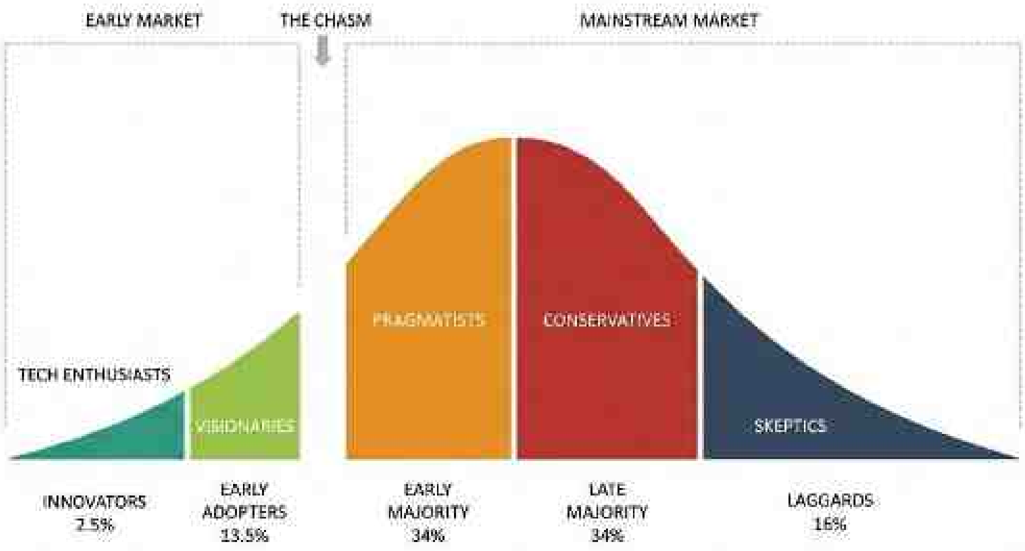
AI Trends 2025

AI agents crossed the chasm



What is Generational?

Generational

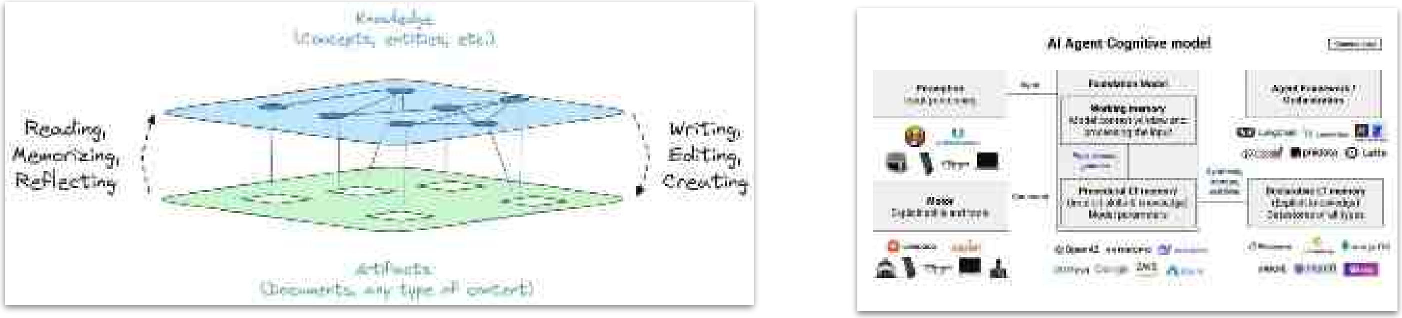
Market & Economic Trends

AI Trends 2024
A story of power laws and the tipping point for agents
AI is driving the stock market
Framework for agentic automation

Company Reports

Power Generative AI With Your Data
databricks
Your data. Your AI. Your future.
Own them all on the new data intelligence platfor
From Generic To Personalized

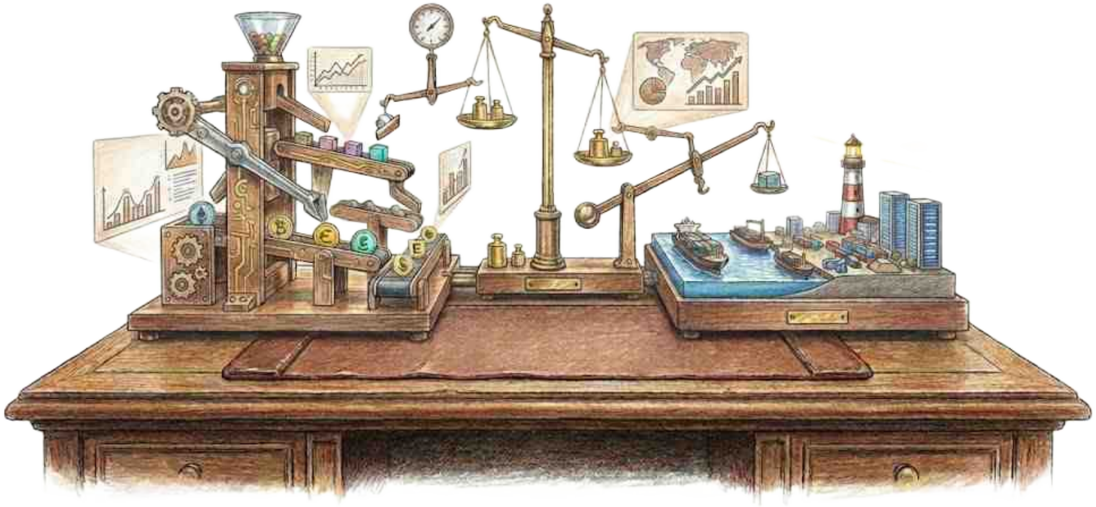
Product & Tech Deep Dives


Knowledge (concepts, entities, etc.)
Reading, Memorizing, Reflecting
Writing, Editing, Creating
Artifacts (Documents, any type of content)
AI Agent Cognitive model
Foundation Model
Working memory
Perception
Action
Agent Frameworks / Orchestration
LangChain
LlamaIndex

Curated Networking & Events

Table of Contents

Section 1: Macroeconomics



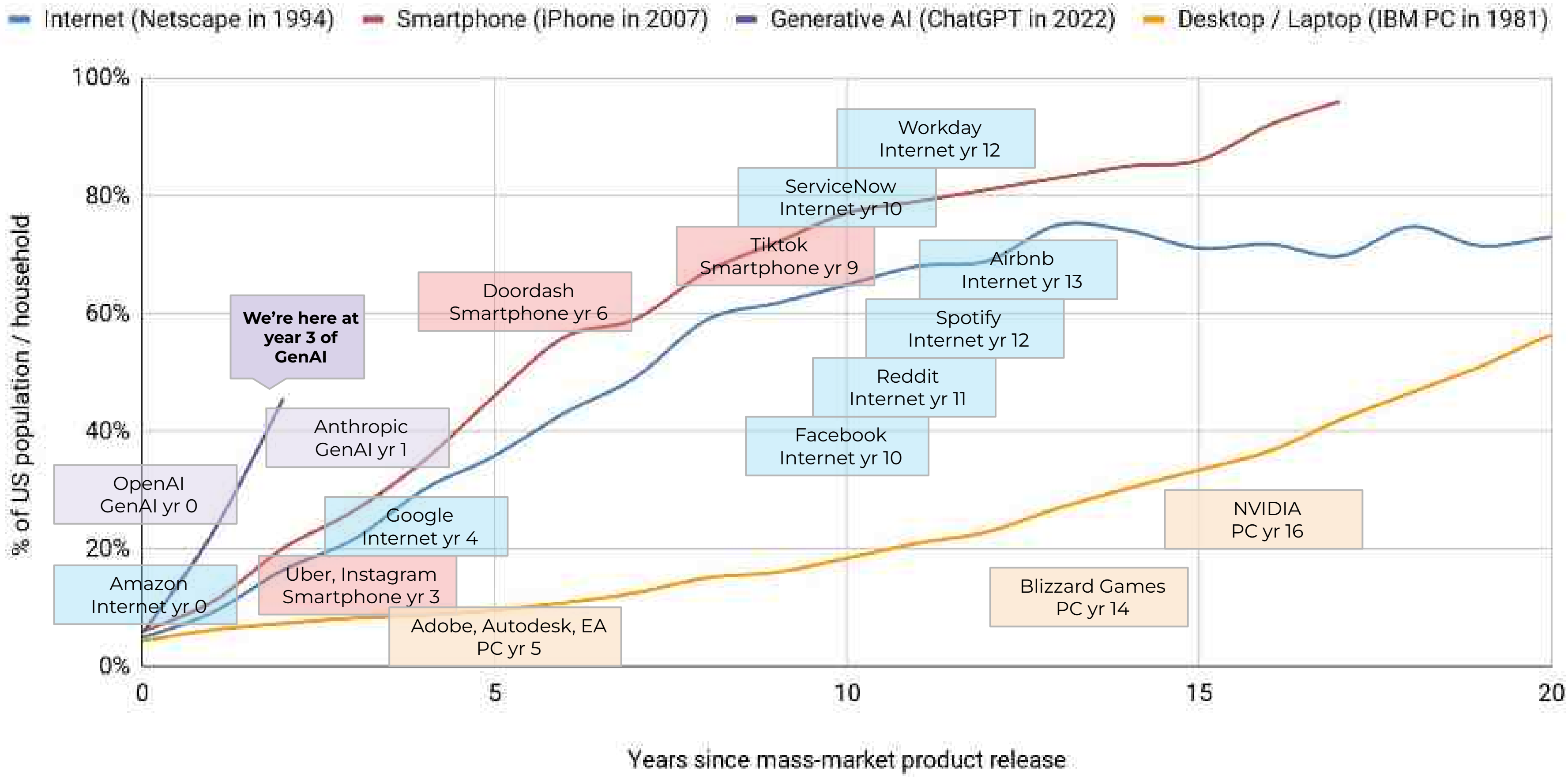
Rapid adoption of generative AI

Generational

Foundation models are the fastest-adopted general-purpose technology (GPT) in history. A GPT refers to a foundational innovation that has broad applications across industries, significantly transforming how we work and live. In just two years, >50% of the U.S. population has used a generative AI product. Each wave of GPT has historically created opportunities for companies to build transformative products, leading to the emergence of iconic businesses. Interestingly, these waves tend to produce two distinct cohorts of leading companies:

- Desktops: Adobe, Blizzard, NVIDIA
- Internet: Amazon, Google, Facebook
- Smartphones: Uber, Doordash, Tiktok

Adoption Pace of General Purpose Technologies in US Households



Source: Generative AI adoption tracker, Statista, Wired, US Census, Comscore, IBIS, World Bank, Author Analysis

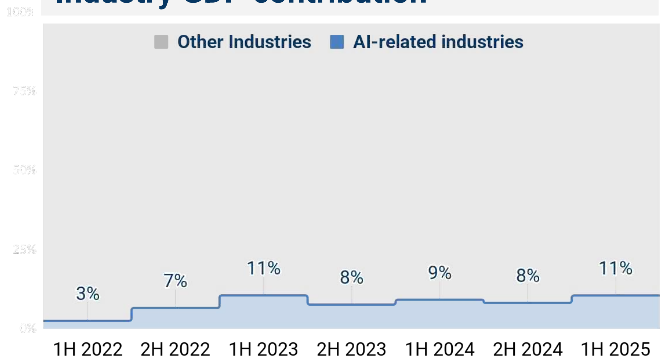
AI macro impact mostly in jobs so far

Generational

Deep dive in the following slides

Gross Domestic Product – AI contributions are overstated

Industry GDP contribution

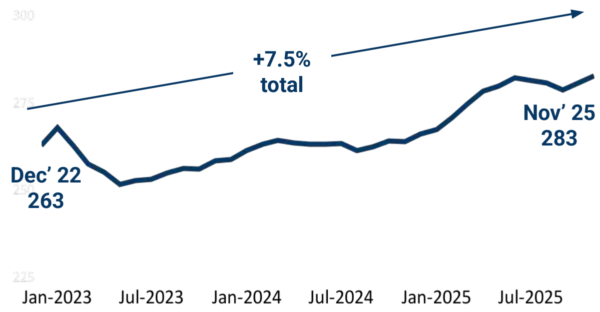


AI is often cited as driving most of US GDP growth in 2025—approximately 1.0–1.2% of 1.6% total growth. However, when accounting for import leakage to Taiwan and other chip-manufacturing countries, domestic impact is smaller.

GDP contribution of AI-related industries has hovered around 8–10% over recent years. While elevated from pre-ChatGPT levels, it remains a modest share of overall GDP.

Inflation – limited national effect, local variation

Household Energy Cost Index

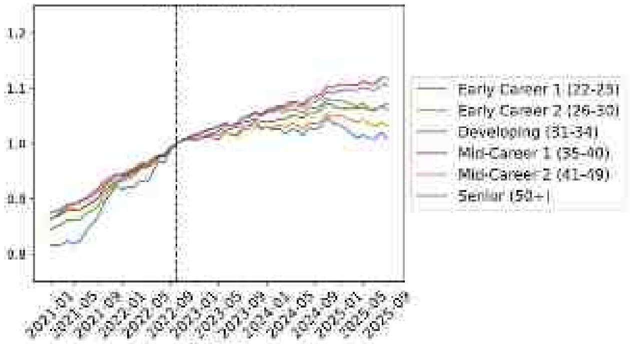


The main drivers of consumer price increases remain non-AI related: housing, transportation, food, and healthcare.

The most AI-relevant CPI component is household energy, which represents only 3% of the index. Household energy prices increased 7.5% since late 2022, but given its small weight, aggregate impact is limited. That said, regions with data center construction—such as the District of Columbia and Virginia—have seen 10–30% year-over-year increases.

Employment – young AI-exposed workers impacted

Employment by Age Group (Normalized)



Broad labor displacement has not materialized. However, evidence suggests young workers in highly automatable roles are affected.

Employment among young workers has declined post-ChatGPT, diverging from historical patterns where employment rises with population and graduation growth.

Additional studies show job postings for AI-exposed roles are down 12% relative to less exposed roles.

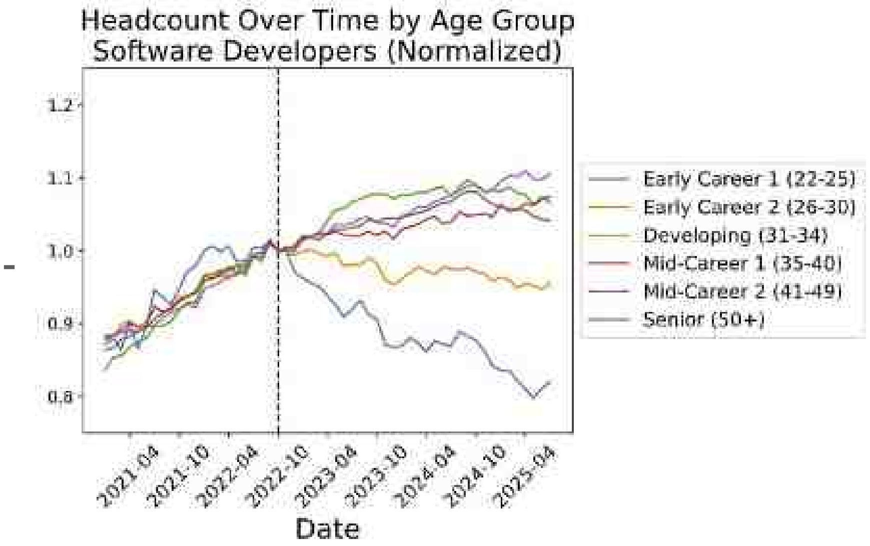
Source: Author analysis, US Bureau Labor of Statistics, US Bureau of Economic Analysis, Stanford

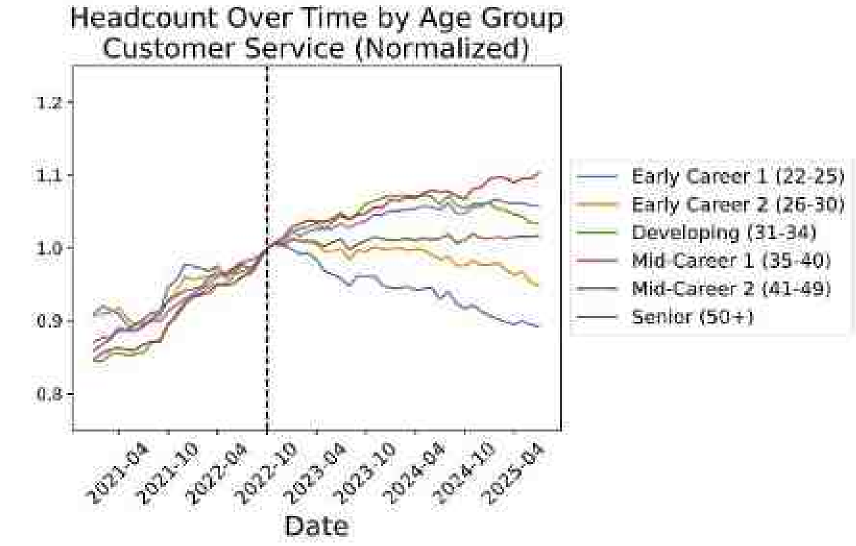
Young workers face steepest AI-driven employment declines

Generational

Payroll data shows that the impact of generative AI is not spread evenly across the workforce. Since late 2022, employment for 22- to 25-year-olds in highly exposed occupations—most notably software development and customer service—has fallen sharply, even as overall job growth continues. By mid-2025, early-career software developers saw employment drop nearly 20% from its peak, while their older peers held steady or grew. A similar divergence is visible in customer service roles.

Occupational Group	Conversation Distribution
Computer and Mathematical	35.9%
Educational Instruction and Library	12.3%
Arts, Design, Entertainment, Sports, and Media	8.2%
Office and Administrative Support	8.2%
Life, Physical, and Social Science	7.2%
Business and Financial Operations	3.0%
Healthcare Practitioners and Technical	2.5%
Management	2.6%
Architecture and Engineering	2.5%
Community and Social Service	2.0%
Production	1.6%
Sales and Related	2.4%
Installation, Maintenance, and Repair	0.5%
Legal	0.7%
Personal Care and Service	0.6%
Building and Grounds Cleaning and Maintenance	0.1%
Construction and Extraction	0.2%
Farming, Fishing, and Forestry	0.4%
Food Preparation and Serving Related	0.4%
Healthcare Support	0.3%
Protective Service	0.2%
Transportation and Material Moving	0.2%

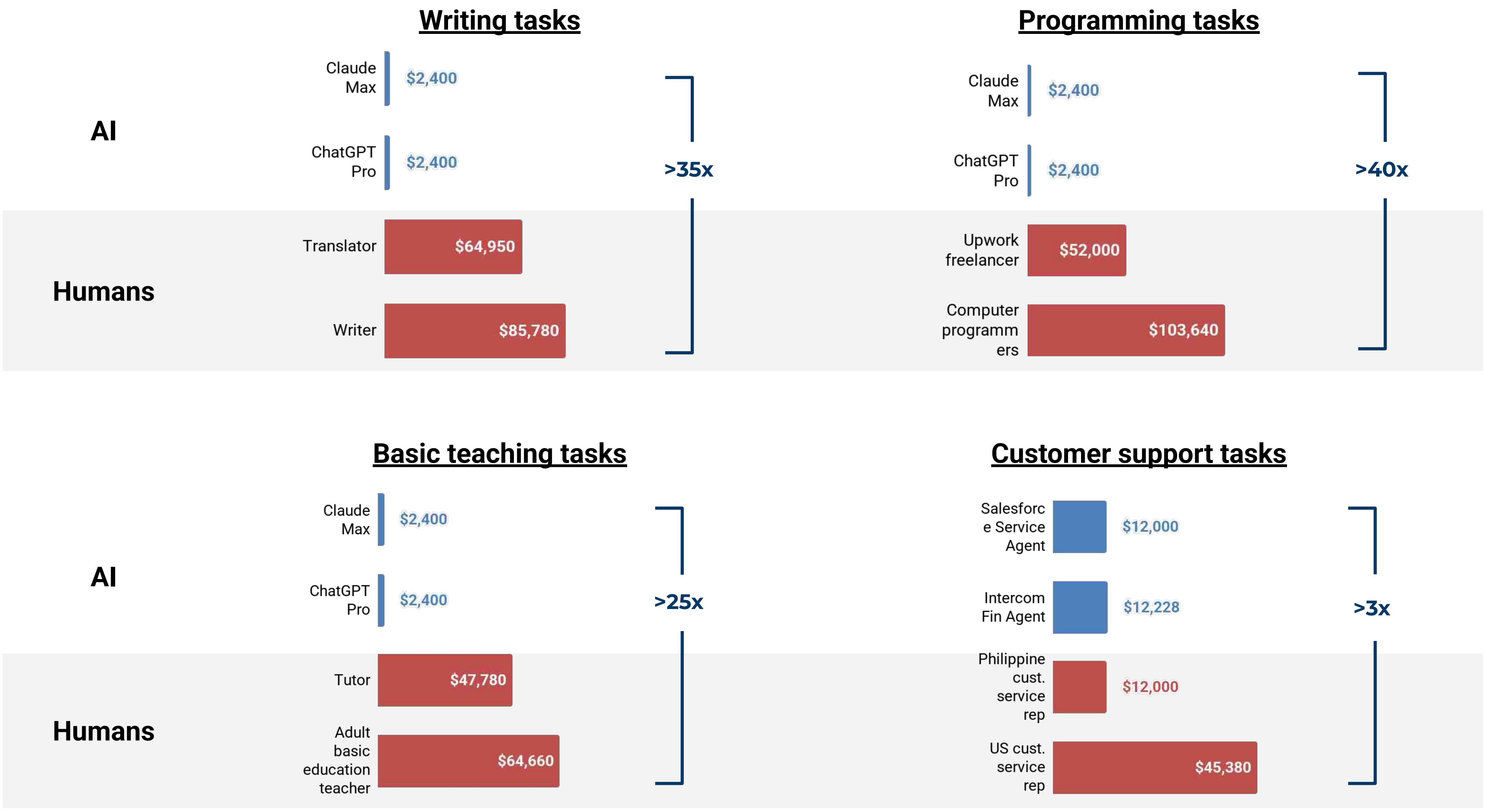




Source: Author analysis, Anthropic, Stanford

AI agents are >20 times cheaper for basic tasks and >3x cheaper for advanced tasks

The low cost of AI tools compared to human labor makes them an easy choice for many. This drives adoption, reshaping industries and how work is done. A common rule is that for companies to adopt a new product, it must be 10x better or 10x cheaper. Leading AI products can already do basic knowledge work tasks for over 20x cheaper with more advanced work at over 3x chapear.

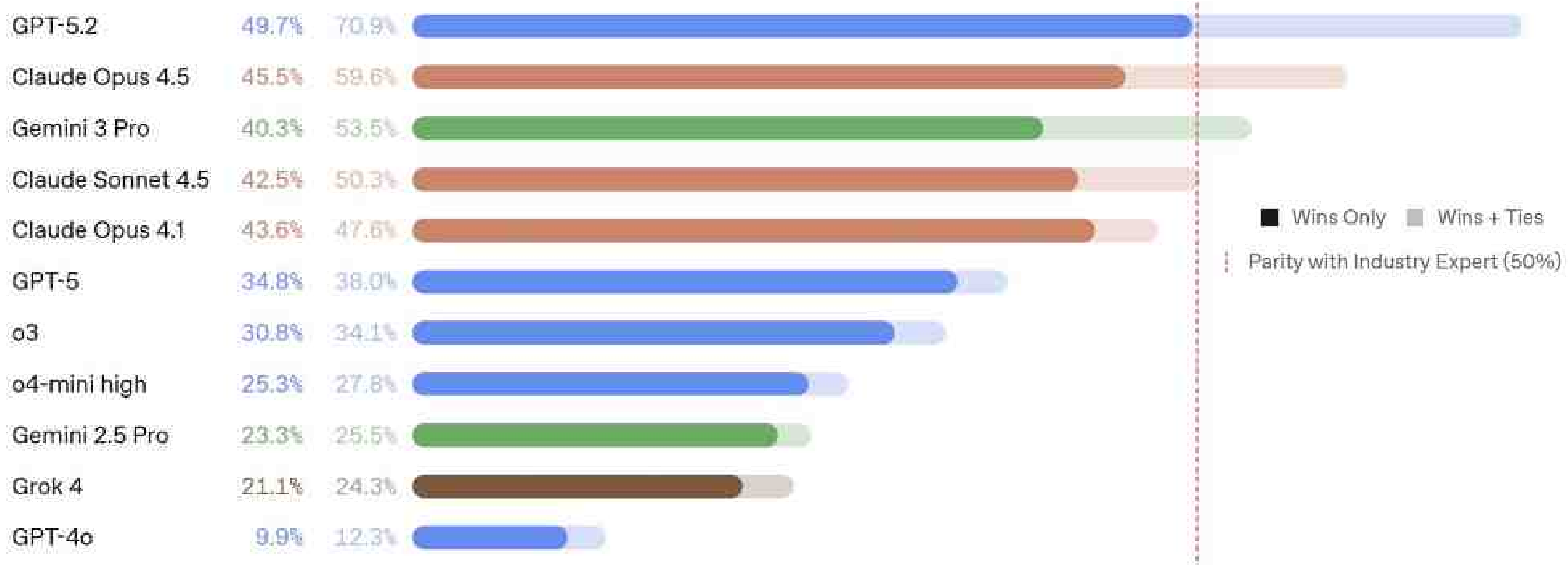


Source: Author analysis, US Bureau Labor of Statistics annual wages, company websites. Freelancer wage assumes 2,080 hours / year. Customer service software assumes 12,000 resolutions / year / agent

AI agents product work better than professionals w/ 14 years of experience

AI quality now matches professional human output for many tasks. This matters because companies won't adopt AI that produces inferior work, regardless of cost savings. The GDPVal benchmark tests whether AI can create complete professional deliverables—like 3D engineering models, financial analyses, and customer service responses—that match work from experts with 14 years of experience. GPT-5.2 matches or beats human experts 70.9% of the time across these tasks, showing AI can replace rather than just assist human workers for specific work outputs.

GDPVal Leaderboard

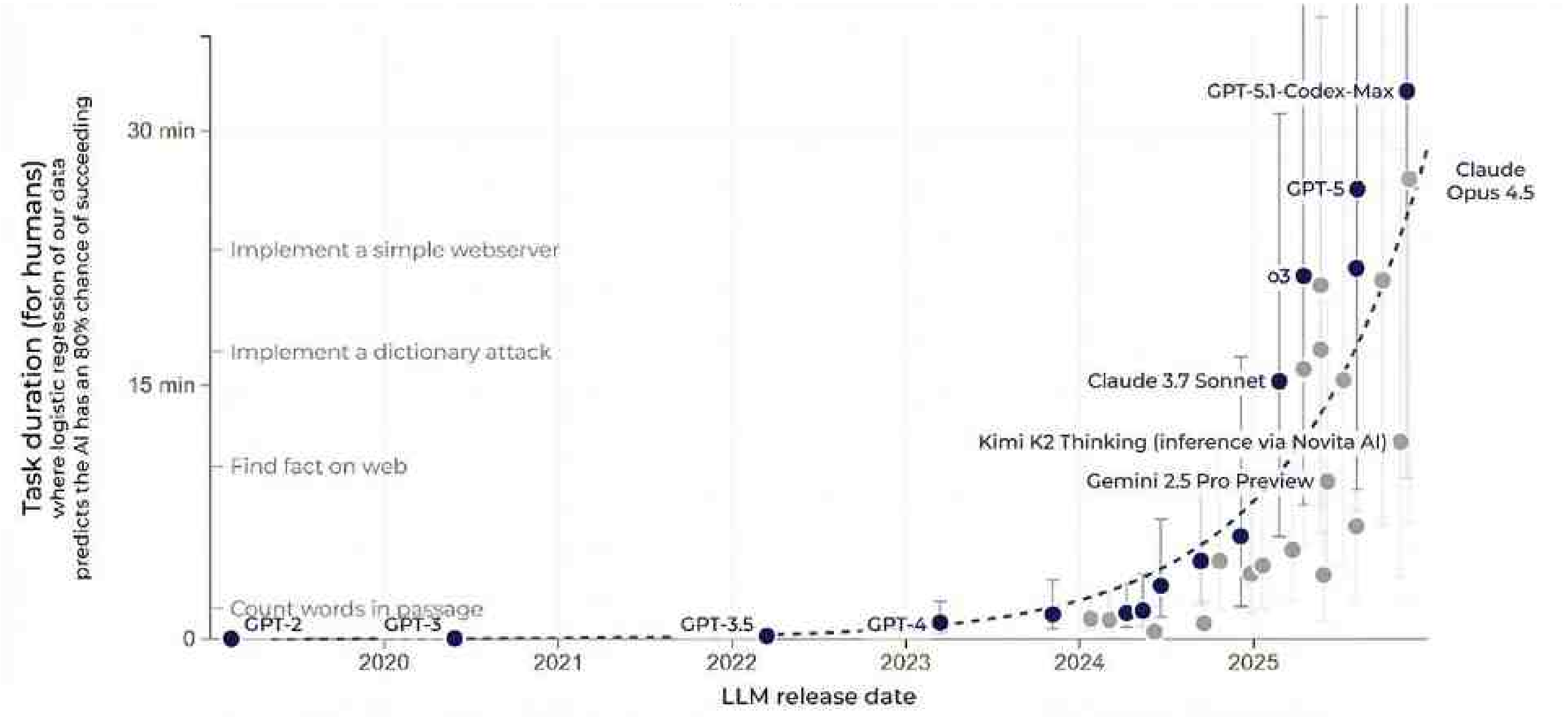


Source: OpenAI

AI agents can work faster and longer than most humans

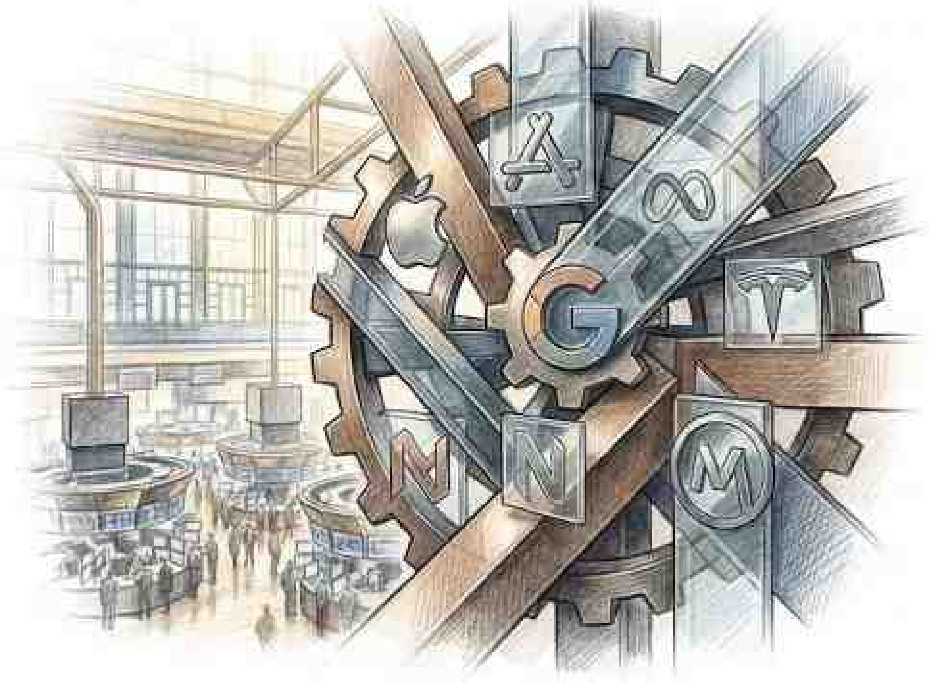
AI can now autonomously complete tasks that take human professionals 30 minutes with 80% reliability—the success rate expected of professional workers. The METR benchmark measures task length by how long human experts need to complete software engineering work, then tests whether AI can finish autonomously. Leading models like Claude Opus 4.5 handle 30-minute tasks reliably, compared to near-zero capability just 2-3 years ago.

