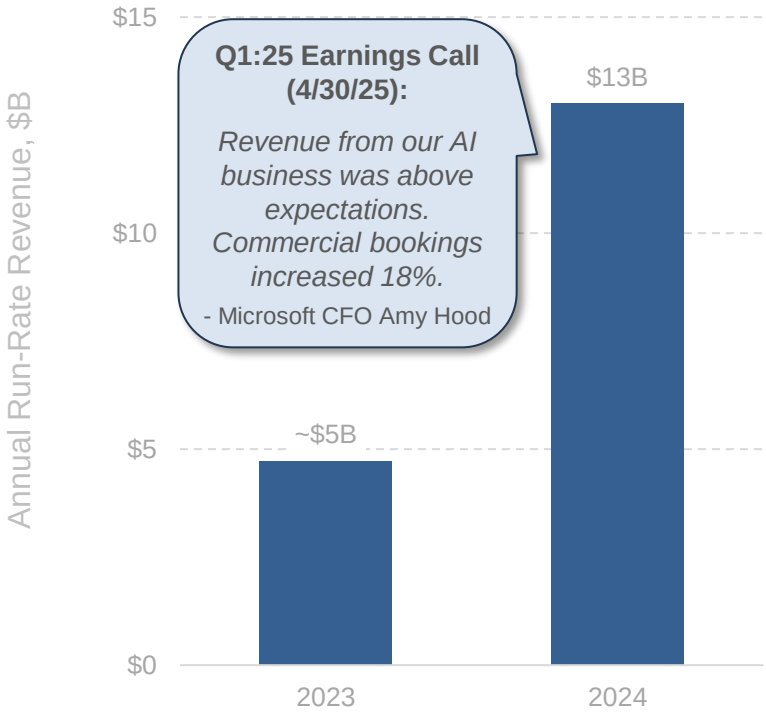


We are innovating across our tech stack and helping customers unlock the full ROI of AI to capture the massive opportunity ahead…

…Already, our AI business has surpassed an annual revenue run rate of $13 billion, up 175% year-over-year.

- Microsoft CEO Satya Nadella, 1/25

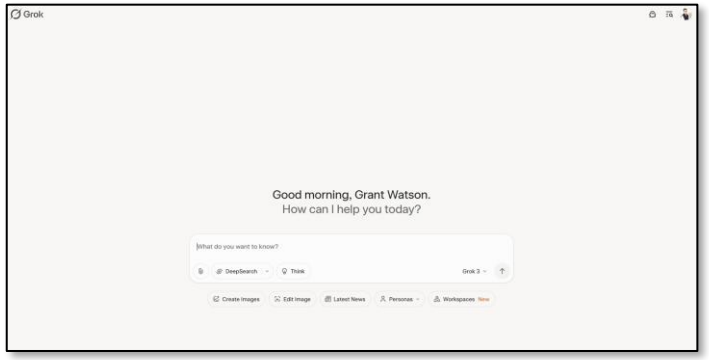
Estimated Microsoft AI Product Annual Run-Rate Revenue ($B)



Note: Microsoft AI revenue likely includes Azure AI services, Microsoft 365 Copilot, GitHub Copilot, Dynamics 365 Copilot, Azure OpenAI Services, and others. Detailed breakdowns not provided on earnings calls. Source: Microsoft Press Release, ‘Microsoft Cloud and AI strength drives second quarter results’ (1/25); & other Microsoft announcements

AI Monetization – Generative Search = xAI Annualized Revenue Up Materially in 2025

xAI: Generative Search, per xAI & *The Wall Street Journal*

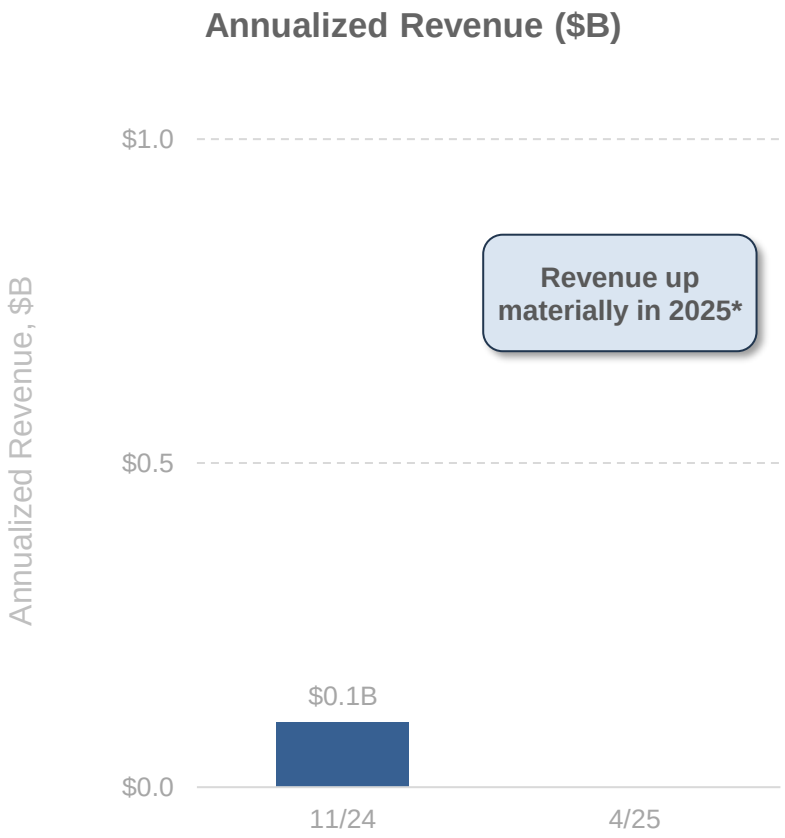


We are pleased to introduce Grok 3, our most advanced model yet: blending strong reasoning with extensive pretraining knowledge. Trained on our Colossus supercluster with 10x the compute of previous state-of-the-art models, Grok 3 displays significant improvements in reasoning, mathematics, coding, world knowledge, and instruction-following tasks.

- xAI Grok 3 Press Release, 2/25

[Grok is a] maximally truth-seeking AI, even if that truth is sometimes at odds with what is politically correct.

- xAI Founder & CEO Elon Musk, 2/25



**Select media reports have xAI revenue being as high as $1B as of 4/25. Source: xAI (2/25); The Wall Street Journal, 'Elon Musk's xAI Startup Is Valued at $50 Billion in New Funding Round' (11/24) (link); CNBC, 'Musk says he's looking to put 'proper value' on xAI during investor call, sources say' (4/25) (link)*

AI Monetization – AI Services = Palantir USA Commercial Customers +65% to 432 Y/Y

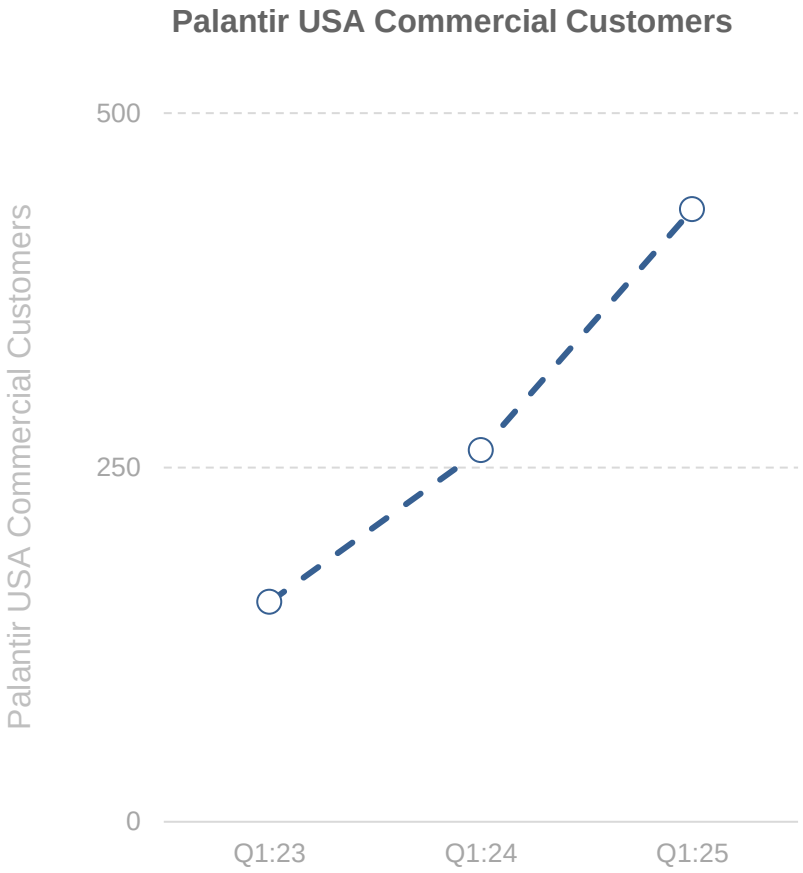
Palantir USA Commercial Customers – Q1:23-Q1:25, per Palantir

We achieved a $1 billion annual run rate in our US commercial business for the first time as AIP [Artificial Intelligence Platform] continues to drive both new customer conversions and existing customer expansions in the US..

- Palantir CFO David Glazer , 5/25

As AI models progress and improve, we continue enabling our customers to maximally leverage these models in production, capitalizing upon the rich context within the enterprise through the Ontology. We remain differentiated in our elite execution to deliver quantified exceptionalism for our customers, ever widening their advantage over the AI have-nots.

- Palantir CRO & Chief Legal Officer Ryan Taylor, 5/25



Source: Palantir

AI Monetization Possibilities – Enterprise =

Horizontal Platform
& / Or
Specialized Software?

To understand where enterprise AI monetization is headed, it helps to ask where software itself is consolidating.

For decades, business software followed a familiar pattern: build a specialized tool, sell it to a narrow user base, and scale up within a vertical. This was the age of vertical SaaS – Toast for restaurants, Guidewire for insurance, Veeva for life sciences...Each tool solved a deep, narrow problem. But with the rise of foundation models and generative AI, others are gunning for these prizes.

Enter the horizontal enterprise platforms – horizontal layers that combines AI-native productivity, search, communication, and knowledge management into one unified interface. Think of it as Slack meets Notion meets ChatGPT, all in one platform.

Horizontal enterprise platforms could usher in a new form of monetization: not by selling siloed software licenses, but by charging for intelligence, embedded throughout the stack. The value shifts from tools to outcomes – from CRMs to automated deal summaries, from service desks to AI-powered resolution flows.

These horizontal capabilities are still early, but they're already being harnessed by incumbents and upstarts alike. Microsoft is integrating Copilot across the stack. Zoom and Canva are layering GenAI into user-facing workflows, while Databricks is infusing GenAI into its data and developer stack. Meanwhile, startups like Glean are betting on AI-first workflows to challenge the suite model…

…But specialist vendors aren’t standing still. If anything, they’re absorbing AI faster – embedding copilots, automating workflows, and fine-tuning models on proprietary industry data. These platforms already have the workflows, the trust, and the structured data that AI thrives on. That gives them a head start in deploying domain-specific intelligence – AI that doesn’t just summarize a meeting, but flags regulatory risks, optimizes pricing in real time, or drafts FDA-compliant documentation. In many cases, their incumbency becomes their advantage: they can roll out AI as a feature, not a product, and monetize it without changing the buying motion.

The next chapter of AI monetization may not be a winner-take-all battle, but a convergence. Horizontal platforms will push breadth, stitching together knowledge across functions; specialists will push depth, delivering AI that speaks the language of compliance, contracts, and customer intent.

The question isn’t whether platforms or specialists win – it’s who can abstract the right layer, own the interface, and capture the logic of work itself. In the AI era, monetization won’t just follow usage – it will follow attention, context, and control.

AI Monetization Possibilities – Enterprise =

Horizontal Platform

& / Or

Specialized Software?

Horizontal Enterprise Platform =

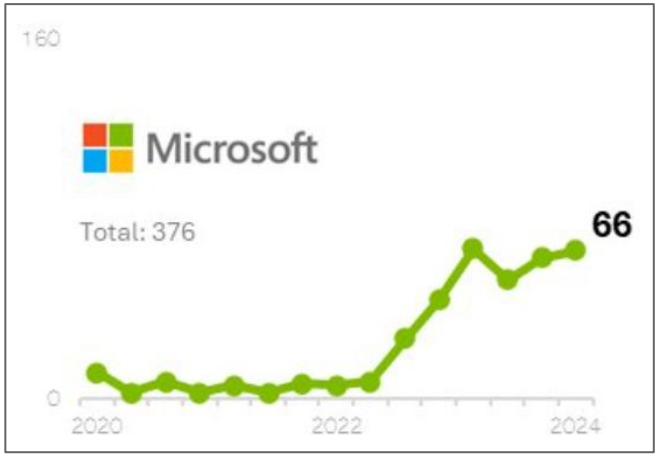
SaaS Incumbents

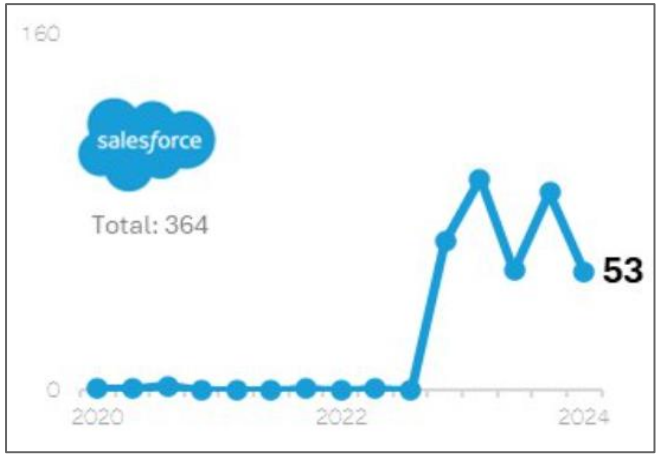
Or

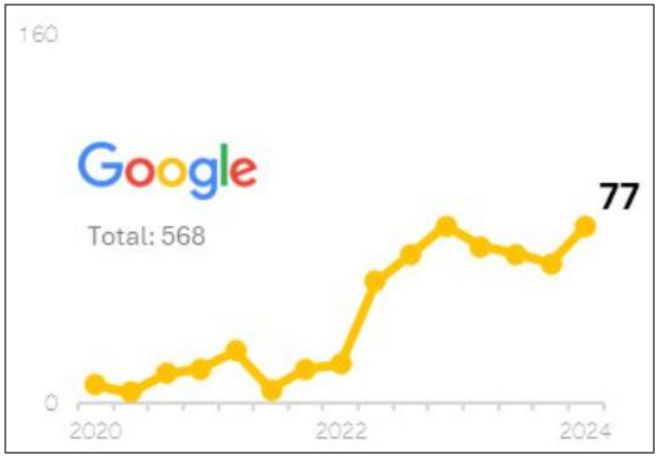
Large Language Model Challengers?

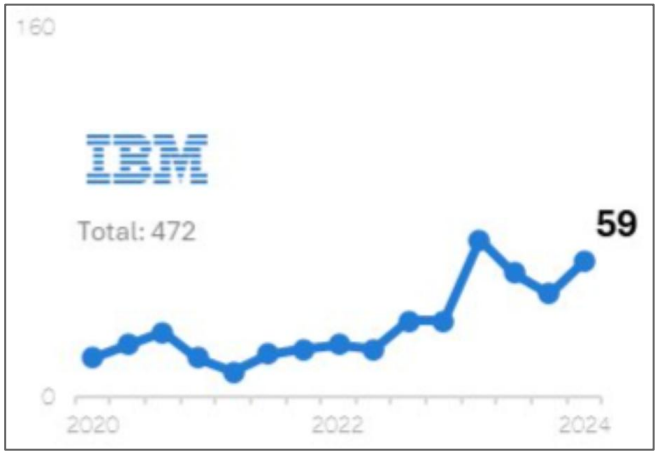
Enterprise SaaS Incumbent AI Rollouts = Broad & Steady Cadence

Number of Mentions of 'AI' on Corporate Earnings Calls – Q1:20-Q1:24, per Uptrends.ai





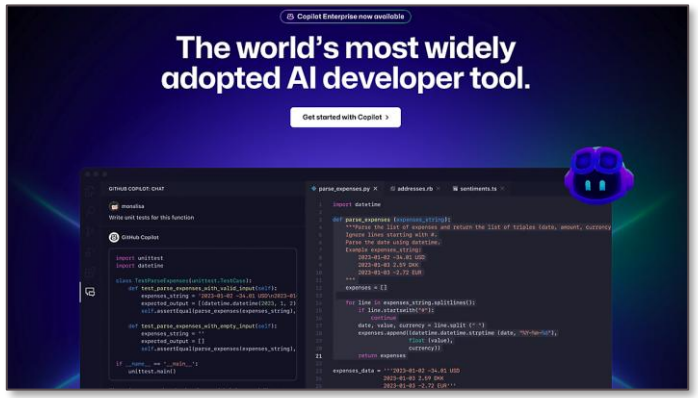




Source: Uptrends.ai (6/24), company announcements & investor filings

Enterprise SaaS Incumbent AI Rollouts = Microsoft GitHub Copilot – 6/22

Microsoft GitHub Copilot – 6/17-6/24, per GitHub, Microsoft & Wells Fargo

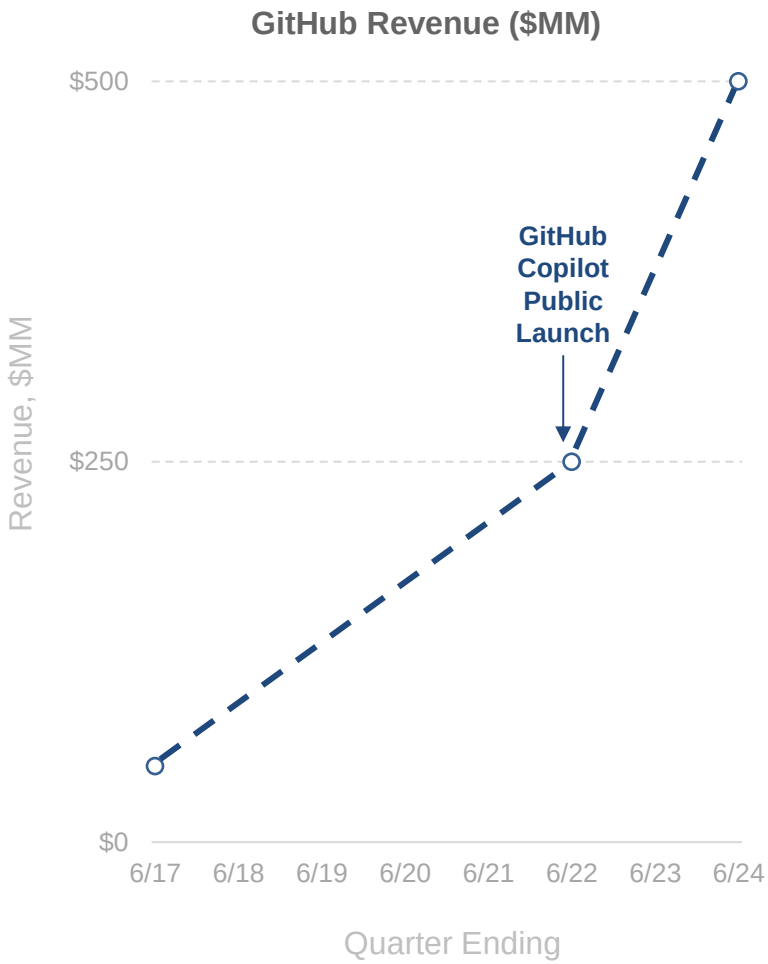


GitHub Copilot is by far the most widely adopted AI-powered developer tool. Just over two years since its general availability, more than 77,000 organizations – from BBVA, FedEx, and H&M, to Infosys and Paytm – have adopted Copilot, up 180% year-over-year.

- Microsoft CEO Satya Nadella, 7/24

We have been delighted by the early response to GitHub Copilot and vs. Code with more than 1 million sign-ups in just the first week post launch. All up, GitHub now is home to 150 million developers, up 50% over the past two years.

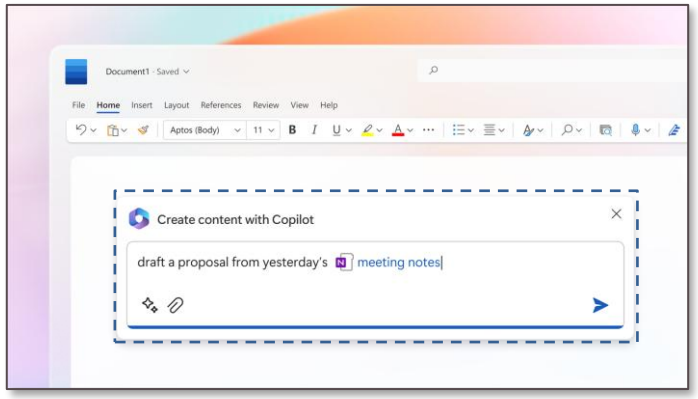
- Microsoft CEO Satya Nadella, 1/25



Note: GitHub revenue is disclosed irregularly; 3 datapoints are from company leadership's disclosures. Public developer launch date shown. GitHub reports annualized revenue; here, we translate this to quarterly revenue. Source: Company announcements & investor filings

Enterprise SaaS Incumbent AI Rollouts = Microsoft 365 Copilot – 3/23

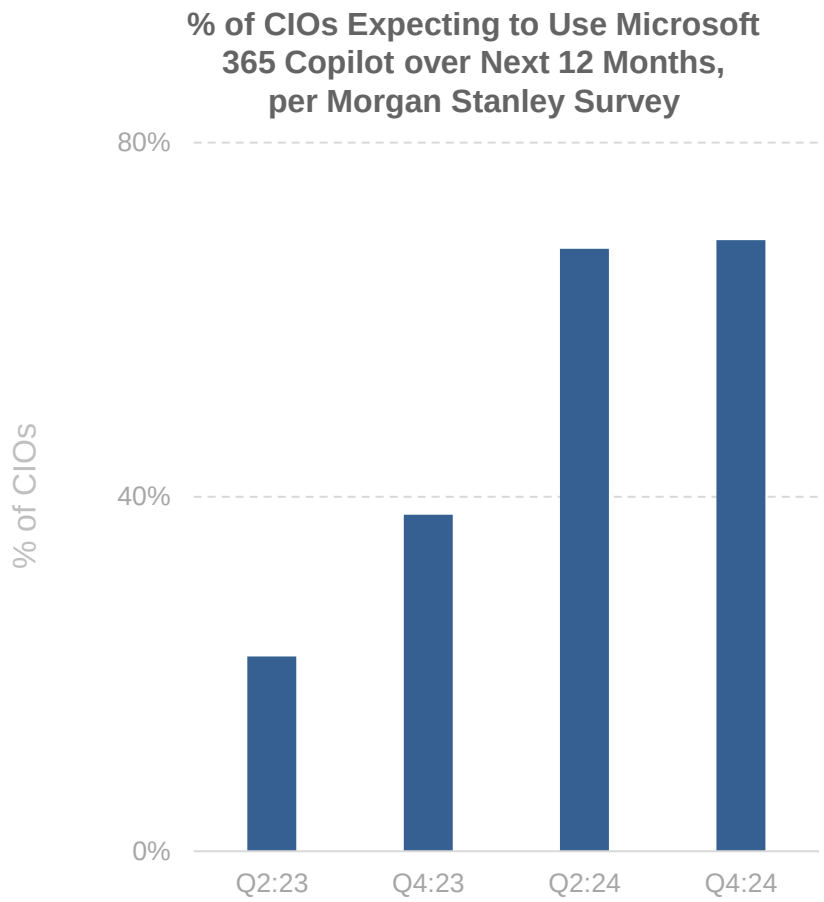
Microsoft 365 Copilot – Q2:23-Q4:24, per Microsoft & Morgan Stanley



We are seeing accelerated customer adoption across all deal sizes as we win new Microsoft 365 Copilot customers and see the majority of existing enterprise customers come back to purchase more seats. When you look at customers who purchased Copilot during the first quarter of availability, they have expanded their seat collectively by more than 10x over the past 18 months. And overall, the number of people who use Copilot daily, again, more than doubled quarter over quarter.

Employees are also engaging with Copilot more than ever. Usage intensity increased more than 60% quarter over quarter, and we are expanding our TAM with Copilot Chat, which was announced earlier this month.

- Microsoft CEO Satya Nadella, 1/25



Note: N=61 CIOs in the USA & EU. Microsoft 365 Copilot was announced in 3/23 but was not made generally available for enterprise customers until 11/23.
Source: Company announcements & investor filings, Morgan Stanley, '4Q24 Preview – Can Microsoft Add Clarity to the AI Monetization Question?' (7/24)

Enterprise SaaS Incumbent AI Rollouts = Adobe Firefly – 3/23

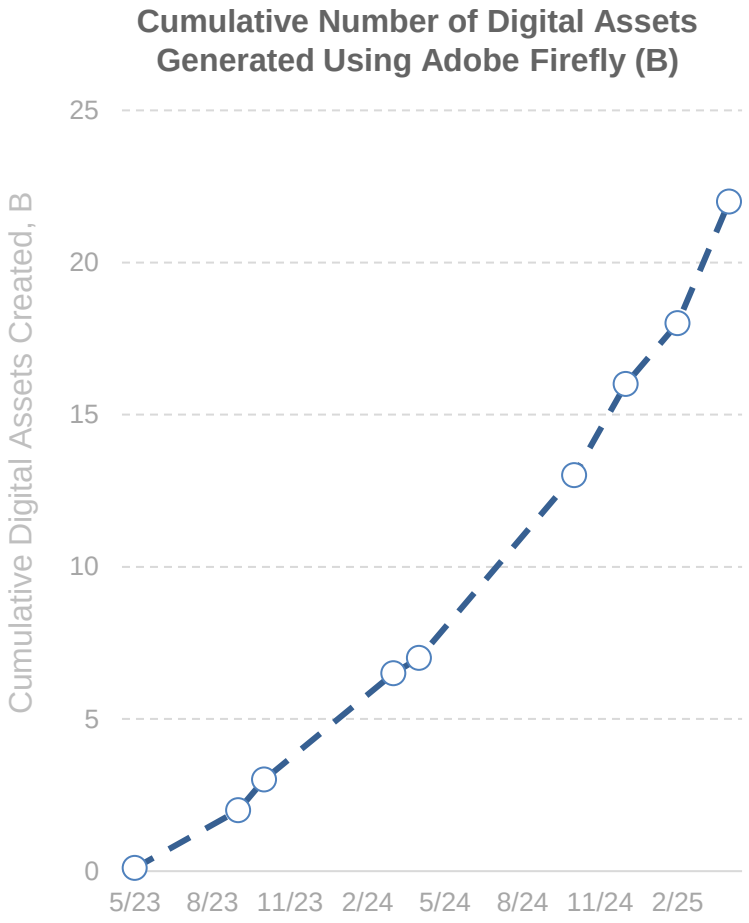
Adobe Firefly – 5/23-4/25, per Adobe

The release of the Adobe FireFly video model in February, a commercially safe generative AI video model, has been very positively received by brands and creative professionals…

…User engagement has been strong with over 90% of paid users generating videos…

…We're delighted with the early interest in these new offerings. Other creative professional and creator highlights include, continued strong adoption of GenAI in our products with Photoshop GenAI monthly active users at approximately 35% and Lightroom GenAI monthly active users at 30%. Users have generated over 20 billion assets with Firefly.

- Adobe President of Digital Media David Wadhwani, 3/25



Note: We assume zero users in the launch month. Adobe Firefly was released as a public beta in March 2023.
Source: Adobe announcements (9/23, 10/23, 3/24, 4/24, 10/24, 12/24, 2/25)

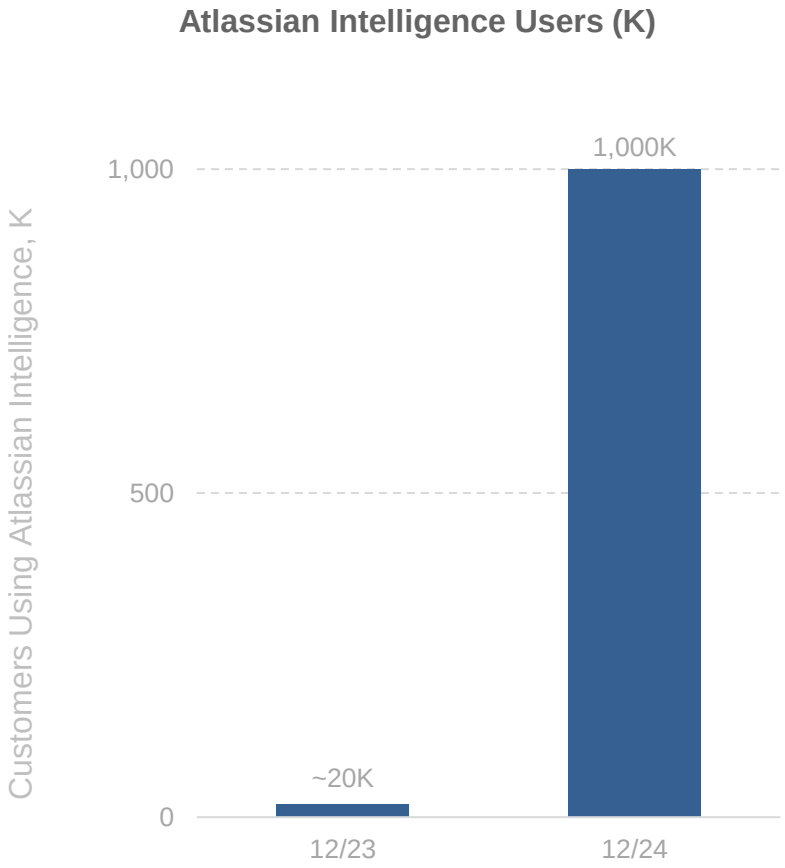
Enterprise SaaS Incumbent AI Rollouts = Atlassian Intelligence – 4/23

Atlassian Intelligence – 12/23-12/24, per Atlassian

Today, more than 1 million monthly active users are utilizing our Atlassian intelligence features to unlock enterprise knowledge, supercharge workflows, and accelerate their team collaboration. These features are clearly delivering value as we've seen a number of AI interactions increase more than 25x year over year…

…Atlassian Intelligence [saw a] 25x improvement in the number of features used over the last year.

- Atlassian Co-Founder & Co-CEO Michael Cannon, 2/25



Note: 12/23 users includes beta users. We assume 20,000 users based on Atlassian's disclosure that 'Nearly 10% of Atlassian's 265,000+ customers have already leveraged Atlassian Intelligence through our beta program.' Source: Atlassian announcements (4/23, 12/23, 12/24)

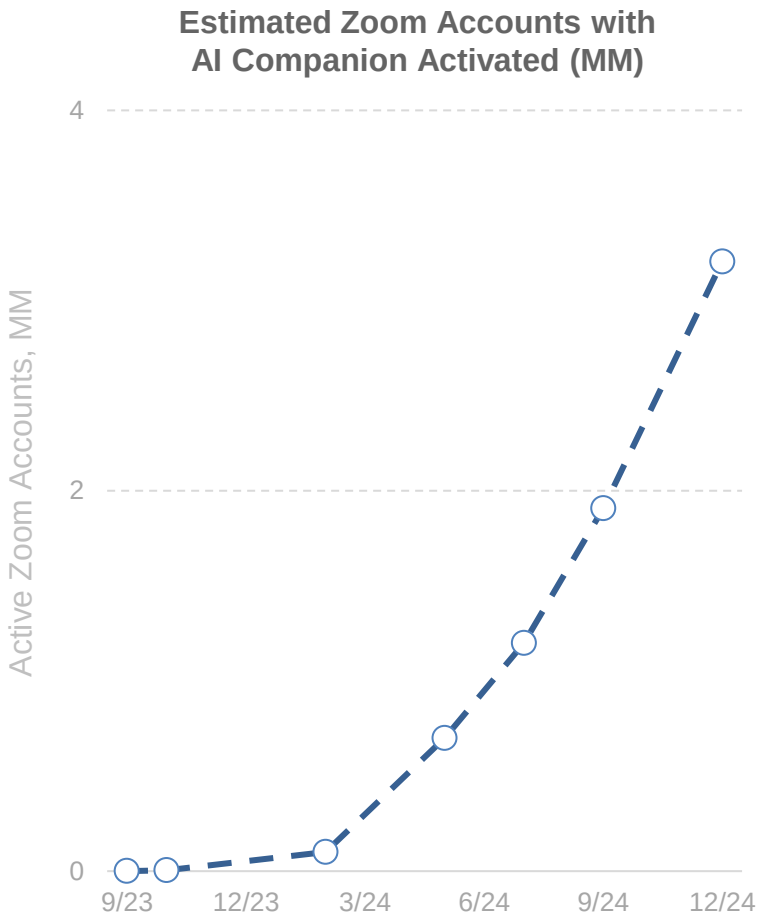
Enterprise SaaS Incumbent AI Rollouts = Zoom AI Companion – 9/23

Zoom AI Companion – 9/23-12/24, per Zoom

Growth in monthly active users of Zoom AI Companion accelerated to 68% quarter over quarter, demonstrating the real value AI is providing customers.
Zoom AI Companion has emerged as a driving force behind our transformation into an AI-first company…

…As part of AI Companion 2.0, we added advanced agentic capabilities, including memory, reasoning, orchestration, and seamless integration with Microsoft and Google services.

- Zoom Founder & CEO Eric Yuan, 2/25



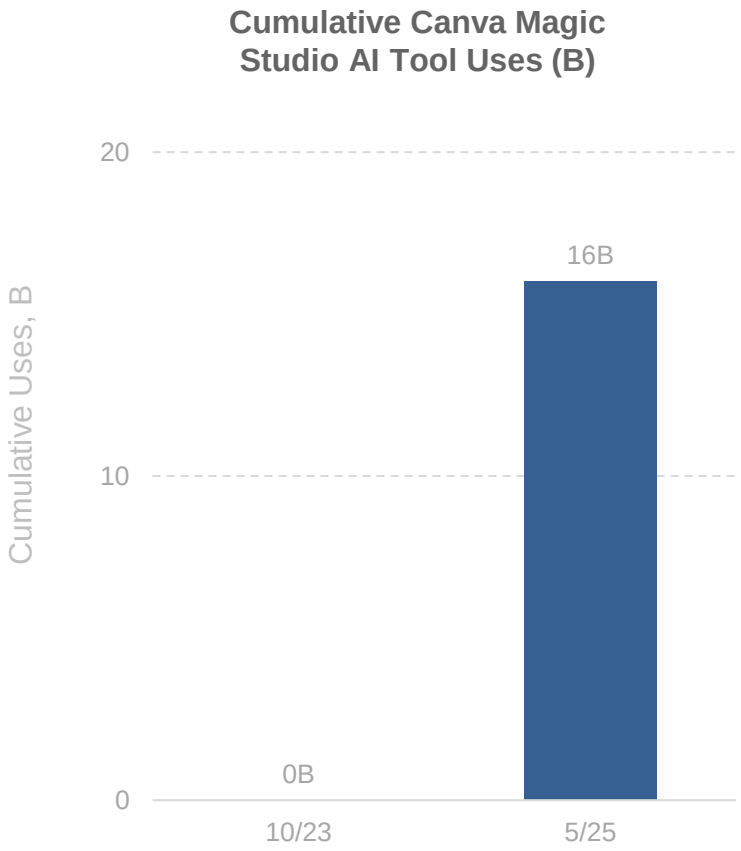
Note: AI Companion MAUs are estimates based on company disclosures. As of 7/30/24, Zoom disclosed they had 1.2MM accounts with AI Companion activated. In Q3 2024, they disclosed 59% Q/Q growth in active accounts; in Q4 2024, they disclosed further 68% Q/Q growth. We assume zero users in the launch month.
Source: Zoom announcements (9/23. 10/23, 2/24, 5/24, 7/24, 9/24, 12/24)

Enterprise SaaS Incumbent AI Rollouts = Canva Magic Studio – 10/23

Canva Magic Studio – 10/23-5/25, per Canva

With Magic Studio there's no need to toggle between multiple AI tools or learn lots of different software – all the best of AI is at your fingertips. Created for the 99% of the world without complex design skills, Magic Studio is jam-packed with easy-to-use AI-powered features across every part of Canva to help you work smarter.
- Canva Press Release, 10/23

Magic Studio is designed to supercharge creativity across our entire community – from enterprise teams to educators and nonprofits. Its easy-to-use AI features are woven into every part of Canva, enabling anyone to spark inspiration, streamline workflows, and scale their content. In fact, our AI tools have been used more than 10 billion times to date.
- Canva Press Release, 10/24



Note: We assume zero users in the launch month. Source: Canva announcements (10/23, 10/24, 5/25)

Enterprise SaaS Incumbent AI Rollouts = Salesforce Agentforce – 9/24

Salesforce Agentforce – 12/24-2/25, per Salesforce

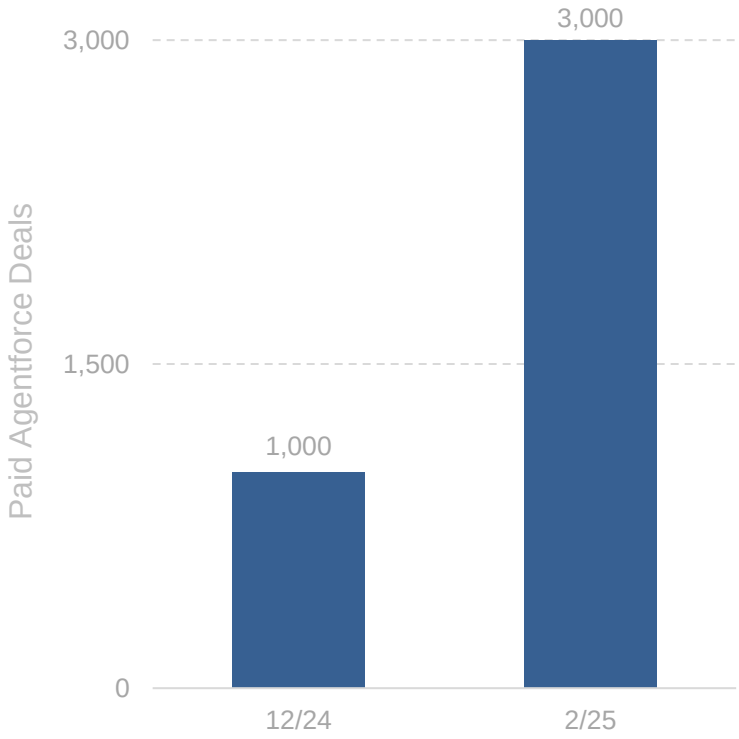
We ended this year with $900MM in Data Cloud and AI ARR. It grew 120% year over year. We've never seen products grow at these levels, especially Agentforce…

…Just 90 days after it went live, we've already had 3,000 paying Agentforce customers who are experiencing unprecedented levels of productivity, efficiency, and cost savings…

…Data Cloud is the fuel that powers Agentforce and our customers are investing in it. And Data Cloud surpassed 50 trillion, that's trillion with a T, records, doubling year over year as customers increase their consumption and investment in our data platform.

- Salesforce Co-Founder & CEO Mark Benioff, 2/25

Number of Paid Agentforce Deals Signed



Note: Agentforce was announced on 9/12/24 but became generally available on 10/29/24. We assume zero users in the launch month.
Source: Salesforce announcements (10/24, 12/24, 2/25)

Horizontal Enterprise Platform =

SaaS Incumbents
Or
Large Language Model Challengers?

OpenAI ChatGPT = Potential Horizontal Enterprise Platform?...

OpenAI = Next-Gen All-in-One Enterprise Platform?

Microsoft Office Suite	OpenAI ChatGPT
Microsoft 365	ChatGPT
9 Applications	1 Application
400MM Paid Users Over 34 Years *1990-2024*	20MM Paid Users Over **2.5 Years** *11/22-4/25*

Source: Microsoft (1/24), Office365 Pros, OpenAI, The Information (4/25) (link)

…OpenAI ChatGPT = Potential Horizontal Enterprise Platform?

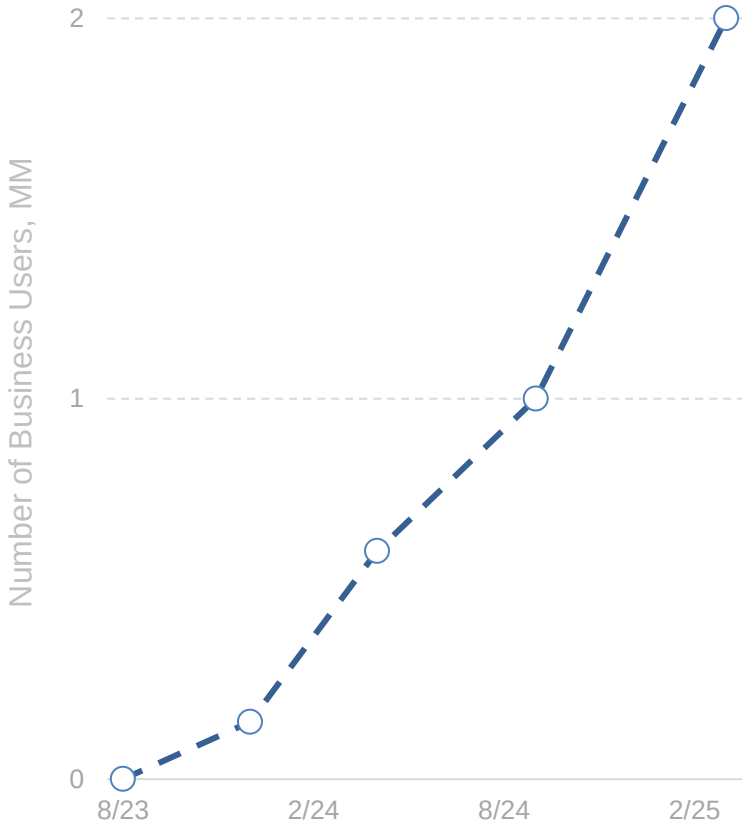
ChatGPT Enterprise – 8/23-3/25, per OpenAI, *Bloomberg*, *Reuters*, & *The Wall Street Journal*

Since ChatGPT's launch just nine months ago, we've seen teams adopt it in over 80% of Fortune 500 companies. We've heard from business leaders that they'd like a simple and safe way of deploying it in their organization. Early users of ChatGPT Enterprise…are redefining how they operate and are using ChatGPT to craft clearer communications, accelerate coding tasks, rapidly explore answers to complex business questions, assist with creative work, and much more.

ChatGPT Enterprise removes all usage caps and performs up to two times faster [vs. ChatGPT Free]…
…ChatGPT Enterprise also provides unlimited access to advanced data analysis, previously known as Code Interpreter.