Time-horizon of software engineering tasks different LLMs can complete 80% of the time



Source: METR

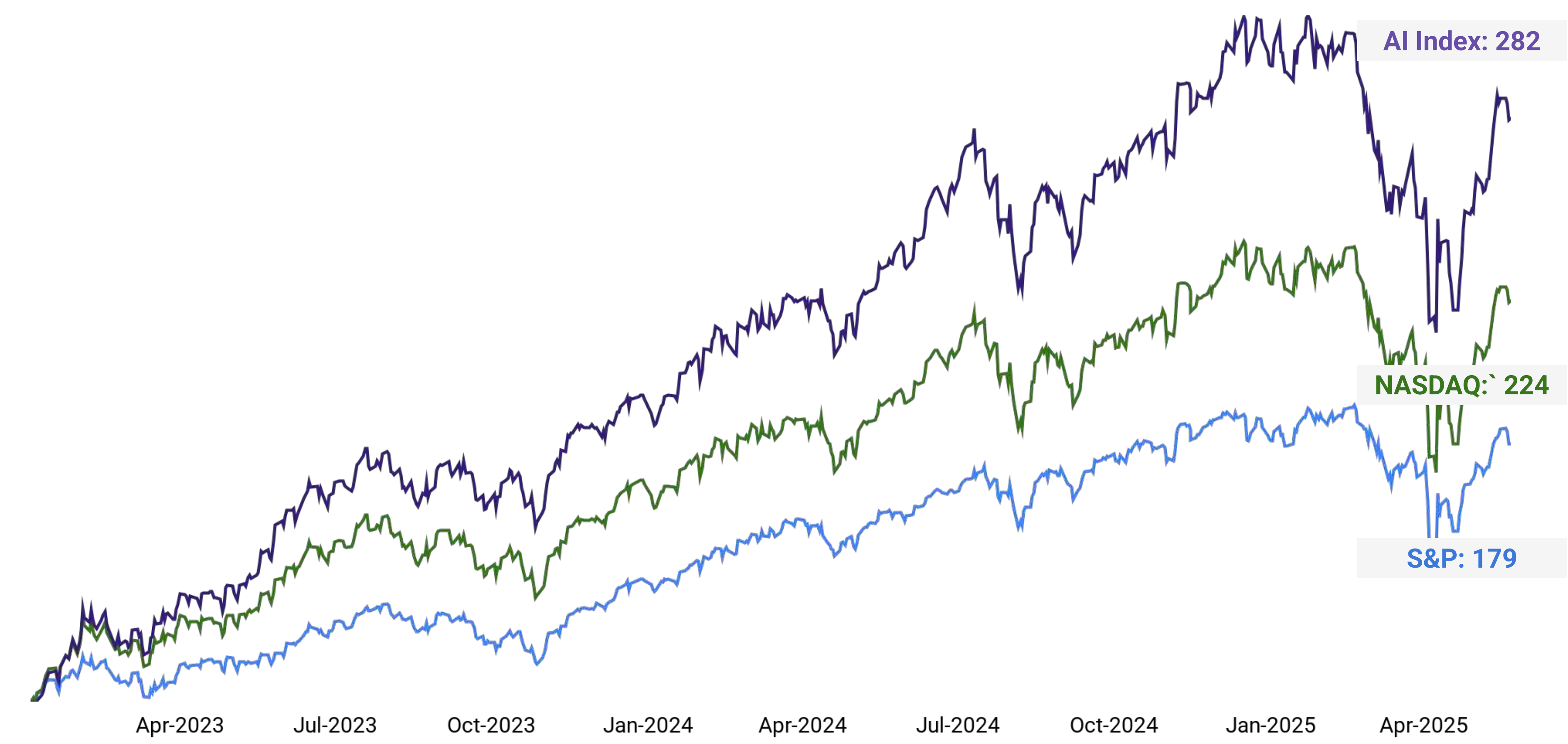
Section 2: Public Markets



AI is driving the stock market

The AI boom kicked off in January 2023, after people had some time over the holidays to process ChatGPT's launch in late November 2022. Since then, AI has fueled stock market gains, with the NASDAQ and S&P 500 returning 124% and 89%, respectively. Meanwhile, the AI Index—a collection of 100 AI-exposed stocks—has soared by 182%. Check out the latest index values at Generational AI Index.

Major Indices (Index = 100 on 1/1/2023)



Source: Google Finance as of 12/19/2025

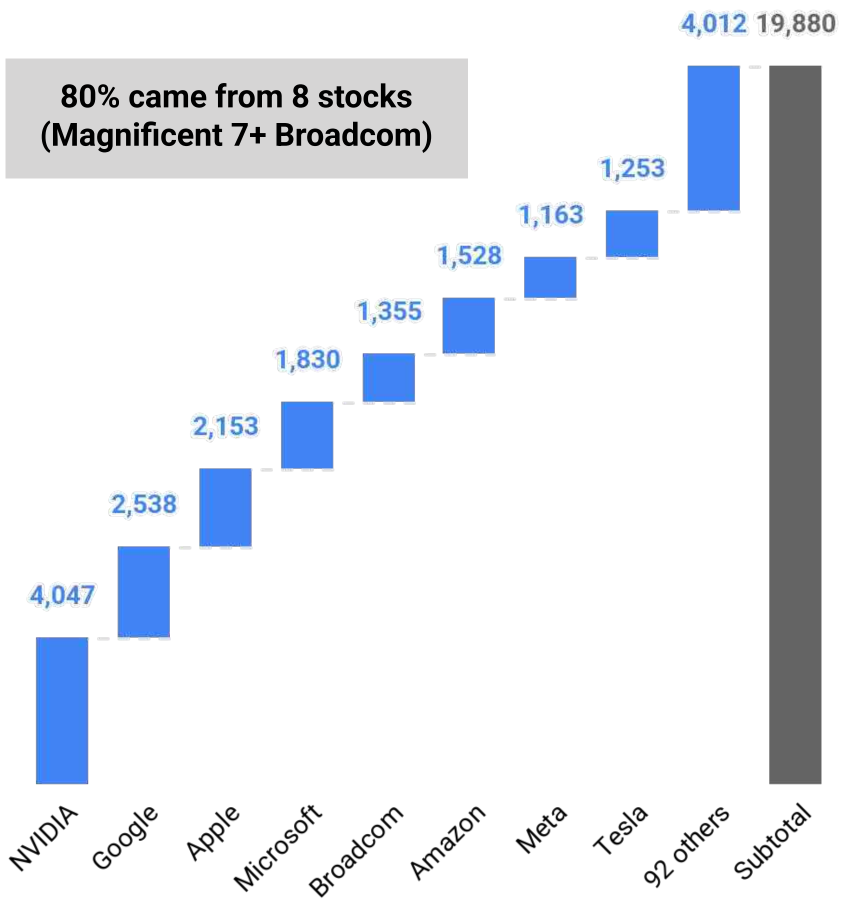
Power law of AI stock market

The AI Index tracks 100 stocks with significant exposure to the AI industry, both positive and negative. While AI has generally driven strong returns, some sectors, like EdTech, have been disrupted by tools like ChatGPT, which offer free "all-in-one" educational solutions. Returns have also been heavily concentrated in a few companies within certain industries, reflecting the power-law dynamics often seen in venture capital. Section 3 takes a closer look at the industries that have benefitted the most.

Returns by Industry

Semiconductors	412%
Internet Services	296%
Hardware	287%
Utilities & Power Generation	260%
Application Software	200%
Banks	172%
E-commerce	169%
Automotive	164%
Biotechnology & Pharmaceuticals	115%
Infrastructure Software	106%
Asset Management & Investment Services	57%
Renewable Energy	40%
Data Center	34%
Consulting & IT Services	8%
Medical Devices & Services	0%
Oil & Gas Equipment & Services	-6%
EdTech	-59%
AI Index	**182%**

Market Cap. gains since 2023



80% came from 8 stocks (Magnificent 7+ Broadcom)

Source: Google Finance as of 12/19/2025

Will demand continue?

Generational

The AI boom is driven by two variables: the best model commands a massive market opportunity—potentially trillions of dollars—and scaling laws still hold, meaning bigger models deliver better performance. This puts model builders in a marathon to build the best frontier model. Both Grok 3 and Grok 4 were released in 2025, costing $220 million and $480 million respectively—just to train. Reports suggest xAI was spending $13 billion per month. OpenAI, Anthropic, and Google models likely cost even more. This race is what drives chip foundries and power utilities to plan their CAPEX.

OpenAI

We are looking at [compute] commitments of about $1.4 trillion over the next 8 years – CEO Sam Altman

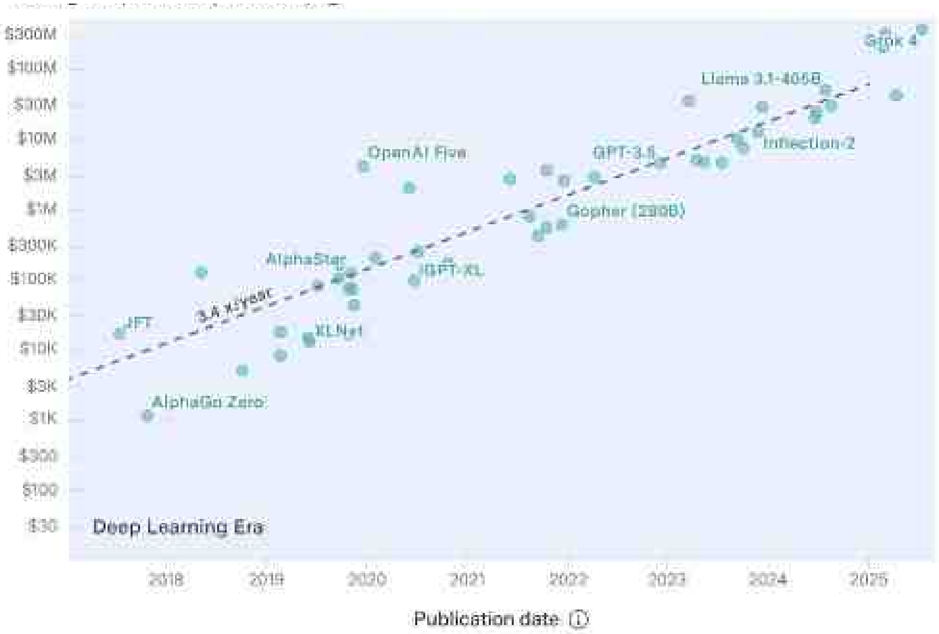
ANTHROP\C

Probably by 2027, [players'] ambitions to build hundred billion dollar clusters. And I think all of that actually will happen. – CEO Dario Amodei

Google

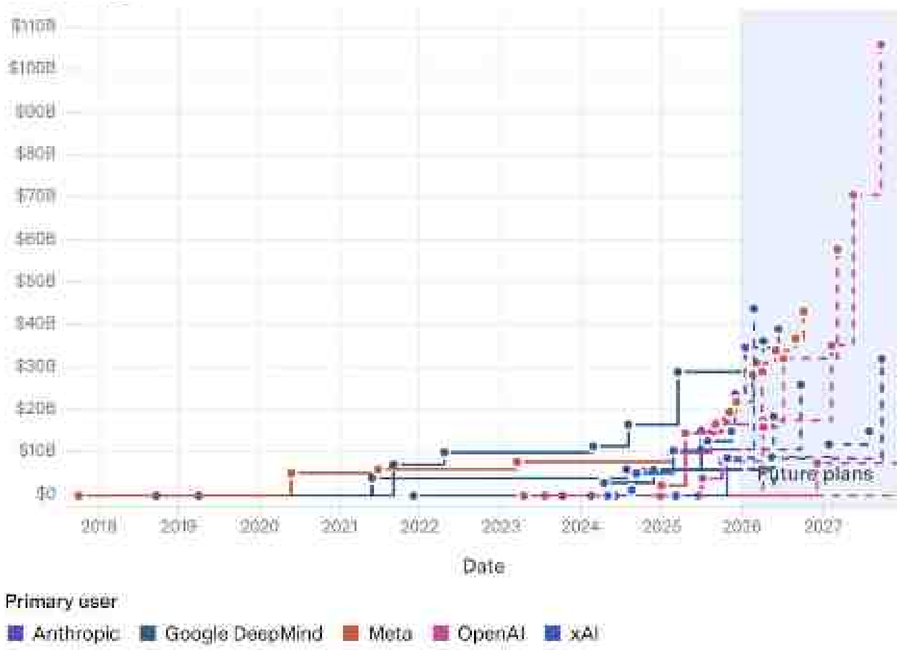
Now we must double [compute] every 6 months..,next 1000x in 4-5 years – Amin Vadhat, Google VP of Infrastructure Nov 2025

Cost of Build Frontier Models



Training compute has been growing **4.6x / year since 2022.** Cost to train frontier models has been growing slightly slower at 3.4x / year since 2022 due to progress in hardware efficiencies

Cost of Existing and Planned Frontier Data Models



Data center investment is scaling exponentially. Planned frontier AI data centers are projected to exceed $100 billion by 2027—up from under $10 billion pre-ChatGPT

Source: Epoch AI, News

Demand will continue: the data

Generational

Both enterprise and consumer adoption continue to trend upward with no signs of slowing. Nearly 3x more S&P 500 companies are now reporting AI-driven revenue or cost impacts compared to early 2023. Meanwhile, over half of US adults have used GenAI—with workplace adoption accelerating fastest, nearly doubling in under a year.

ESTĒE LAUDER

AI has driven a 31% increase in ROI from our North American media campaigns, enabling faster decisionmaking and stronger real-time market responsiveness.

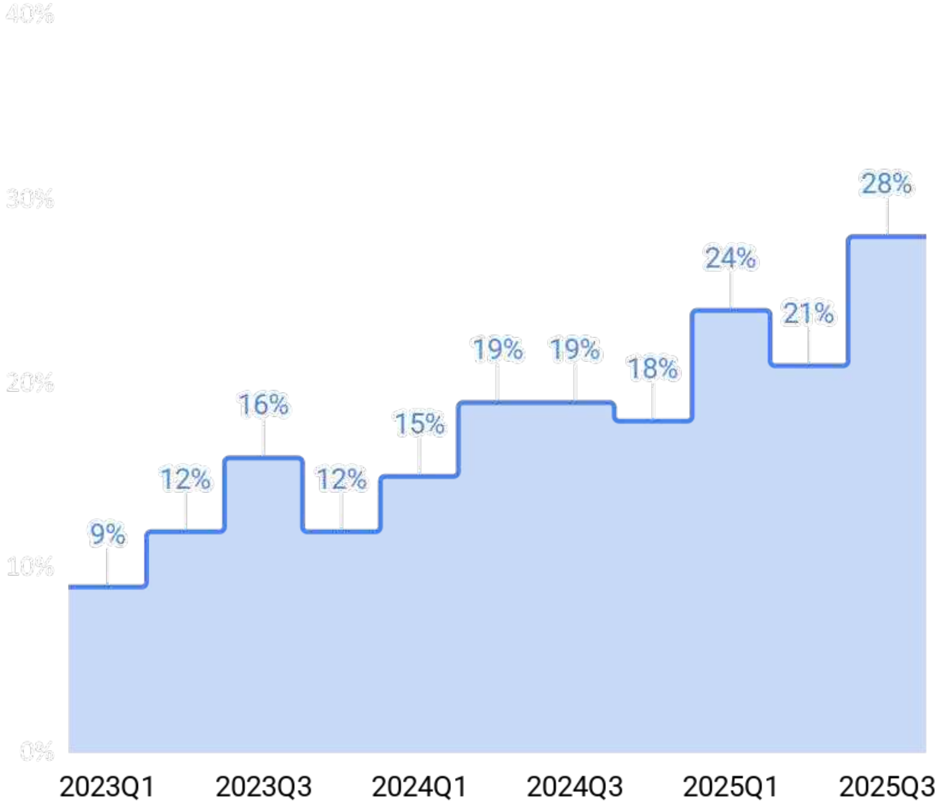
Walmart

We've used GenAI to improve our product catalog. Without the use of GenAI, this work would have required nearly 100 times the current head count to complete in the same amount of time.

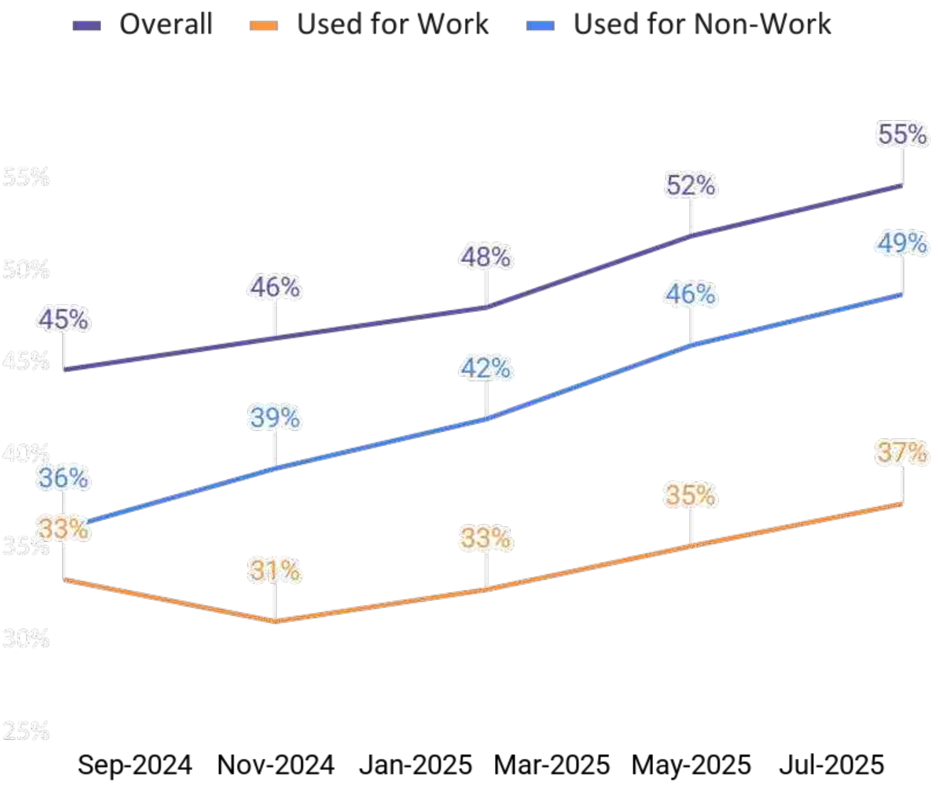
CHIPOTLE

We saw with the AI tool about a 46%, 47% uplift in engagement through that welcome journey, so that informed what we are now calling the win-back journey.

% of S&P500 companies providing AI revenue & cost benefits



% US population adopting GenAI



Source: Morgan Stanley, Federal Reserve Bank of St. Louis, News

Demand will continue: a simple analogy

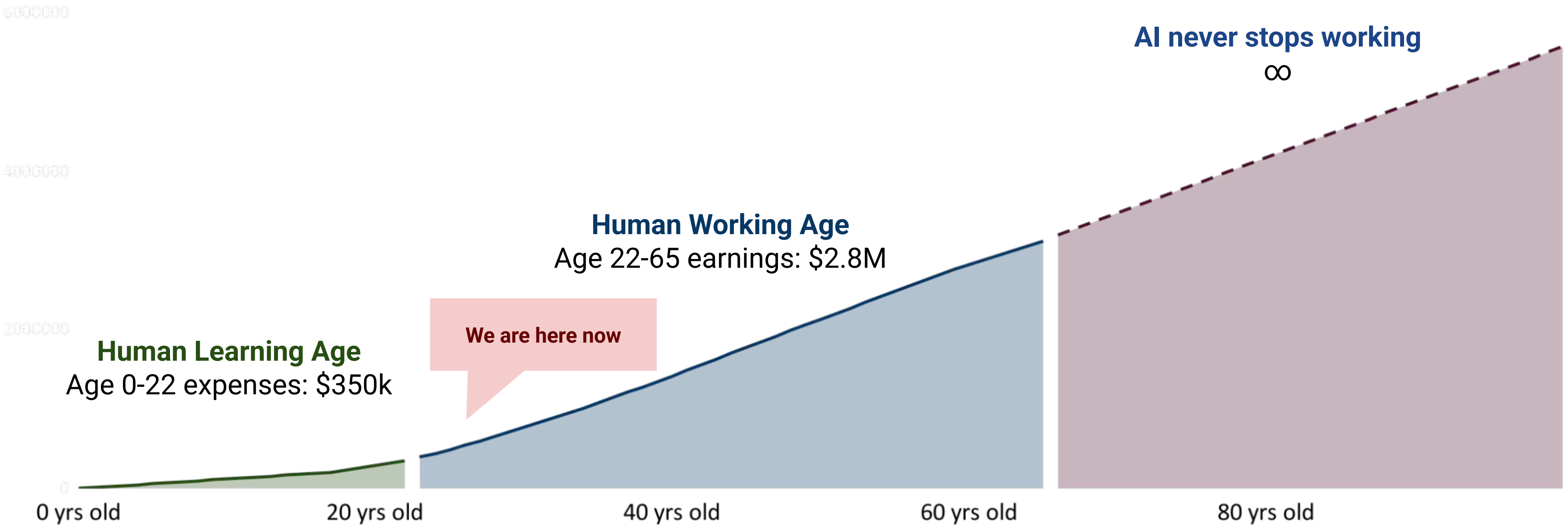
Generational

From 2022 to 2025, AI went through its 'learning phase'—building foundation models through massive training investments. That work is done. Frontier models now match or exceed professional-level performance: GPT-5.2's output is preferred or equivalent to professionals with 14 years of experience.

Now we're entering the working phase. The economic parallel is intuitive: humans invest roughly $350,000 over 20 years in education, then generate $2.8 million in lifetime earnings over 40+ years of work. AI is following the same pattern—intensive upfront training costs, followed by value generation through deployment.

Today's models focus primarily on knowledge work—writing, analysis, coding, research. But the next frontier is already visible: autonomous vehicles are deployed on roads today, warehouse robots are operational at scale, and humanoid robots are on the horizon. Like a professional in their first years of work, AI is in the early stages of its productive phase—starting with cognitive tasks, soon expanding to physical ones.

Cumulative gross economic value of expenses and earnings in a lifetime



Source: US Bureau Labor of Statistics

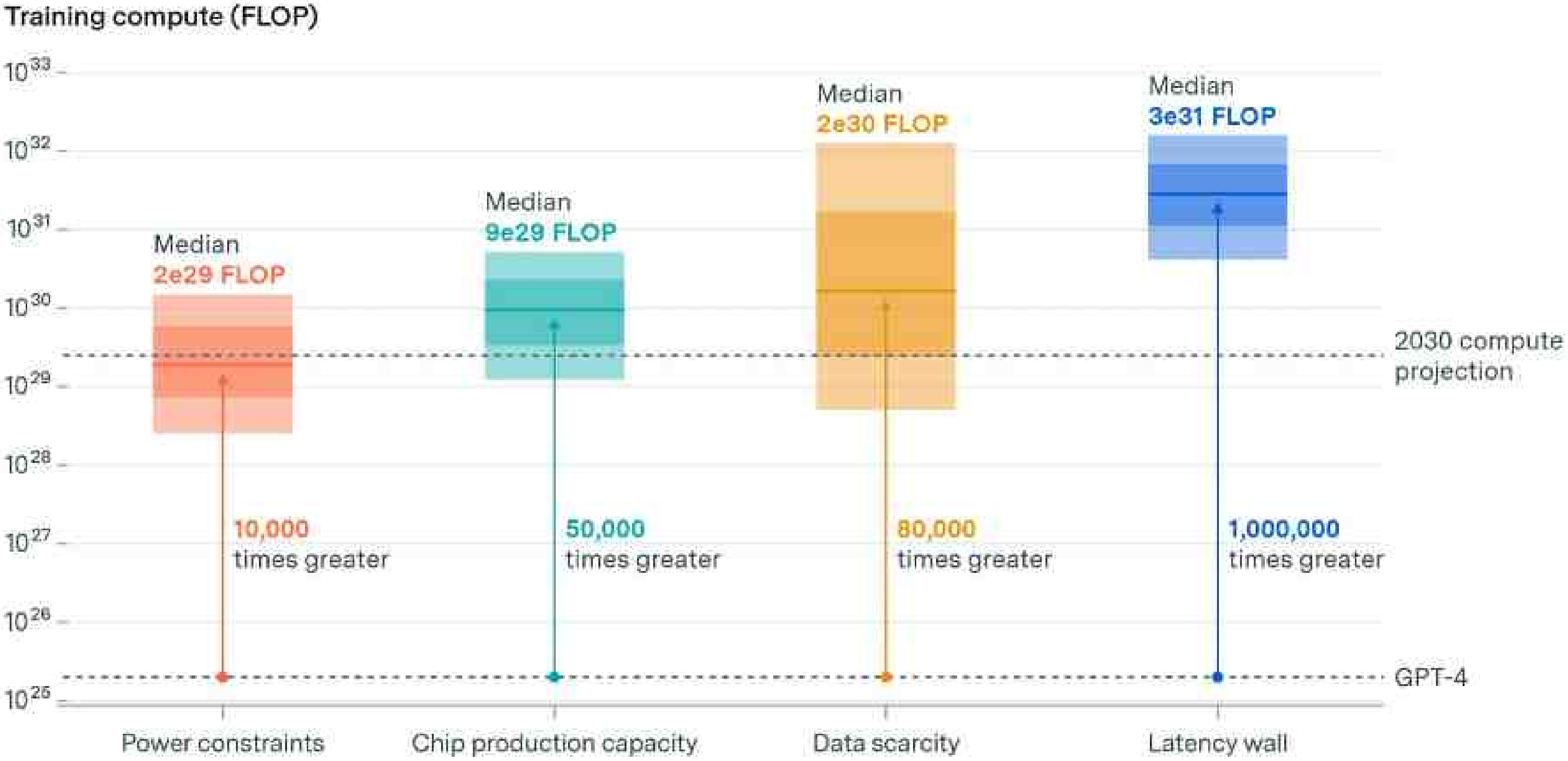
Energy is the biggest blocker to scaling

Generational

AI training could scale 10,000x by 2030 relative to GPT-4—but infrastructure constraints will determine the pace. Epoch AI analyzed four potential bottlenecks and found that power is the tightest by far: it limits scaling to 10,000x, while chip manufacturing could support 50,000x, data availability 80,000x, and network latency up to 1,000,000x.

Power is the binding constraint. Single-campus data centers will reach 1–5 GW, while distributed networks could access 2–45 GW. This requires $100B+ in infrastructure buildout over 3–5 years across power generation, transmission, and grid interconnection. The supply-demand imbalance is most acute in energy—expect tight market dynamics across the power value chain through 2030.

Constraints to scaling training runs by 2030



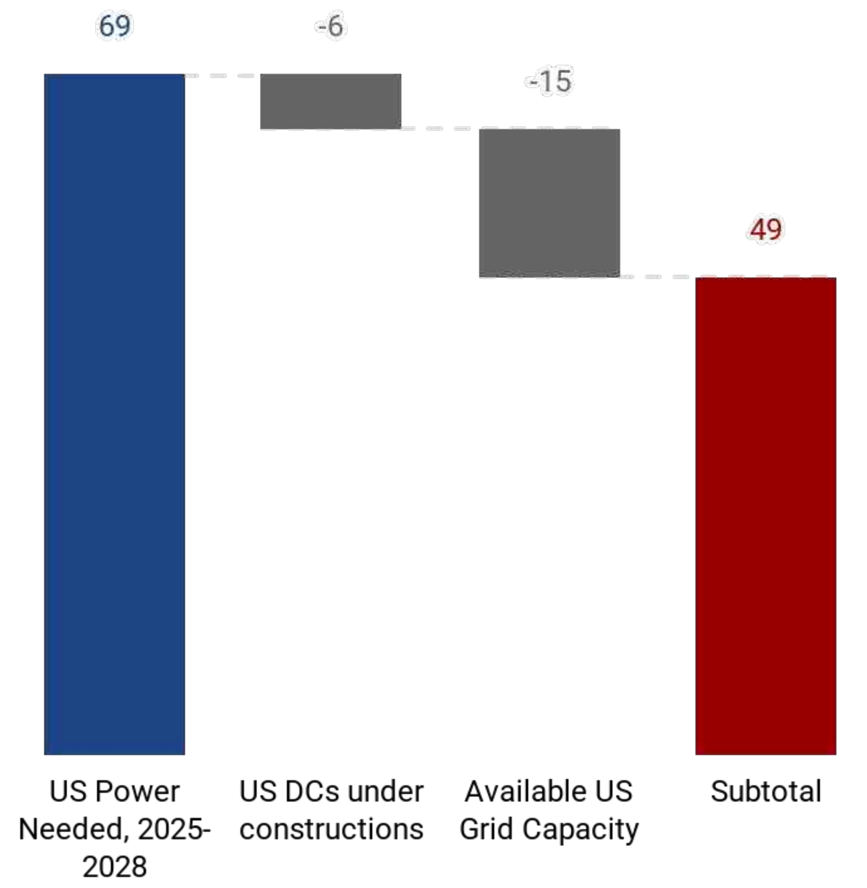
Source: EpochAI

Energy shortage

Generational

The push for AGI models is driving unprecedented demand for energy, outpacing the ability of utilities to expand capacity. As hyperscale data centers scale operations, the resulting energy gap could limit the growth of AI and other power-intensive applications. With the U.S. grid already constrained, energy shortages will outstrip chip shortages as the critical bottleneck.

US Data Center energy supply-demand gap (GW)



Gartner

The explosive growth of hyperscale data centers for GenAI applications is creating an unprecedented demand for power. This demand is outpacing utilities' ability to expand their capacity, leading to potential shortages that could restrict the growth of AI and other power-intensive applications from 2026 onward
– Gartner

DIGITALBRIDGE

We started talking about this over two years ago at the Berlin Infrastructure Conference when I told the investor world, we're running out of power in five years. Well, I was wrong about that. We're kind of running out of power in the next 18 to 24 months.
– Digital Bridge earnings call Apr 2024

DIGITAL REALTY

Over the past few weeks, we've seen several examples of the lengths that some hyperscalers will go to reserve enough power for their fast-growing compute requirements.