- ChatGPT Enterprise Release Statement, 8/23

Number of ChatGPT Business Users (MM)
(Includes Enterprise / Team / Education)



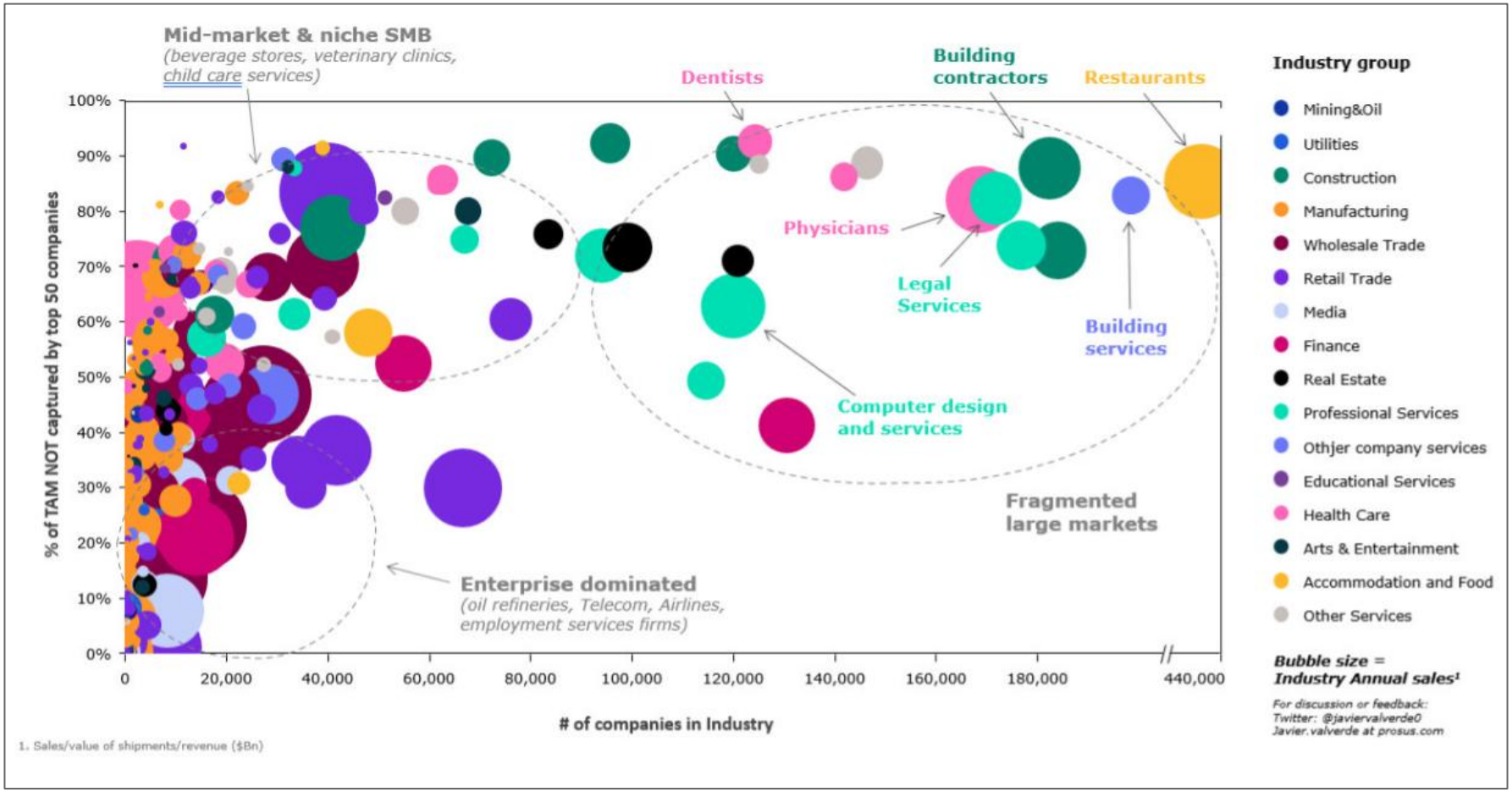
Note: We assume zero users in the launch month. Source: OpenAI announcements (12/23, 4/24, 9/24, 3/25), Bloomberg (4/24), Reuters (9/24), The Wall Street Journal (3/25)

AI Monetization Possibilities – Enterprise =

Horizontal Platform
& / Or
Specialized Software?

AI Monetization – Enterprise = Specialized Software Opportunities in Fragmented Markets, per Prosus

USA Industries by Number of Companies & Market Share – 2024, per Prosus



Source: Prosus, 'The Timeless Appeal of Vertical SaaS' (3/24)

AI-Enabled Specialized Software @
Large Service Industries =

Growing Very Quickly…

Software Engineering
Product Development
Healthcare
Legal
Customer Service
Financial Services

AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Software Engineering

Specialized AI – Software Engineering (Code Editor) = Anysphere Cursor AI ARR @ $1MM to $300MM in Twenty-Five Months

Anysphere Cursor AI – 3/23-4/25, per Anysphere

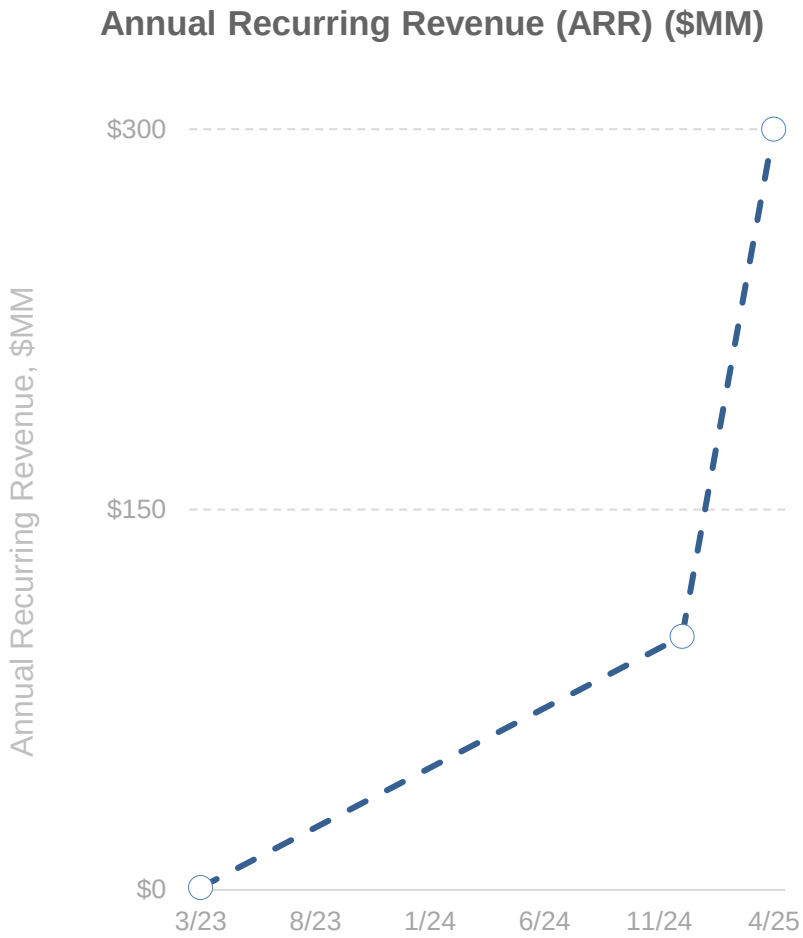
Something beautiful is happening to code…our aim with Cursor is to continue to lead this shift, by building a magical tool that will one day write all the world's software…

…Already, in Cursor, hours of hunting for the right primitives are being replaced by instant answers. Mechanical refactors are being reduced to single 'tabs.' Terse directives are getting expanded into working source. And thousand-line changes are rippling to life in seconds.

- Anysphere Press Release (8/24)

…We're delighted to report that Cursor is now used by millions of programmers as their editor of choice. Our proprietary models now generate more code than almost any LLMs in the world and edit over a billion characters per day.
Our business is large and fast growing, having exceeded $100MM in recurring revenue.

- Anysphere Team (8/24 & 1/25)



Note: Cursor launched in 4/23. We show 3/23 as the first datapoint with an assumed $0 in ARR. Source: Cursor / Anysphere (8/24, 11/24 & 1/25), Anysphere Co-Founder & CEO Michael Truell via Lenny's Newsletter, 'The rise of Cursor: The $300M ARR AI tool that engineers can't stop using' (5/1/25)

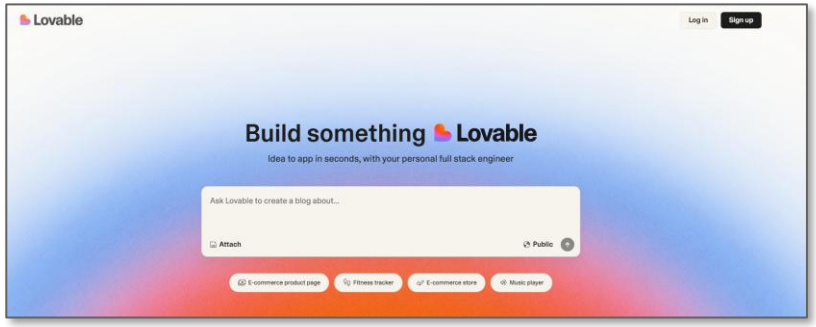
AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Product Development

Specialized AI – Product Development (No-Code Product-Building) = Lovable ARR +13x to $50MM in Five Months

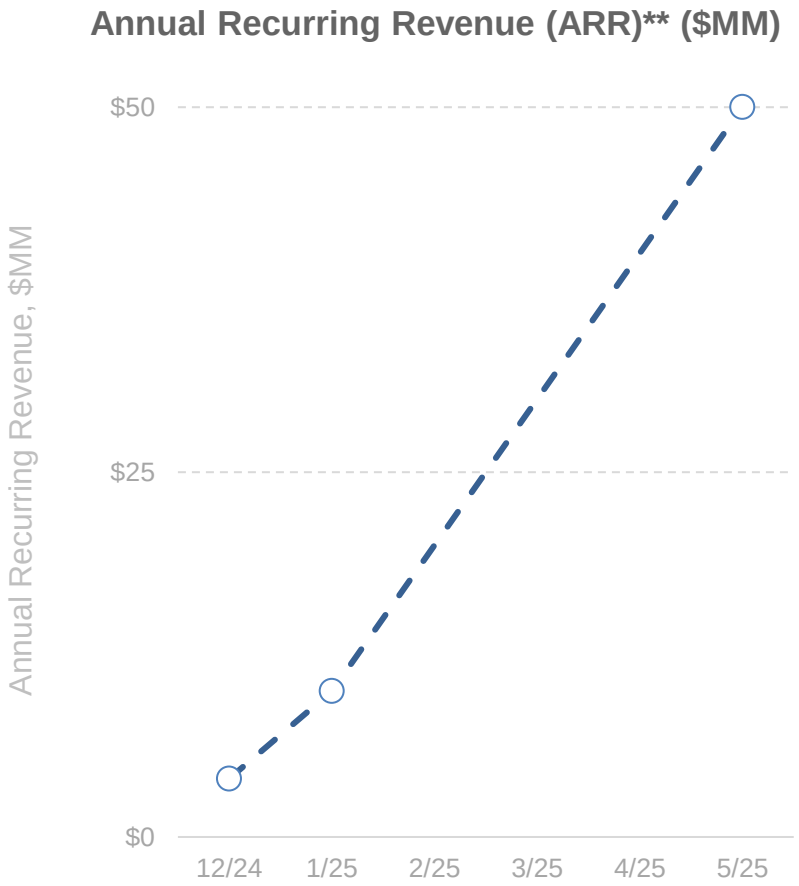
Lovable – 12/24-5/25



Note: Lovable is an AI-powered application development platform that enables users to create full-stack web applications by describing their ideas in natural language. The platform translates these descriptions into functional applications, handling frontend and backend code generation, database integration, and deployment.

The opportunity here is immense. We are on the verge of a paradigm shift where the barriers to building software-based products disappear. Now, anyone can become an entrepreneur, launch a product and build a business in minutes.

- Frederik Cassel, Creandum, 'Backing Lovable: Move Fast and Make Things,' 2/25*



**Per Creandum website. **From Lovable Co-Founder & CEO Anton Osika's LinkedIn posts & podcast appearances. Source: Lovable (5/25), Creandum (2/25)*

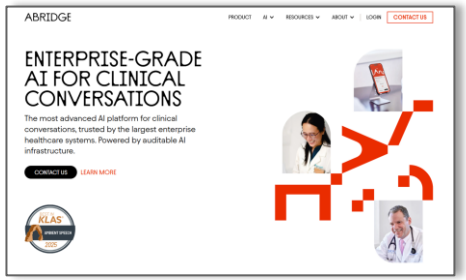
AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Healthcare

Specialized AI – Healthcare (Clinical Conversations) = Abridge @ $50MM to $117MM CARR in ~Five Months

Abridge – 10/24-3/25, per Abridge & *The Information*



Yazdi Bagli, Kaiser's EVP of IT and enterprise business services, said he believes [Kaiser Permanente's] Abridge partnership is one of the largest generative AI deployments in health care…

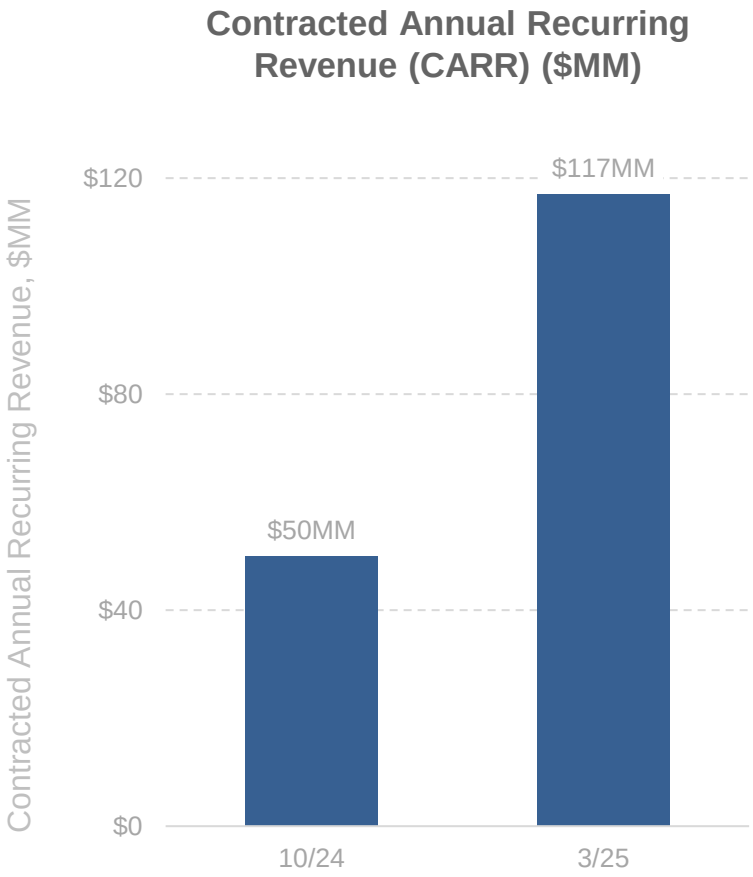
…The national rollout includes more than 25,000 doctors and clinicians, 40 hospitals, and north of 600 medical offices…

…The feedback from doctors has been effusive:
'It saved my marriage.' And:
'You'd have to take it away from my cold, dying hands.'

- Fortune Magazine (2/25)

We are incredibly proud of our partnership with Kaiser – where a majority of Kaiser doctors are using Abridge to summarize patient visits, with over 10 million completed to date. As one of our earliest deployments, it is a great example of how we are building alongside our many hospital partners and helping them grow with Abridge.

- Abridge CFO Sagar Sanghvi (5/25)



Note: 3/25 figure is quoted as being as of Q1:25. We conservatively assume this maps to 3/25. Abridge's CARR goes live within weeks of contracting. Source: Abridge (12/24 & 5/25), Fortune (2/25), The Information (10/24 & 5/25) (link & link)

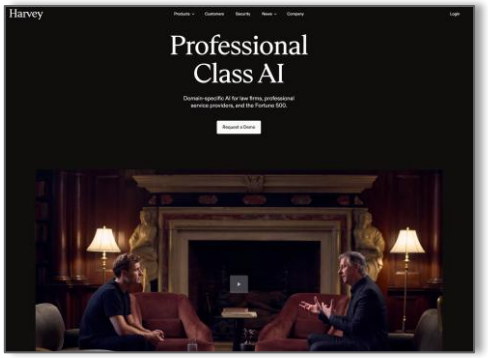
AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Legal

Specialized AI – Legal (Workflows) =
Harvey @ $10MM to $70MM ARR in Fifteen Months, per *The Information & Business Insider*

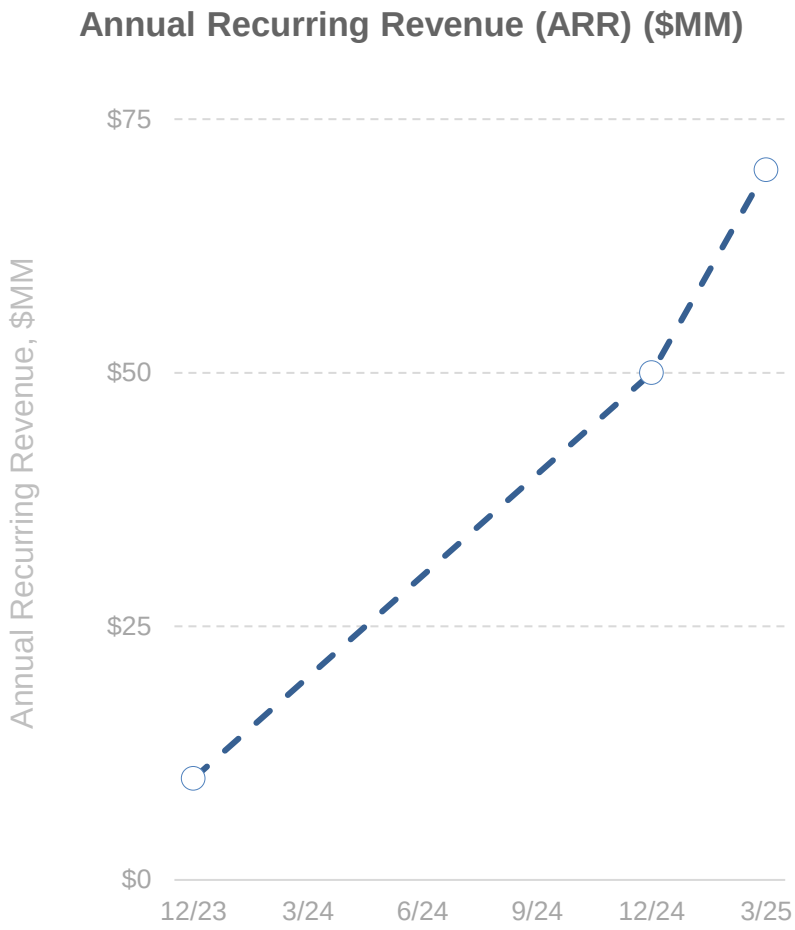
Harvey – 12/23-4/25, per *The Information & Business Insider*



In 2024, we saw 4x annual recurring revenue (ARR) growth and expanded from 40 customers to 235 customers in 42 countries, including the majority of the top 10 USA law firms.

We've also seen the legal and professional services industry shift faster than ever before. Lawyers are adopting technology at an unprecedented rate, centuries-old firms are experimenting with new business models, and enterprises are driving significant savings with AI-enabled workflows. The pace of change will only accelerate in 2025.

- Harvey Co-Founder & CEO Winston Weinberg & Co-Founder & President Gabe Pereyra (2/25)



Source: Harvey (2/25), The Information estimates (1/25) (link, link), & Business Insider (5/25) (link)

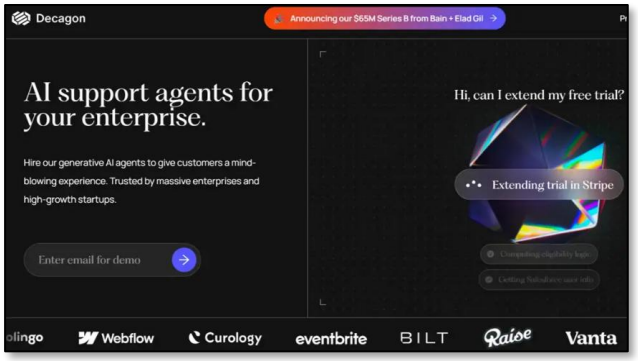
AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Customer Service

Specialized AI – Customer Service (AI Support Agents) = Decagon @ ~$1MM to $10MM ARR in One Year

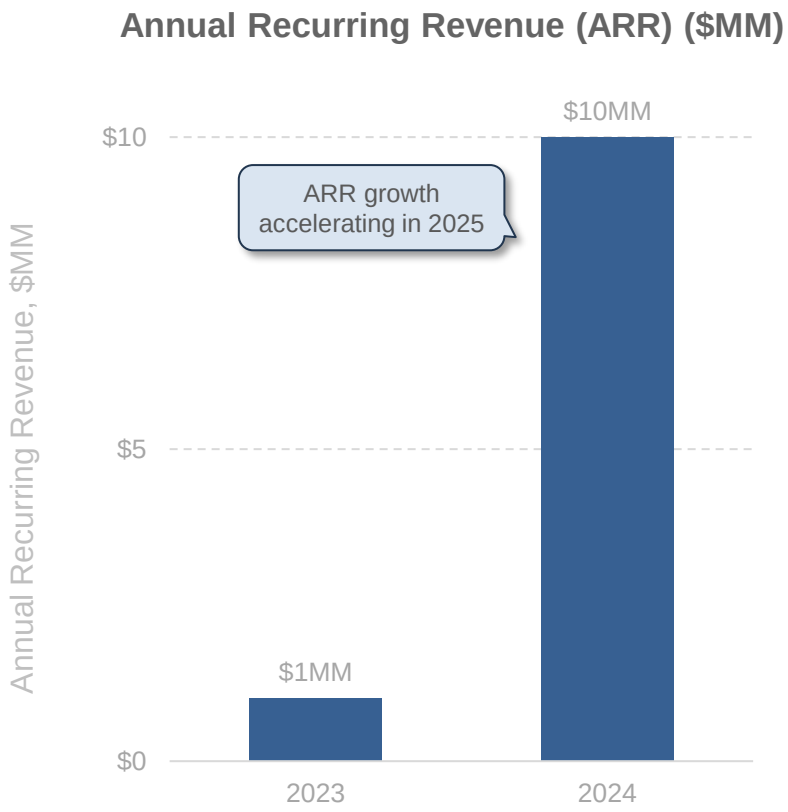
Decagon – 2023-2024, per Decagon



AI is often seen as destroying jobs, but at Decagon, we believe the opposite. Our AI agents are enhancing jobs, not replacing them…

…In a few years, every company will have AI agents running their customer experiences. Customer support staff are no longer fielding routine tasks; they are now becoming AI managers – configuring, training and overseeing the AI agents that handle repetitive work.

- Decagon Co-Founder & CEO Jesse Zhang (10/24)



Note: Source: Decagon (12/23, 10/24, 12/24)

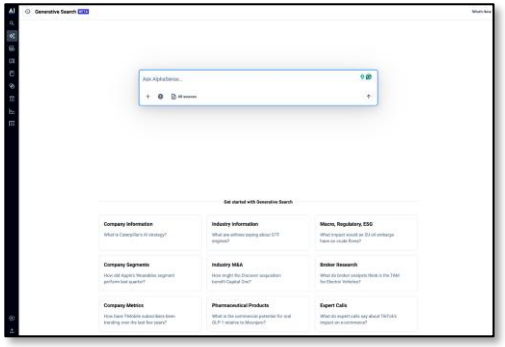
AI-Enabled Specialized Software Companies @ Large Service Industries =

Growing Very Quickly…

Financial Services

Specialized AI – Financial Services (Research & Analysis) = AlphaSense @ ~$150MM to ~$420MM in Two Years

AlphaSense – 2022-2024, per AlphaSense

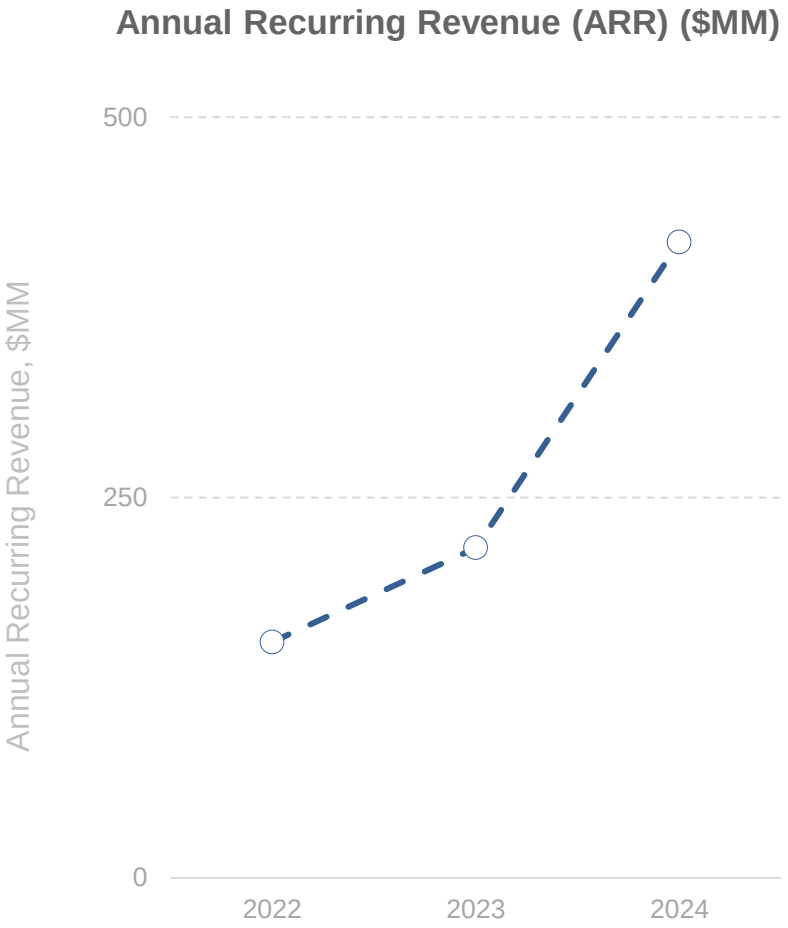


We are at a tipping point where AI-driven insights are no longer a luxury but a necessity – every company's market value is the sum of the decisions it makes.

Surpassing $400 million in ARR and our rapid growth are clear signals that businesses are recognizing the transformative power of our end-to-end market intelligence platform.

As we scale, our focus remains on product and technology innovation, ensuring we deliver high-value solutions and cutting-edge AI and smart workflow capabilities to our customers.

- AlphaSense Co-Founder & CEO Jack Kokko (3/25)



Source: AlphaSense (3/25)

Next AI Use Case Frontiers =

Broad & Varied

Next AI Use Case Frontiers = Broad & Varied

Next AI Use Case Frontiers – 5/25

Highlights = Pages 246-247

Medical Discovery & Development

Precision Manufacturing

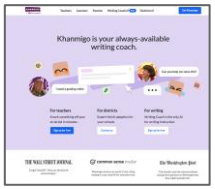
Multi-Purpose Robotics

Autonomous Scientific Research

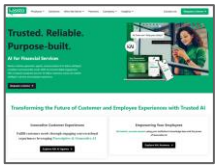
Supply Chain Optimization

Cybersecurity & Threat Detection

Personalized Education



Autonomous Finance



Environmental & Climate Monitoring

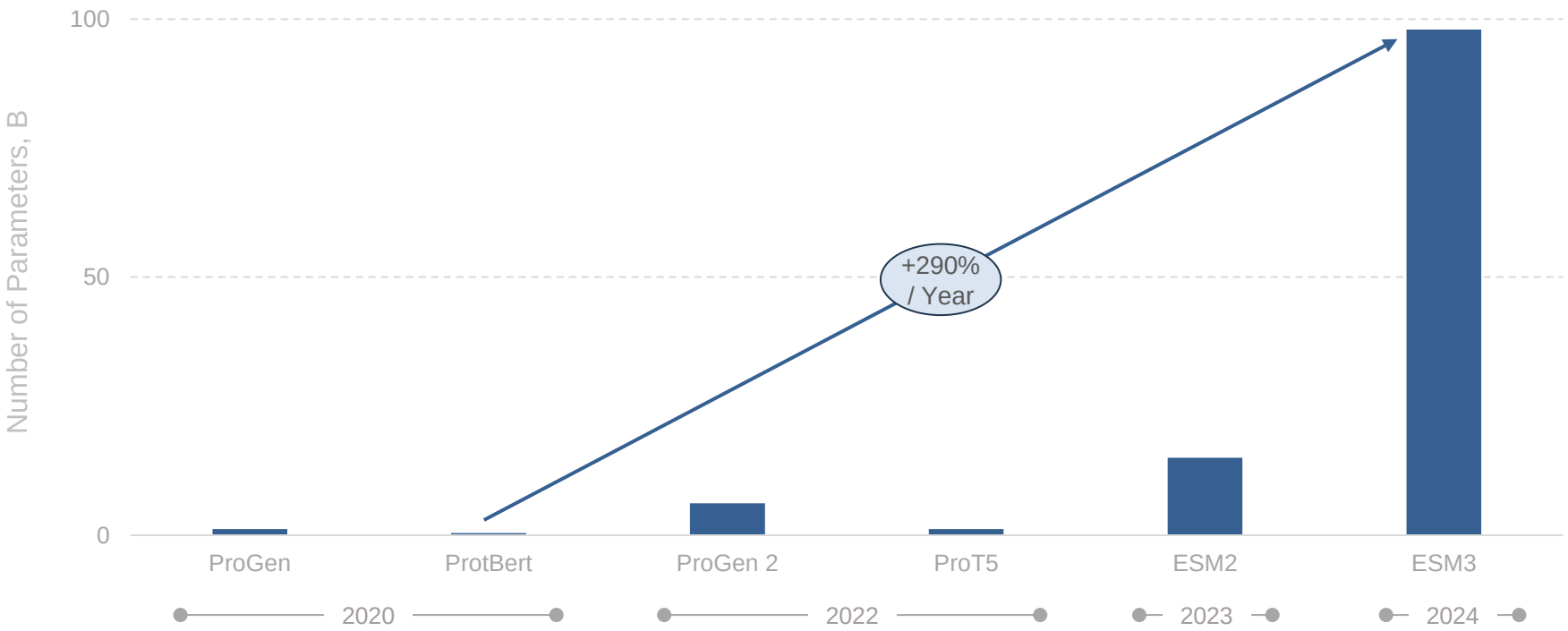
Energy Grid Management

Note: List is not comprehensive. Source: Drug Development & Discovery = Insilico; Precision Manufacturing = Landing AI; Multi-Purpose Robotics = Figure AI; Autonomous Scientific Research = IBM's RoboRXN; Supply Chain Optimization = o9 Solutions; Cybersecurity & Threat Detection = Vectra AI; Personalized Education = Khanmigo; Autonomous Finance = Kasisto; Environmental & Climate Monitoring = ClimateAI; Energy Grid Management = Uplight; BOND analysis

Next AI Use Case Frontier – Protein Sequencing = Model Size +290% Annually to 98 Billion Parameters Over Four Years

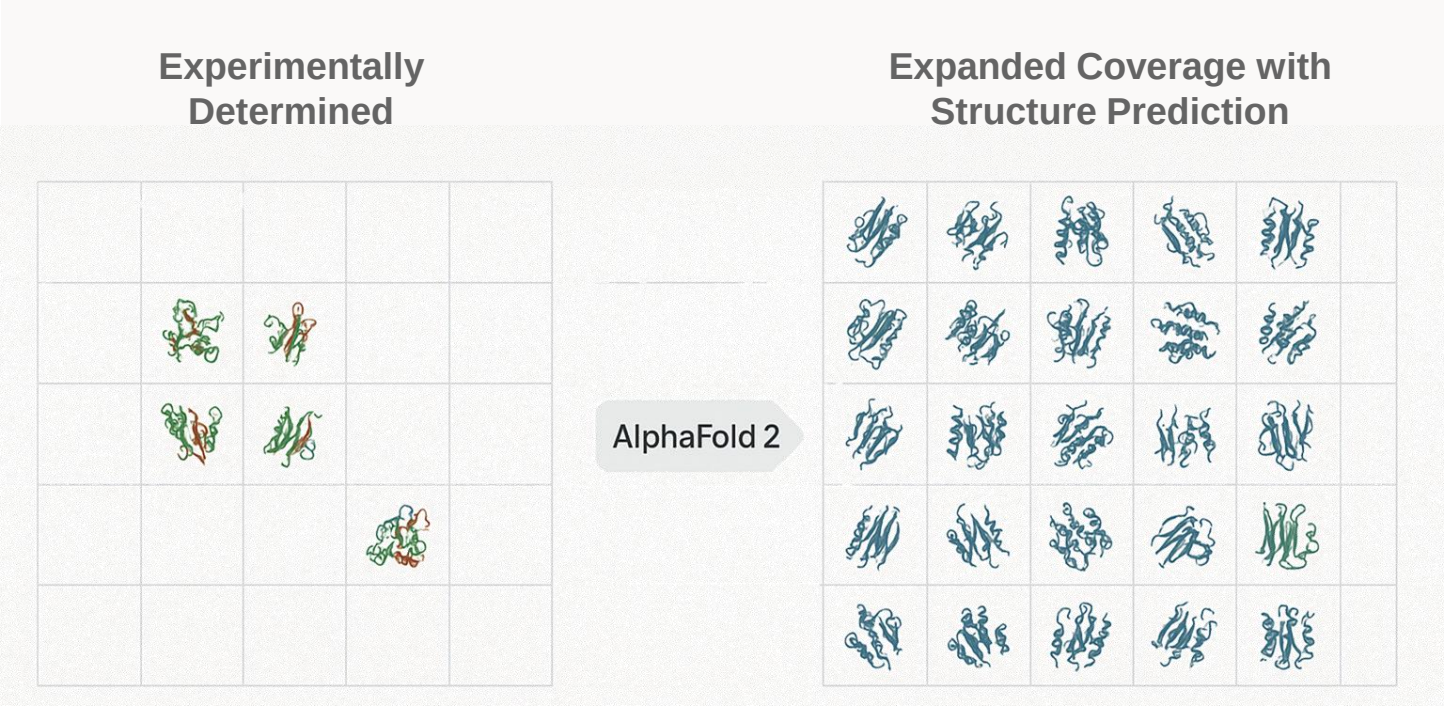
Size of Major Protein Sequencing Models (B Parameters) – 2020-2024, per Stanford RAISE Health

Per Stanford HAI (4/25): *The past year has witnessed remarkable progress in AI models applied to protein sequences. Large-scale machine learning models have improved our ability to predict protein properties, accelerating research in structural biology and molecular engineering…These AI-driven approaches have transformed protein science by minimizing reliance on costly, time-intensive experimental methods, enabling rapid exploration of protein function and design.*



Note: List of models may not be comprehensive.
Source: Stanford RAISE Health via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

Next AI Use Case Frontier – Protein Sequencing = Synthetically Generated Protein Data Yields 1,000x Expansion via AlphaFold



214,121
Protein Structures
in PDB (2024)

214MM
Predicted Protein Structures
in AFDB (2024)

Note: AFDB predicted protein structure counts may be higher as of year-end 2024. Source: Google DeepMind, RCSB Protein Data Bank (2024)

Outline

1. **Seem Like Change Happening Faster Than Ever?**
   Yes, It Is
2. **AI User + Usage + CapEx Growth** =
   Unprecedented
3. **AI Model Compute Costs High / Rising + Inference Costs Per Token Falling** =
   Performance Converging + Developer Usage Rising
4. **AI Usage + Cost + Loss Growth** =
   Unprecedented
5. **AI Monetization Threats** =
   Rising Competition + Open-Source Momentum + China's Rise
6. **AI & Physical World Ramps** =
   Fast + Data-Driven
7. **Global Internet User Ramps Powered by AI from Get-Go** =
   Growth We Have Not Seen Likes of Before
8. **AI & Work Evolution** =
   Real + Rapid

AI Monetization Threats

=

<u>Rising Competition</u>

+

Open-Source Model Momentum

(& China's Rise)

Rising Competition =

AI Model Releases

On the back of Google's 'Attention is All You Need' Transformers research paper in 2017,
the first wave of 'modern AI' (read: LLMs) focused on text: models such as OpenAI's GPT-3 and Meta's Llama-1
showed that teaching computers to finish sentences at scale could unlock broad reasoning abilities.
Yet human communication is rarely text-only, and often not even text-first.
Images, audio, video, and sensor readings carry context that words alone miss,
so researchers at the same companies –
and peers like Google, Anthropic, and xAI, among others –
began extending language models to handle additional signals.

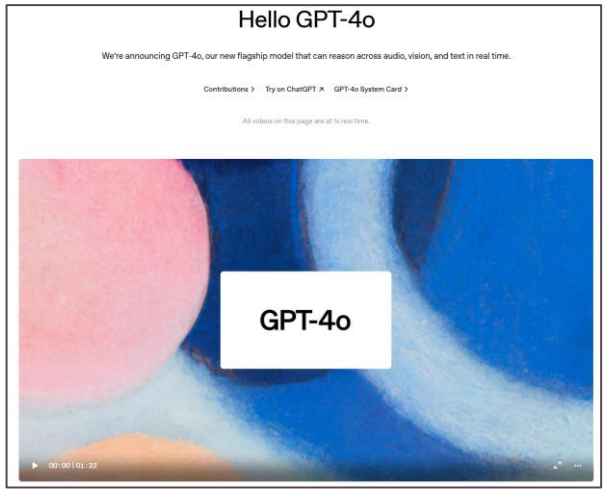
Multimodal AI models are the result. They embed text, pictures, sound, and video
into a shared representation and generate outputs in any of those formats.
A single query can reference a paragraph and a diagram, and the model can respond
with a spoken summary or an annotated image – without switching systems.
Each new modality forces models to align meaning across formats rather than optimize for one.

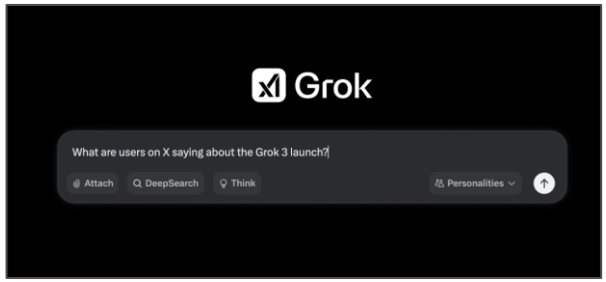
The path to this capability unfolded stepwise: OpenAI's CLIP paired vision and language in 2021;
Meta followed with ImageBind in 2023 and Chameleon in 2024;
and by 2024-2025, frontier systems such as GPT-4o, Claude 3, and Chameleon had become fully multimodal.
Each new modality forced the models to align meaning across formats rather than optimize for one.

The payoff is practical.
A field engineer can aim a phone camera at machinery and receive a plain-language fault diagnosis;
a clinician can attach an X-ray to a note and get a structured report draft;
and an analyst can combine charts, transcripts, and audio clips in a single query.
Compared with text-only models, multimodal systems cut context switching,
capture richer detail, and enable applications –
quality control, assistive tech, content creation – where visual or auditory information matters as much as words.

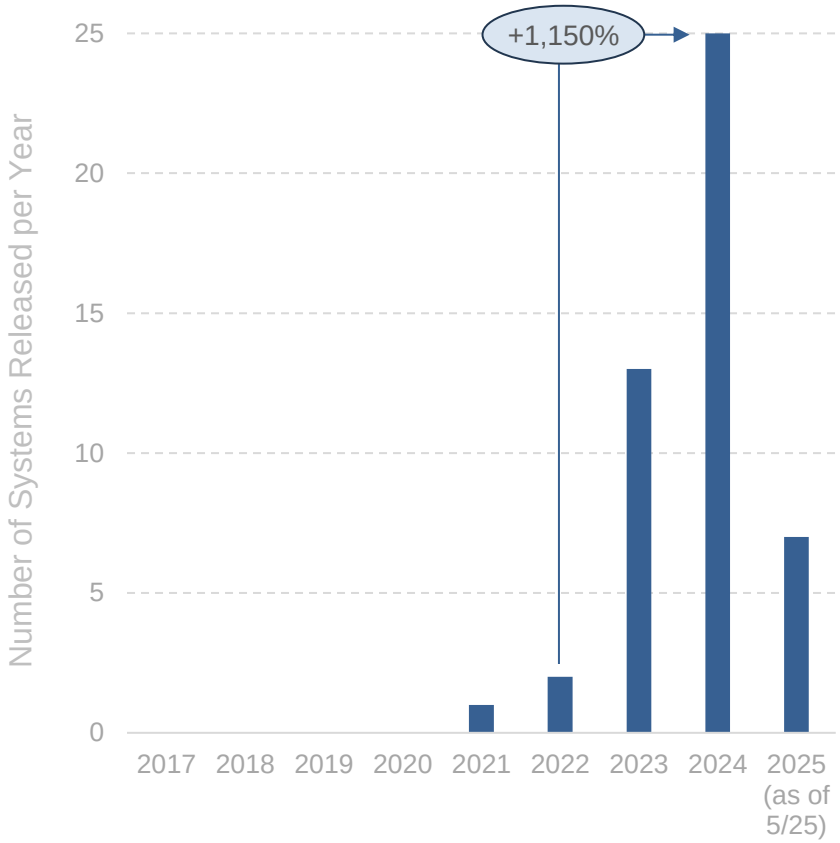
Large-Scale AI *Multimodal** Model Competition = +1,150% Rise in Models Released Over Two Years, per Epoch AI

Multimodal Models – Examples





Large-Scale** Multimodal Models – Releases

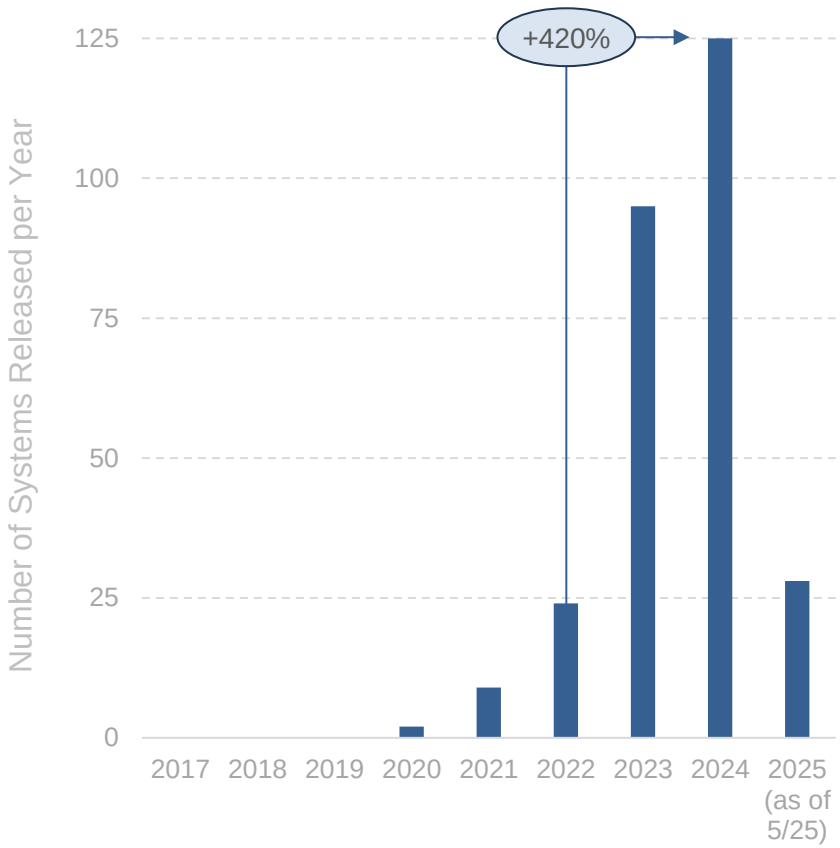


*A multimodal AI model is one that can process and integrate multiple types of data, e.g., text, images, audio, or video, to understand and generate outputs across different modalities.
**Epoch AI defines large-scale as models where their training compute is confirmed to exceed 10^{23} floating-point operations. An AI system can operate in more than one domain and may be double-counted across pages. Source: Epoch AI via Our World in Data (4/25), OpenAI, DeepSeek, Google

Large-Scale AI *Language* Model Competition = +420% Increase in Models Released Over Two Years, per Epoch AI

Language Models – Examples

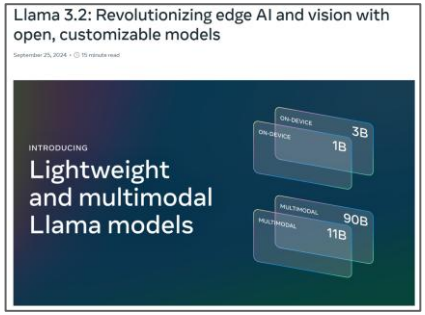
Large-Scale* Language Models – Releases



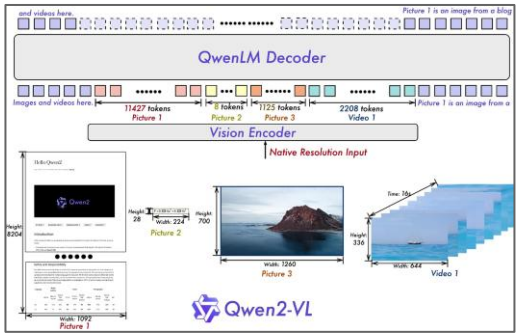
**Epoch AI defines large-scale as models where their training compute is confirmed to exceed 10^{23} floating-point operations. An AI system can operate in more than one domain and may be double-counted across pages. Many models shown are multimodal. Source: Epoch AI via Our World in Data (4/25), OpenAI, DeepSeek, Google*

Large-Scale AI *Vision* Model Competition = +109% Increase in Models Released Y/Y, per Epoch AI