We've seen a deal to reactivate 3-mile island [Microsoft]; another hyperscaler [Google] partnering with an existing utility to develop small modular reactors; and the third [Amazon], executing power purchase agreements to purchase nuclear energy for multiple SMRs that have yet to be built. Yet each of these plants is still years away from beginning to generate power, underscoring the value of lower capacity blocks today and perhaps for the next several years.
– Digital Realty earnings call Oct 2024

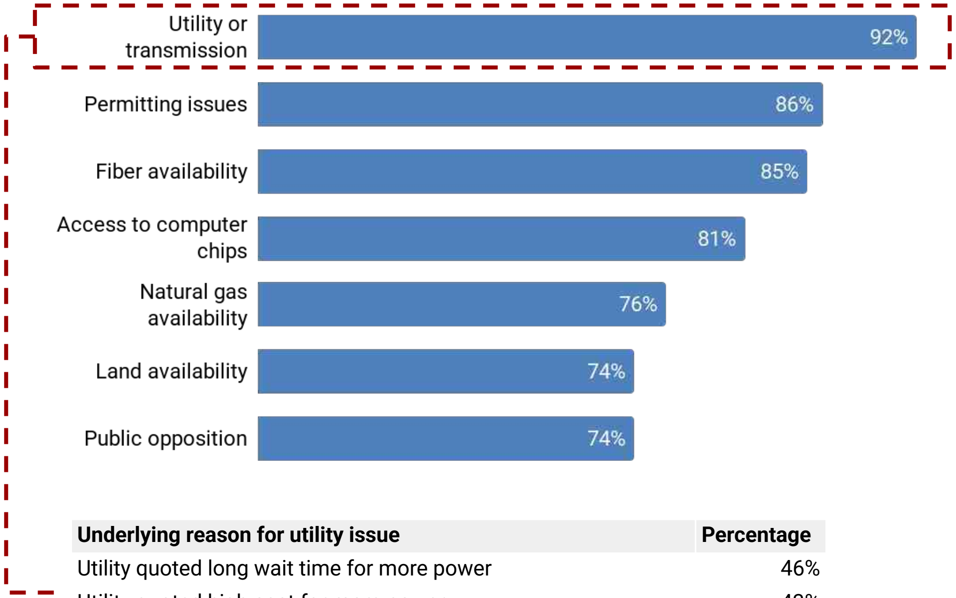
Source: Gartner, BCG, Morgan Stanley, Public Filings, News

Data centers are desperate for energy

Generational

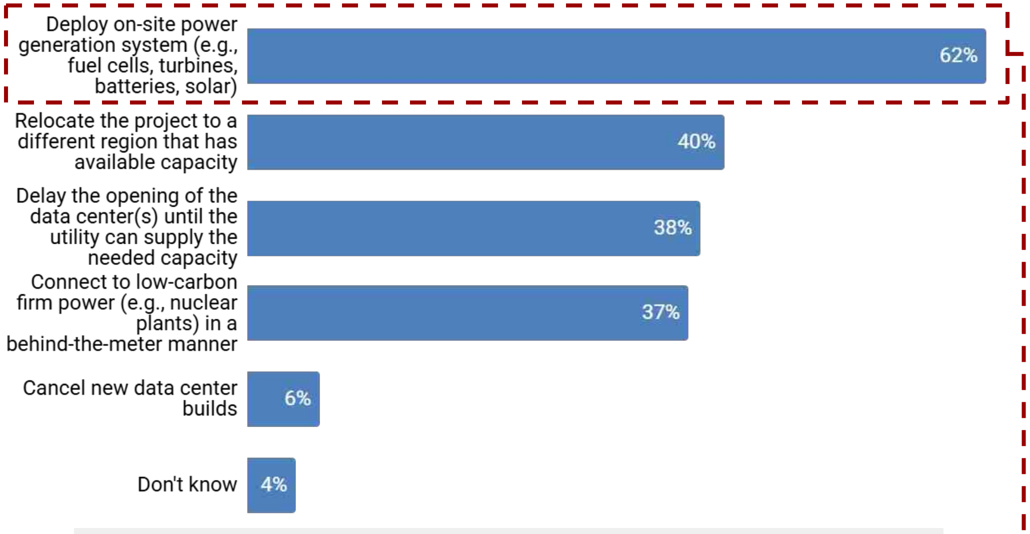
A survey of 149 senior data center industry professionals shows how tight the energy market is. 92% cited utility capacity as a barrier and 44% facing 4+ year wait times, the industry is making a decisive pivot. Six in ten operators now plan on-site power generation as their primary contingency—marking a fundamental shift in how AI infrastructure gets built. When X AI built Colossus 1, Elon Musk brought in 35 semi-trailer-sized gas turbines to power the data center. The need for energy was so great that Elon Musk bought an power plant overseas to ship it to the US to power Colossus 2

What's slowing down data center projects?



Underlying reason for utility issue	Percentage
Utility quoted long wait time for more power	46%
Utility quoted high cost for more power	43%
Long lead time for on-site power generation equipment	42%
Utility power has unpredictable energy rates	42%
Utility power reliability is inadequate	32%
Utility has given no written promise on delivery time	32%
Site layout issues (e.g., space, can't run distribution lines)	27%
Not running into any challenges	15%
Utility does not enable sufficient renewable energy procurement	13%
Not applicable / Don't know	8%

How are data centers addressing the utility issue?



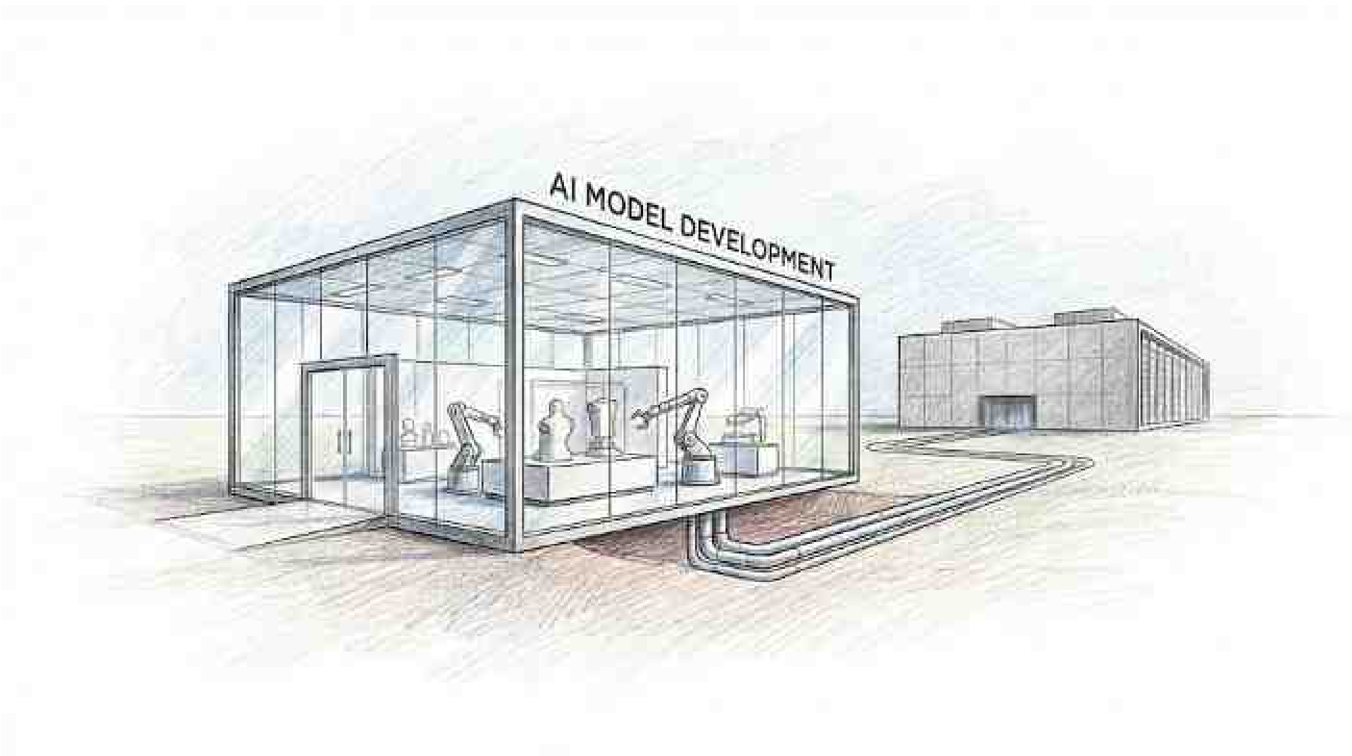
Top On-Site Deployment Model	Description
Bridge to power	When utility arrives, utility becomes prime power
Island power	The data center never connects to the utility
Bridge to power	When utility arrives, utility becomes backup power

Source: Schneider Electric, Inside Climate News

Gas turbines at Colossus

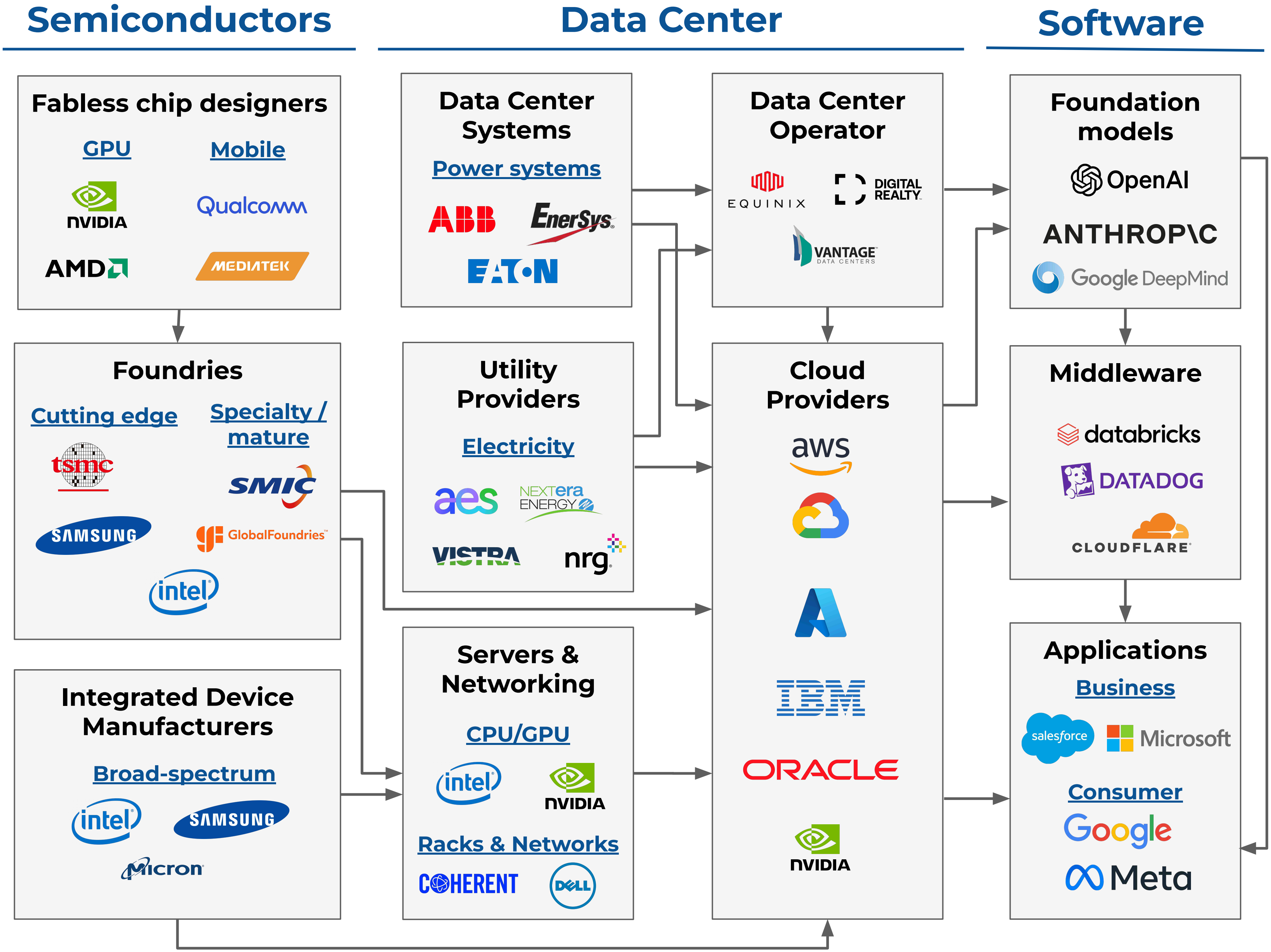
Rows of gas turbines around Colossus 1

Section 3: Industry Deep Dives



High-level AI Value Chain

Generational



Semiconductors

Fabless chip designers

GPU
- NVIDIA
- AMD

Mobile
- Qualcomm
- MediaTek

Foundries

Cutting edge
- tsmc
- Samsung
- intel

Specialty / mature
- SMIC
- GlobalFoundries

Integrated Device Manufacturers

Broad-spectrum
- intel
- Samsung
- Micron

Data Center

Data Center Systems

Power systems
- ABB
- EnerSys
- EATON

Utility Providers

Electricity
- aes
- NextEra Energy
- VISTRA
- nrg

Servers & Networking

CPU/GPU
- intel
- NVIDIA

Racks & Networks
- COHERENT
- DELL

Data Center Operator

- EQUINIX
- Digital Realty
- Vantage Data Centers

Cloud Providers

- aws
- IBM
- ORACLE
- NVIDIA

Software

Foundation models

- OpenAI
- ANTHROPIC
- Google DeepMind

Middleware

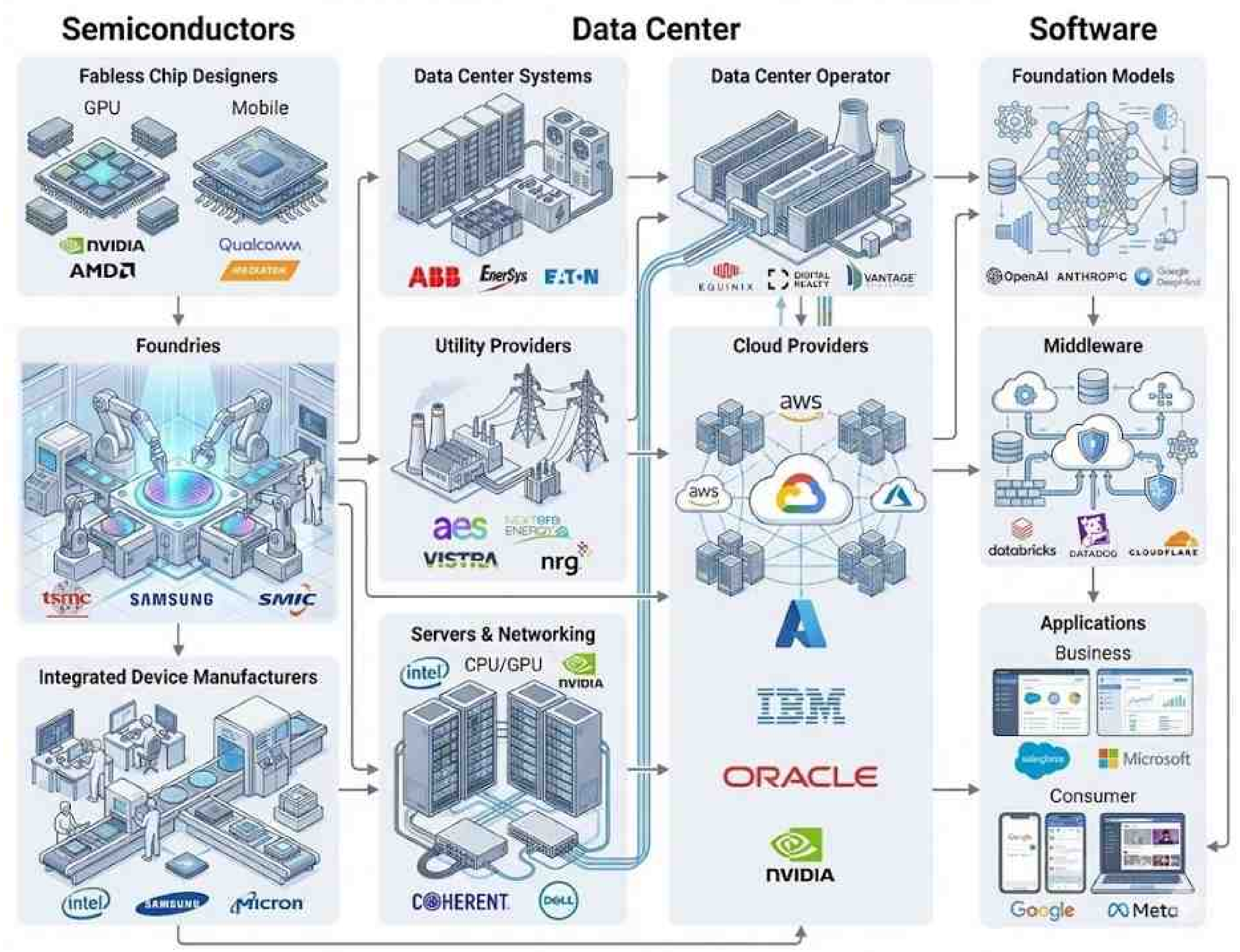
- databricks
- DATADOG
- CLOUDFLARE

Applications

Business
- salesforce
- Microsoft

Consumer
- Google
- Meta

High-level AI Value Chain (🍌Version)



AI Data Center Cost

AI data centers are the backbone of today's AI. Built to handle the intense computing demands of generative AI training and inference, their main cost drivers are compute capacity (GPUs, networking) and electricity. It's no surprise that companies in these sectors have seen their stock prices soar.

CAPEX

IT Equipment (GPUs, Networking, Storage) – Bulk of the investment. High-end GPUs, networking gear, and storage systems power AI workloads.

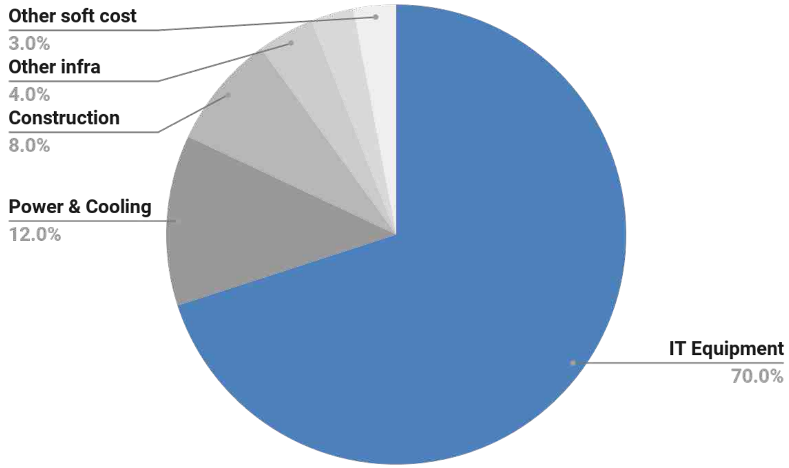
Power and Cooling – Essential infrastructure to power up and cool down high-performance hardware.

Construction – Building the physical facility to house all the equipment.

Other Infra (Security, Monitoring) – Security systems and monitoring equipment to safeguard operations.

Land – Cost of acquiring land for the data center site.

Other Soft Costs (Engineering, Design) – Engineering and design services to plan and optimize the facility.



OPEX

Electricity – Biggest operational cost. Power-hungry GPUs and cooling systems drive energy bills sky-high.

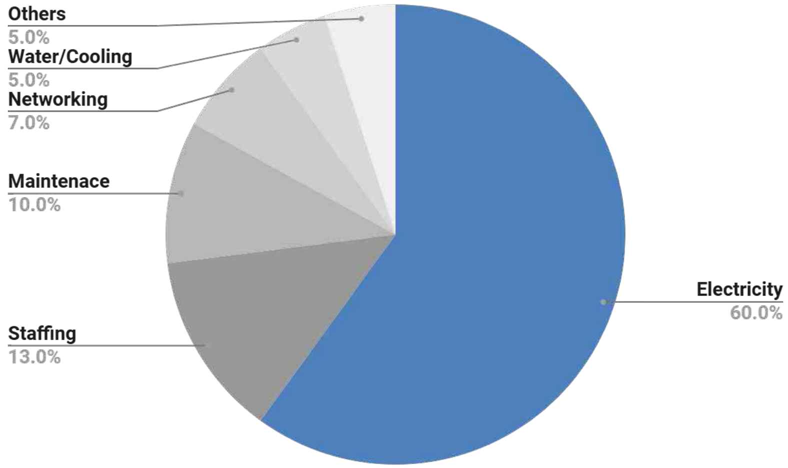
Staffing – Salaries for skilled staff who run and maintain the data center.

Maintenance – Upkeep of IT equipment and facilities to ensure smooth operations.

Networking – Costs for bandwidth, connectivity, and network management.

Water/Cooling – Expenses for water usage and cooling operations.

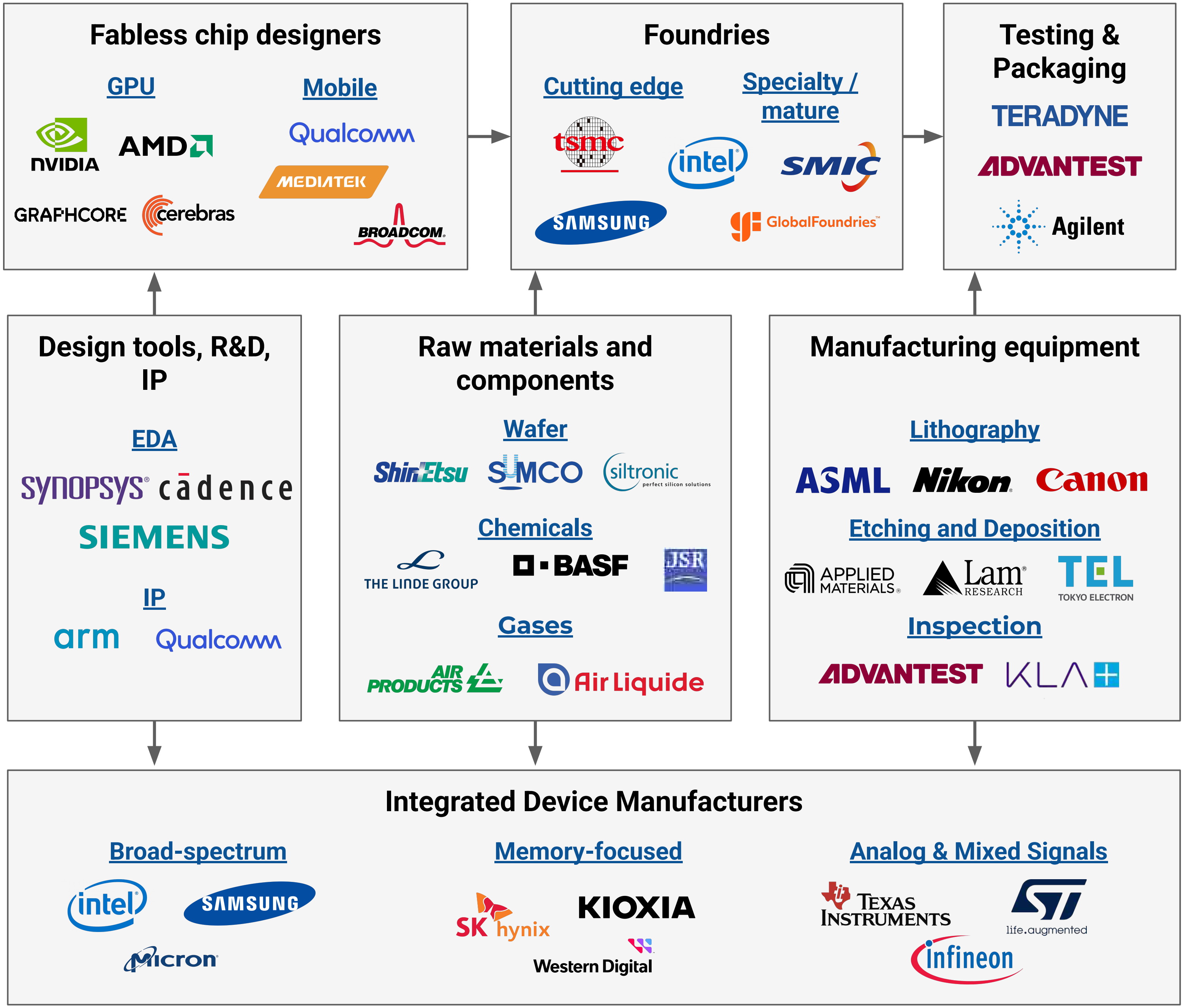
Others – Miscellaneous operational expenses like insurance, taxes, and compliance.



Source: IDC, Bain, Author analysis

Semiconductor Value Chain

Generational



Fabless chip designers

GPU: NVIDIA, AMD, GRAPHCORE, cerebras

Mobile: Qualcomm, MEDIATEK, BROADCOM

Foundries

Cutting edge: tsmc, intel, SAMSUNG

Specialty / mature: SMIC, GlobalFoundries

Testing & Packaging

TERADYNE, ADVANTEST, Agilent

Design tools, R&D, IP

EDA: SYNOPSYS, cadence, SIEMENS

IP: arm, Qualcomm

Raw materials and components

Wafer: ShinEtsu, SUMCO, siltronic perfect silicon solutions

Chemicals: THE LINDE GROUP, BASF, JSR

Gases: AIR PRODUCTS, Air Liquide

Manufacturing equipment

Lithography: ASML, Nikon, Canon

Etching and Deposition: APPLIED MATERIALS, Lam RESEARCH, TEL TOKYO ELECTRON

Inspection: ADVANTEST, KLA

Integrated Device Manufacturers

Broad-spectrum: intel, SAMSUNG, Micron

Memory-focused: SK hynix, KIOXIA, Western Digital

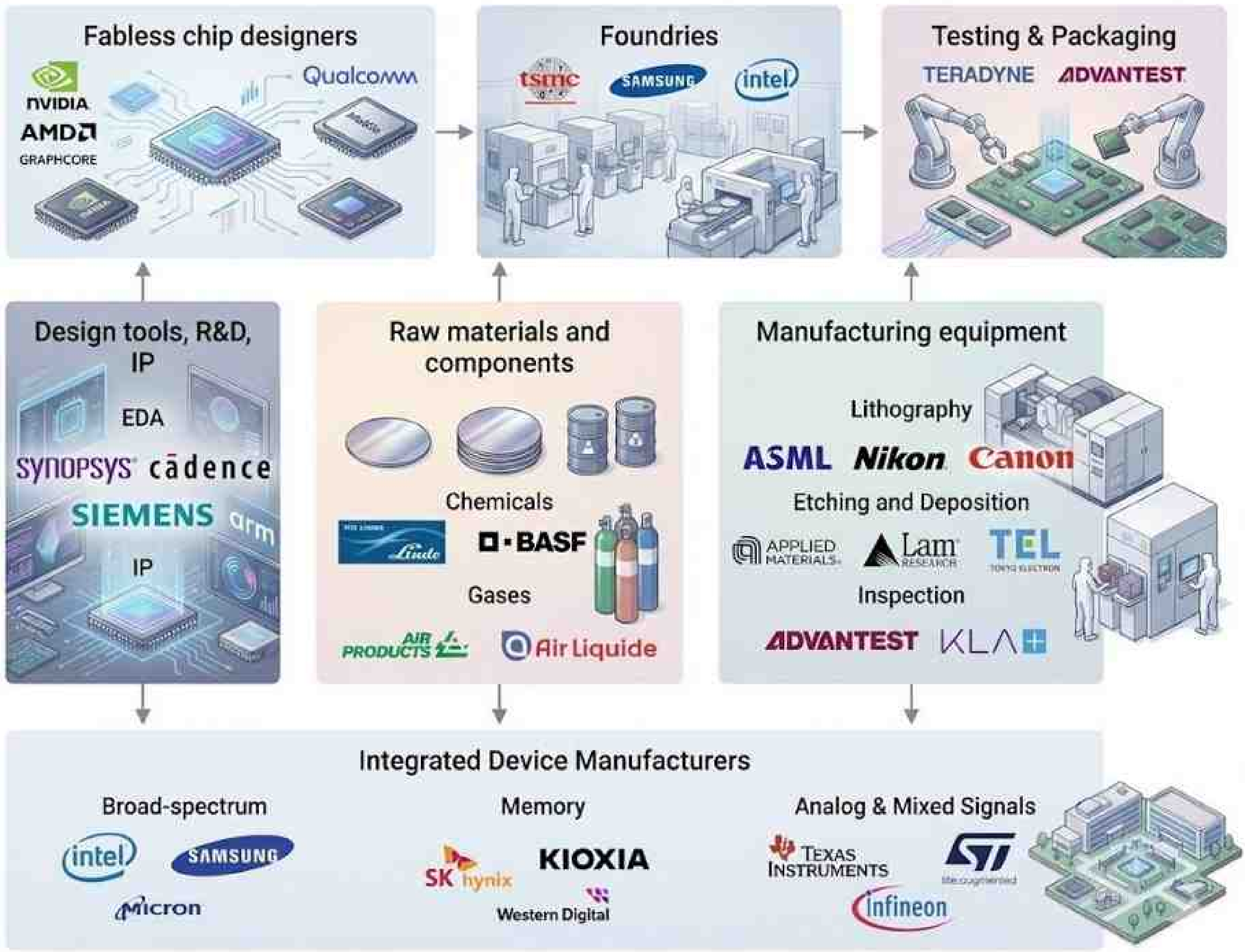
Analog & Mixed Signals: TEXAS INSTRUMENTS, ST life.augmented, infineon

Fabless model
In-house chip design with outsourced manufacturing

Supplier layer
A complex supply chain for every component of a chip and the machines that make them

Integrated model
Vertically integrated chip development

Semiconductor Value Chain (🍌Version)



Fabless chip designers

Foundries

Testing & Packaging

Design tools, R&D, IP

EDA

IP

Raw materials and components

Chemicals

Gases

Manufacturing equipment

Lithography

Etching and Deposition

Inspection

Integrated Device Manufacturers

Broad-spectrum

Memory

Analog & Mixed Signals

Semiconductor Value Chain description

Generational

Design tools, R&D, IP
This category includes the intellectual foundation of the semiconductor industry—Electronic Design Automation (EDA) tools, semiconductor intellectual property (IP), and research and development activities. Companies here provide the software, reference architectures, and building-block designs that chip designers use to create advanced semiconductor products. They reduce complexity, accelerate time-to-market, and enable increasingly sophisticated chip designs.

Raw Materials and Components
This segment covers the essential inputs—wafers, chemicals, and gases—that semiconductor manufacturers require to produce chips. High-purity materials and specialized chemical compounds underpin the entire fabrication process. Without these carefully sourced and refined inputs, it would be impossible to achieve the precision and quality standards that modern chips demand.

Manufacturing Equipment
This category includes the machinery and tools used by foundries and integrated device manufacturers (IDMs) to fabricate chips. Lithography systems, etching and deposition tools, and inspection equipment ensure that transistor features are patterned with nanoscale precision. These cutting-edge machines define the technical limits of chip complexity and performance.

Foundries
Foundries specialize in semiconductor manufacturing. They typically do not design their own chips, focusing instead on producing devices for fabless companies. By offering advanced process nodes, foundries provide the manufacturing muscle needed to turn chip designs into physical products. They ensure that fabless chip designers can access state-of-the-art fabrication without maintaining their own production lines.

Fabless Chip Designers
These companies create chip designs in-house but outsource manufacturing. They focus on architecture, functionality, and performance rather than running their own fabrication facilities. By working closely with foundries, fabless designers can quickly scale production and adapt to the latest process technologies without capital-intensive investments in manufacturing plants.

Semiconductor Value Chain description

Generational

Testing & Packaging
After fabrication, chips must be tested, packaged, and made ready for integration into end-user devices. Specialized test equipment and packaging techniques ensure chip reliability, functionality, and durability. These services represent the final step of the production process before chips reach system integrators and device manufacturers.

Integrated Device Manufacturers
IDMs control the entire chain from design to fabrication, testing, and packaging within a single organization. They combine the roles of fabless designers and foundries under one roof, enabling tighter integration and more direct control over the supply chain. This vertically integrated approach can streamline innovation, improve quality, and reduce time-to-market.

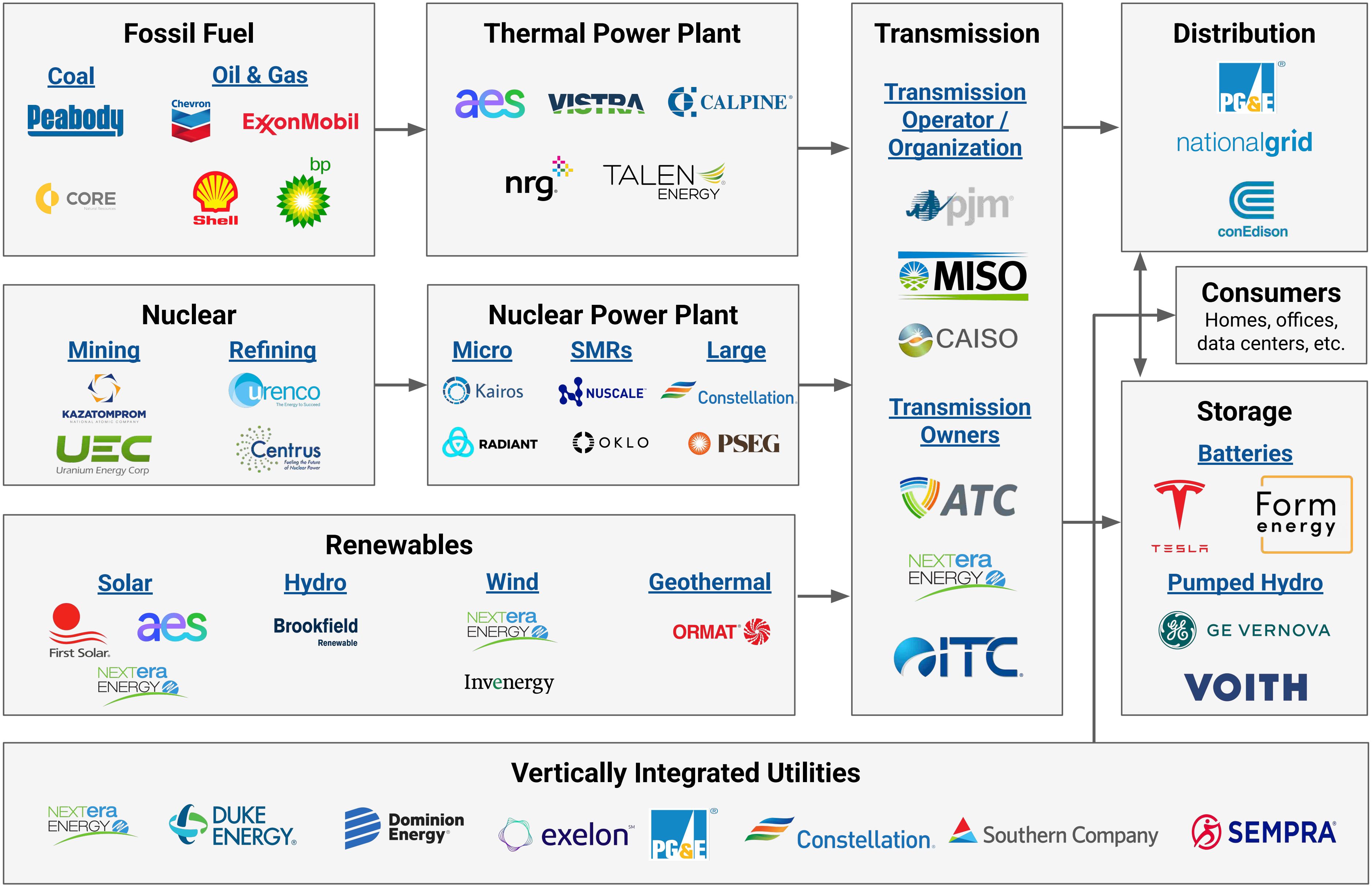
How they are connected
Each category in the semiconductor value chain feeds into the next. Design tools and IP fuel the work of fabless designers and IDMs, who rely on raw materials and specialized equipment to turn concepts into chips. Foundries convert these designs into silicon, while testing and packaging providers ensure product readiness. IDMs bridge all stages internally, while fabless designers and foundries collaborate to achieve similar outcomes through partnership. Together, this interconnected system supports the complex, global process of semiconductor innovation, production, and delivery.

Electricity Value Chain

Generational

Generation
Fuel sources converted to electricity. Mix of utility-owned plants and independent power producers selling into wholesale markets.

Distribution
Bulk transmission and local delivery to end users. Grid interconnection is the critical bottleneck.



Fossil Fuel

Coal
- Peabody
- CORE

Oil & Gas
- Chevron
- ExxonMobil
- Shell
- bp

Thermal Power Plant

- aes
- VISTRA
- CALPINE
- nrg
- TALEN ENERGY

Nuclear

Mining
- KAZATOMPROM
- UEC Uranium Energy Corp

Refining
- urenco
- Centrus

Nuclear Power Plant

Micro
- Kairos
- RADIANT

SMRs
- NUSCALE
- OKLO

Large
- Constellation
- PSEG

Renewables

Solar
- First Solar
- aes
- NEXTera ENERGY

Hydro
- Brookfield Renewable

Wind
- NEXTera ENERGY
- Invenergy

Geothermal
- ORMAT

Transmission

Transmission Operator / Organization
- pjm
- MISO
- CAISO

Transmission Owners
- ATC
- NEXTera ENERGY
- ITC

Distribution

- PG&E
- nationalgrid
- conEdison

Consumers

Homes, offices, data centers, etc.

Storage

Batteries
- TESLA
- Form energy

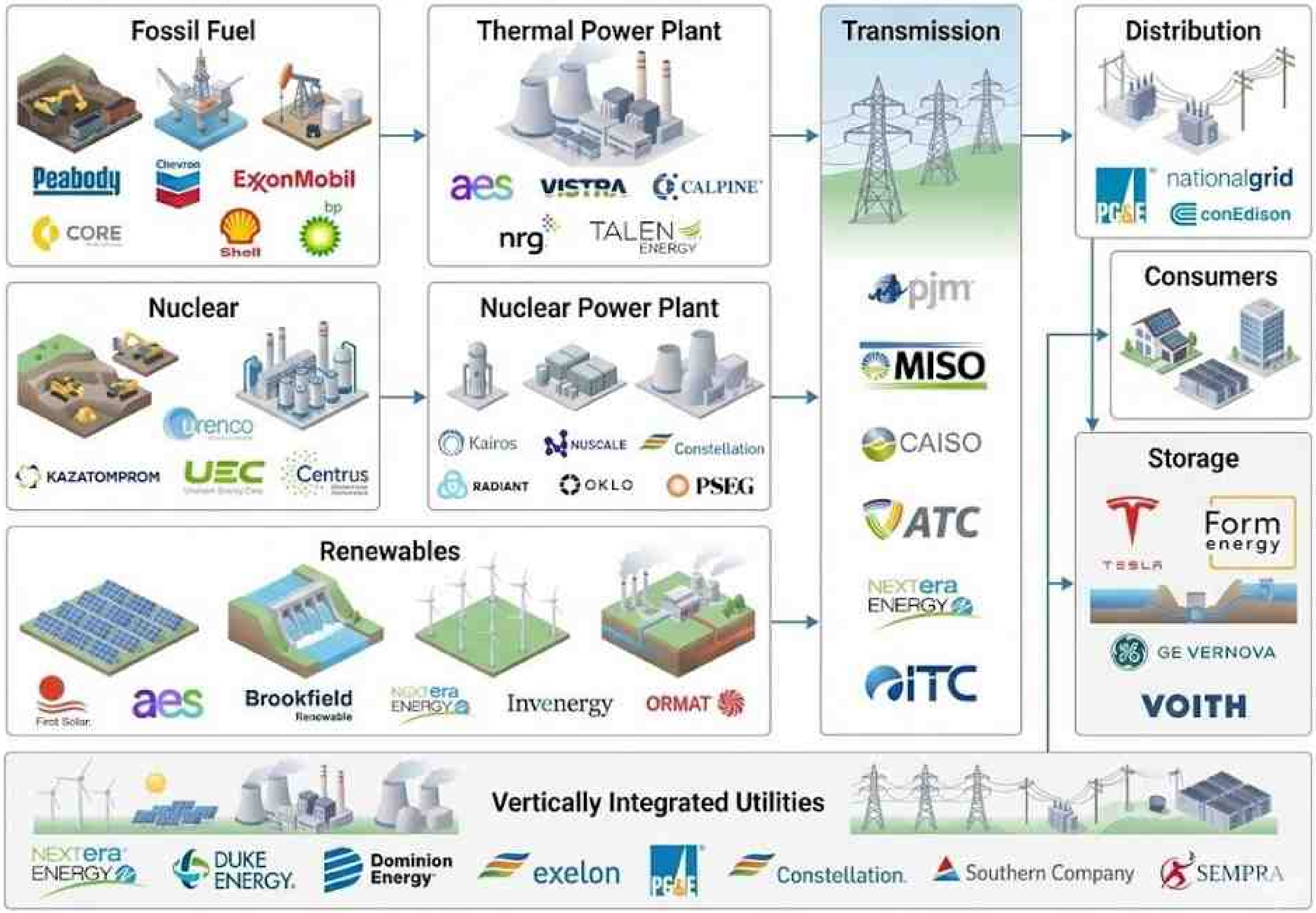
Pumped Hydro
- GE VERNOVA
- VOITH

Vertically Integrated Utilities

- NEXTera ENERGY
- DUKE ENERGY
- Dominion Energy
- exelon
- PG&E
- Constellation
- Southern Company
- SEMPRA

Electricity Value Chain (🍌Version)

Generational



Electricity Value Chain description

Fossil Fuel
This category includes the extraction, processing, and transport of coal, natural gas, and oil used to generate electricity. Natural gas dominates at approximately 43% of U.S. electricity generation, while coal continues to decline at around 16%. Companies here operate wells, mines, pipelines, and storage facilities that supply thermal power plants with the fuel they need. The availability and price of these commodities directly influence electricity costs and grid reliability.

Nuclear
Nuclear fuel begins with uranium mining and proceeds through enrichment, where the concentration of fissile U-235 is increased from 0.7% to 3-5% for use in reactors. Enrichment is a strategic chokepoint—Russia controls roughly 40% of global capacity, creating supply chain risk for Western markets. Advanced reactors and SMRs require high-assay low-enriched uranium (HALEU), which barely exists outside Russian production. These upstream constraints will shape the pace of new nuclear deployment.

Renewables
Renewables convert natural energy flows—sunlight, wind, water, and geothermal heat—directly into electricity without a separate fuel input. Unlike fossil fuels and nuclear, the "source" and "generation" steps are effectively combined; a solar panel or wind turbine is both the resource and the power plant. This simplicity reduces fuel cost risk but introduces intermittency, making grid integration and storage critical. Renewables currently account for approximately 20% of U.S. generation and represent the fastest-growing segment.

Thermal Power Plant
Thermal power plants burn fossil fuels to produce steam, which drives turbines to generate electricity. Natural gas combined-cycle plants offer efficiency and flexibility, while peaker plants provide rapid response during demand spikes. Coal plants, though declining, still contribute baseload power in some regions. These facilities are owned by utilities and independent power producers (IPPs) who sell electricity into wholesale markets.

Nuclear Power Plant
Nuclear power plants use controlled fission reactions to generate heat, producing steam that drives turbines. Large reactors (1,000+ MW) form the backbone of the existing fleet, providing reliable baseload power with zero carbon emissions. Small modular reactors (SMRs) and microreactors represent the emerging frontier—factory-built, scalable, and attractive for data center applications. Hyperscalers are signing power purchase agreements directly with nuclear operators to secure firm, carbon-free electricity.

Electricity Value Chain description

Transmission

Transmission moves bulk electricity over high-voltage lines from power plants to population centers. Independent System Operators (ISOs) and Regional Transmission Organizations (RTOs) coordinate flow and operate wholesale energy markets, though they do not own the physical infrastructure. Transmission owners—utilities, independent companies, and investors—maintain the lines, towers, and substations. This segment represents the primary bottleneck in the energy transition: over 2,600 GW of generation capacity sits in interconnection queues, with five-year average wait times and only 19% of projects reaching commercial operation.

Distribution

Distribution delivers electricity the final mile from substations to homes, businesses, and industrial facilities. Local utilities operate lower-voltage networks of lines, transformers, and smart meters that serve their regulated territories. Electrical equipment suppliers provide the switchgear, transformers, and power management systems critical to reliability. For data centers, distribution infrastructure and equipment availability can determine whether a facility gets energized on schedule.

Storage

Storage absorbs excess electricity when supply exceeds demand and discharges it when the grid needs power. Batteries—primarily lithium-ion today, with long-duration technologies emerging—dominate new deployments and enable renewable integration. Pumped hydro remains the largest installed base globally but faces geographic constraints for new development. Storage sits between generation and distribution with bidirectional flow, smoothing the intermittency of renewables and providing grid stability services.

Vertically Integrated Utilities

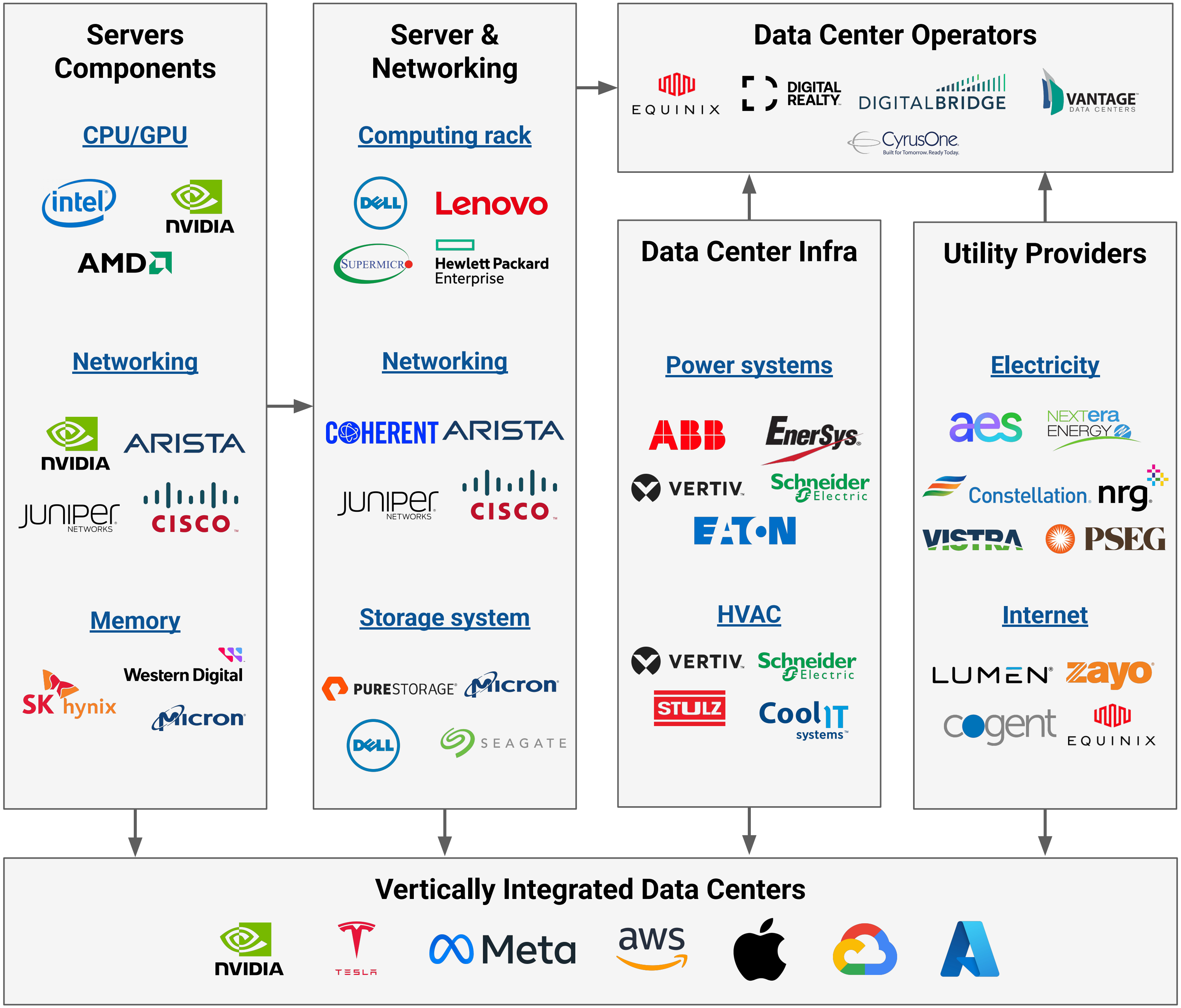
Vertically integrated utilities own and operate assets across the entire value chain—from power plants through transmission lines to local distribution networks. They function as regulated monopolies within their service territories, earning guaranteed returns in exchange for reliability obligations. These companies provide one-stop accountability for electricity delivery but face pressure to accommodate independent generators, renewable developers, and large customers seeking direct supply arrangements.

How they are connected

Each segment of the electricity value chain feeds into the next. Fossil fuels and nuclear materials supply power plants, while renewables generate electricity directly from natural flows. Thermal and nuclear plants convert these inputs into bulk power, which transmission networks carry across regions. Distribution systems deliver electricity locally, and storage balances supply and demand across all stages. Vertically integrated utilities span this entire process internally, while independent generators, transmission operators, and storage providers collaborate to achieve similar outcomes through market mechanisms. Together, this interconnected system keeps electricity flowing reliably from diverse sources to millions of end users—including the data centers driving AI infrastructure growth.

Data Center Value Chain

Generational



Servers Components

CPU/GPU

intel, NVIDIA, AMD

Networking

NVIDIA, ARISTA, JUNIPER NETWORKS, CISCO

Memory

SK hynix, Western Digital, Micron

Server & Networking

Computing rack

DELL, Lenovo, SUPERMICRO, Hewlett Packard Enterprise

Networking

COHERENT, ARISTA, JUNIPER NETWORKS, CISCO

Storage system

PURESTORAGE, Micron, DELL, SEAGATE

Data Center Operators

EQUINIX, DIGITAL REALTY, DIGITALBRIDGE, VANTAGE DATA CENTERS, CyrusOne Built for Tomorrow. Ready Today.

Data Center Infra

Power systems

ABB, EnerSys, VERTIV, Schneider Electric, EATON

HVAC

VERTIV, Schneider Electric, STULZ, CoolIT systems

Utility Providers

Electricity

aes, NEXTera ENERGY, Constellation, nrg, VISTRA, PSEG

Internet

LUMEN, zayo, cogent, EQUINIX

Vertically Integrated Data Centers

NVIDIA, TESLA, Meta, aws

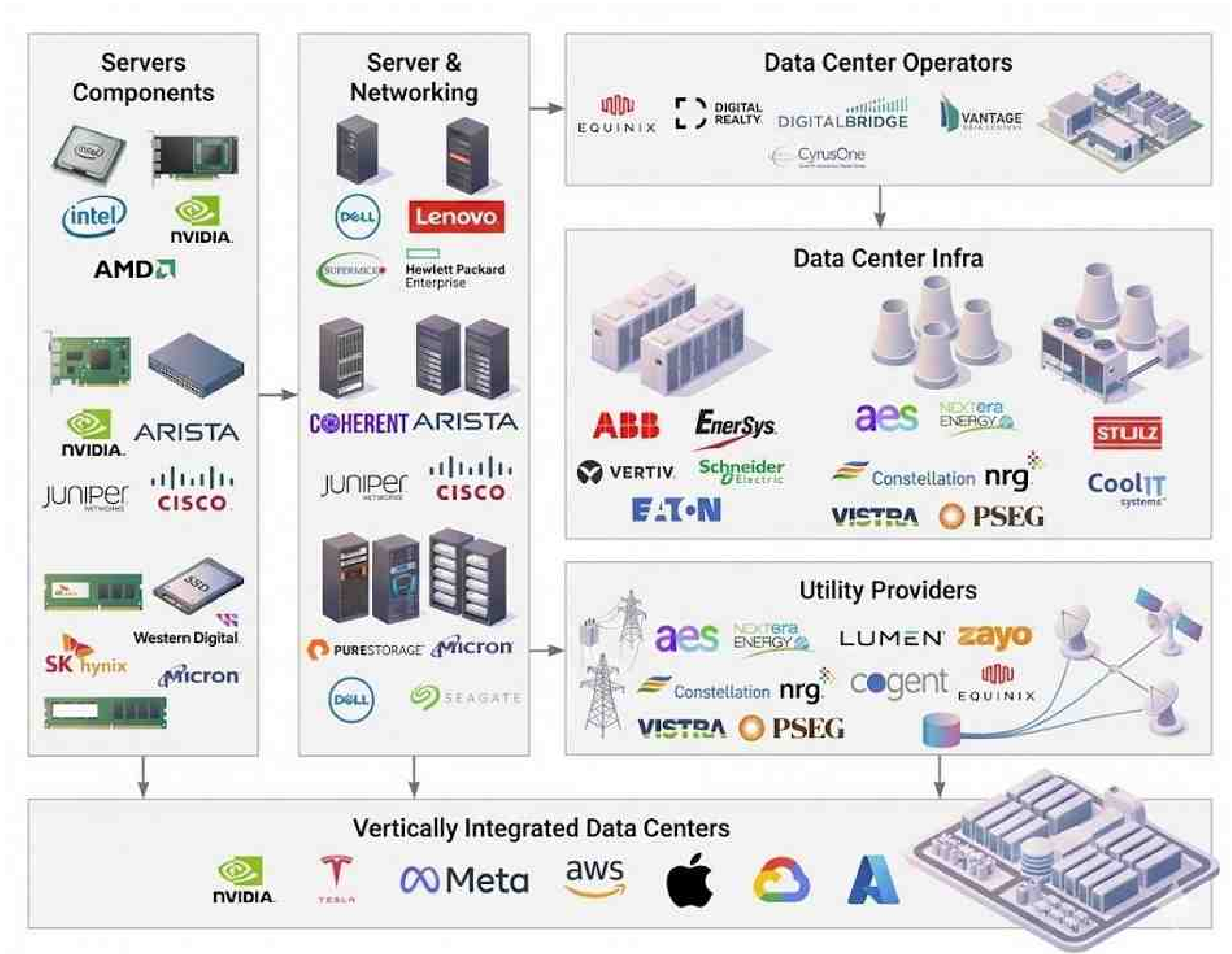
Independent data center operators
Orgs that own and manage large-scale data center facilities for customers

Supplier layer
A complex supply chain for every component of a chip and the machines that make them

Vertically integrated players

Data Center Value Chain (🍌Version)

Generational



Servers Components

Server & Networking

Data Center Operators

Data Center Infra

Utility Providers

Vertically Integrated Data Centers

Data Center Value Chain description

Generational

Server Components
This category represents the fundamental building blocks of computing systems—processors (CPU, GPU), networking interfaces, and memory modules. These components are sourced from various manufacturers and then integrated into servers and other IT equipment.

Server & Network
Here, the raw server components are assembled into full computing racks, storage arrays, and networking infrastructures. This layer integrates the core technology elements into functional IT units that can handle processing, storage, and data transfer tasks.

Data Center Infrastructure
Beyond the IT hardware, this includes the supporting physical environment—power distribution, cooling (HVAC), and other facility systems. Robust infrastructure ensures that servers and networking equipment operate within optimal conditions, maintaining performance, reliability, and energy efficiency.

Data Center Operators
These organizations run the data center facilities. They manage the physical space, provide maintenance, ensure security, and often offer services such as colocation and connectivity. They bring together the infrastructure layer with the server and network layer to deliver a managed environment to end users.

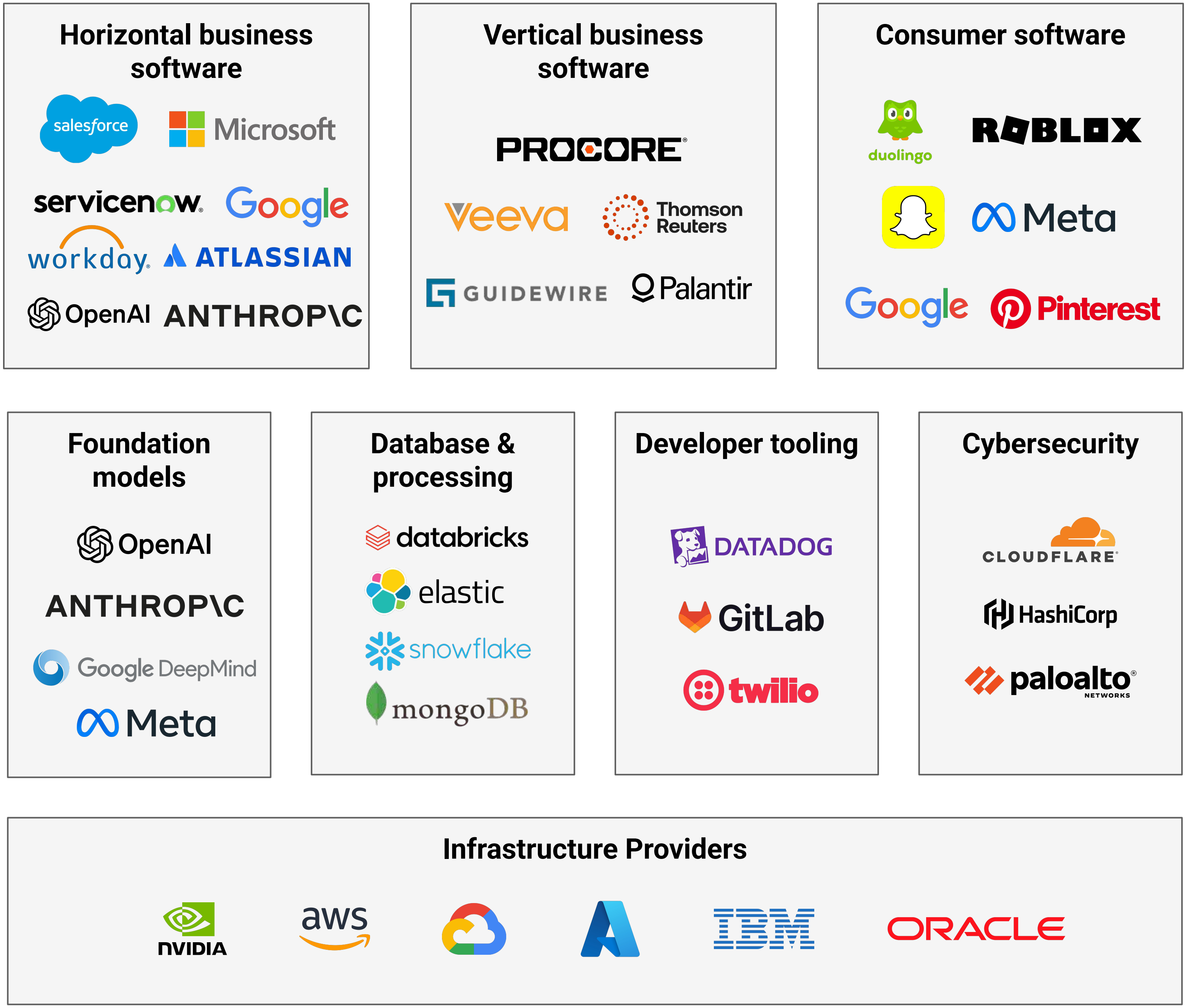
Utility Providers
Suppliers of electricity, water, and connectivity networks. They deliver the essential resources that keep the data center operational, such as consistent power and stable, high-speed internet connections.

Vertically Integrated Data Centers
Companies that control the entire stack, from chip design and server assembly to the management of large-scale data centers. By integrating vertically, they streamline supply chains, optimize performance at every layer, and maintain tighter control over costs and innovation cycles.

How they are connected
The server components feed into assembled server and network solutions, which require reliable data center infrastructure to function effectively. Data center operators bring together these hardware and infrastructure layers, managing the environment that supports users' computing needs. Utility providers ensure the continuous flow of critical resources. Finally, vertically integrated data centers stand apart by managing and optimizing the entire stack—from silicon to facility operations—creating a closed loop that can enhance efficiency, quality control, and innovation across the whole ecosystem.