Vision Models* – Examples

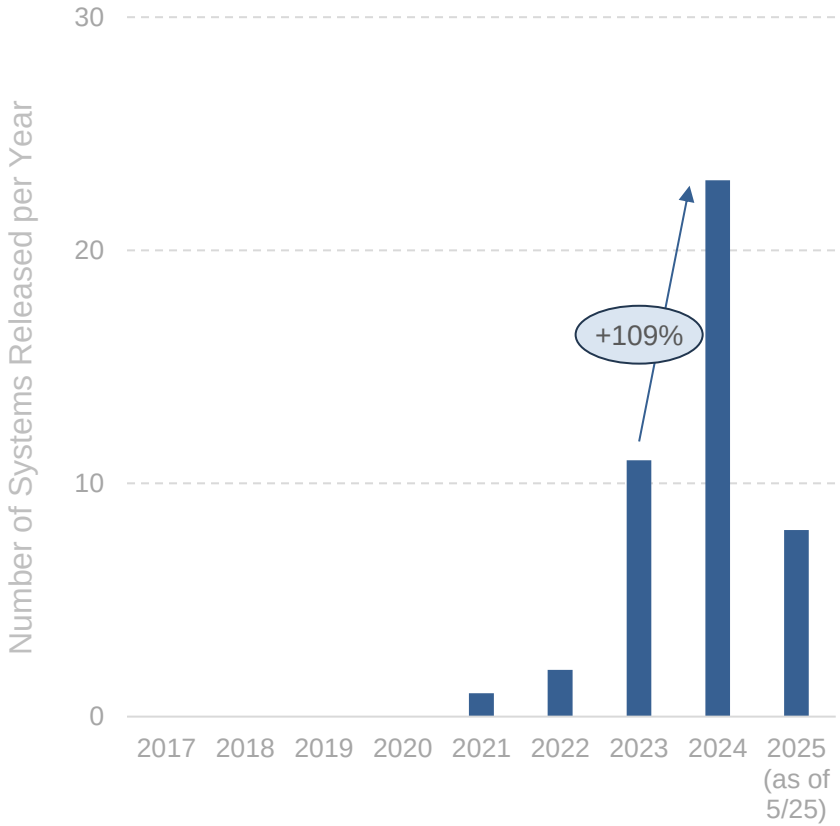


Meta Llama 3.2 – 9/24



Qwen2-VL – 12/24

Large-Scale* Image Models – Releases

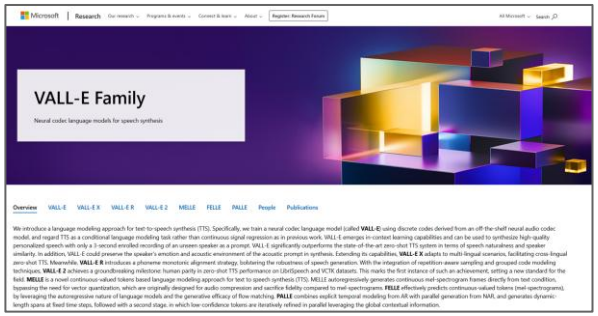


**Epoch AI defines large-scale as models where their training compute is confirmed to exceed 10^{23} floating-point operations. An AI system can operate in more than one domain and may be double-counted across pages. Many models shown are multimodal. Source: Epoch AI via Our World in Data (4/25), Meta, Alibaba*

Large-Scale AI *Speech / Audio* Model Competition = +367% Increase in Models Released Y/Y, per Epoch AI

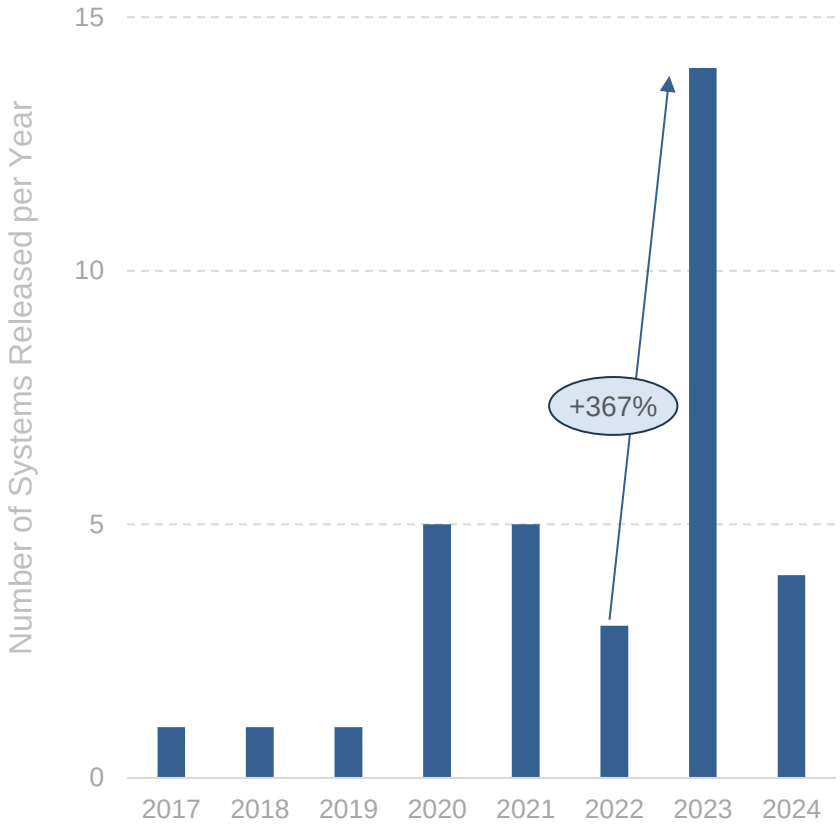
Speech / Audio Models – Examples

OpenAI GPT 4o Speech – 5/24



Microsoft VALL-E – 1/23

Speech / Audio Models – Releases



Note: An AI system can operate in more than one domain and may be double-counted across pages. Includes models without verified training compute. Many models shown are multimodal. Source: Epoch AI (5/25), Microsoft (1/23), OpenAI (5/24), Amazon, Pinterest

Large-Scale AI *Video* Model Competition = +120% Increase in Models Released Y/Y, per Epoch AI

Video Models – Examples

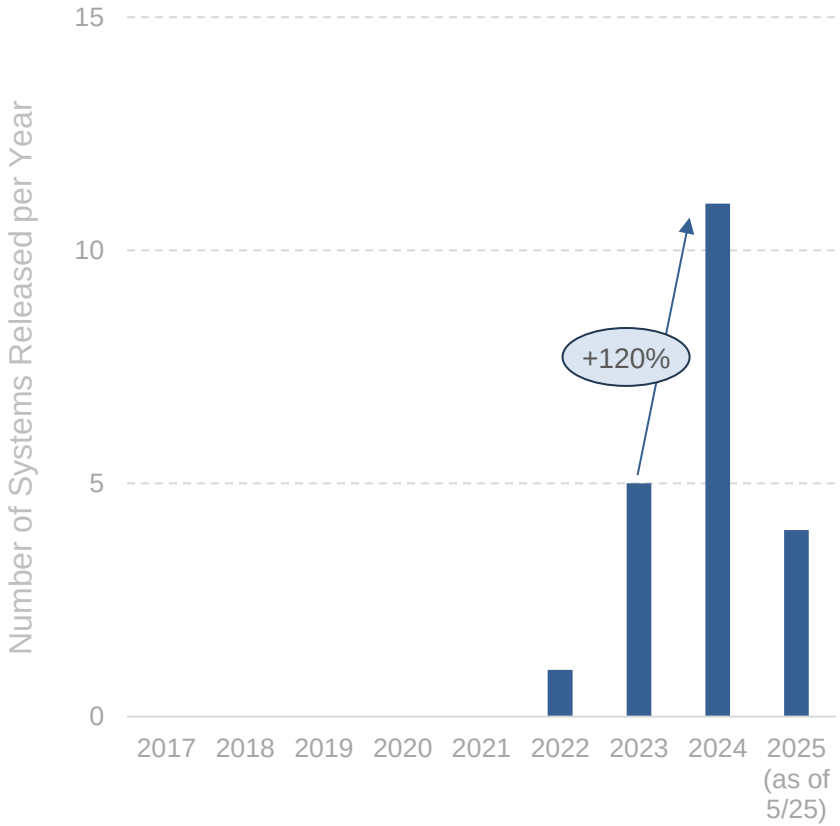
OpenAI Sora – 12/24

Amazon Nova Reel – 12/24

According to academic studies, 50% of the human brain is wired for visual processing. The ability for users to explore their interest visually and take action on them… is particularly relevant for Gen Z… who have been raised on an internet of visual content across images and video.

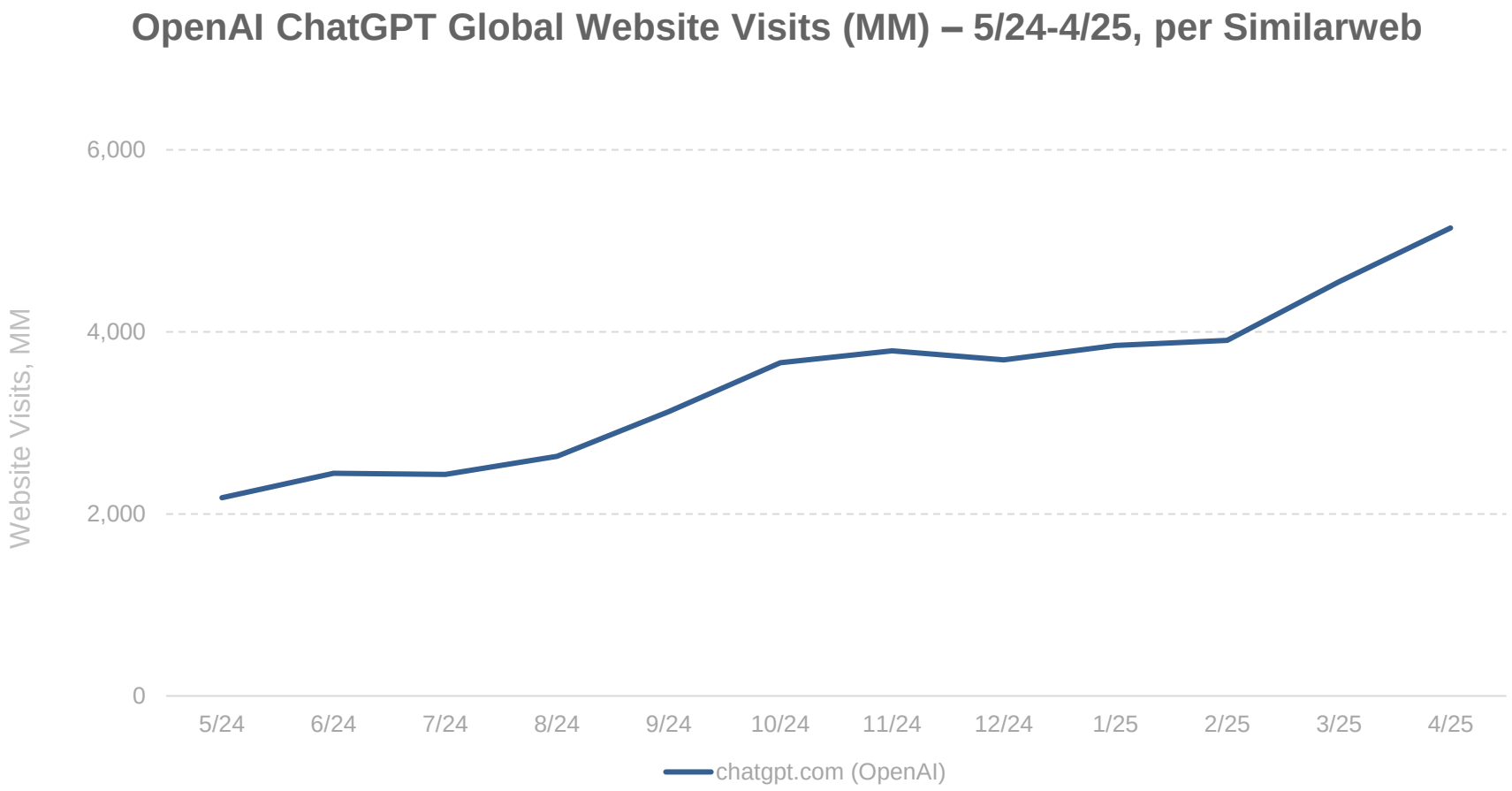
- Pinterest CEO Bill Ready (5/25)

Large-Scale* Video Models – Releases



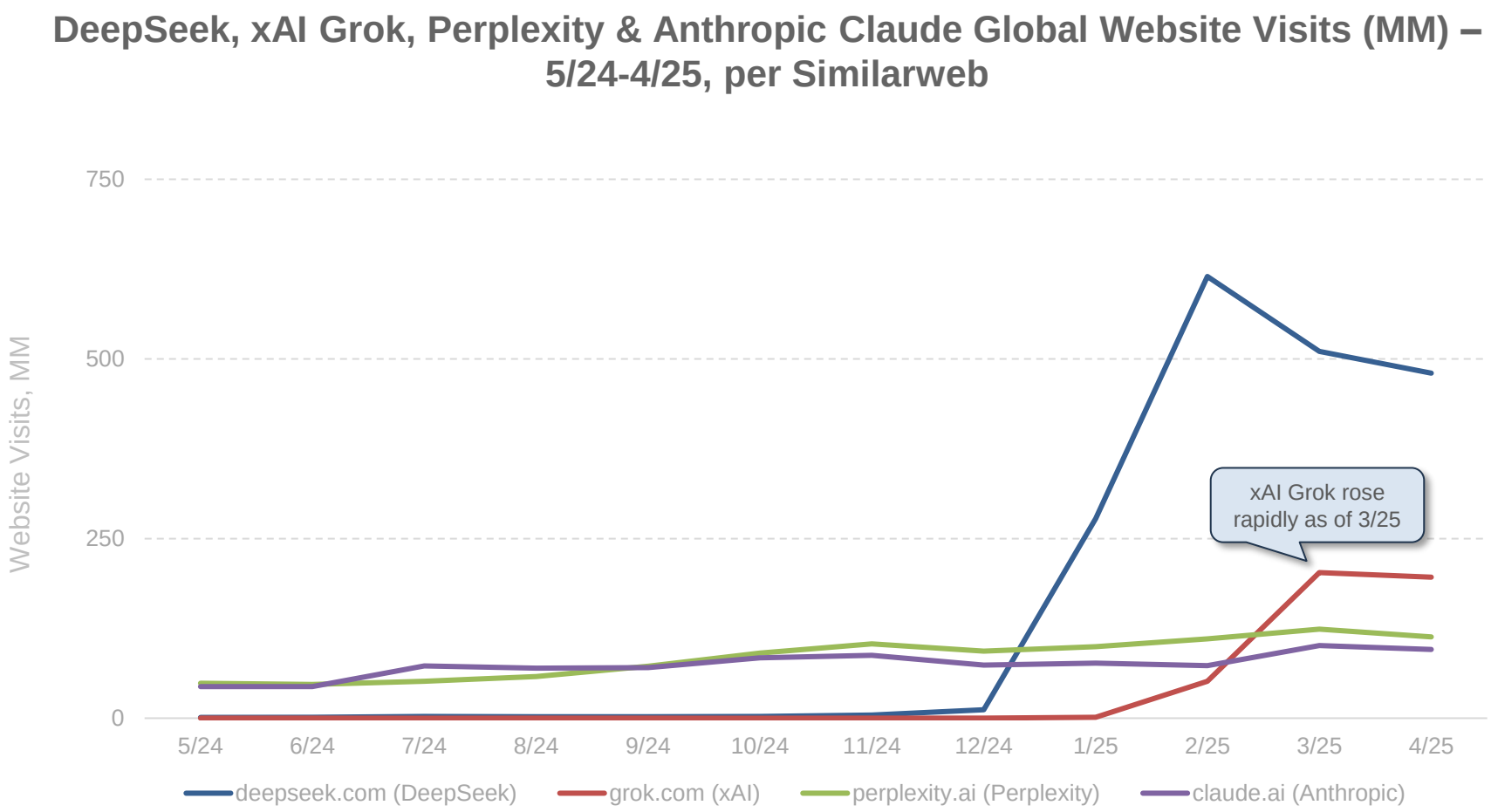
**Epoch AI defines large-scale as models where their training compute is confirmed to exceed 10^{23} floating-point operations. An AI system can operate in more than one domain and may be double-counted across pages. Many models shown are multimodal. Source: Epoch AI via Our World in Data (4/25), OpenAI, Amazon, Pinterest, Pinterest*

LLM Competition – Website Visits = OpenAI ChatGPT Biggest @ 5.1B Site Visits…



Note: Includes desktop & mobile (non-app) website visits. China data may be subject to informational limitations due to government restrictions. Source: Similarweb (5/25)

…LLM Competition – Website Visits = DeepSeek & xAI Grok Also Rising @ 196-480MM Visits Each



Note: Includes desktop & mobile (non-app) website visits. China data may be subject to informational limitations due to government restrictions. Source: Similarweb (5/25)

LLM Competition – Product Releases During Week of May 19, 2025 = It Wasn't Just Google's Annual I/O Conference

Select AI Product Announcements – 5/19/25-5/23/25, per Google, Microsoft, Anthropic & OpenAI

Gemini

- Gemini Live camera & screen sharing
- Project Mariner computer use
- Updated Gemini 2.5 Flash
- Gemini 2.5 Pro
- Native audio output for 2.5 Flash & Pro Previews
- Thinking Budgets for Gemini 2.5 Pro
- Deep Think
- Project Astra capabilities
- Gemini in Chrome
- Deep Research improvements
- Gemini Agent Mode
- Google AI Pro Subscription
- Google AI Ultra Subscription
- Google Beam
- Google Meet speech translation
- Personalized Smart Replies
- Jules
- Imagen 4
- Veo 3
- Lyria 2
- Flow TV
- Project Moohan
- Glasses with Android XR
- 'Try on' experiment
- Agentic checkout
- Gemini interactive quizzes
- Canvas Create menu
- LearnLM integration into Gemini 2.5
- SDK support for Model Context Protocol (MCP) definitions in Gemini API
- Gemini Diffusion
- SynthID Detector
- Conversational tutor prototype
- Google Live API audiovisual input & native audio out dialogue
- Gemma 3n
- AI studio enhancements
- Android Studio Journeys
- Android Studio Version Upgrade Agent
- Wear OS 6 Developer Preview
- Gemini Code Assist
- New Firebase features
- Google AI Edge Portal
- Google Vids
- Enhanced Audio Overviews
- Sparkify experiment

Copilot

- Magentic-UI
- Copilot Studio multi-agent orchestration
- GitHub Copilot asynchronous functioning
- Azure AI Foundry expansion
- NLWeb
- Model Context Protocol (MCP) integration
- Entra Agent ID
- SQL Server 2025
- Windows Subsystem for Linux Open-Source
- GitHub Copilot Chat Extension
- Aurora AI-Powered Weather Forecasting

ANTHROP\C

- Claude Opus 4
- Claude Sonnet 4

OpenAI

- Acquisition of io

Note: Announcements include products that were made immediately-available and forthcoming products. List is non-exhaustive. Source: Google Microsoft, Anthropic, OpenAI (5/25)

AI Monetization Threats

=

Rising Competition

+

Open-Source Model Momentum

(& China's Rise)

To understand where AI model development is headed, it helps to examine how two distinct approaches – closed-source and open-source – have evolved and diverged.

In the early days of modern machine learning (2012-2018), most models were open-source, rooted in academic and collaborative traditions.
But as AI systems became more powerful and commercially valuable, and as development shifted from academia to industry, a parallel movement emerged – around 2019 (when GPT-2 launched with restricted weights), the development of proprietary (closed-source) models, motivated by proprietary interests, competitive advantage, and safety concerns.

Closed models follow a centralized, capital-intensive arc. These models – like OpenAI's GPT-4 or Anthropic's Claude – are trained within proprietary systems on massive proprietary datasets, requiring months of compute time and millions in spending. They often deliver more capable performance and easier usability, and thus are preferred by enterprises and consumers, and – increasingly – governments. However, the tradeoff is opacity: no access to weights, training data, or fine-tuning methods.

What began as a research frontier became a gated product experience, served via APIs, licensed to enterprises, and defended by legal and commercial firewalls. Now, the AI race is coming full circle.
As LLMs mature – and competition intensifies – we are seeing resurgence of open-source models owing to their lower costs, growing capabilities, and broader accessibility for developers and enterprises alike.
These are freely available for anyone to use, modify, and build upon, and thus are generally preferred by early-stage startups, researchers / academics, and independent developers.

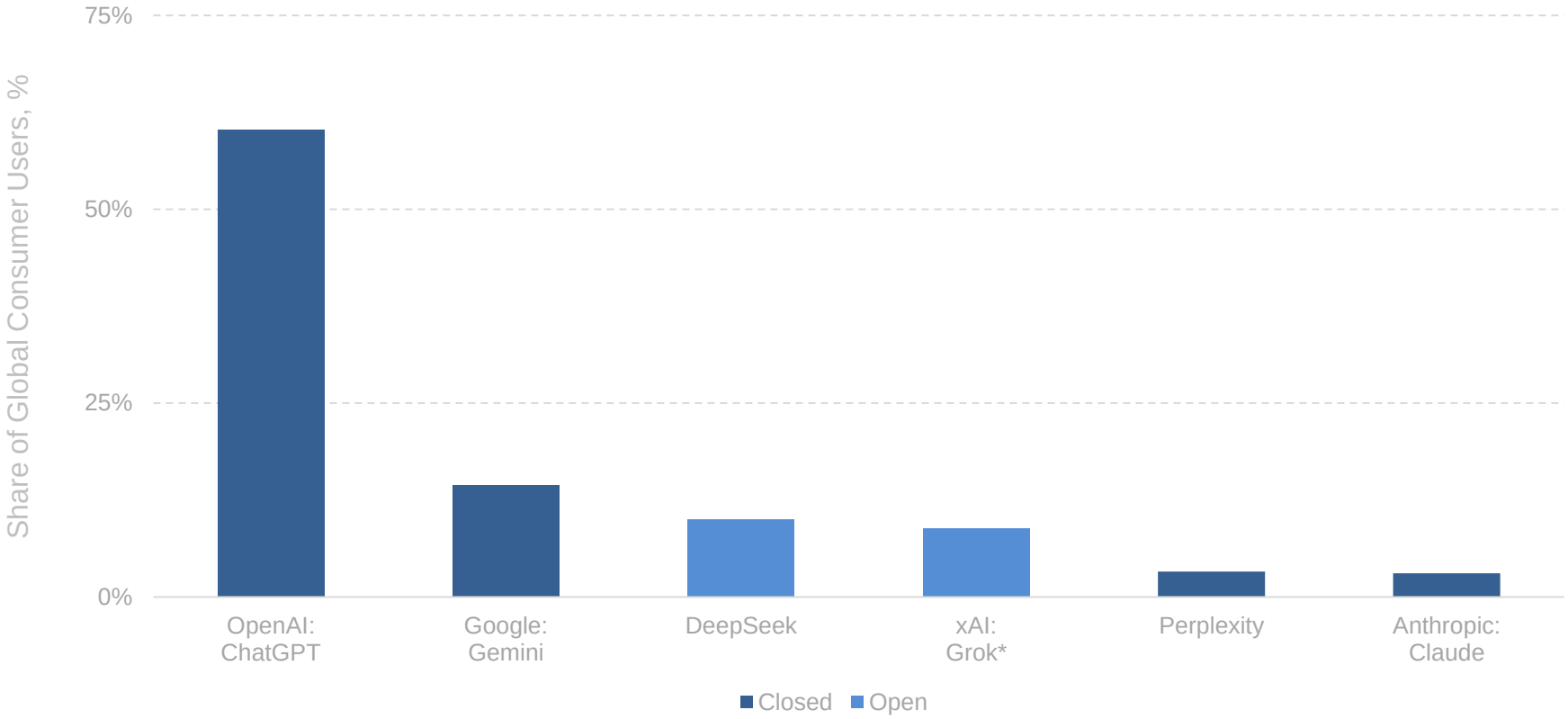
Platforms like Hugging Face have made it frictionless to download models like Meta's Llama or Mistral's Mixtral, giving startups, academics, and governments access to frontier-level AI without billion-dollar budgets.
Open-source AI has become the garage lab of the modern tech era: fast, messy, global, and fiercely collaborative.
And China (as of Q2:25) – based on the number of large-scale AI models* released – is leading the open-source race, with three large-scale models released in 2025 – DeepSeek-R1, Alibaba Qwen-32B and Baidu Ernie 4.5**.