Software Stack

Generational



Horizontal business software

Vertical business software

Consumer software

Application layer
Software that end users interact with directly, such as business applications or consumer services

Foundation models

Database & processing

Developer tooling

Cybersecurity

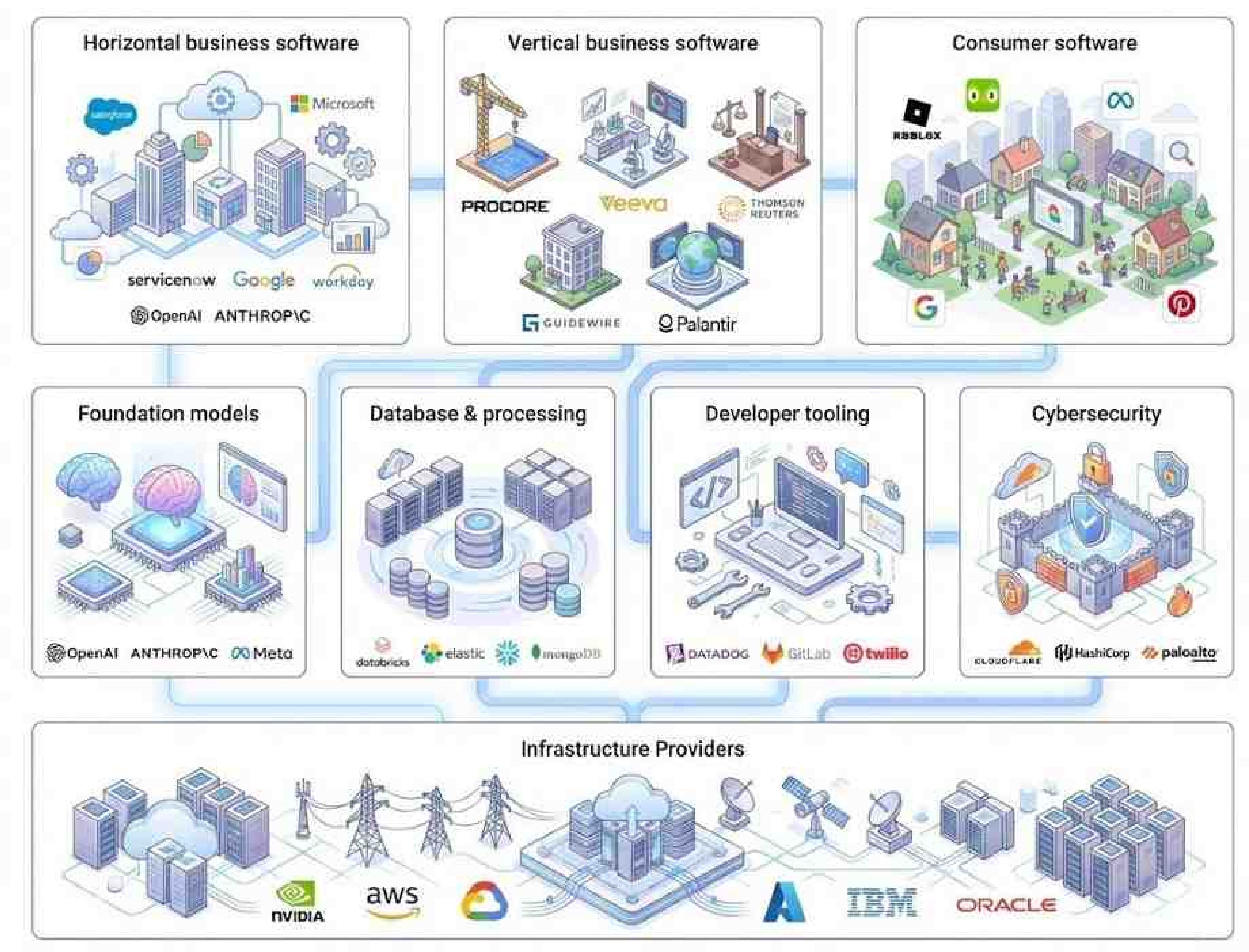
Middleware layer
Underlying services and tools that support and enhance the applications running above them.

Infrastructure Providers

Infrastructure layer
Foundational computing, storage, and networking resources that ensures software is reliable and scalable

Software Stack (🍌 Version)

Generational



Software Stack description

Generational

Horizontal business software
General-purpose solutions that address core business functions across industries—like customer relationship management, human resources, and productivity tools. These tools are broadly applicable, helping organizations streamline operations, collaborate effectively, and manage workflows regardless of their specific sector.

Vertical business software
Industry-specific applications tailored to the unique processes and regulatory requirements of particular fields, such as construction, healthcare, or financial services. By focusing deeply on niche needs, these solutions provide features and compliance measures that generic tools may not offer, helping companies in specialized domains optimize performance and reduce complexity.

Consumer software
End-user-facing applications designed for personal use, entertainment, or communication. These products cater to individual consumers, offering intuitive interfaces and engaging experiences, ranging from social media platforms to language-learning apps and online games.

Foundation models
Advanced AI and machine learning frameworks that serve as building blocks for intelligent applications. These models—often trained on vast datasets—provide capabilities like language understanding, image recognition, or recommendation engines that developers can integrate into business or consumer-facing software to add sophisticated features without reinventing core AI components.

Database & processing
Data management and analytics platforms that handle the storage, retrieval, and transformation of information. These technologies ensure that applications have fast, secure, and scalable access to the data they need. They also support real-time insights, analytics, and machine learning workloads that power intelligent decision-making.

Software Stack description

Developer tooling
A suite of software development, testing, deployment, and monitoring tools that streamline the production of applications. From version control and continuous integration to performance monitoring and debugging utilities, these tools help developers build reliable software more efficiently, ensuring that the layers above can evolve quickly and respond to user needs.

Cybersecurity
Solutions and services that protect infrastructure, data, and applications from cyber threats. They ensure data integrity, maintain user trust, and support regulatory compliance by safeguarding networks, systems, and software against attacks, breaches, and other security vulnerabilities.

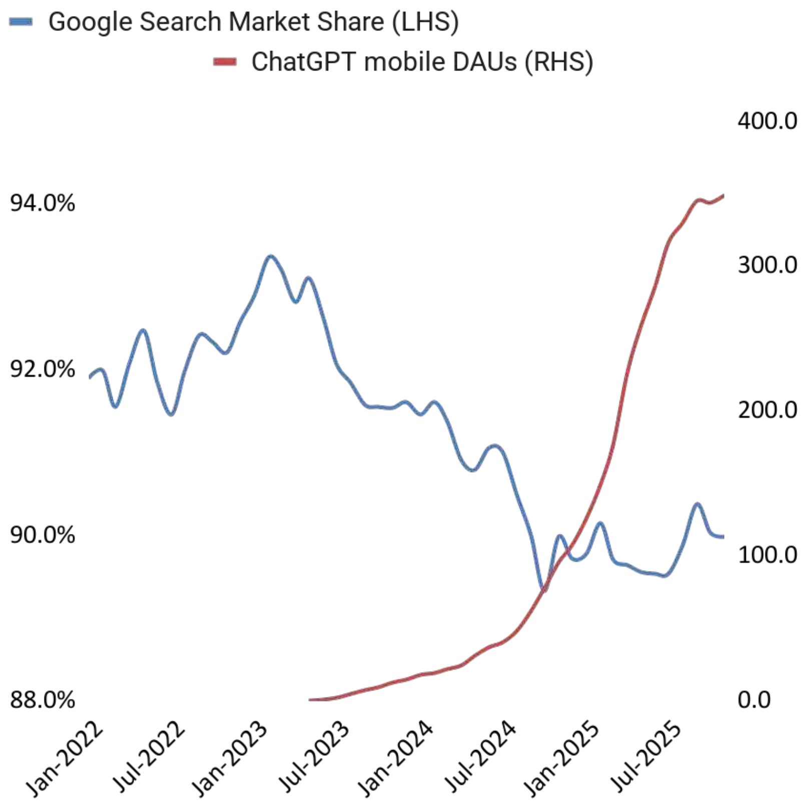
Infrastructure providers
Cloud, hardware, and platform vendors that supply the foundational compute, storage, and networking resources. These providers give applications and middleware the raw power and global reach they need to operate reliably and at scale. By handling the complexity of physical servers, data centers, and connectivity, infrastructure providers free other layers to focus on software innovation and user experience.

How they are connected
The software stack layers build upon one another. Infrastructure providers deliver the core computing and networking resources. On top of this, middleware services—such as databases, AI models, developer tools, and security solutions—provide essential capabilities that streamline software creation, enhance functionality, and ensure safety. Finally, application layers leverage this foundation to deliver tailored business solutions or consumer experiences. Each gray box category contributes its specialization to create a cohesive, scalable ecosystem that supports modern digital applications end-to-end.

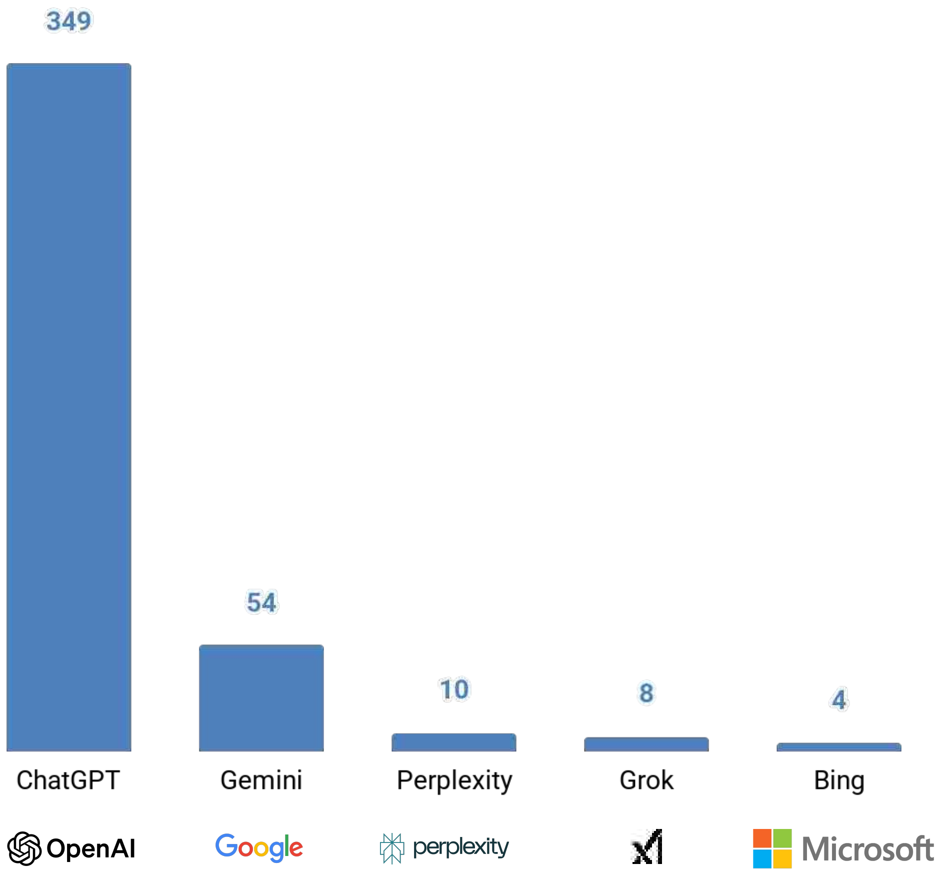
Google still is king of search

For over a decade, "Googling" has been the default consumer behavior online, but signs of disruption are emerging. Google's search market share has dipped to ~90% as OpenAI's ChatGPT gain traction. Every 1% loss in market share represents ~$31 billion in value (1% of Google's market cap). In 2024, Google kept losing market share but reversed its trajectory with its Gemini app and GenAI-enabled search experiences (AI overviews and AI mode). Gemini quickly overtook Perplexity as the 2nd leading dedicated GenAI app.

Google list market share but stabilized at 90%



GenAI mobile DAUs in millions



Source: Statcounter, Sensor Tower

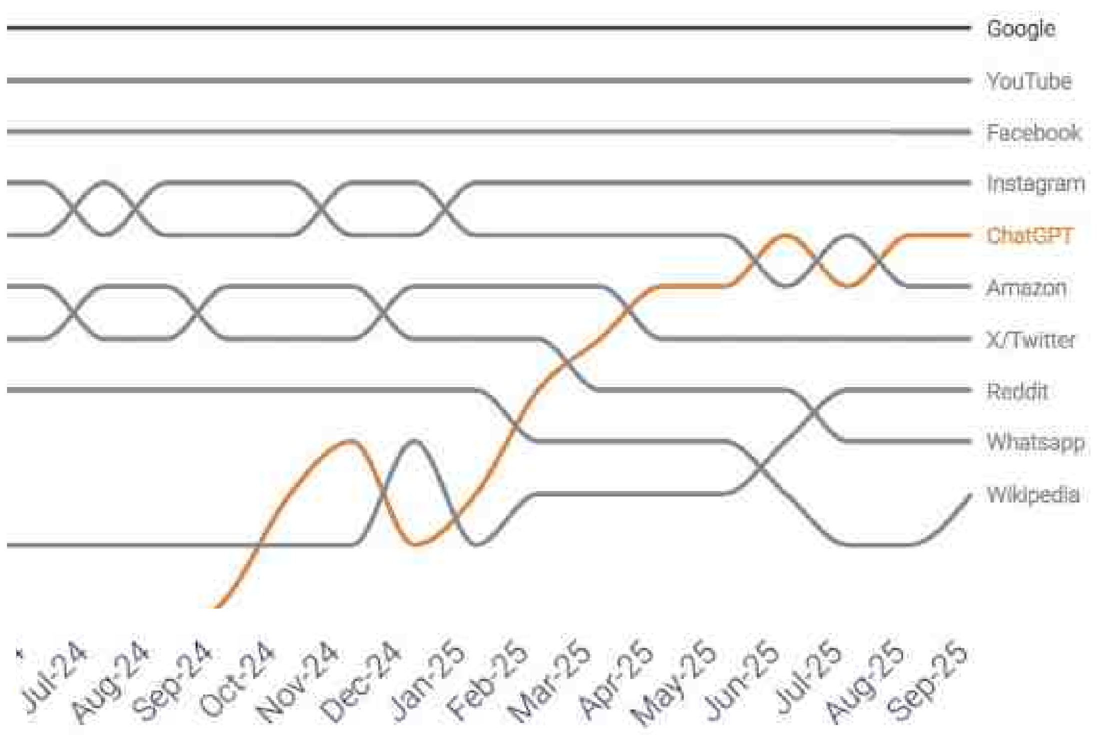
Internet services is about user attention

Generational

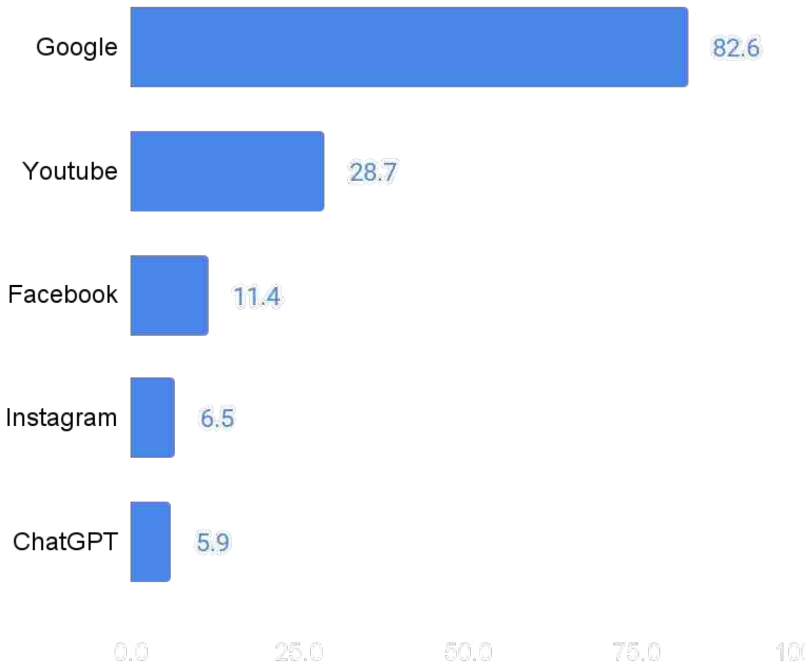
The internet services business model largely revolves around capturing users' attention to display ads and encourage purchases. The overall structure of the web has remain largely the same with just ChatGPT rising to the top 5. Other call outs:

- **Google is #1**: Google has as many visits as the other top 10 websites combined
- **Social platforms make up a vast majority of other top web properties:** YouTube, Facebook, Instagram, X, Reddit, Whatsapp
- **ChatGPT rocketship**: OpenAI's ChatGPT, only three years old, has already become a top 5 web property

Ranking of Top Websites by Monthly Visits



Website Monthly Visits in billions



Source: Similarweb, eMarketer, Author analysis, Company filings, News. Ad revenue based on 2023 numbers

Internet services is all about attention

Generational

While ChatGPT rose to become a top 5 website (note that a company can have multiple web properties, hence they're ranked sixth below), the internet market is still about ads and attention. Properties that capture more attention have more ad revenue.

ChatGPT has potential to generate billions in ad revenue, especially if it inserts itself in the commerce / purchasing path. You can see the difference between how Google and Amazon monetizes based on where they are in the purchasing process – users are still researching on Google (38% market share relative to 53% time spent) vs those on Amazon have high purchasing intent (9% market share relative to 2% time spent)

Top 10 companies based on web traffic		Avg monthly visit Sep-Nov' 25 (millions)	*Visits share*	Monthly time spent (million hrs)	*Time spent share*	Has ads business?	Ad revenue ($ billions)	*Ad revenue share*
Google	Google (incl. Youtube)	114,867	*27%*	23,949	*53%*	✓	297	*38%*
Meta	Meta (FB, Insta, Whatsapp)	21,690	*5%*	3,419	*8%*	✓	200	*25%*
Microsoft	Microsoft (LinkedIn, Github)	11,313	*3%*	1,385	*3%*	✓	16	*2%*
yahoo!	Yahoo	7,741	*2%*	923	*2%*	✓	1	*0%*
amazon	Amazon (incl. Twitch)	7,435	*2%*	737	*2%*	✓	71	*9%*
OpenAI	OpenAI (incl. ChatGPT)	6,322	*2%*	654	*1%*	Not yet	0	*0%*
X	X	4,320	1%	911	2%	✓	2	*0%*
	Reddit	3,910	*1%*	368	*1%*	✓	2	*0%*
	Wikipedia	3,560	*1%*	192	*0%*		0	*0%*
	Tiktok	2,960	*1%*	226	*0%*	✓	22	*3%*
Yandex	Yandex	2,840	1%	395	1%	✓	4.9	*1%*

Source: Similarweb, eMarketer, Author analysis, Company filings, News. Ad revenue based on 2025 Q3 numbers

Ads coming to ChatGPT

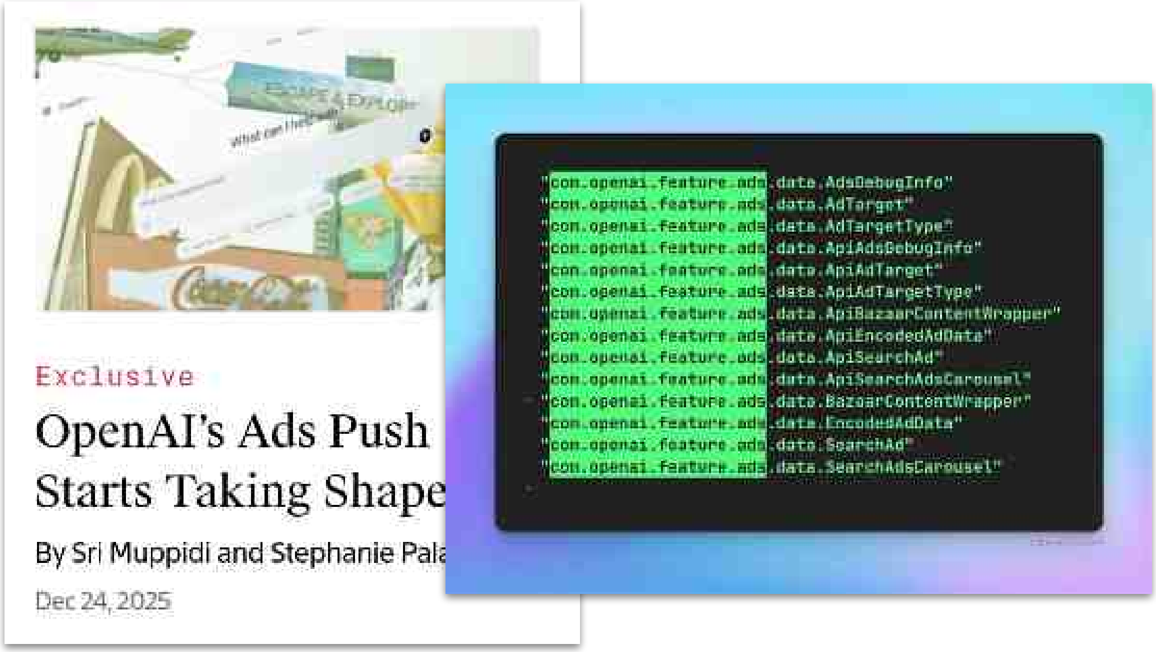
Generational

With over 900 million weekly active users, the ad opportunity might be too much to resist. So its no surprise that there's recent evidence that OpenAI is actively exploring advertising for ChatGPT, with code discovered in the Android app beta revealing references to "ads feature" and "search ads carousel," and internal mockups obtained by The Information showing formats ranging from sponsored content woven into responses to sidebar placements.

As Sam Altman himself said, the primary challenge of serving ads in ChatGPT is trust: unlike search engines where users expect a mix of paid and organic results, ChatGPT users expect definitive answers, making any perception of advertiser influence potentially problematic. The use case also differs from social media platforms—ChatGPT users arrive with specific intent rather than passive scrolling, making ad integration more complex.

Perplexity's experience offers a cautionary data point: after launching ads in November 2024 with partners like Indeed and Whole Foods, the company generated just $20,000 in advertising revenue. By October 2025, Perplexity stated they are not taking any new advertisers effectively pausing the initiative after advertisers cited limited scale, inability to measure ROI, and broader consumer skepticism around AI search results.

News suggesting ads are coming soon to ChatGPT



Exclusive

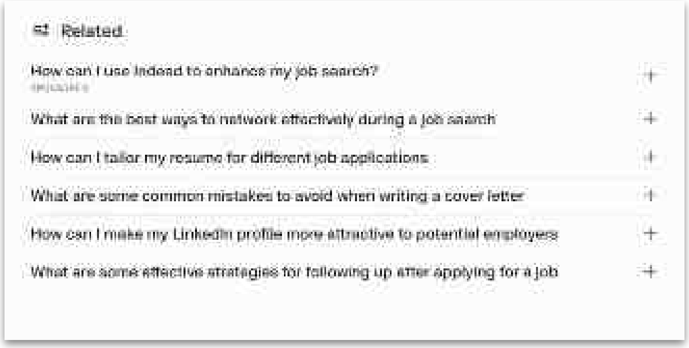
OpenAI's Ads Push Starts Taking Shape

By Sri Muppidi and Stephanie Pala

Dec 24, 2025

```
"com.openai.feature.ads.data.AdsDebugInfo"
"com.openai.feature.ads.data.AdTarget"
"com.openai.feature.ads.data.AdTargetType"
"com.openai.feature.ads.data.ApiAdsDebugInfo"
"com.openai.feature.ads.data.ApiAdTarget"
"com.openai.feature.ads.data.ApiAdTargetType"
"com.openai.feature.ads.data.ApiBazaarContentWrapper"
"com.openai.feature.ads.data.ApiEncodedAdData"
"com.openai.feature.ads.data.ApiSearchAd"
"com.openai.feature.ads.data.ApiSearchAdsCarousel"
"com.openai.feature.ads.data.BazaarContentWrapper"
"com.openai.feature.ads.data.EncodedAdData"
"com.openai.feature.ads.data.SearchAd"
"com.openai.feature.ads.data.SearchAdsCarousel"
```

Example Perplexity ad



Related

How can I use Indeed to enhance my job search? +

What are the best ways to network effectively during a job search +

How can I tailor my resume for different job applications +

What are some common mistakes to avoid when writing a cover letter +

How can I make my LinkedIn profile more attractive to potential employers +

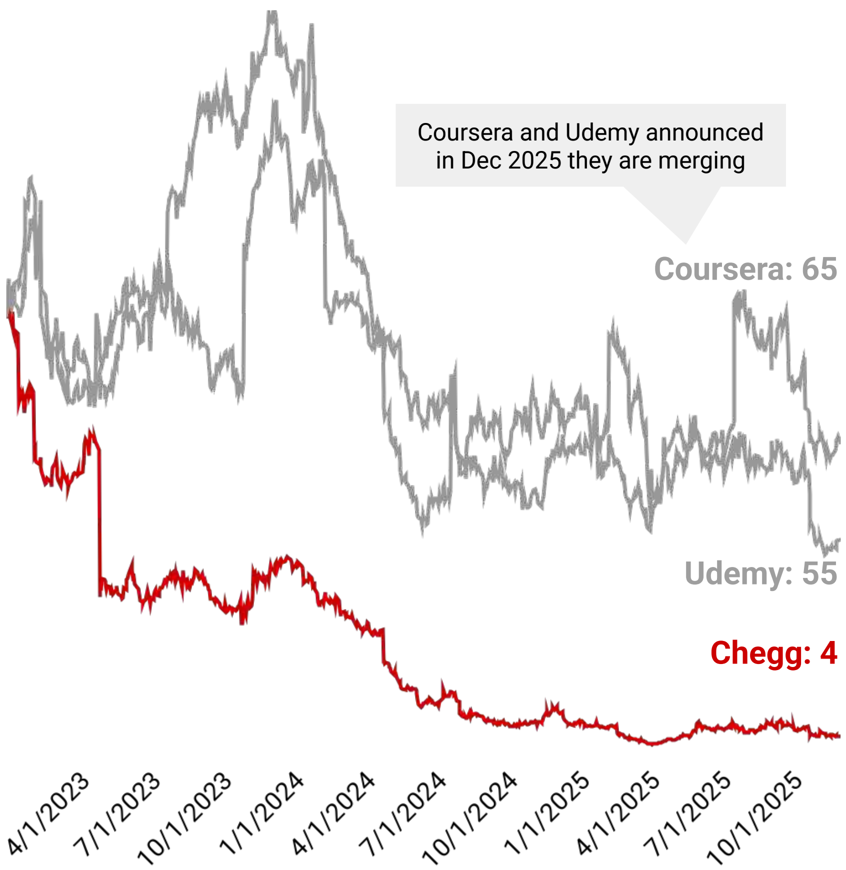
What are some effective strategies for following up after applying for a job +

Source: The Information, Company websites

EdTech is disrupted

Tools like ChatGPT gives students 24/7 access to tutoring and support, making education more accessible and flexible. Educators are also leveraging these tools to automate administrative tasks, allowing them to focus more on student interaction.

EdTech stocks (Index = 100 on 1/1/2023)



Coursera and Udemy announced in Dec 2025 they are merging

Source: Google Finance as of 12/19/2025

Chegg Chegg's Journey

Chegg (NYSE:CHGG) primarily generates revenue through its subscription services, offering academic support like homework help, tutoring, writing assistance, and math problem-solving.

Key company metrics

- Revenue: Down 58% from Q1 2023 ($188M → $78M)
- Stock: Down 99% from peak, 97% from Jan 2023
- Subscribers: Down ~50% (5.1M → ~2.5M)
- Traffic: Down 50% from Google

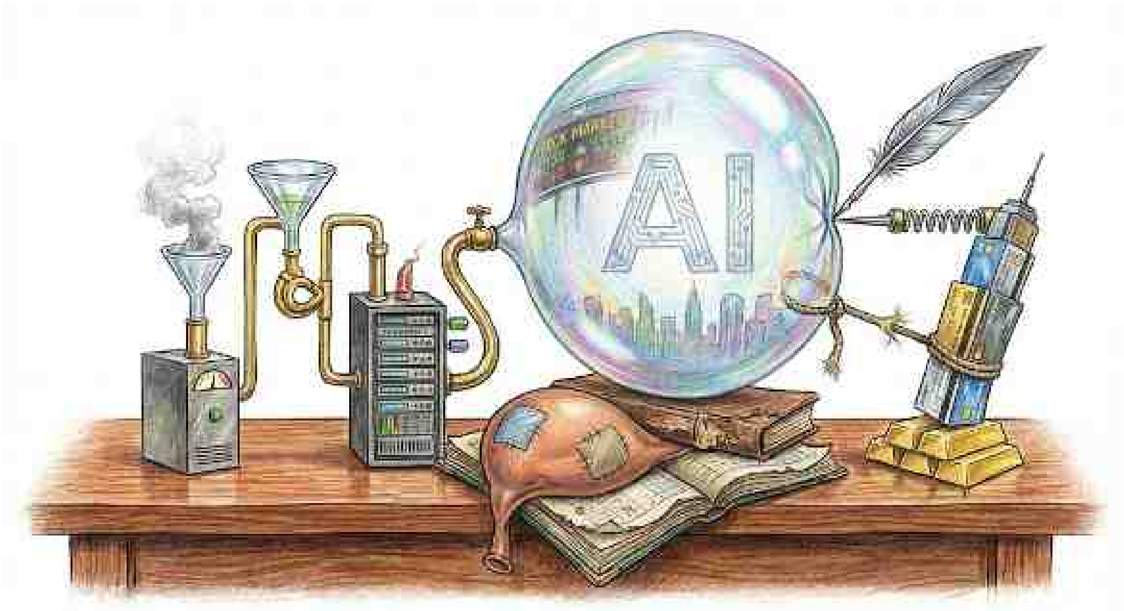
Earnings call quotes

"In the first part of the year, we saw no noticeable impact from ChatGPT on our new account growth. However, since March we saw a significant spike in student interest in ChatGPT."
— Chegg CEO in May 2023

"Across our industry, there has been a continued increase in the adoption of free and paid generative AI products... students are increasingly turning to generative AI for academic support to homeworking exams. "
– Chegg CEO in Nov 2024

"It's clear that the rise of AI and the subsequent negative impact on traditional sources of traffic have disrupted almost every direct-to-consumer industry... Our Google traffic dropped by 50%."
— Chegg CEO in Nov 2025

Section 4: Are we in an AI Bubble?

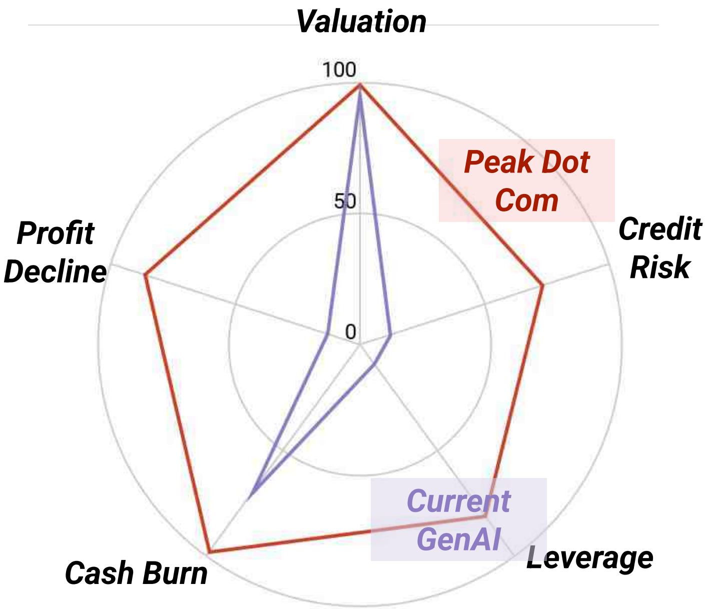


We are not in a AI bubble yet. Macro fundamentals look healthy

A bubble occurs when asset prices detach from fundamental value and then rapidly collapse. The problem: we can only confirm a bubble in hindsight. To assess risk in real-time, we look for leading indicators that preceded past crashes. Goldman Sachs identified five macro signals that peaked before the dot-com bust: extreme valuations, declining profits, elevated credit risk, corporate cash burn, and rising leverage. These moved in sequence—valuations ran hot first, then fundamentals deteriorated, then credit markets cracked.`

Applying this framework to AI in 2025, four of five indicators remain healthy. Corporate profits are strong, credit spreads are tight, the tech sector runs a financial surplus, and leverage sits well below dot-com peaks. The sole warning sign is valuations at the 95th percentile—approaching but not yet at the 99th percentile extreme of March 2000. Valuations alone don't cause crashes; they require deteriorating fundamentals or tightening credit as a catalyst. Neither is present today.

Macro Fundamentals Percentile Rank Since 1990	Dot Com Mid Cycle 1997 Q3	Peak Dot Com 2000 Q1	GenAI Today 2025 Q3	Current Status
Valuation (P/E ratio)	81	99	95	🔴Hot
Profit Decline (Inverse profit as % of GDP)	54	81	9	🟢Healthy
Credit Risk (IG credit spread)	1	73	12	🟢Healthy
Cash Burn (financial deficit as % of GDP)	72	98	70	🟢Healthy
Leverage (debt-to-profit ratio)	31	86	13	🟢Healthy

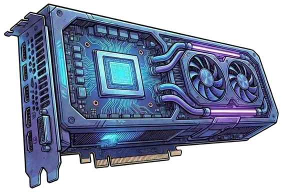


Source: Goldman Sachs, Author analysis

Bubbles drive productive investments but GPUs are different, they depreciate fast

Speculative overinvestment can yield long-term economic value. The railroad mania of the 1840s left behind track that carried freight for a century. The telecom bubble of the late 1990s laid 80+ million miles of fiber optic cable—85% sat "dark" for years after the bust. Global Crossing and WorldCom went bankrupt, but the fiber remained.Google, Amazon, and universities later acquired this dark fiber cheaply. YouTube's model—free video streaming—was only viable because the telecom bust made bandwidth cheap. The bubble front-loaded infrastructure that would have been difficult to build otherwise.

The AI bull case extends this logic: Even if data centers are overbuilt, won't the next generation inherit cheap compute? No – Fiber and GPUs have opposite depreciation dynamics. New tech made fiber more valuable over time. It makes aging GPUs worthless.



1990s Fiber		*2020s Data Center*
The physical medium (glass in the ground)	**Where value lives**	Mostly in the GPUs itself
Makes them more valuable. When wavelength-division multiplexing arrived, the same glass carried 16-96x more data.	**What new tech does**	Makes them obsolete. The new NVIDIA Blackwell chips delivers 3-5x more throughput than Hopper per watt.
Swap endpoint equipment (multiplexers), keep the fiber. Several generations of technology development without overhauling the backbone.	**Upgrade path**	Replace the entire chip. No endpoint swap extends a GPU's competitive life.
Irrelevant—fiber is passive, consumes no power.	**Power economics**	Critical – data centers are power-constrained. New GPU generate multiple more compute on same power.
25-30 years physical. Economic value increases with tech progress	**Asset lifespan**	2-4 years physical. Value decreases with each new generation
Google, Amazon built the cloud on cheap dark fiber acquired from bankrupt telecoms	**Post-bust utility**	TBD – but obsolete GPUs may become e-waste, not latent capacity for the next generation

Source: Author analysis

Generational

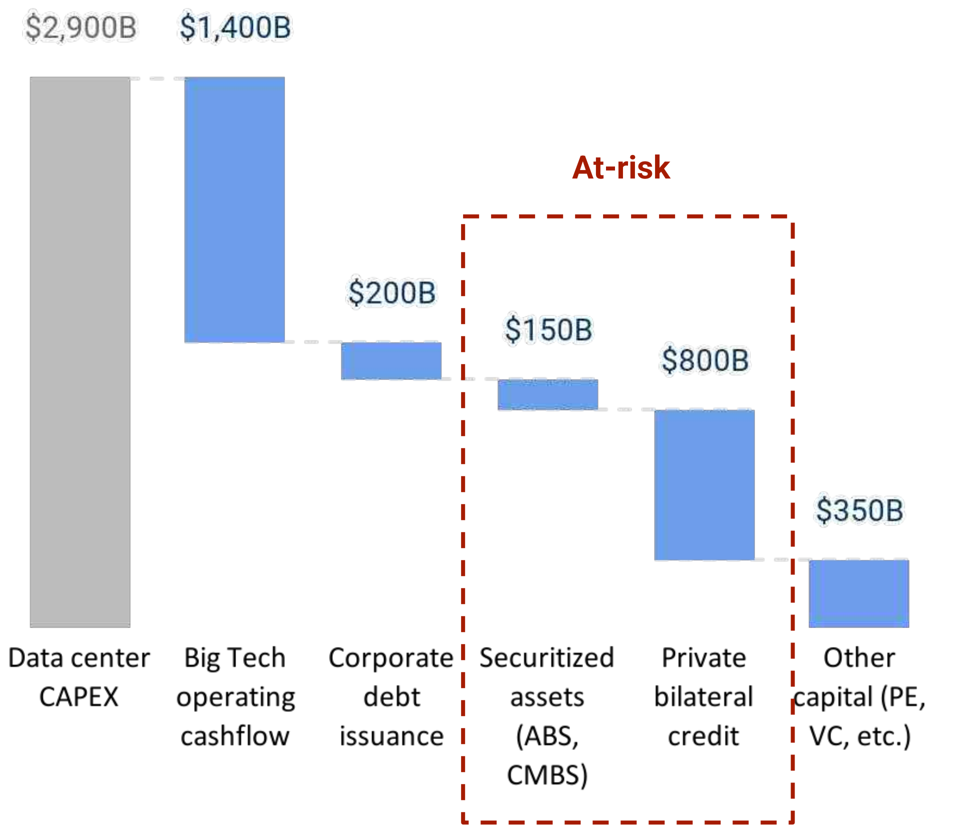
Demand for AI is real. But rapid GPU depreciation creates credit risk

Unlike speculative booms, the AI buildout is backed by tangible demand—record-low data center vacancy and massive pre-leasing commitments. The fragility lies not in demand, but in the capital stack assembled to meet it: specifically, reliance on external financing secured against rapidly depreciating assets.

Big Tech cash flows, while massive, can only fund about half of projected data center capex through 2028. The remaining $1.5 trillion must come from external sources, with private credit providing the largest—and highest-risk—tranche. Big Tech' cash reserves have been drawn down (from 29% cash-to-asset ratio in 2021 to 15% in 2025), making external capital a necessity.

Demand	Supply
🟢Healthy	🟡 At Risk
Record low 1.6% data center vacancy. 74% of new data center capacity is pre-leased	$1.5T external financing for data centers needed by 2028. $800M from untested private credit
Real enterprise adoption and revenue	Loans assume 5-6 year asset life for a rapidly depreciating GPU
Rapid consumer adoption (ChatGPT >900M WAUs)	New data centers face 1-4 years of delay in getting power (discussed in section 2)

Data Center CAPEX by 2028 and Financing Sources



Source: Author analysis

GPU Economics: Testing the 5-6 Year Depreciation Assumption

Generational

Previous slides established that AI demand is real, but GPUs depreciate fundamentally differently than past infrastructure. Lenders have extended $800B+ in private credit assuming 5-6 year asset lives. The table tests that assumption.

We model a single GPU using CoreWeave-like financing (85% debt, 10% rate, 5-year amortizing) and compare Market prices (current), Breakeven (debt + OPEX), and Equilibrium ($/TFLOPS parity when supply normalizes).

At today's market prices, the debt structure works—all generations show 23-48% margins. The risk emerges when compute abundance arrives. Current prices sit 77% above equilibrium for A100 and 41% for H100, a shortage premium that compresses as B200 scales and newer chips enter the market.

At equilibrium, the 5-year assumption holds for H100 and newer (15-48% margins). It fails for A100—equilibrium lands 26% below breakeven. Operators holding A100 debt need the shortage to persist for the full loan term or consider recapping to finance newer GPUs. If compute abundance arrives first, credit losses follow.

Illustrative 1 GPU "Data Center"	A100	H100	B200	Notes
GPU Overview				
Availability	Q2 2020	Q3 2022	Q1 2025	
FP16 TFLOPS	312	990	2,250	
TDP (Watts)	400	700	1,000	
CAPEX & Debt Service				
GPU Purchase Price	$15,000	$30,000	$42,000	
Total CAPEX	$27,273	$54,545	$76,364	GPU is 55% of DC CAPEX
Debt Service	$6,115	$12,231	$17,123	85% debt capital, 10% rate, 5yr amortization
OPEX				
Electricity Cost/yr	$252	$442	$631	TDP × $0.08/kwh × 1.2 PUE × 6,570 hrs
Total OPEX/yr	$420	$736	$1,051	Electricity is 60% of OPEX
Price Comparisons at 75% utilization				
Market Price	$1.29	$3.29	$5.29	Lambda Labs Dec 2025
Total Expenses / Breakeven Price	$0.99	$1.97	$2.77	Debt Service + OPEX
Equilibrium Price	$0.73	$2.33	$5.29	$/TFLOPS parity with B200
Cash Margins (Non-GAAP)				
At Market Price	23%	40%	48%	
At Equilibrium Price	-36%	15%	48%	

Comparable to Coreweave's debt structure

GPU market prices are elevated due to shortage in compute capacity

Source: Author analysis

Market Is Already Pricing In Credit Risk

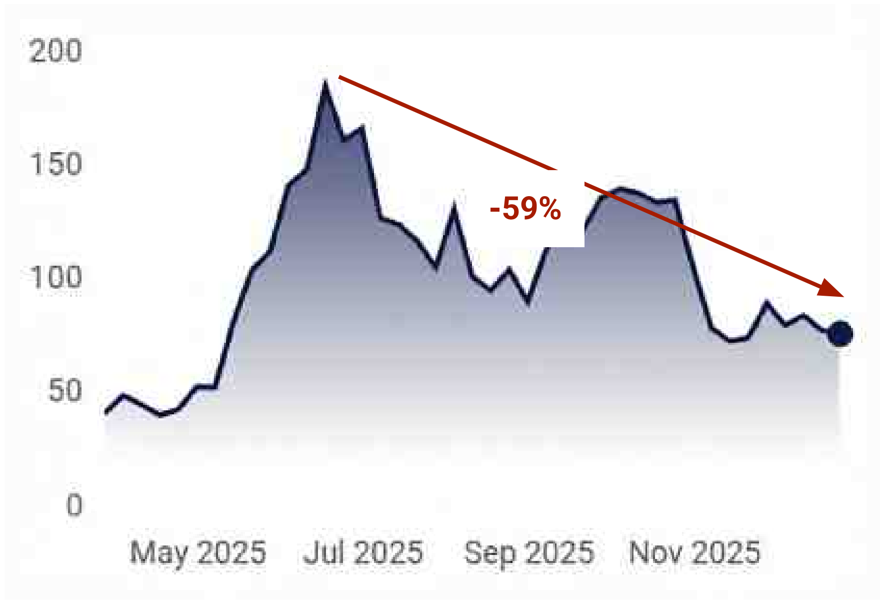
While not yet a full-blown crisis, credit and equity markets are showing clear signs of stress specifically related to this risk. These are not broad market trends, but targeted signals of concern around the data center build out.

CoreWeave's 60% decline reflects the collision of GPU economics with leveraged growth: $8B in annual cash burn, $310M quarterly interest expense, and hardware that depreciates faster than debt pays down. Oracle's bond market tells the same story—CDS spreads at 2009 crisis levels despite investment-grade ratings, with $248B in off-balance-sheet data center commitments yet to hit the books.

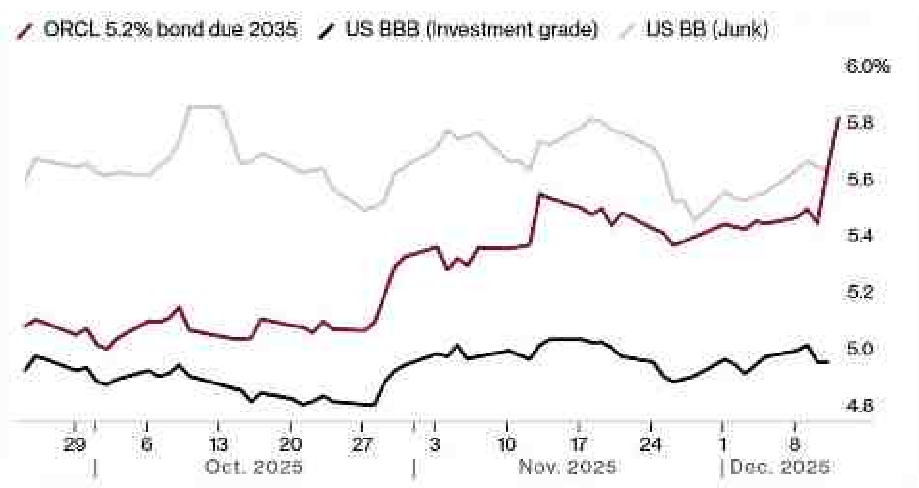
I am also deeply concerned about the "speculative" data center market. We foresee a significant financing crisis in 2027-2028 for speculative landlords. – Alexander Davis, CEO of Disruptive VC, Groq's largest investor

Capital investments required for AI computing represent significantly higher financial risks than previous tech cycles... [We view] potential overbuilding and technical obsolescence as key credit risks – Moody's May commentary

Coreweave stock price down ~60% from peak



Oracle bonds are trading like junk bonds



Source: Author analysis, Bloomberg, Google Finance

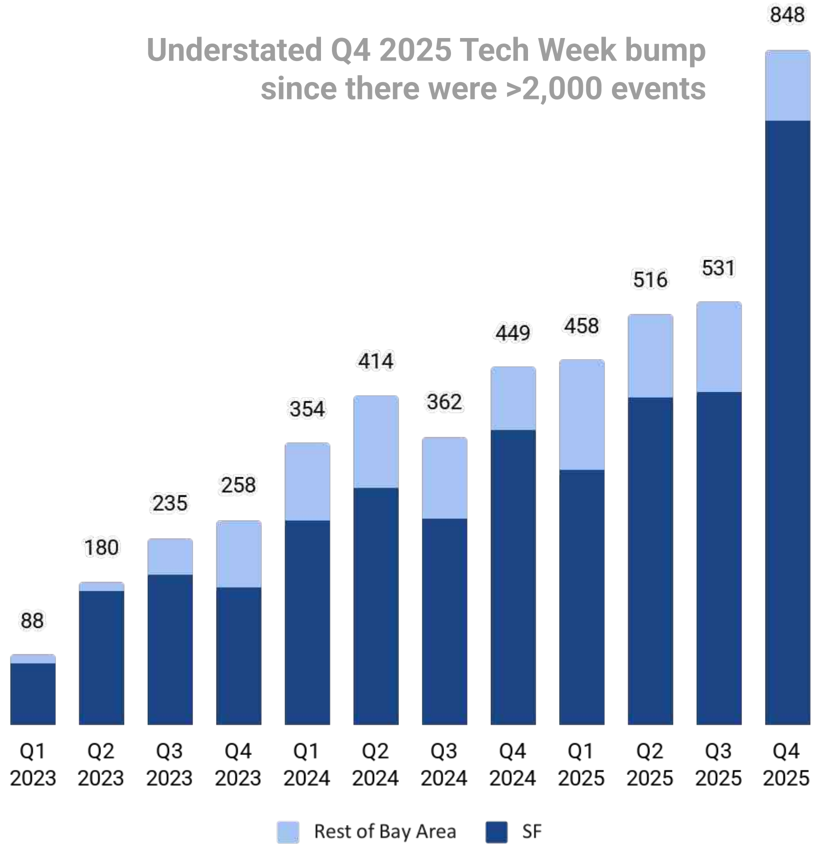
Section 5: Private Markets

San Francisco is still the center of AI

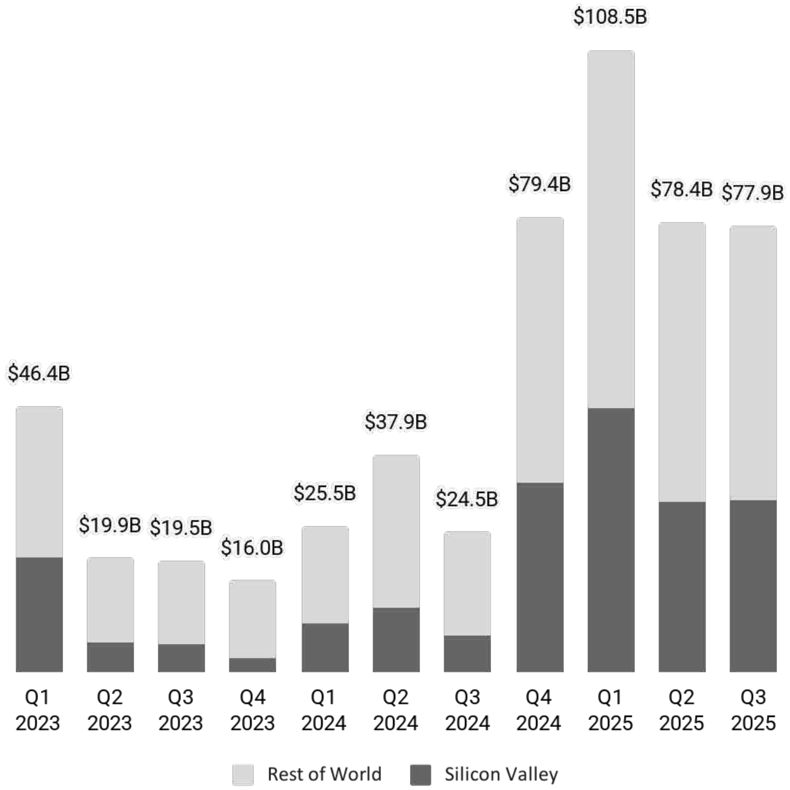
The San Francisco Bay Area, often called "Cerebral Valley," remains the epicenter of the AI world, with its unique mix of hacker houses, startups, and coworking spaces driving innovation. In 2024, the region hosts over almost 10 AI events daily (excluding remote events), up from just three per day a few years ago. San Francisco city itself hosts almost 90% of these events.

When it comes to funding, Silicon Valley's dominance is clear, attracting 40% of global AI venture dollars year-to-date. This blend of community and capital cements its status as the go-to hub for AI.

>10 AI events happening everyday with 90% in SF city



40% of 2025 YTD AI funding goes to SF Bay Area Startups

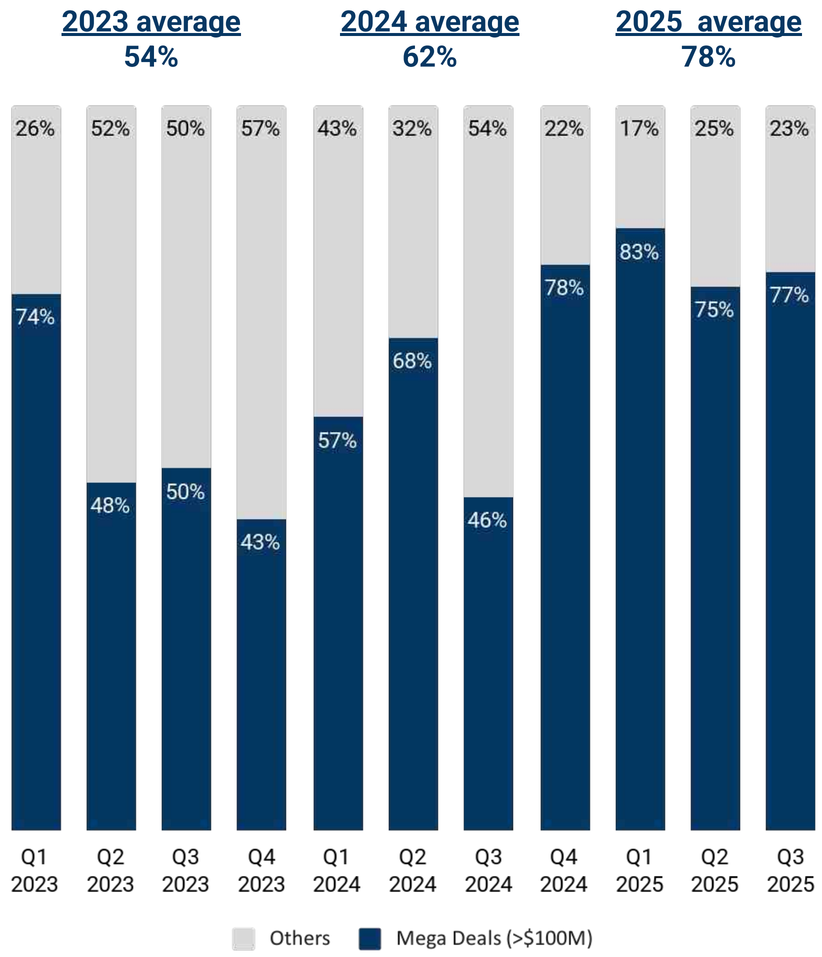


Source: Cerebral Valley, CB Insights, Author analysis

AI funding gets even more concentrated

The AI startup ecosystem is becoming increasingly concentrated. In 2025 YTD, mega deals (>$100M deals) took up 80% of the funding. This shift reflects investors placing larger bets on fewer companies, pushing valuations higher. However, this concentration raises concerns about valuation bubbles and the cascading risks if some of these companies fail. The scale of self-dealing has also increased making some investors wary.

~80% of AI funding goes to large funding rounds



Private AI companies that have raised the most in 2025

Company	Description	Latest Valuation	2025 Funding
OpenAI	Frontier AI company	$500.0B	$41.0B
Anthropic	Frontier AI company	$350.0B	$32.5B
Scale	Data provider	$30.2B	$14.8B
xAI	Frontier AI company	$75.0B	$12.5B
Project Prometheus	AI lab	N/A	$6.2B
Anysphere	Coding agents	$29.3B	$3.2B
Anduril	Defense tech	$30.5B	$2.5B
Groq	Cloud infrastructure	$6.9B	$2.3B
Mistral AI	AI lab	$13.7B	$2.0B
Reflection AI	AI lab	$8.0B	$2.0B
Safe Superintelligence	AI lab	$32.0B	$2.0B
Thinking Machines	AI lab	$12.0B	$2.0B

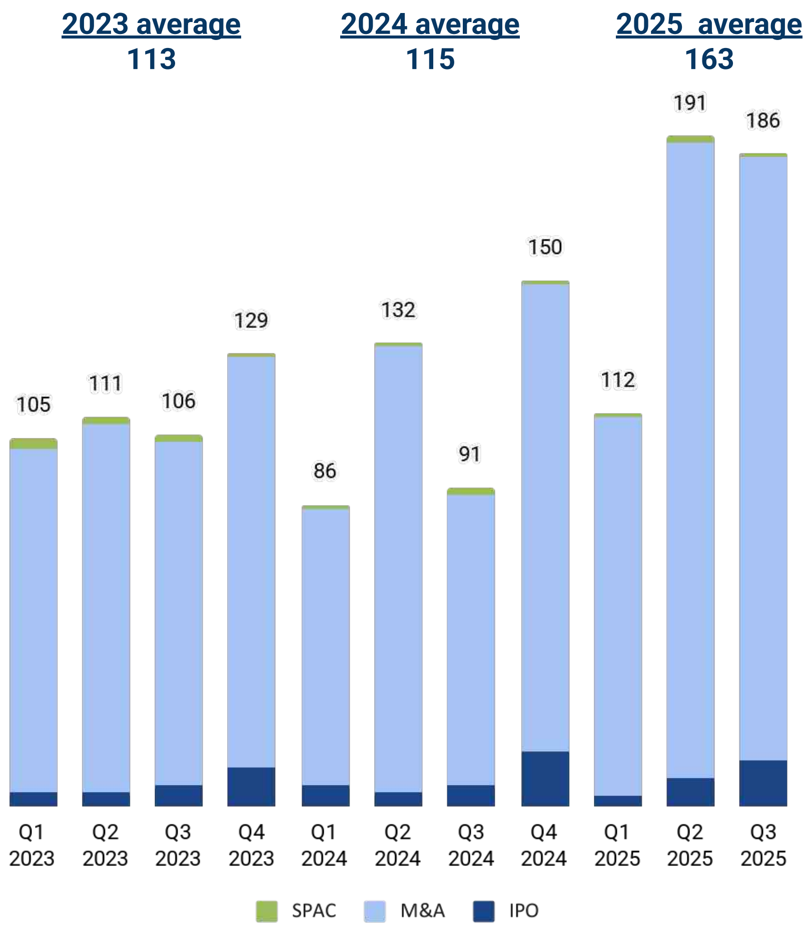
Source: Cerebral Valley, CB Insights, Author analysis

Rich exits are happening, mostly through acquisitions

Generational

AI exit activity increased 44% from 2023 to 2025 (113 → 163 average quarterly deals). But the buyer profile tells a different story. The largest exits aren't IPOs creating independent companies—they're consolidation into Big Tech. Google ($32B Wiz), NVIDIA ($20B Groq), OpenAI ($6.5B io), Softbank ($6.5B Ampere). The "acqui-hire" structure—licensing IP and hiring talent without formal acquisition—has become standard playbook to sidestep antitrust review.

~80% of AI funding goes to large funding rounds



Largest AI exits both IPO & M&A as of 12/30/2025

Target Company	Acquirer/IPO	Round Valuation
Wiz	Google	$32.0B`
Coreweave	IPO	$23.0B
Groq	NVIDIA	$20.0B
io	OpenAI	$6.5B
Ampere	Softbank	$6.5B
Figure	IPO	$5.3B
Moveworks	ServiceNow	$2.9B
Pattern	IPO	$2.5B
Windsurf	Google	$2.4B
Manus	Meta	~$2B
AI21	NVIDIA (rumor)	$2-3B
Weights & Biases	CoreWeave	$1.7B
HeartFlow	IPO	$1.5B
Sana Labs	Workday	$1.1B
Statsig	OpenAI	$1.1B

Source: CB Insights, Author analysis

L&A is the M&A tactic to avoid antitrust reviews

Generational

Licensing and acquihiring (L&A) has emerged as Big Tech's go-to strategy for acquiring generative AI companies. Instead of full acquisitions, firms like Microsoft, Amazon, and Google have opted for quasi-acquisitions of companies like Inflection, Adept, and Character.ai. Through this approach, they hire key AI talent and pay licensing fees to investors, gaining access to expertise, datasets, and GPU capacity while sidestepping regulatory scrutiny. This method compensates investors while avoiding the antitrust reviews and legal complexities of traditional M&A.

- Inflection co-founder Mustafa Suleyman is now the CEO of Microsoft AI
- Adept co-founder David Luan now heads the AGI Lab in Amazon
- Character.ai co-founder Noam Shazeer now co-leads Google Gemini

For a deep dive, read Unpacking Big Tech's quasi-acquisitions of GenAI companies and $200M AI Talent

Acquirer	Acquiree	Employees Hired	License Terms	Deal Terms	Acquiree financing	Acquiree last valuation
Microsoft	Inflection AI	~100% of 70 employees	Non-exclusive license to AI models	$620 million for the license, additional $30 million to waive legal rights related to mass hiring. Investors get 1.1-1.5x return	$1.6B	$4.0B
Amazon	Adept AI	~66% of 100 employees	Non-exclusive license to AI models and dataset	Adept receives ~$25 million. investors get ~1x return (~$414 million distributed)	$414M	$1.0B
Google	Character AI	~21% of 140 employees	Non-exclusive license to AI models	$2.5 billion. Investors get 2.5-12.5x return	$193M	$1.0B
Meta	Scale AI	<1% of 1,400 employees (~10)	Non-exclusive license to data-labeling tech and datasets	$14.3 billion for 49% non-voting stake. Investors get ~2x return	$1.6B	$14.0B
Google	Windsurf	~20% of 150 employees	Non-exclusive license to technology	$2.4 billion in licensing fees + compensation. Investors get ~1.6x return	$243B	$1.25B
NVIDIA	Groq	~90% of 550 emloyees	Non-exclusive license to technology	$20 billion in licensing fees + compensation. Investor payout detaills TBD.	$1.8B	$6.9B

Source: Media reports

AI companies are redefining revenue benchmarks

Generational

a16z analysis of hundreds of AI companies over 18 months reveals structural shifts: consumer AI now outpaces enterprise in Year 1 ARR ($4.2M vs $2.1M median), the old $1M "best in class" is now the lower end, and top performers see growth accelerate through Year 1 rather than slow.

Key Patterns

The "Great Expansion" in retention
Pre-AI consumer apps saw 30-40% Y1 revenue retention as "best in class." Leading AI companies now see >100% retention—users spend more over time via usage-based pricing, and bring tools into workplaces where they can expense them. Cohorts appreciate rather than depreciate.

Consumer-to-enterprise "pull" is accelerating
Consumer AI tools are getting "yanked into enterprise faster than ever"—employees adopt individually then bring to teams. Companies no longer choose B2C or B2B; they serve both. Canva took 7 years to launch Teams; in 2025, such delays are no longer viable.

Speed is becoming a moat
Many raise Series A before 12 months of revenue (8-9 mo median). The gap between good and exceptional is widening. Enterprise CIOs cite faster innovation rate as the primary reason they prefer AI-native vendors over incumbents.

Consumer brands translate to enterprise demand
Strong consumer products (ChatGPT, ElevenLabs, Cursor) drive enterprise adoption—CIOs note decisions driven by "employees loving the product." This dual-market pull has led to much faster growth than prior eras. The most important enterprise companies of the AI era may begin as consumer products.