The split has consequences. Open-source is fueling sovereign AI initiatives, local language models, and community-led innovation.
Closed models, meanwhile, are dominating consumer market share and large enterprise adoption.
We're watching two philosophies unfold in parallel – freedom vs. control, speed vs. safety, openness vs. optimization – each shaping not just how AI works, but who gets to wield it.

**Large-scale AI models = Models with training compute confirmed to exceed 10^{23} floating point operations.*
***To be made open-source as of 6/30/25, per Baidu.*

Closed vs. Open-Source Models – Monthly Active Users (MAUs) = Closed Models Dominating With Consumers, per YipitData

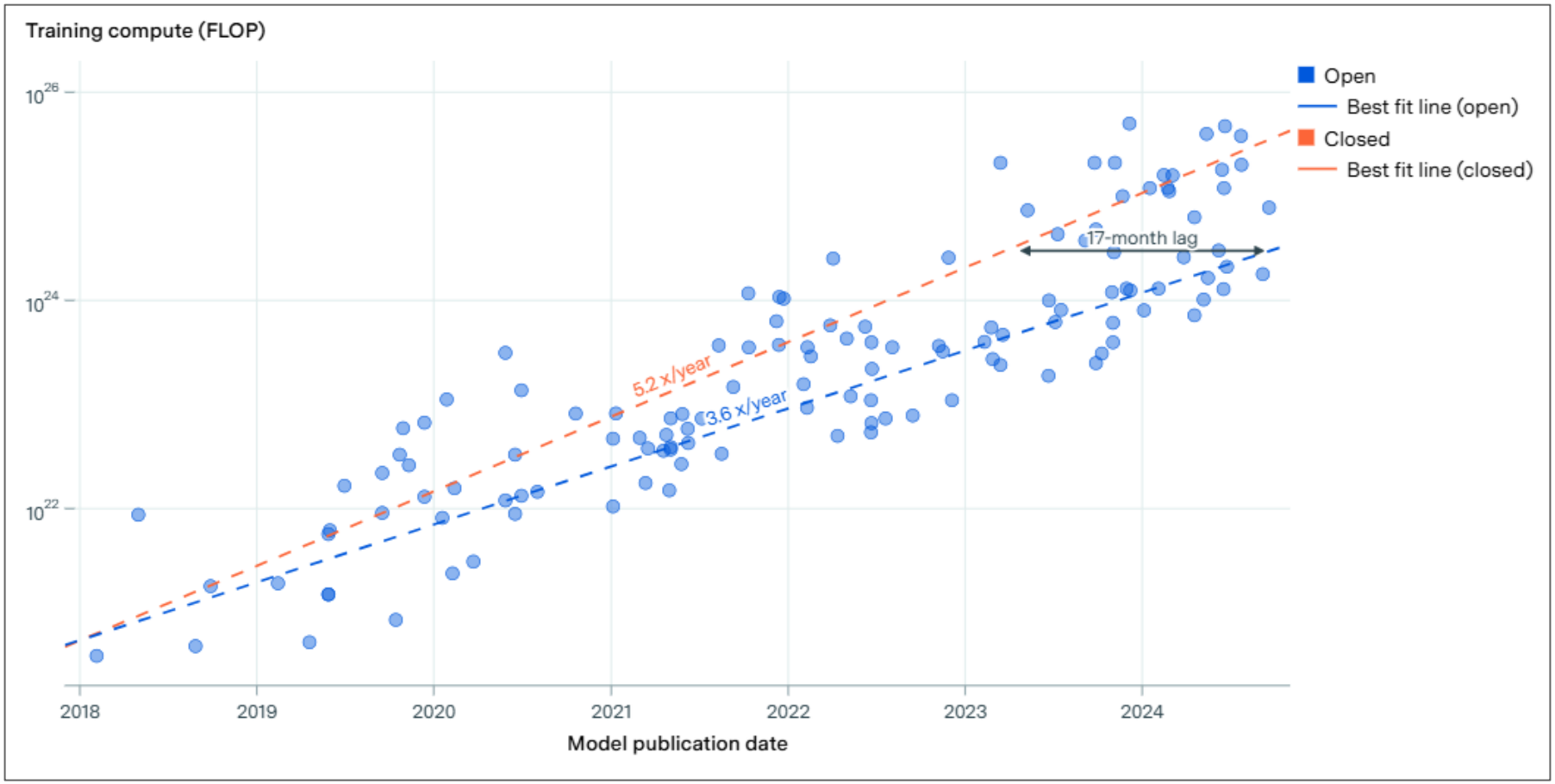
Estimated Share of Global Monthly Active Users (MAUs) Across Six Leading LLMs – 4/25, per YipitData



**xAI open-sourced the Grok-1 base model in March 2024, but newer versions and full chatbot features remain proprietary. Note: Data is a subset of global internet users and absolute user data will be understated; however, given that the panel is globally-representative (with limitations on China-specific data), relative comparisons / trends are informative. Desktop users only. Figures calculate the number of users on a given platform, divided by the number of users on all platforms combined. Figures are non-deduped (i.e., users using multiple platforms may be counted twice). Data measures several million global active desktop users' clickstream data. Data consists of users' web requests & is collected from web services / applications, such as VPNs and browser extensions. Panel is globally-representative (with limitations on China-specific data). Users must have been part of the panel for 2 consecutive months to be included. Source: YipitData (5/25)*

Closed vs. Open-Source Models – Compute Investment = Closed Models Higher, per Epoch AI

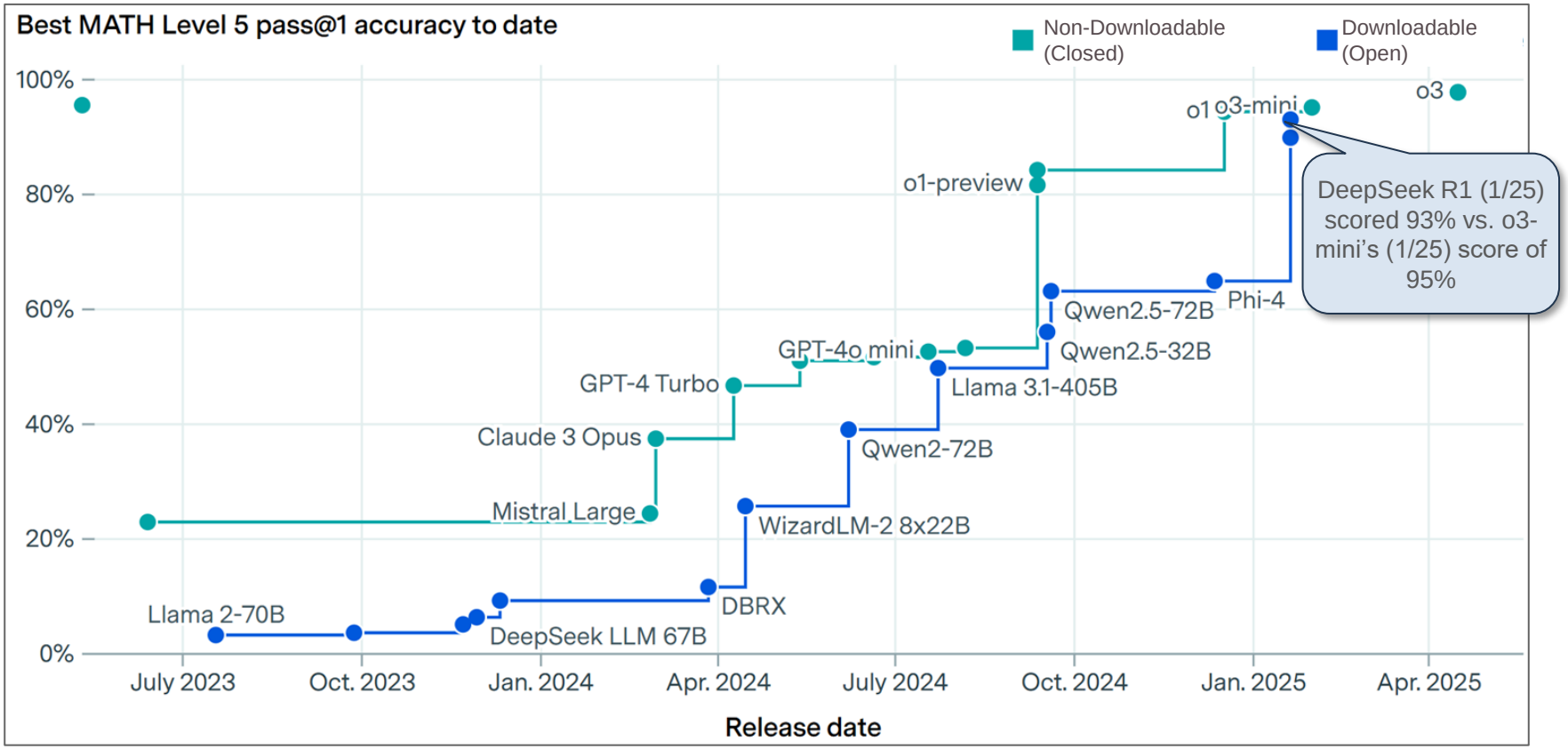
Training Compute Resources for Open vs. Closed LLMs – 2/18-9/24, per Epoch AI



Source: Epoch AI (11/24)

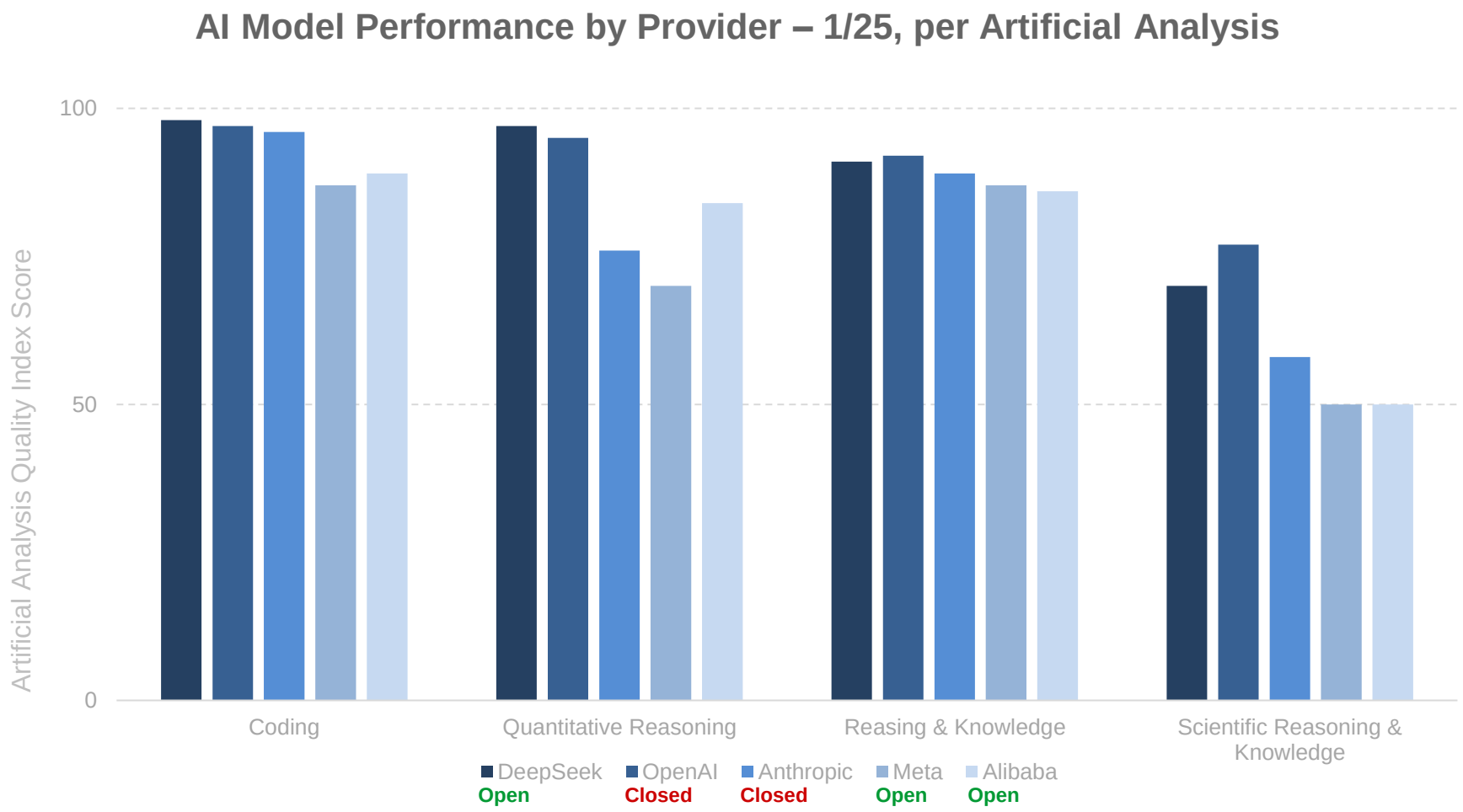
Closed vs. Open-Source Models – Performance = Gap Closing…China Rising, per Epoch AI…

Performance on MATH Level 5 Test, Open vs. Closed LLMs by Year Released – 6/23-4/25, per Epoch AI



Note: MATH Level 5 pass@1 refers to the accuracy of an AI model on the MATH benchmark, a dataset of high school competition-level mathematics problems. Level 5 indicates the most challenging problems in the benchmark. 'pass@1' measures whether the model correctly solves the problem on its first attempt. Source: Epoch AI (5/25)

…Closed vs. Open-Source Models – Performance = Gap Closing…China Rising, per Artificial Analysis



Note: Scores are out of 100. The models for each company that are measured: for OpenAI, o1; for Alibaba, Qwen 2.5 72B; for Meta, Llama 3.1 405B; for Anthropic, Claude 3.5 Sonnet. The tests used are HumanEval, MATH-500, MMLU and GPQA Diamond. Source: Artificial Analysis via NBC News, 'Why DeepSeek is different, in three charts' (1/25)

Rising Performance of Open-Source Models

+

Falling Token Costs

=

Explosion of Usage by **Developers Using AI**

Closed-source models – like GPT-4, Claude, or Gemini –
have dominated usage among consumers and large enterprises,
largely because of their early performance advantage, ease of use, and broader awareness.
These models came bundled in clean, productized interfaces and offered reliable outputs with minimal setup.

For enterprises, they promise security and ease-of-use for non-technical employees.
For consumers, they came with name recognition, fast onboarding, and polished UX.
That combination has kept closed models at the center of the AI mainstream.

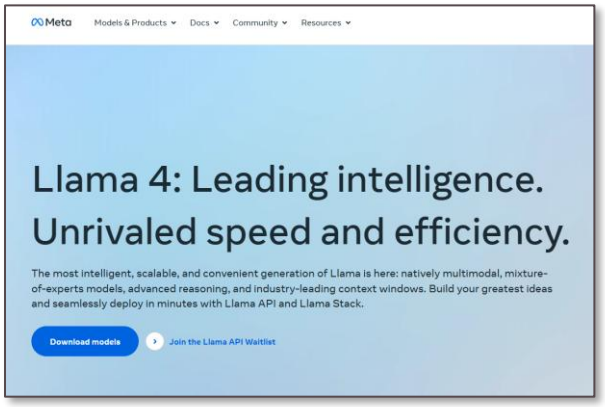
But performance leadership is no longer a given. Open-source models are closing the gap – faster than many expected – and doing so at a fraction of the cost to users. Models like Llama 3 and DeepSeek have demonstrated competitive reasoning, coding, and multilingual abilities, while being fully downloadable, fine-tunable, and deployable on commodity infrastructure.

For developers, that matters. Unlike enterprise buyers or end-users,
developers care less about polish and more about raw capability, customization, and cost efficiency.
And it is developers – more than any other group –
who have historically been the leading edge of AI usage.
The recent trend appears increasingly clear: more developers are gravitating toward low-cost,
high-performance open models, using them to build
apps, agents, and pipelines that once required closed APIs.

Time will tell if that advantage scales beyond the developer ecosystem.
Many open-source tools still lack the brand power, plug-and-play user experience (UX),
and managed services that drive adoption among consumers and large organizations.
But as the cost-performance ratio of open models continues to improve –
and if the infrastructure to support them becomes more turnkey –
those advantages could start to spread beyond the developer community.

Developer AI Model Activity = +3.4x Increase in Downloads of Meta Llama in Eight Months

Meta Llama – 8/24-4/25, per Meta Platforms

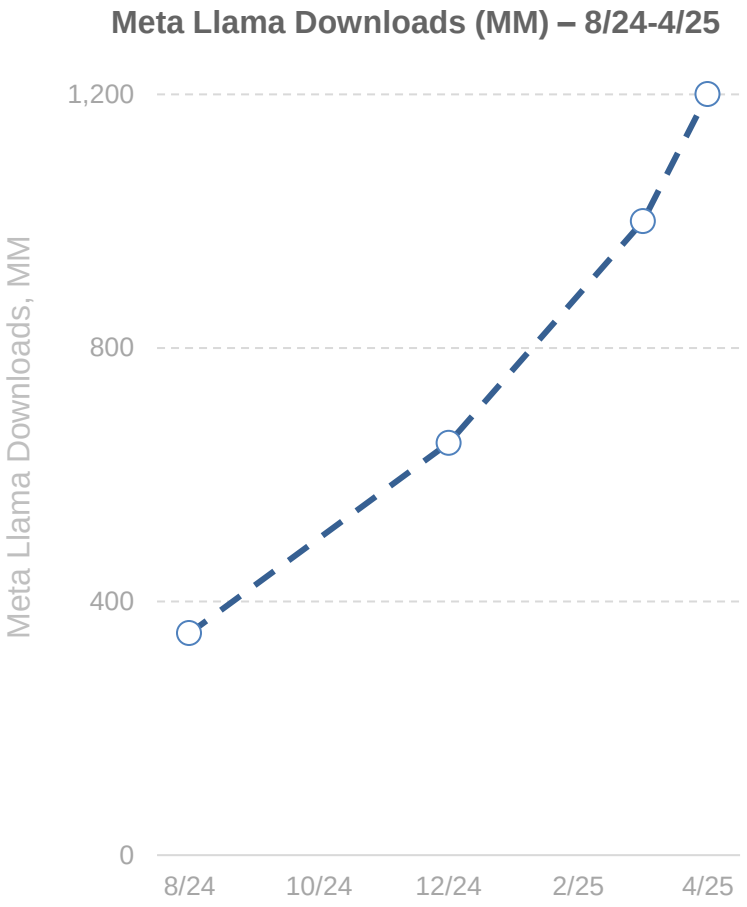


I predicted that 2025 was going to be the year that open source became the largest type of model that people are developing with, and I think that's probably going to be the case. That's kind of how we're thinking about this overall.

- Meta Platforms CEO Mark Zuckerberg, 5/25

The groundswell of support for Llama has been awesome. We announced ten weeks ago a billion downloads after the release of Llama 4. In just ten weeks, that number is now 1.2. And if you look at Hugging Face (where the downloads are happening), what's cool is that most of these are derivatives. We have thousands of developers contributing.