Select Examples

CURSOR	**$1M to $100M ARR in 1 year**
Lovable	**$240K to $7M ARR in 1 year**
together.ai	**$1M to $26M ARR in 1 year**
Gamma	**$4M to $30M ARR in 1 year**
HeyGen	**$400K to $19M ARR in 1 year**

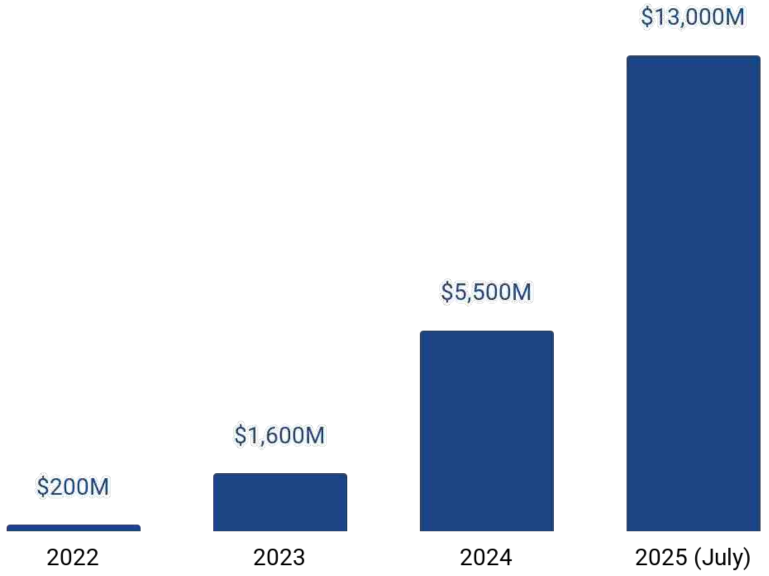
Year 1 benchmarks	Pre-AI	AI Era
Enterprise startup	$1M ARR	**$2.1M ARR**
Consumer startup	Delayed	**$4.2M ARR**
Time to series A	18-24 months	**8-9 months**
Net revenue retention	30-40%	**>100%**

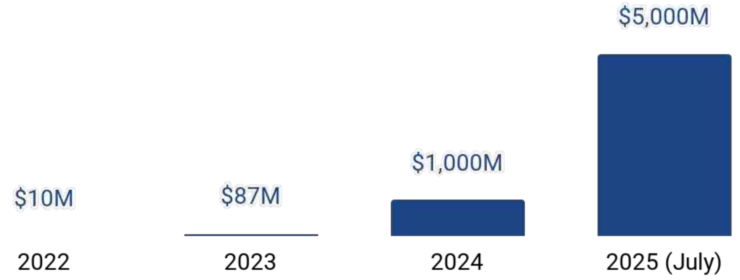
Source: a16z

OpenAI & Anthropic are generational firms that only commercialized <5 years ago

OpenAI and Anthropic highlight two distinct paths in the foundation model space. OpenAI's growth is consumer-driven, with almost 1 billion weekly active ChatGPT users. Anthropic, on the other hand, focuses on enterprise use case with a 40% share of the API market vs OpenAI's 27%. Their respective valuations would already in the top 25 of S&P500 companies.

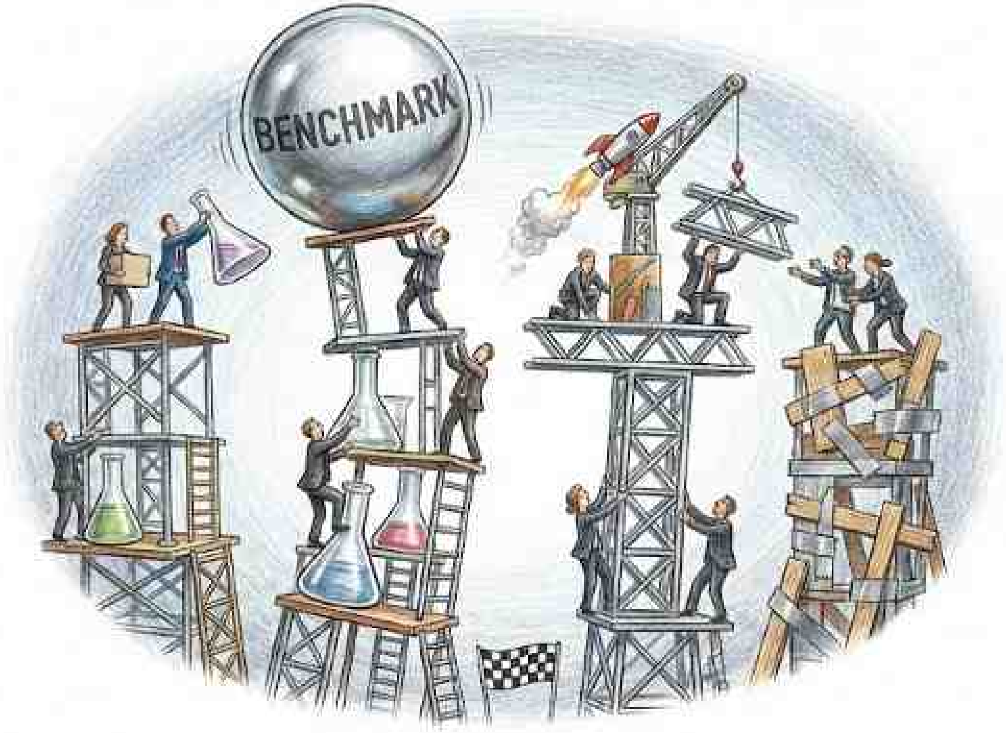
	OpenAI OpenAI is consumer-driven	ANTHROP\C Anthropic is enterprise-driven
Annualized Run rate July 2025	$13B	$5B
Capital raised	$65.1B 5.0x ARR ratio	$46.5B 9.3x ARR ratio
Valuation	$500B (secondaries) Equivalent to being in top 20 of S&P500 39x ARR ratio	$350B Equivalent to being in top 25 of S&P500 70x ARR ratio





Source: Sarcra

Section 6: AI Startup Benchmarks

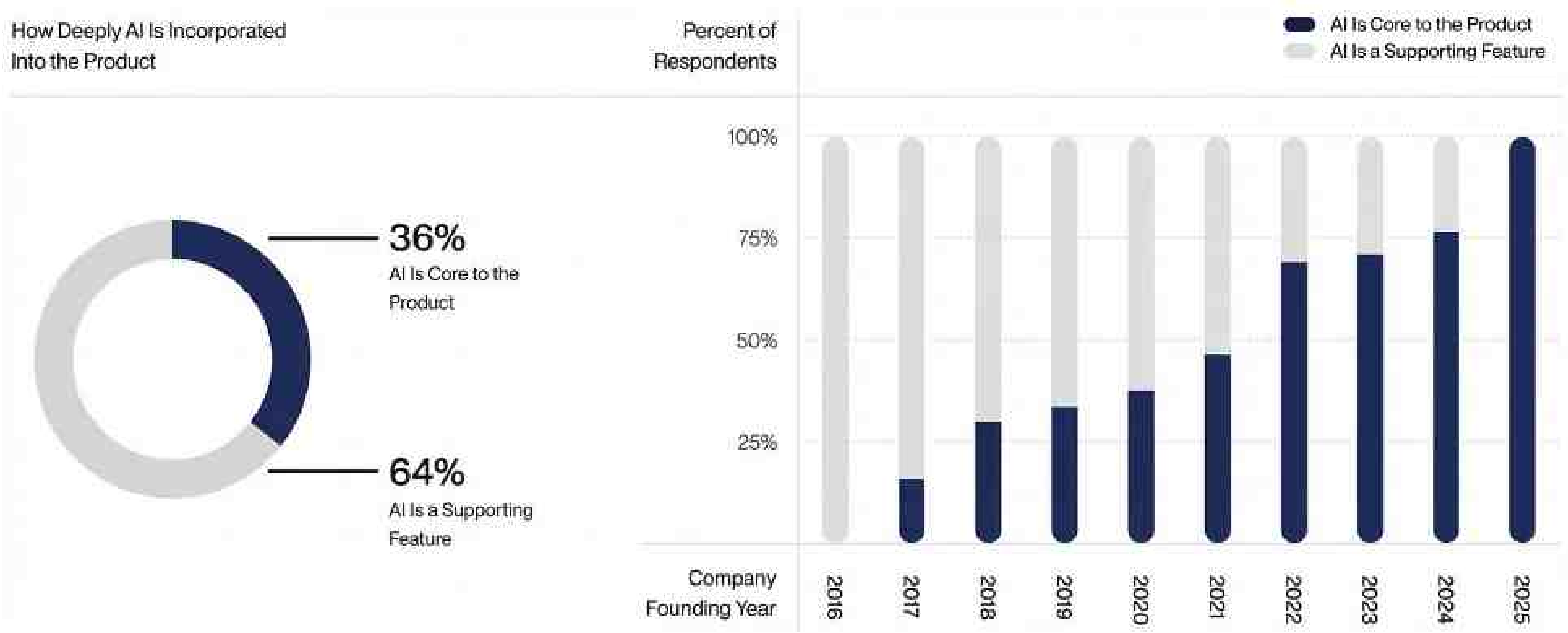


Complete Shift in Last 10 Years to AI Centricity

Generational

ChatGPT's 2022 launch significantly changed how deeply companies incorporate AI in their products. Of respondents whose companies were founded in 2016, none indicated that AI was core to their product.Fast forward ten years and that has completely flipped — AI is core to 100% of products for companies founded in 2025. The progression is clear and steady over the last decade towards an AI-driven product strategy

AI Product Incorporation

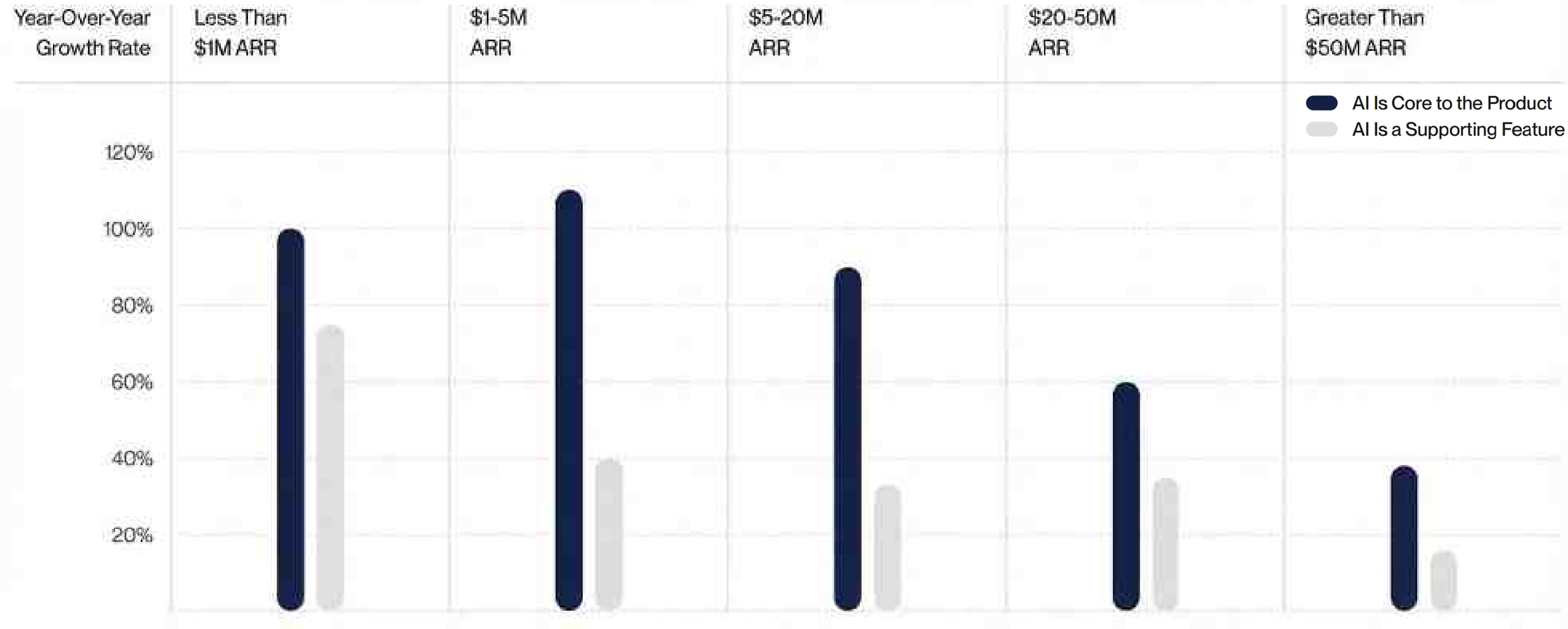


Source: High Alpha

Embedding AI Deeply in Product Drives Outsized Growth

SaaS companies with AI deeply incorporated into their products outperform peers across all ARR bands, growing twice as fast as those with AI as a supporting product feature. The performance gap is most significant in the $1-5 million ARR cohort, where AI differentiation drives 70% faster growth. The gap narrows with scale, but growth rates remain higher.

Growth Rate Based On How Deeply Ai Is Incorporated Into The Product Offering

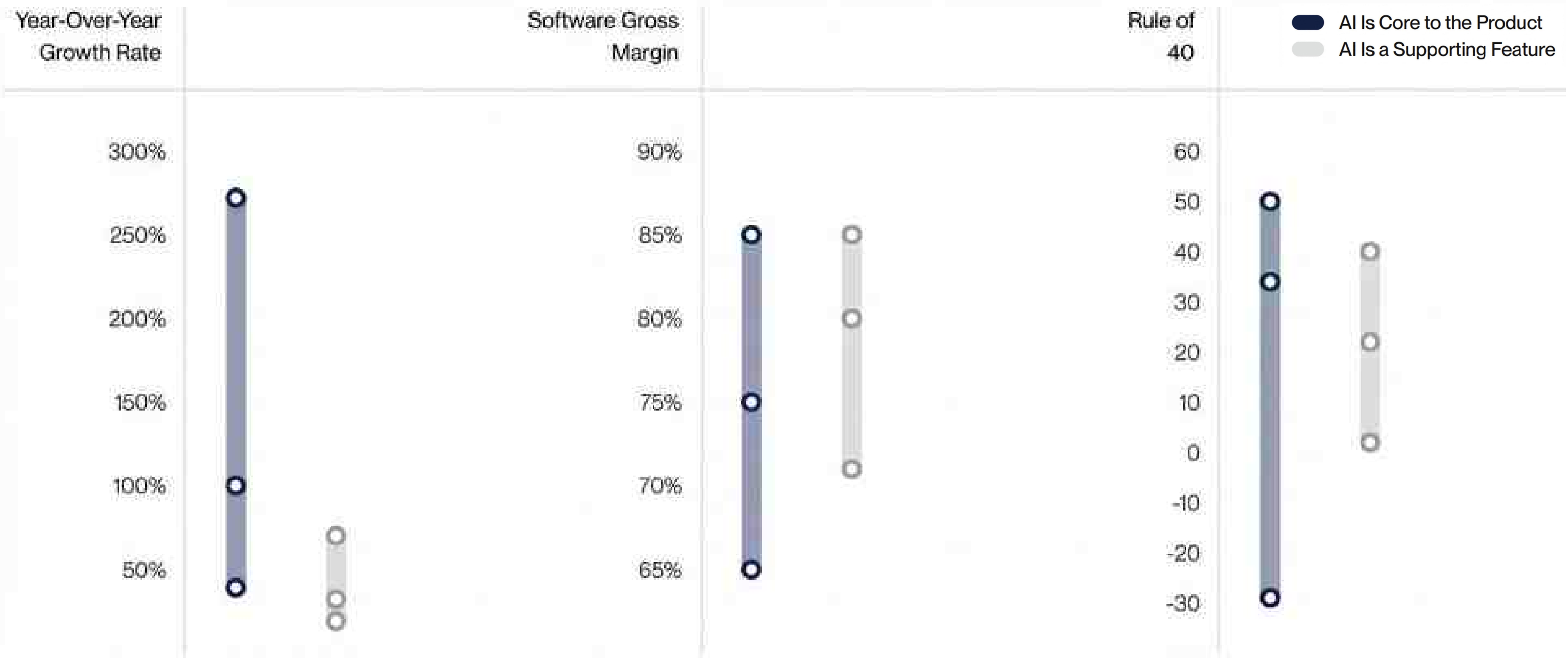


Source: High Alpha

Although Tradeoffs Exist, Growth Far Outweighs Lower Gross Margins

Companies with AI at the core of their products are growing over three times faster than peers, but with roughly five points lower in median gross margin. The tradeoff highlights that AI-first products drive outsized growth and Rule of 40 performance, even as delivery costs rise. Surprisingly, median gross revenue retention and net revenue retention were identical between companies where AI is core to the product and those where AI is a supporting feature.`

Performance When AI Is Core To The Product Vs. Supporting Feature

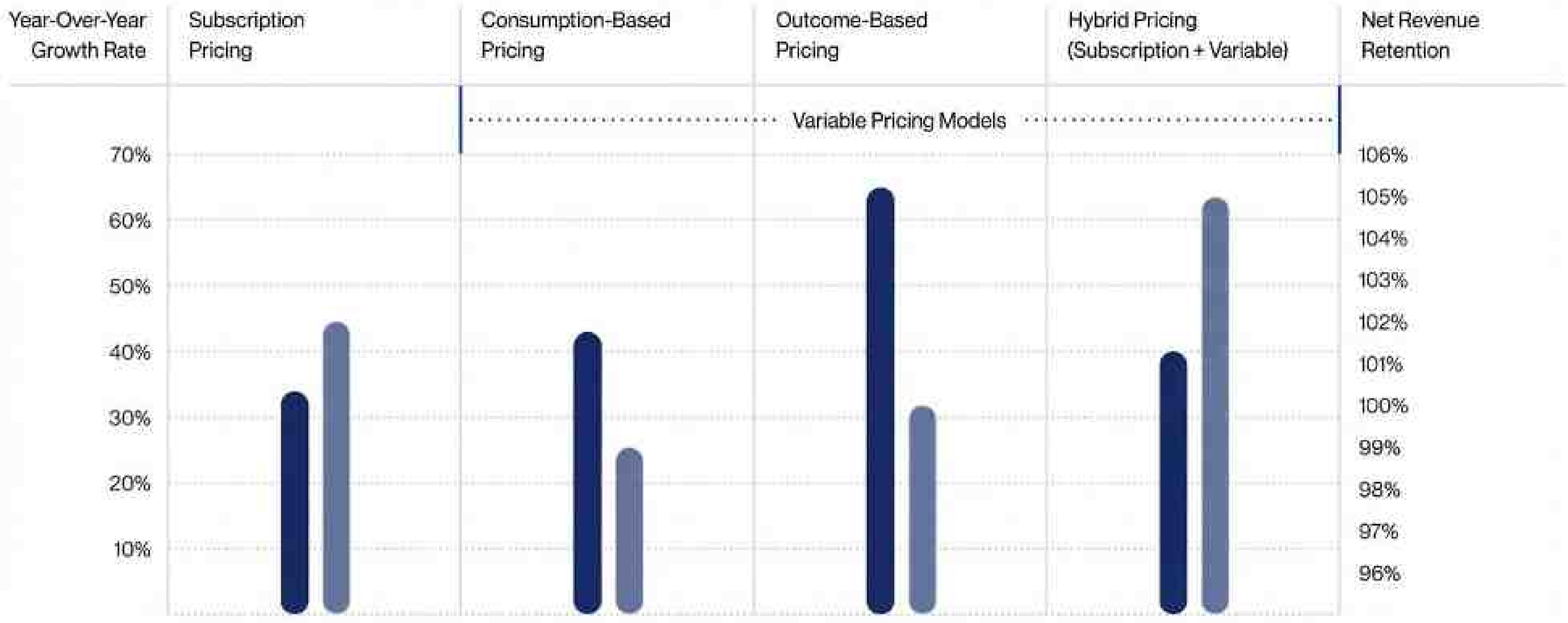


Source: High Alpha

Overall Pricing Model Has Outsized Impact on Growth and Retention

Subscription pricing models have historically dominated SaaS pricing, but recently there's been a shift to more variable pricing models like consumption and outcome-based pricing. These variable pricing models are helping drive faster growth among respondents. Hybrid pricing models show the strongest net revenue retention, likely because they provide the best combination of stability and stickiness in the subscription component while monetizing the value that comes from increased usage via the variable component.

Primary Pricing Model By Performance Metrics

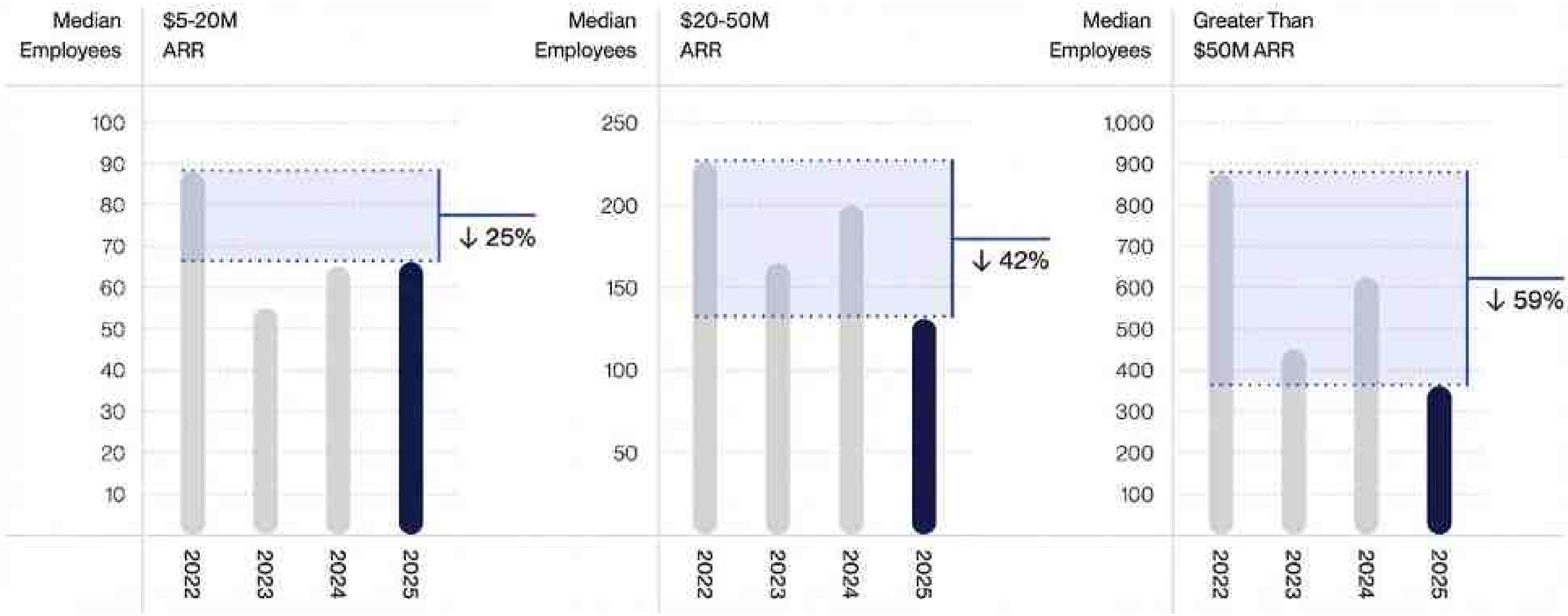


Source: High Alpha

Later-Stage Companies Are Operating With Leaner Teams

Companies beyond $5 million ARR have experienced a substantial decrease in employee count over the past four years. These changes make the businesses more resilient and efficient, requiring less outside capital to sustain operations.

Median Employees By ARR Band

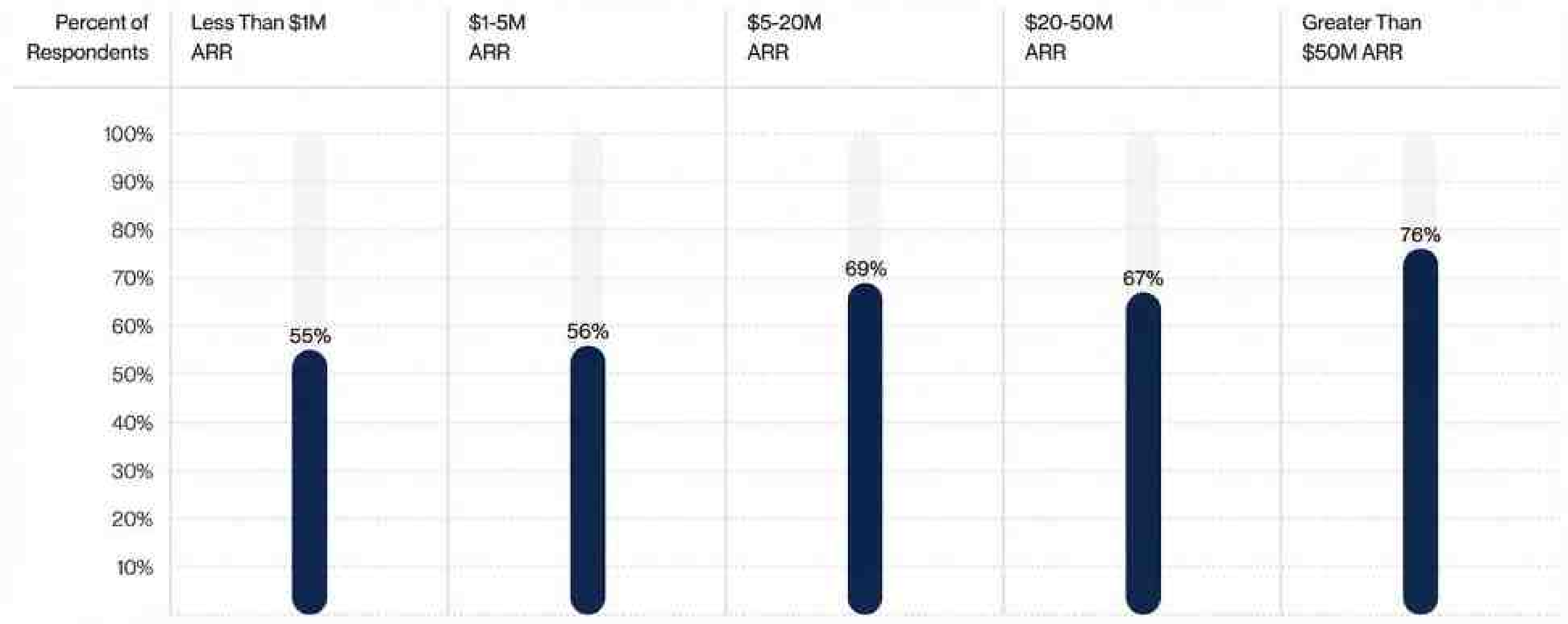


Source: High Alpha

Over Half of Companies Confirm AI Spurred Headcount Reduction

Larger companies indicated they had reduced headcount due to AI at a higher rate than smaller companies. Larger organizations are converting AI productivity gains into real efficiency while smaller ones were constructed with AI productivity baked in.

Percent Of Companies That Reduced Headcount Due To AI

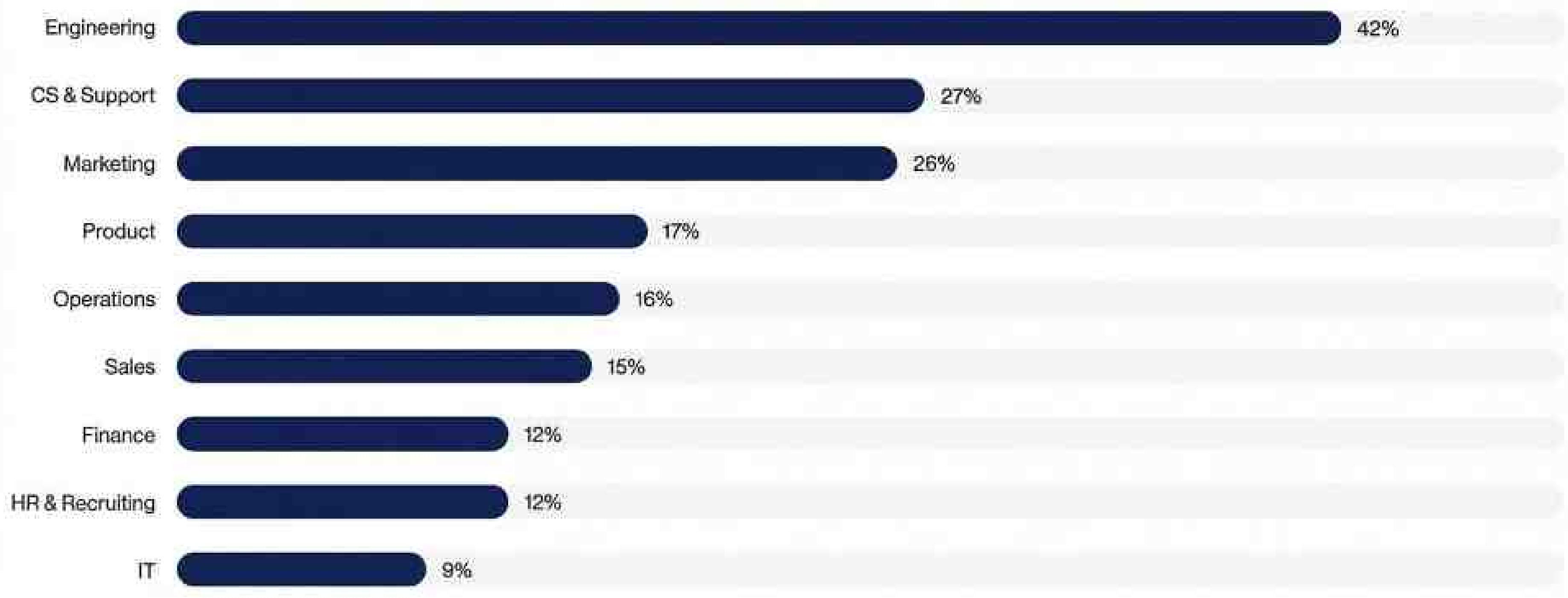


Source: High Alpha

Engineering Tops the List of AI-Driven Headcount Reductions

Engineering headcount reductions due to AI are significantly higher than any other department. Back office functions like Finance, HR, and IT have been less impacted by reductions at this point. Early AI products focused on coding, customer support, and content generation likely contribute significantly to driving headcount reductions in the top three functional areas: engineering, customer success & support, and marketing.

Percent Of Companies Reducing Headcount Due To AI By Functional Area



Source: High Alpha

Section 7: Scaling Laws & Models

Original scaling = pretraining compute

Generational

OpenAI's 2020 paper introduced scaling laws, showing that model performance improves as dataset size and model size increase. Pretraining compute effectively captures this scaling behavior since compute is determined by data, model size, and training strategy. The study tested dataset sizes from 22M to 22B tokens (a 1,000x increase) and model sizes from 1K to 1B parameters (a 1,000,000x increase), demonstrating that the relationship between compute and performance holds empirically.