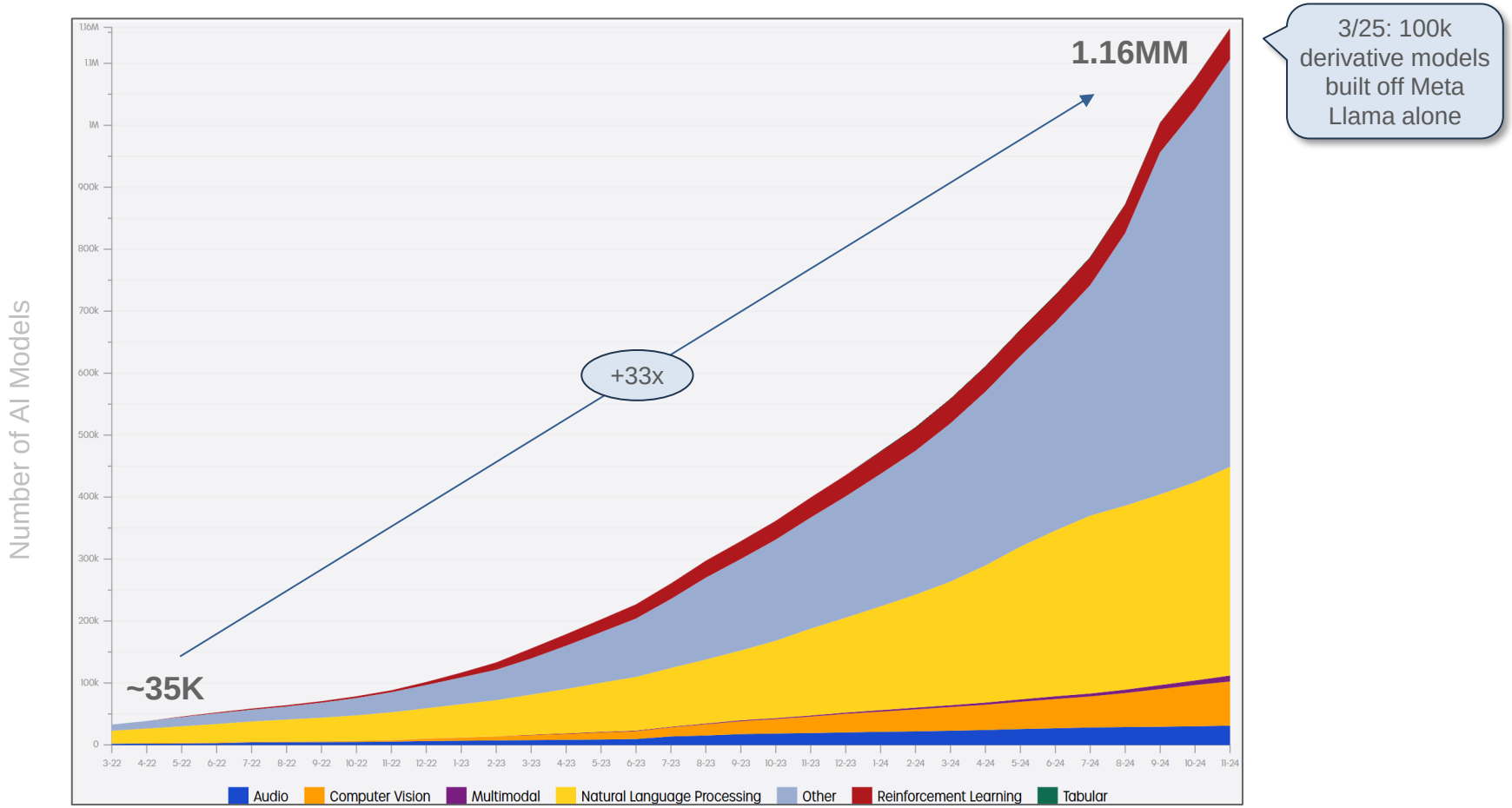
- Meta Platforms Chief Product Officer Chris Cox, 5/25



Note: 12/24 disclosure counted downloads of Llama and its derivatives. Source: Meta Platforms (8/24, 12/24, 3/25, 4/25), Stratchery podcast (5/25)

Developer AI Model Activity = +33x Increase in AI Models on Hugging Face – 11/24 vs. 3/22

AI Models Available from Hugging Face – 3/22-11/24, per Hugging Face



Note: Hugging Face is an online platform that hosts and shares machine learning models, datasets, and tools – commonly used to access, test, and deploy AI models, including large language models. It has become a central hub for the open-source AI community. May include open-source and closed models. Source: Hugging Face (5/25), Meta (3/25)

AI Monetization Threats

=

Rising Competition

+

Open-Source Model Momentum

(& China's Rise)

As noted on page 8, Meta CTO Andrew Bosworth referred to the current state of AI as *our space race* and *the people we're discussing, especially China, are highly capable…*

In this context, it is important to remember what the stakes of the Space Race were: proving which political system could innovate faster and win the world's trust in the process. Coming out on top in the Space Race played a role in enhancing USA's strategic deterrence and cementing the primacy of western democratic values. The AI 'space race,' also has the potential to reshape the world order.

China certainly knows these stakes. Back in 2015, 'Made in China 2025,' a new Chinese government initiative to shift the country from low-cost to high-value manufacturing in critical industries, seemed decades away. Fast forward to today, and China has dramatically accelerated its capabilities in these strategic sectors like robotics, electrification, and 'information technology' – best expressed by world-class artificial intelligence.

Chinese AI capabilities now underpin nationally strategic areas such as battlefield logistics, target recognition, cyber operations, and autonomous decision-making platforms. In 2025, Chinese state media highlighted the integration of AI into non-combat support functions (e.g., military hospitals), while the Ministry of Science and Technology reinforced its commitment to 'indigenous innovation' in strategic technologies.

The implications of Chinese AI supremacy would be profound.

As OpenAI's Sam Altman noted in a July 2024 *Washington Post* Op-Ed, *If [authoritarian regimes] manage to take the lead on AI, they will force U.S. companies and those of other nations to share user data, leveraging the technology to develop new ways of spying on their own citizens or creating next-generation cyberweapons to use against other countries.*

…Meanwhile, alongside AI, broader economic trade tensions between the USA and China continue to escalate, driven by competition for control over strategic technology inputs. China, for now, remains the dominant global supplier of 'rare earth elements' – materials essential to advanced electronics, defense systems, and clean energy infrastructure – an imbalance that the USA is working hard to counter. Simultaneously, the USA has prioritized the reshoring of semiconductor manufacturing, supported by the CHIPS and Science Act, and bolstered its partnerships with allied nations (including Japan, South Korea and the Netherlands) to reduce reliance on Chinese supply chains.

Taiwan continues to play a pivotal role in this dynamic. Despite American invention of core semiconductor technology like transistors and EUV lithography, it is Taiwan's TSMC – the world's most advanced semiconductor foundry – that drives global semiconductor production and is therefore central to both countries' strategic calculations.

It has taken a long time for the USA to wake up, but after two decades of inaction, both political parties are calling loudly for change. While each has taken a different approach (export controls in the Biden administration, economic nationalism and reshoring in the Trump administration), the move towards treating cutting-edge technology development as a core part of the national interest is a welcome adjustment.

As Senators John Cornyn and Mark Warner noted in 2020 regarding semiconductors,
America's innovation in semiconductors undergirds our entire innovation economy…unfortunately, our complacency has allowed our competitors – including adversaries – to catch up.

However, despite these measures, American intellectual property remains at risk; per OpenAI,
We know PRC (China) based companies – and others – are constantly trying to distill the models of leading US AI companies…it is critically important that we are working closely with the US government to best protect the most capable models from efforts by adversaries and competitors to take US technology.

What is clear, however, is that the American tone about Chinese technology has morphed since the early 2000s enthusiasm around China's entry into the World Trade Organization (WTO). AI, semiconductors, and critical minerals, and technology developments in general, are no longer viewed solely as economic or technology assets – they represent strategic levers of national resilience and geopolitical power, core to both the USA and China.

Public Market Capitalization Leader
Tells of Last Thirty Years =

Extraordinary USA Momentum…
China Rising

Global Public Market Capitalization Leaders – May, 2025 = 83% (25 of 30) USA-Based…

Global Public Companies Ranked By Market Capitalization – 5/15/25, per Capital IQ

Rank 2025	Company	HQ Country	Sector	Market Cap ($B)
1	Microsoft	USA	Software / AI	$3,368B
2	NVIDIA	USA	Semis / AI	3,288
3	Apple	USA	Hardware / AI	3,158
4	Amazon	USA	Internet / AI	2,178
5	Alphabet (Google)	USA	Internet / AI	1,997
6	Saudi Aramco	Saudi Arabia	Energy	1,686
7	Meta Platforms (Facebook)	USA	Internet / AI	1,619
8	Tesla	USA	Auto / AI	1,104
9	Broadcom	USA	Semis / AI	1,094
10	Berkshire Hathaway	USA	Finance	1,093
11	TSMC	Taiwan	Semis / AI	856
12	Walmart	USA	Consumer Products	771
13	JP Morgan Chase	USA	Finance	743
14	Visa	USA	Finance	678
15	Eli Lilly	USA	Healthcare	658
16	Tencent	China	Software / AI	591
17	Mastercard	USA	Finance	529
18	Netflix	USA	Internet / AI	501
19	Exxon Mobil	USA	Energy	468
20	Costco Wholesale	USA	Consumer Products	448
21	Oracle	USA	Hardware / AI	447
22	Procter & Gamble	USA	Consumer Products	381
23	Home Depot	USA	Consumer Products	376
24	Johnson & Johnson	USA	Consumer Products	360
25	SAP	Germany	Software / AI	343
26	Bank of America	USA	Finance	334
27	ICBC	China	Finance	330
28	AbbVie	USA	Healthcare	321
29	Coca-Cola	USA	Consumer Products	308
30	Palantir	USA	Software / AI	302

Source: Capital IQ (as of 5/15/25)

…Global Public Market Capitalization Leaders – December, 1995 = 53% (16 of 30) USA-Based

Global Public Companies Ranked By Market Capitalization – 12/31/95, per Bloomberg

Rank 1995	Company	HQ Country	Sector	Market Cap ($B)
1	Nippon Telegraph	Japan	Telco	$128B
2	General Electric	USA	Industrials	120
3	AT&T	USA	Telco	103
4	Exxon	USA	Energy	100
5	Coca-Cola	USA	Consumer Products	94
6	Merck	USA	Healthcare	81
7	Toyota	Japan	Automotive	79
8	Roche	Switzerland	Healthcare	78
9	Altria	USA	Consumer Products	75
10	Industrial Bank of Japan	Japan	Finance	71
11	MUFG Bank	Japan	Finance	68
12	Sumimoto Mitsui	Japan	Finance	66
13	Fuji Bank	Japan	Finance	64
14	Dai-Ichi Kangyo Bank	Japan	Finance	61
15	UFJ Bank	Japan	Finance	59
16	Novartis	Switzerland	Healthcare	57
17	Procter & Gamble	USA	Consumer Products	57
18	Johnson & Johnson	USA	Consumer Products	55
19	Microsoft	USA	Software	52
20	Walmart	USA	Consumer Products	51
21	IBM	USA	Hardware / Software	51
22	DirecTV	USA	Media	49
23	Intel	USA	Hardware	47
24	BP	United Kingdom	Energy	46
25	Nestle	Switzerland	Consumer Products	45
26	Mobil	USA	Energy	44
27	PepsiCo	USA	Consumer Products	44
28	AIG	USA	Finance	44
29	Shell	United Kingdom	Energy	44
30	Sakura Bank	Japan	Finance	43

Source: Bloomberg (as of 5/15/25)

Over the past thirty years (1995 to 2025), just six companies remained on the top 30 most highly valued **publicly traded global companies** – Microsoft / Walmart / Exxon Mobil / Procter & Gamble / Johnson & Johnson / Coca-Cola.

New entrants are NVIDIA / Apple / Amazon / Alphabet (Google) / Saudi Aramco / Meta Platforms (Facebook) / Tesla / Broadcom / Berkshire Hathaway / TSMC / JP Morgan Chase / Visa / Eli Lilly / Tencent / Mastercard / Netflix / Costco Wholesale / Oracle / Home Depot / SAP / Bank of America / ICBC / AbbVie / Palantir.

In 1995, USA had 53% (16 of 30) of the most valuable companies and 83% (25 of 30) in 2025.
Japan came next with 9, now 0.
Switzerland followed with 3, now 0. UK had 2, now 0.

In 2025, new geographic entrants include
China with 2 and Saudi Arabia / Taiwan / Germany with 1 each.

Global Public Technology Market Cap Leaders – May, 2025 = 70% (21 of 30) USA-Based…

Global Technology Companies Ranked By Market Capitalization – 5/15/25, per Capital IQ

Rank 2025	Company	HQ Country	Sector	Market Cap ($B)
1	Microsoft	USA	Software / AI	$3,368B
2	NVIDIA	USA	Semis / AI	3,288
3	Apple	USA	Hardware / AI	3,158
4	Amazon	USA	Internet / AI	2,178
5	Alphabet (Google)	USA	Internet / AI	1,997
6	Meta Platforms (Facebook)	USA	Internet / AI	1,619
7	Tesla	USA	Auto / AI	1,104
8	Broadcom	USA	Semis / AI	1,094
9	TSMC	Taiwan	Semis / AI	856
10	Tencent	China	Software / AI	591
11	Netflix	USA	Internet / AI	501
12	Oracle	USA	Hardware / AI	447
13	SAP	Germany	Software / AI	343
14	Palantir	USA	Software / AI	302
15	ASML	Netherlands	Semis / AI	300
16	Alibaba	China	Internet / AI	281
17	Salesforce	USA	Software / AI	279
18	T-Mobile	USA	Telco	273
19	Samsung	S. Korea	Hardware / AI	268
20	Cisco	USA	Semis / AI	256
21	IBM	USA	Hardware / AI	243
22	China Mobile	China	Telco	241
23	Reliance	India	Telco	216
24	ServiceNow	USA	Software / AI	214
25	Intuitive Surgical	USA	Health Tech	201
26	AT&T	USA	Telco	197
27	Siemens	Germany	Hardware / AI	194
28	Uber	USA	Internet / AI	189
29	AMD	USA	Semis / AI	186
30	Intuit	USA	Software / AI	185

Source: Capital IQ (as of 5/15/25)

…Global Public Technology Market Cap Leaders – December, 1995 = 53% (16 of 30) USA-Based

Global Technology Companies Ranked By Market Capitalization – 12/31/95, per Bloomberg

Rank 1995	Company	HQ Country	Sector	Market Cap ($B)
1	Nippon Telegraph	Japan	Telco	$128B
2	AT&T	USA	Telco	103
3	Microsoft	USA	Software	52
4	IBM	USA	Hardware / Software	51
5	Intel	USA	Hardware	47
6	BellSouth	USA	Telco	43
7	HP	USA	Hardware	43
8	GTE	USA	Telco	42
9	BT	United Kingdom	Telco	34
10	Panasonic	Japan	Hardware	34
11	SingTel	Singapore	Telco	34
12	Motorola	USA	Hardware	34
13	Hitachi	Japan	Hardware	33
14	Verizon	USA	Telco	29
15	Toshiba	Japan	Hardware	26
16	Peraton	USA	Software / Hardware	25
17	Nynex	USA	Telco	24
18	Sony	Japan	Hardware	22
19	Cisco	USA	Hardware	21
20	Fujitsu	Japan	Hardware	20
21	PCCW	Hong Kong	Telco	20
22	NEC	Japan	Software	19
23	Oracle	USA	Hardware	18
24	MCI	USA	Telco	18
25	Sharp	Japan	Hardware	18
26	TelMex	Mexico	Telco	17
27	KDDI	Japan	Telco	17
28	US West	USA	Telco	17
29	Cable & Wireless	USA	Telco	16
30	Telekom Malaysia	Malaysia	Telco	16

Source: Bloomberg (as of 5/15/25)

Over the past thirty years (1995 to 2025), just five companies remained on the top 30 most highly valued **publicly traded global technology companies** – Microsoft / Oracle / Cisco / IBM / AT&T.

New entrants are NVIDIA / Apple / Amazon / Alphabet (Google) / Meta Platforms (Facebook) / Tesla / Broadcom / TSMC / Tencent / Netflix / SAP / Palantir / ASML / Alibaba / Salesforce / T-Mobile / Samsung / China Mobile / Reliance / ServiceNow / Intuitive Surgical / Siemens / Uber / AMD / Intuit.

In 1995, USA had 53% (16 of 30) of the most valuable tech companies and 70% (21 of 30) in 2025.
In 1995, Japan had 30% (9 of 30) of the top tech companies and 0 in 2025.
UK / Singapore / Hong Kong / Mexico / Malaysia had 1, now 0.

In 2025, new geographic entrants include China with 3, Germany with 2, Taiwan with 1, Netherlands with 1, South Korea with 1 & India with 1.

Note that while Taiwan has only one company on the list – TSMC – the company produces 80%-90% of the world's most advanced semiconductors and 62%+ of global semiconductors as of Q2:24, per The Center for Strategic & International Studies & Counterpoint Research.

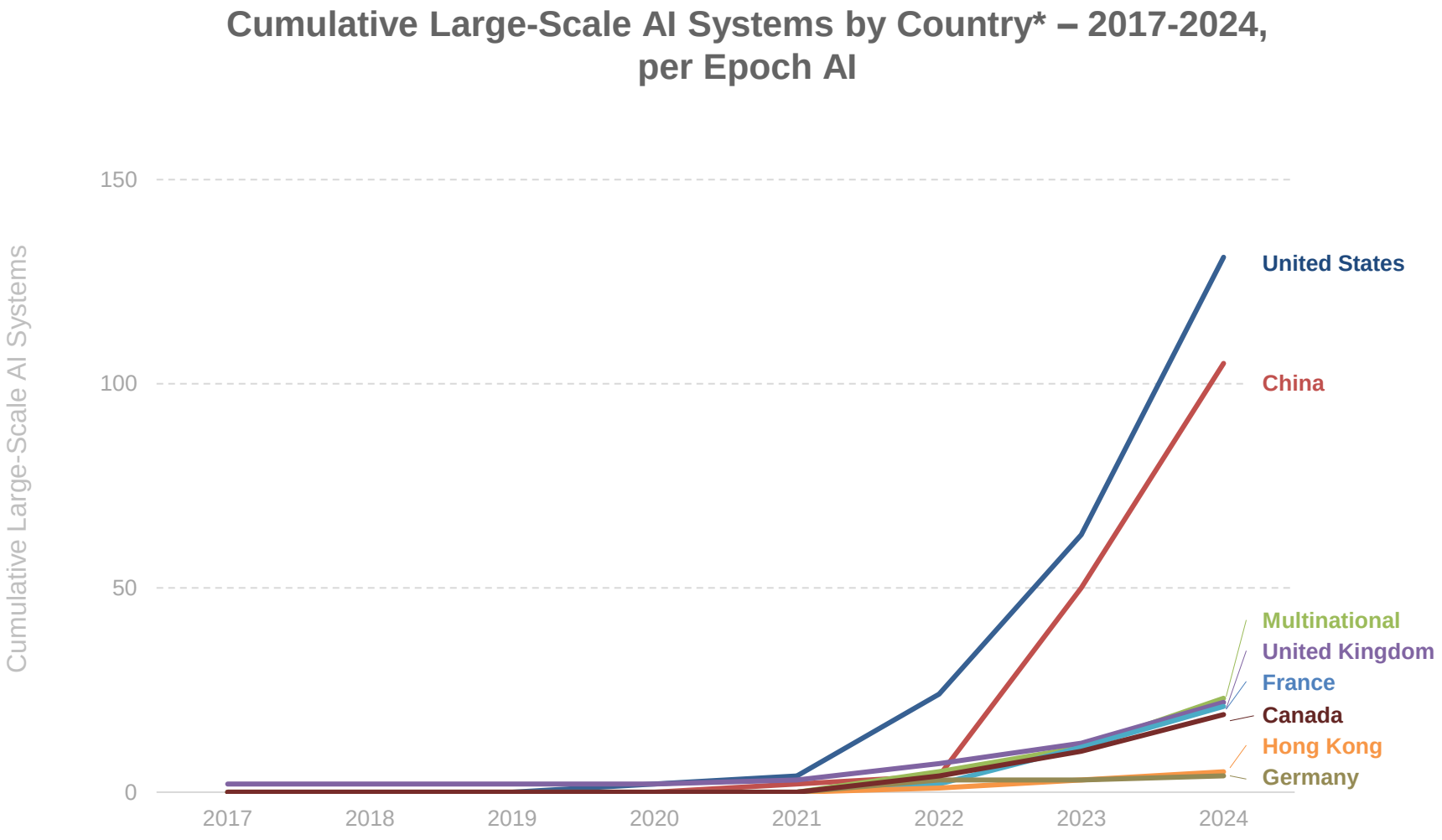
It's stunning how much can change in a generation…
the emergence of internet connectivity was foundational to most of the new adds.
The emergence of AI will have the same type of effect over the next three decades, but likely faster.

Source: Center for Strategic & International Studies, 'A Strategy for The United States to Regain its Position in Semiconductor Manufacturing' (2/24); Counterpoint Research, 'Global Semiconductor Foundry Market Share: Quarterly' (3/25)

USA vs. China in Technology =

China’s AI Response Time
Significantly Faster vs. Internet 1995

AI Large Language Model (LLM) Leadership = USA & China Outpacing Rest of World (RoW), per Epoch AI



Hong Kong is a Special Administrative Region (SAR) of China, not an independent country. Note: Epoch AI defines AI models as 'large-scale' when their training compute is confirmed to exceed 10^{23} floating-point operations. Source: Epoch AI via Our World In Data (5/25)

China AI = Rapid Relevance…
DeepSeek R1 – 1/20/25…

Reuters

World Business Markets Sustainability Legal Breakingviews Technology Investigations More

DeepSeek narrows China-US AI gap to three months, 01.AI founder Lee Kai-fu says

By Liam Mo and Kane Wu

March 25, 2025 12:08 AM PDT · Updated 20 hours ago

The Deepseek logo is seen in this illustration taken Jan. 27, 2025. REUTERS/Dado Ruvic/Illustration/File Photo Purchase Licensing Rights

BEIJING/HONG KONG, March 25 (Reuters) - China has narrowed the AI development gap with the United States to just three months in some areas, because firms such as DeepSeek have worked out how to use chips and apply algorithms more efficiently, the CEO of Chinese startup 01.AI Lee Kai-fu said.

Lee, a prominent figure in the global artificial intelligence space and a former head of Google China, told Reuters that startup DeepSeek revealed that China had pulled ahead in areas such as infrastructure software engineering

We believe that as the economy develops, China should gradually become a contributor instead of freeriding. In the past 30+ years of the IT wave, we basically didn't participate in real technological innovation. We're used to Moore's Law falling out of the sky, lying at home waiting 18 months for better hardware and software to emerge. That's how the Scaling Law is being treated…

What we see is that Chinese AI can't be in the position of following forever. We often say that there is a gap of one or two years between Chinese AI and the United States, but the real gap is the difference between originality and imitation. If this doesn't change, China will always be only a follower – so some exploration is inescapable.

- DeepSeek CEO Liang Wenfang, 11/24

Source: Reuters, 'DeepSeek narrows China-US AI gap to three months, 01.AI founder Lee Kai-Fu says' (3/25); China Talk Media (11/24)

…China AI = Rapid Relevance…
Alibaba Qwen 2.5-Max – 1/29/25…

Mashable

Meet Alibaba's Qwen 2.5, an AI model claiming to beat both DeepSeek and OpenAI's ChatGPT

Mashable

Wed, January 29, 2025 at 9:11 AM PST · 1 min read

Photo Illustration by Thomas Fuller/SOPA Images/LightRocket via Getty Images

Alibaba, the Chinese tech giant, released a new version of its AI model and made big claims — notably that it outperforms OpenAI's ChatGPT and the newly ascending DeepSeek.

Qwen2.5-Max outperforms DeepSeek V3 in benchmarks such as Arena-Hard, LiveBench, LiveCodeBench, and GPQA-Diamond, while also demonstrating competitive results in other assessments, including MMLU-Pro.

Our base models have demonstrated significant advantages across most benchmarks, and we are optimistic that advancements in post-training techniques will elevate the next version of Qwen2.5-Max to new heights.

The scaling of data and model size not only showcases advancements in model intelligence but also reflects our unwavering commitment to pioneering research. We are dedicated to enhancing the thinking and reasoning capabilities of large language models through the innovative application of scaled reinforcement learning.

- Alibaba Qwen 2.5 Press Release, 1/25

Source: Mashable, 'Meet Alibaba's Qwen 2.5, an AI model claiming to beat both DeepSeek and OpenAI's ChatGPT' (1/25); Alibaba (1/25)

…China AI = Rapid Relevance…
Baidu Ernie 4.5 Turbo – 4/25/25

Baidu launches new AI model amid mounting competition

By **Reuters**

April 24, 2025 9:05 PM PDT · Updated 12 days ago

People walk near a Baidu logo at the company headquarters in Beijing, China April 23, 2021. REUTERS/Florence Lo/File Photo Purchase Licensing Rights

BEIJING, April 25 (Reuters) - Chinese search engine giant Baidu (9888.HK) launched its latest artificial intelligence model, Ernie 4.5 Turbo, on Friday, as it ramps up AI product offerings amid mounting competition in the Chinese AI market.

It also launched its new reasoning model Ernie X1 Turbo.

ERNIE 4.5 Turbo is the newest member of the flagship ERNIE foundation model family. Imagine an AI that's not just smart, but also affordable and versatile. Here's why it's turning heads:

- Multimodal Prowess: It excels in handling text, images, and even videos, making it a Swiss Army knife for developers.

- Cost-Effectiveness: Priced at just RMB 0.8 per million tokens for input and RMB 3.2 for output, it's 80% cheaper than its predecessor – and a fraction of the cost of leading competitors. It costs only 40% of DeepSeek V3 and just 0.2% of GPT-4.5.

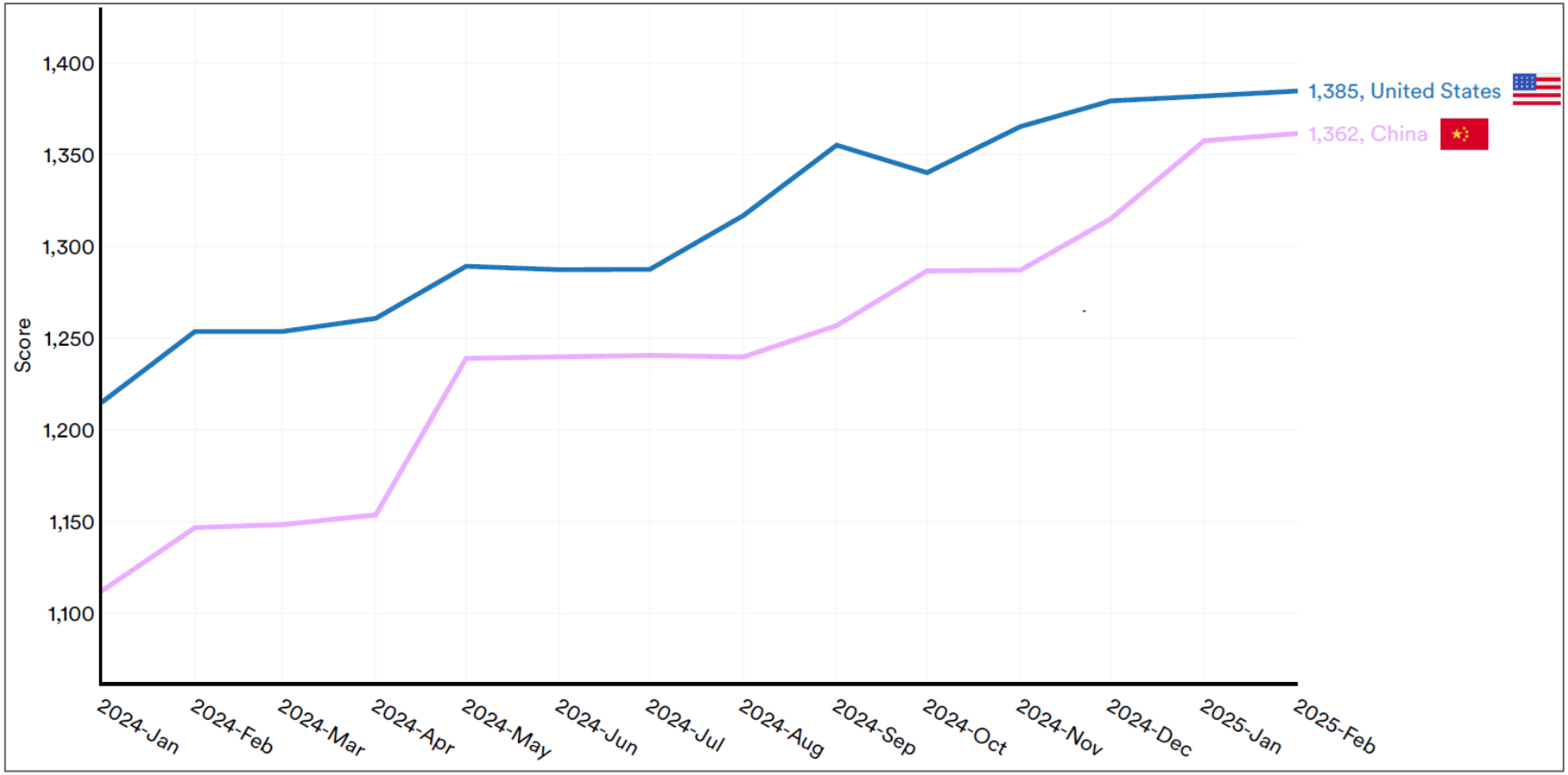
- High Performance: Benchmark tests show it matches GPT-4.1 and outperforms GPT-4o in most multimodal tasks – delivering high-impact results with every run.

- Baidu Post on X, 4/24/25

Source: Reuters, 'Baidu launches new AI model amid mounting competition' (4/24/25); Baidu via X, 'Supercharging AI Innovation with More Powerful and More Affordable New Models' (4/24/25)

China AI =
LLM Performance Catching Up to USA Models, per Stanford HAI…

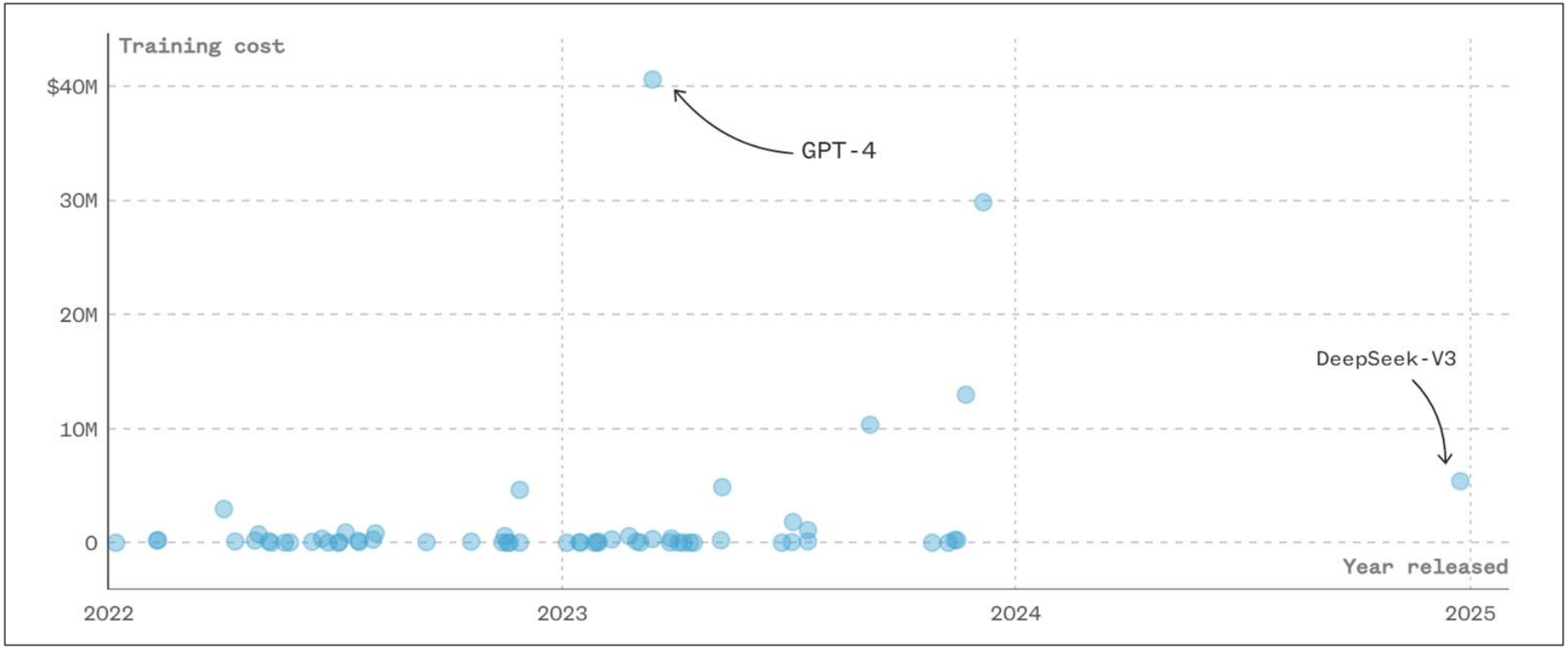
Performance of Top-Scoring USA vs. Chinese AI Model on LMSYS Chatbot Arena – 1/24-2/25, per Stanford HAI & LMSYS



Note: The LMSYS Chatbot Arena is a public website where people compare two AI chatbots by asking them the same question and voting on which answer is better. The results help rank how well different language models perform based on human judgment. Only the highest-scoring model in any given month is shown in this comparison.
Source: LMSYS via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

…China AI =
LLMs Achieving Performance with Lower Training Costs, per Epoch AI…

LLM Training Cost by Year Released – 2022-2024, per Epoch AI & NBC News



Source: Epoch AI via NBC News, 'Why DeepSeek is Different, in Three Charts' (1/25)

…China AI =
LLMs Increasingly Powered by Local Semiconductors…

Huawei delivers advanced AI chip 'cluster' to Chinese clients cut off from Nvidia

CloudMatrix 384 sent to data centres serving Chinese tech companies as they seek to fill gap created by US export controls

Huawei is accelerating its chip development to help Chinese tech companies compete with global peers © Tobias Schwarz/AFP/Getty Images

Zijing Wu in Hong Kong and **Eleanor Olcott** in Beijing

Published APR 29 2025

60

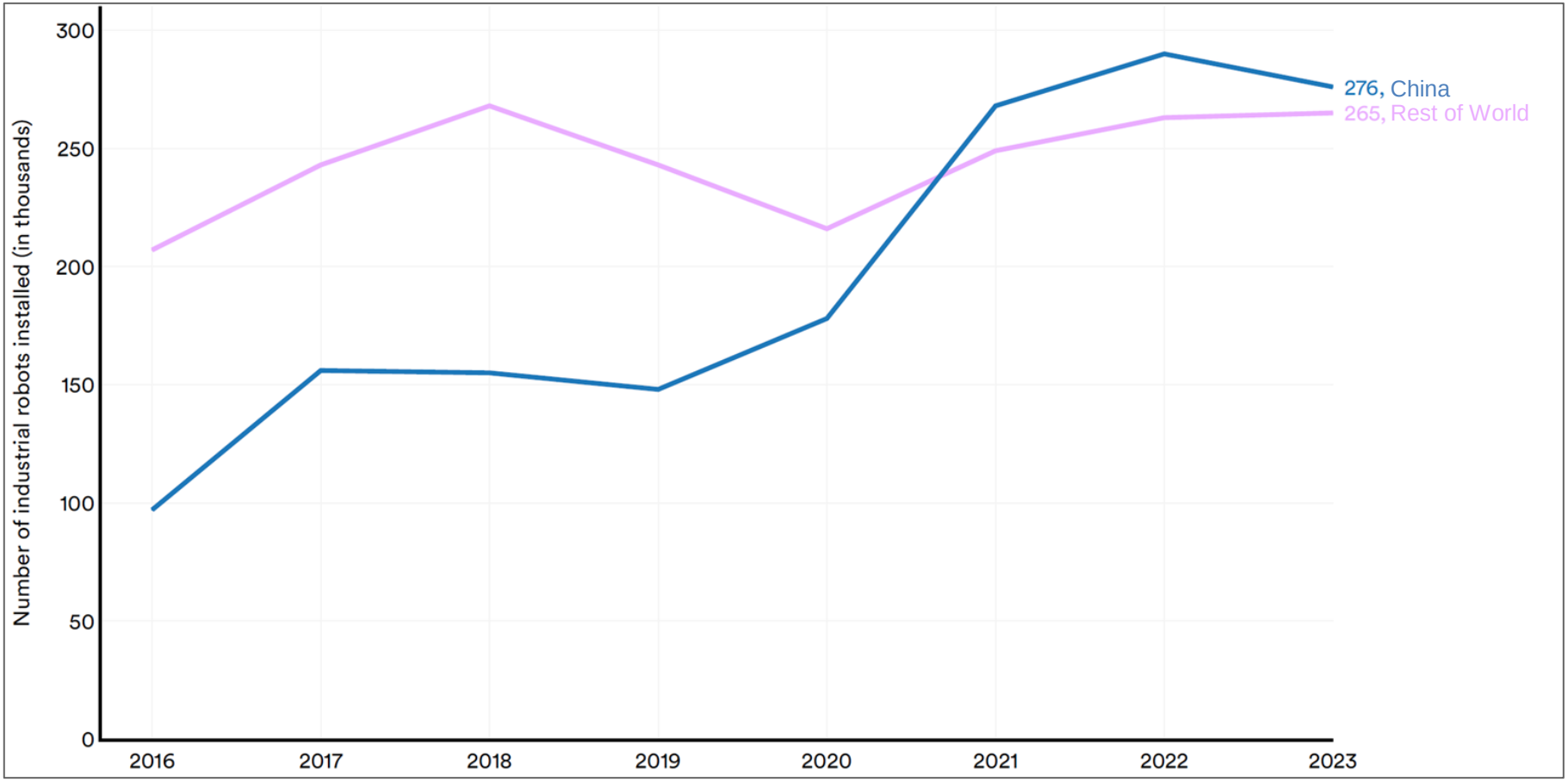
Huawei has started the delivery of its advanced artificial intelligence chip 'cluster' to Chinese clients who are increasing orders after being cut off from Nvidia's semiconductors because of Washington's export restrictions…

- Financial Times, 4/29/25

Source: Financial Times, 'Huawei delivers advanced AI chip 'cluster' to Chinese clients cut off from Nvidia' (4/29/25)

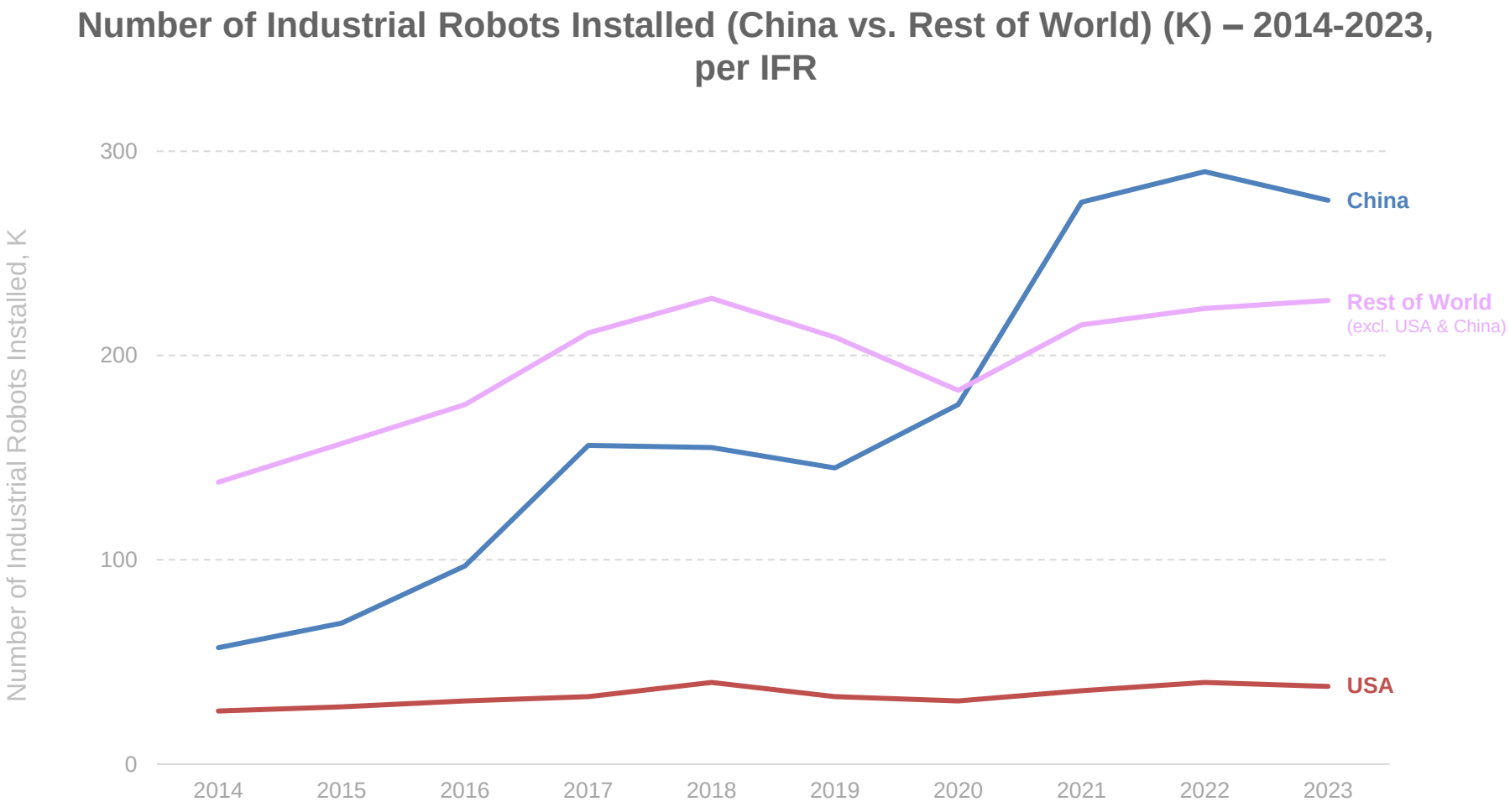
…China AI =
Industrial Robot Installed Base Higher vs. Rest of World…

Number of Industrial Robots Installed (China vs. Rest of World) (K) – 2023, per IFR



Source: International Federation of Robotics (IFR) (2024) via Nestor Maslej et al., 'The AI Index 2025 Annual Report,' AI Index Steering Committee, Stanford HAI (4/25)

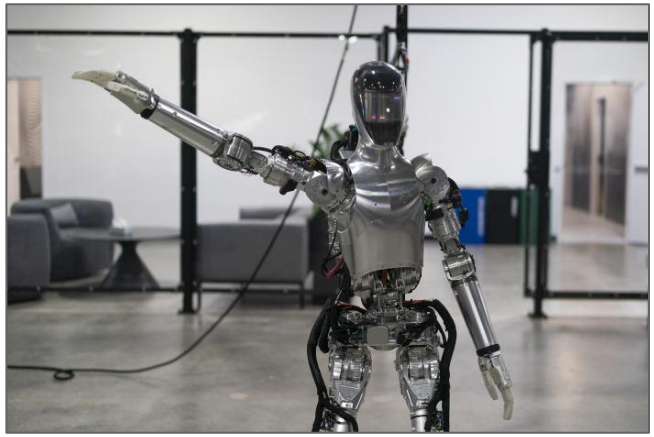
…China AI =
Industrial Robot Installed Base Higher vs. Rest of World

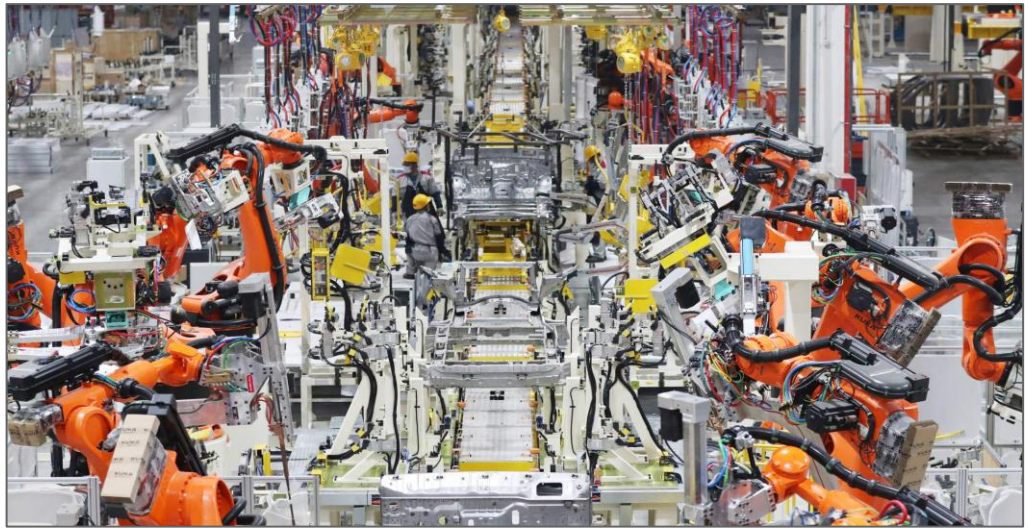


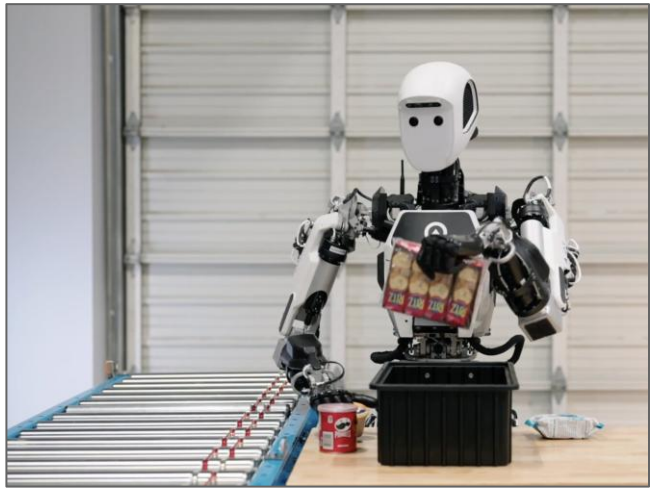
Source: International Federation of Robotics (IFR) (2024)

Robots – Industrial & Humanoid = Creating New Data @ New Scale

Images of Industrial & Humanoid Robots, per *The Wall Street Journal*







Source: The Wall Street Journal (2/18, 5/22, 9/22, 5/25)