The original scaling laws charts below illustrate this: test loss, which represents inverted model performance, improves as compute scales (lower test loss is better). A simple rule of thumb for compute-optimal training runs:

- *Compute Flops = 6 * Data tokens * Model parameters*
- *Model parameters = 2.7 * Data tokens*

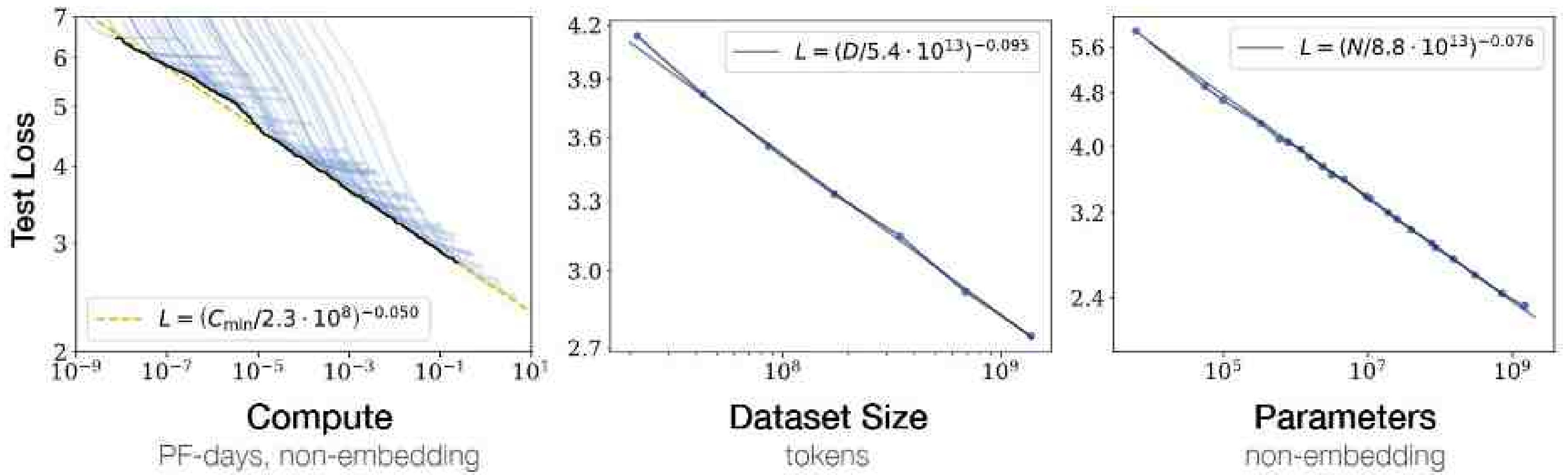


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Source: *Scaling Laws for Neural Language Models*

Economic benchmarks guide ROI

Generational

For years, AI progress was measured by academic tests—math olympiads, coding challenges, and reasoning puzzles. But these benchmarks often disconnected from what actually drives the economy. In 2025, a new wave of evaluations emerged from Mercor, OpenAI, and Scale AI, each asking a more practical question: Can AI do economically valuable work?

These benchmarks ground the AI hype in real data. Rather than testing abstract intelligence, they measure whether AI can produce deliverables that professionals get paid to create—and at what quality. The results offer a clearer picture of AI's true ROI and its near-term impact on the workforce: significant augmentation potential

Research Institution	MERCOR	OpenAI	scale
Benchmark	**AI Productivity Index**	**GDPval**	**Remote Labor Index**
Research Question	Can AI do the job of an elite junior professional?	Can AI do tasks across the broad economy?	Can AI fully automate freelance work?
Test	400 tasks from investment bankers (Goldman), consultants (McKinsey), lawyers, and doctors. Real deliverables: financial models, legal memos, patient diagnoses. Each takes 1-8 hrs for a human.	1,320 tasks across 44 occupations (software devs, nurses, engineers, accountants) in 9 GDP-driving sectors representing $3T in U.S. wages. Built by professionals with 14 yrs avg experience.	240 real freelance projects from Upwork across 23 domains (game dev, architecture, video editing). Total: $144K in value, 6,000+ human hours. AI must complete the entire project end-to-end with no human help.
Metric	% of expert grading criteria satisfied (29 criteria avg per task). 100% = "analyst in a box."	Win rate: how often expert graders prefer AI output over human output in blind comparison.	Automation rate: % of projects AI completed to client-acceptable quality.
Result	GPT-5: 64% (Law: 70%, Banking: 60%). Up from GPT-4o's 36% one year earlier. No model production-ready.	Claude Opus 4.1: 49% win rate vs. human experts. GPT-5: 41%. AI completes tasks ~100x faster/cheaper, but needs oversight.	Manus: 2.5%. Of $144K in project value, top AI earned just $1,720.
Takeaway	Rapid progress on elite professional tasks, but still not production-ready without human oversight.	AI is faster and cheaper than humans, but still needs a human in the loop.	Full automation is nowhere close. 97.5% of real freelance work still needs humans.

Source: Mercor, OpenAI, Scale AI

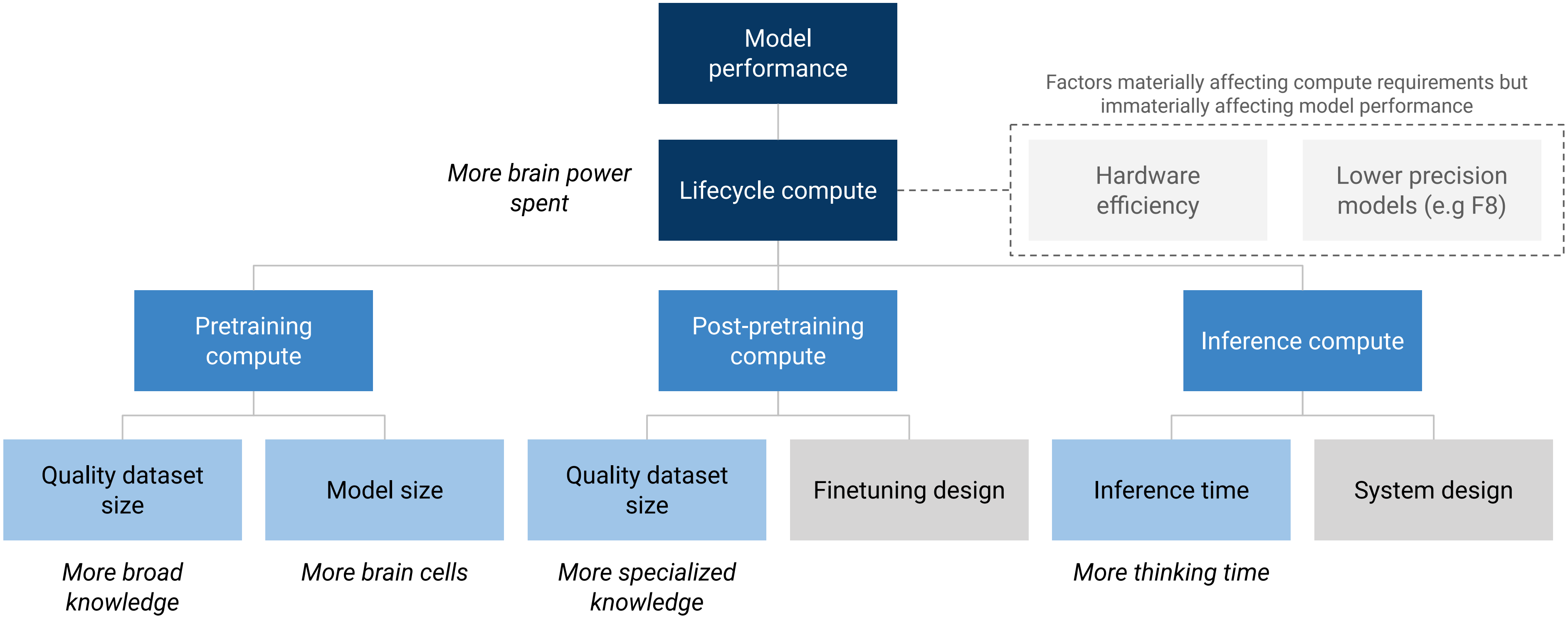
Scaling today = lifecycle compute

Generational

The original scaling law paper was published in 2020 when models were much simpler. At the time, instruction-tuned GPT models didn't exist in production, let alone RAG (retrieval-augmented generation) or agentic systems. Today, there are far more factors—"knobs"—that influence compute, from inference time optimizations in reasoning models to the design of self-reflecting agents.The diagram below provides a simplified view of scaling laws today. Blue boxes represent knobs controlled by foundation model providers, while gray boxes are determined by downstream developers based on their system design. I won't delve into the latter since it depends on specific use cases.

Discussions about the scaling "wall" can be confusing because terms are often used inconsistently. For example:

- **What is model performance?** In the original paper, performance was measured by how well the model predicted the next word. Today, it's about how well the model performs across a range of practical, real-world tasks.
- **Plateauing vs. supply constraints:** When people talk about limits, are they referring to performance plateauing or constraints in resources? For instance, with data: does more data fail to improve performance, or is it simply that new training data is running out?

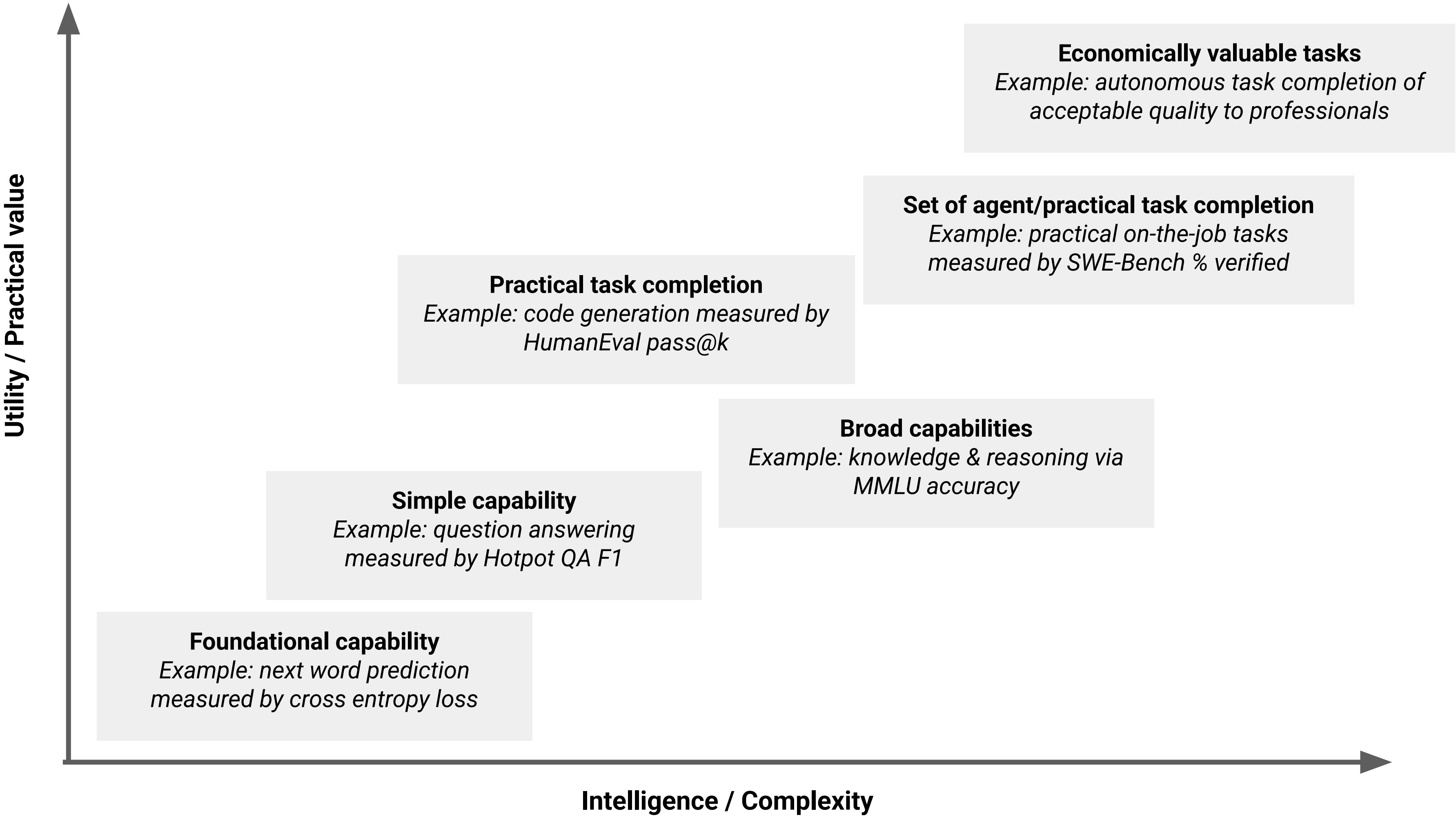


Source: Author analysis

Scaling laws – what are we measuring?

The metrics used to measure scaling laws have evolved alongside AI models. In the original neural scaling law paper, performance was measured by next-word prediction—a simple, foundational task. Today, performance metrics focus on practical utility, such as task completion, reasoning, and agentic capabilities.

The diagram illustrates this progression, with foundational capabilities at one end and broad, real-world applications at the other. While earlier metrics prioritized theoretical benchmarks, the focus has shifted to how effectively models can deliver value in practical, real-world scenarios.



Economically valuable tasks
Example: autonomous task completion of acceptable quality to professionals

Set of agent/practical task completion
Example: practical on-the-job tasks measured by SWE-Bench % verified

Practical task completion
Example: code generation measured by HumanEval pass@k

Broad capabilities
Example: knowledge & reasoning via MMLU accuracy

Simple capability
Example: question answering measured by Hotpot QA F1

Foundational capability
Example: next word prediction measured by cross entropy loss

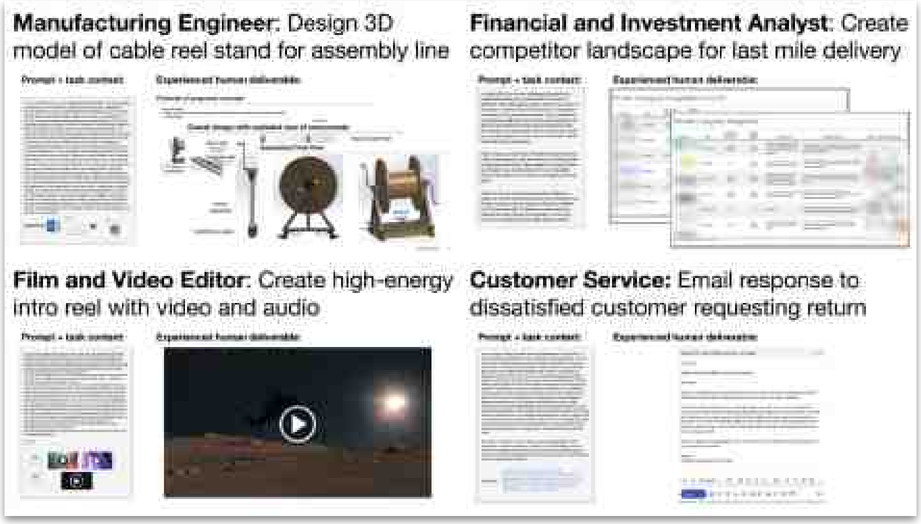
Source: Author analysis

AI matches humans on real work output

Generational

GDPVal benchmarks AI against human experts with an average of 14 years of experience. Tasks (see right) include 3D CAD modeling for manufacturing, financial competitor analysis, video editing, and customer service responses. GPT-5.2 matches or beats the human experts 70.9% of the time.

This benchmark tests whether AI can produce professional-grade work products, not just assist with them. A model generates a complete 3D assembly fixture design or financial analysis deck, and evaluators compare it to what an experienced professional produced. The results show AI producing equivalent output for significantly lower cost and faster turnaround on these specific task types.



Manufacturing Engineer: Design 3D model of cable reel stand for assembly line

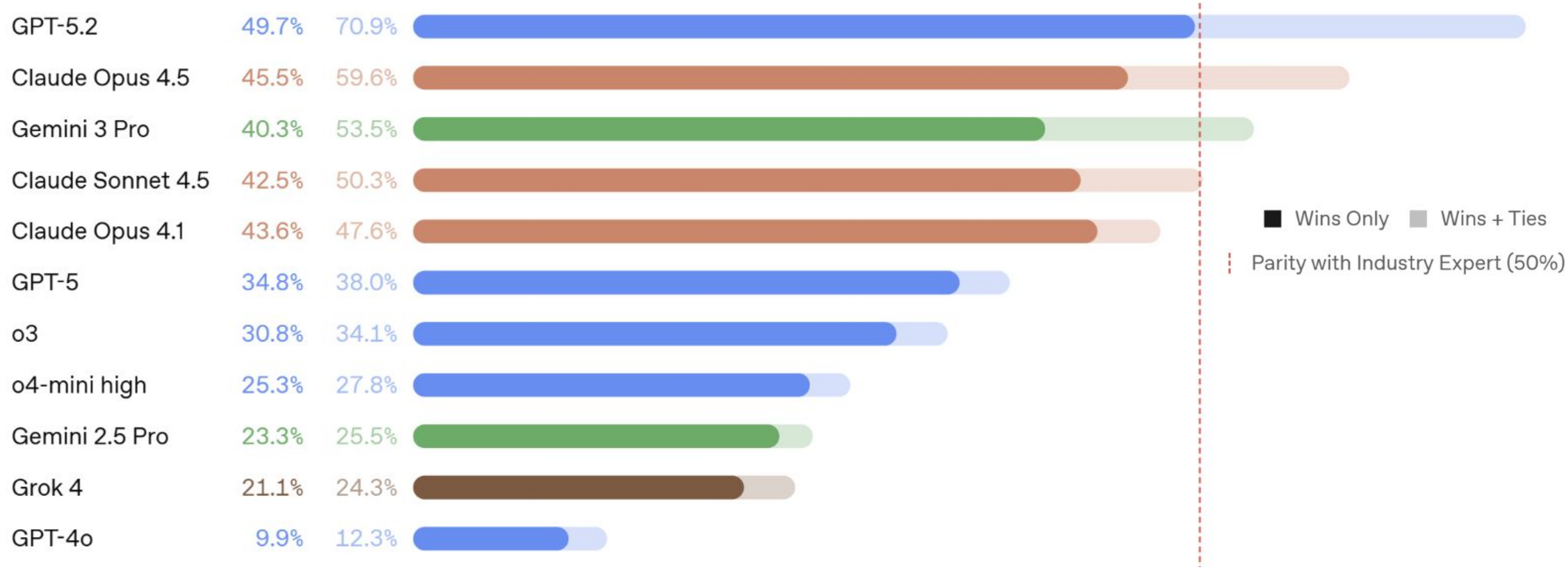
Financial and Investment Analyst: Create competitor landscape for last mile delivery

Film and Video Editor: Create high-energy intro reel with video and audio

Customer Service: Email response to dissatisfied customer requesting return

Prompt + task context
Experienced human deliverable
Prompt + task context
Experienced human deliverable
Prompt + task context
Experienced human deliverable
Prompt + task context
Experienced human deliverable

GDPVal Leaderboard



Source: OpenAI

Experts are still needed to train models

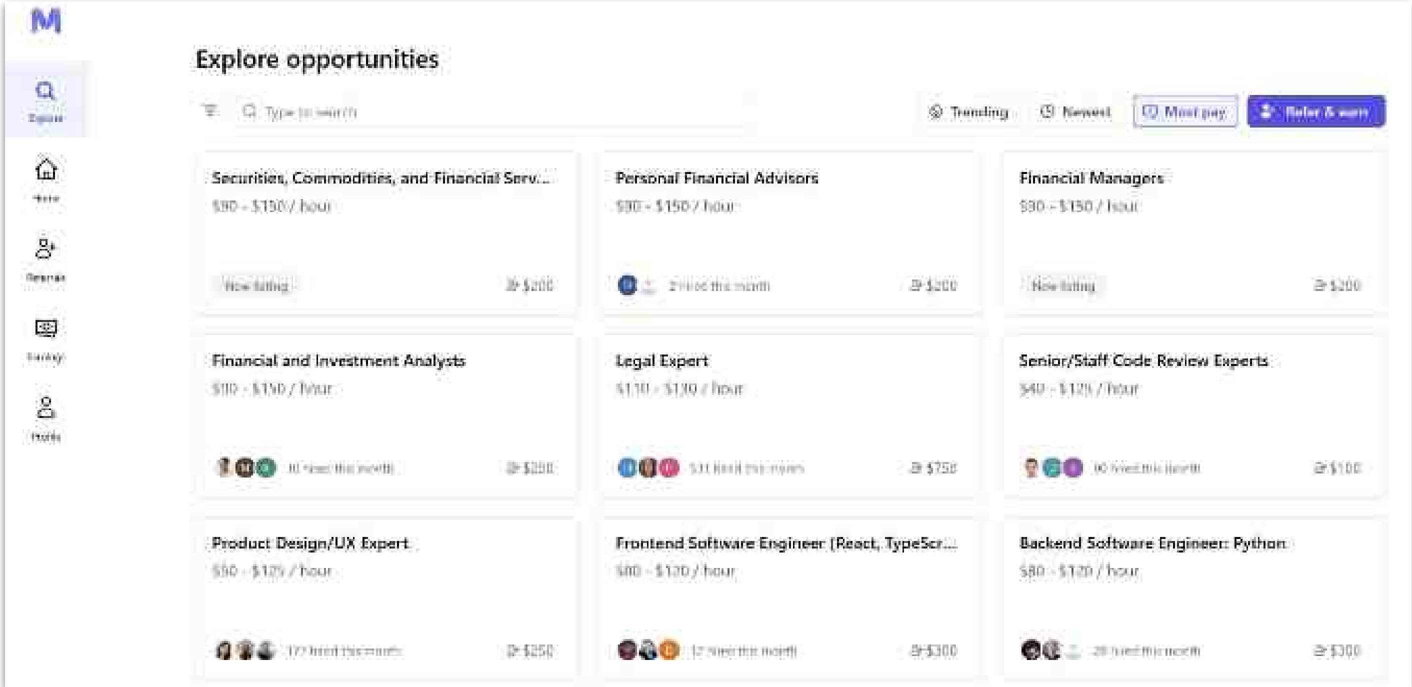
Generational

AI model training has evolved from relying on general crowdsourced data to requiring specialized expert knowledge. Post-training techniques—particularly reinforcement learning from human feedback (RLHF) and supervised fine-tuning—now demand domain-specific expertise that average data annotators cannot provide. Expert data costs $5-$20 per preference point versus $0.01 for synthetic data, but frontier labs still depend on human experts for critical post-training stages.

How Expert Data Works:

- Task Creation: Experts design problems that challenge AI models in their domain (e.g., PhD mathematicians create competition-level math problems, medical researchers design clinical reasoning scenarios)
- Solution Generation: Experts provide detailed, step-by-step solutions demonstrating expert-level reasoning—not just final answers but the thought process
- Evaluation & Ranking: Experts evaluate AI-generated responses for correctness, approach quality, and domain accuracy—teaching models to distinguish good from mediocre reasoning
- Rubric Development: Experts create assessment frameworks for what constitutes high-quality outputs in their domain

As Nathan Lambert (AI researcher) notes: "While all frontier labs still rely on human data for parts of their post-training pipeline, AI can be substituted at most stages"—but for breakthrough capabilities, expert knowledge remains irreplaceable.



Explore opportunities

Securities, Commodities, and Financial Serv...
$90 - $150/ hour
New listing ⇒ $200

Personal Financial Advisors
$90 - $150 / hour
2 hired this month ⇒ $200

Financial Managers
$30 - $150 / hour
New listing ⇒ $200

Financial and Investment Analysts
$110 - $150 / hour
10 hired this month ⇒ $200

Legal Expert
$110 - $130 / hour
531 hired this month ⇒ $750

Senior/Staff Code Review Experts
$40 - $125 / hour
90 hired this month ⇒ $100

Product Design/UX Expert
$50 - $125 / hour
172 hired this month ⇒ $250

Frontend Software Engineer (React, TypeScr...
$80 - $120 / hour
12 hired this month ⇒ $300

Backend Software Engineer: Python
$80 - $120 / hour
28 hired this month ⇒ $300

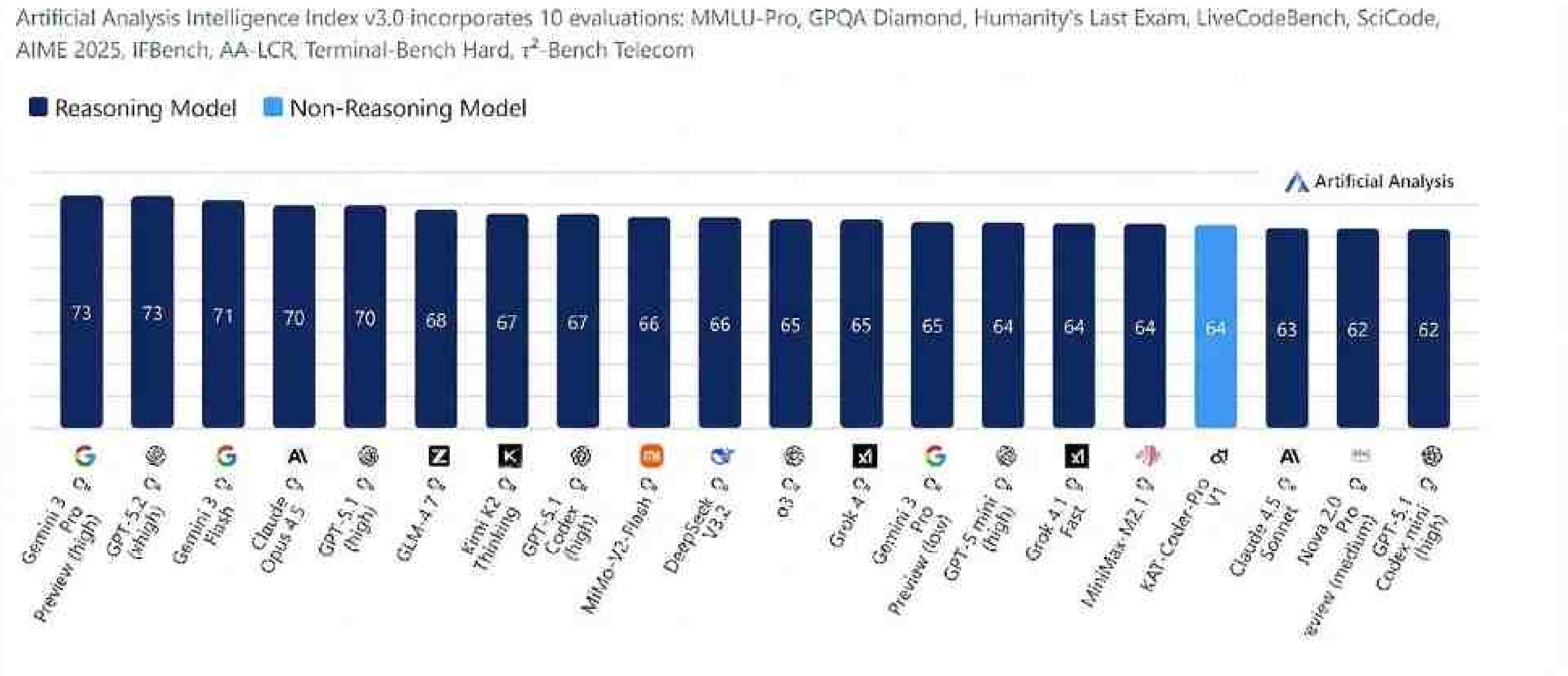
Source: Mercor

Reasoning models dominate

Generational

Reasoning models mark a step change in AI architecture—the shift from instant pattern-matching to deliberative problem-solving. Before o1's September 2024 launch, models generated responses immediately; reasoning models spend compute "thinking," testing approaches, and self-correcting before answering. This unlocked PhD-level science, competition mathematics, and complex multi-step tasks previously impossible. The industry converged within 15 months—every major provider adopted reasoning. The chart shows the result: 19 of top 20 models use reasoning architecture (purple) while traditional models (blue) are largely displaced. The adoption speed signals the industry recognized this as fundamental infrastructure, not incremental improvement.

Artificial Analysis Intelligence Index by Model Type



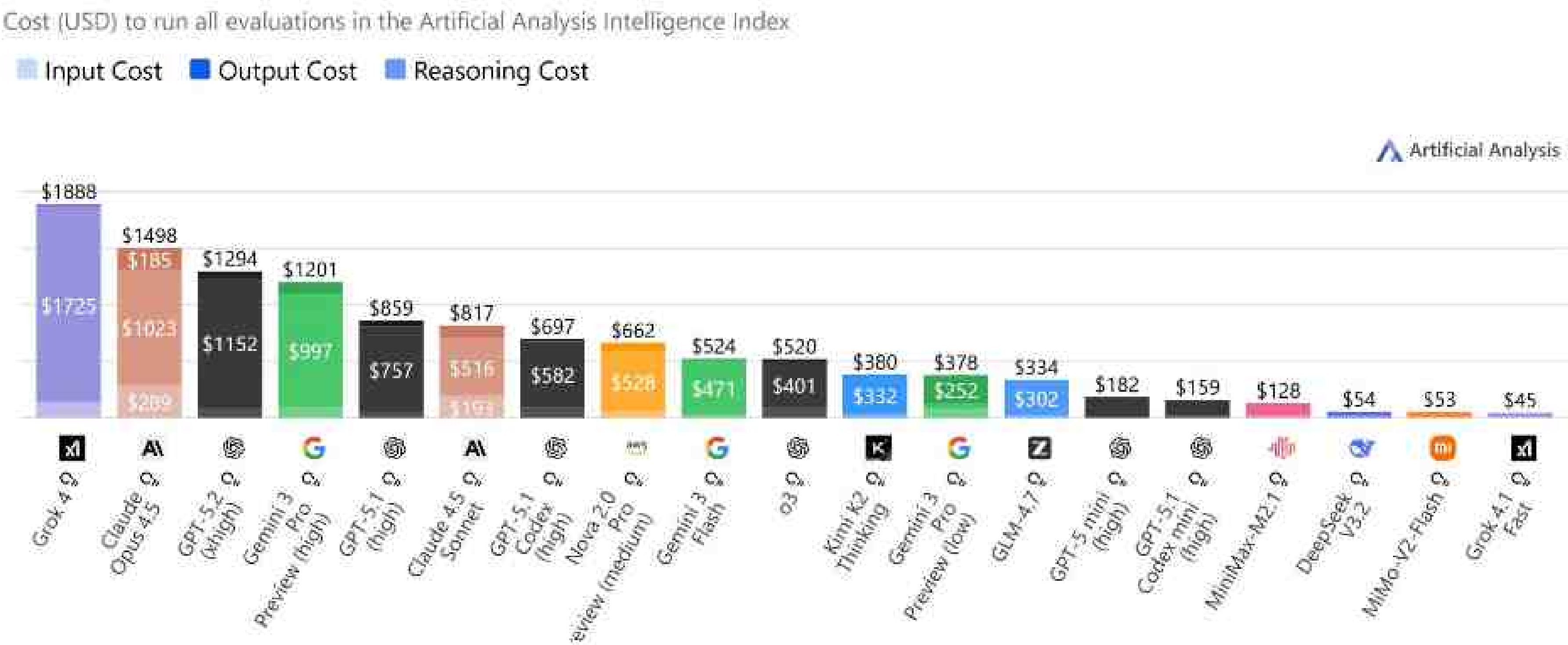
Source: Artificial Analysis, Author analysis

Intelligence comes at a cost

Generational

Reasoning models reversed the AI cost curve. Per-token prices fell 1,000x since GPT-3, but reasoning multiplies total costs 4-6x through hidden "thinking" tokens billed as expensive output. At maximum settings, 80% of tokens are invisible reasoning work before the final answer. The chart illustrates this economic reality: premium reasoning models (Grok 4, Claude 4.5, GPT-5) cost $1,200-$1,900 to evaluate versus under $100 for lighter models. This reshapes AI economics—organizations now optimize for "cost per solved problem" rather than "cost per token." The capability unlocks tasks impossible with traditional models, but the infrastructure footprint is fundamentally different: reasoning can require 100x more compute than single-pass inference.

Cost to Run Artificial Intelligence Index

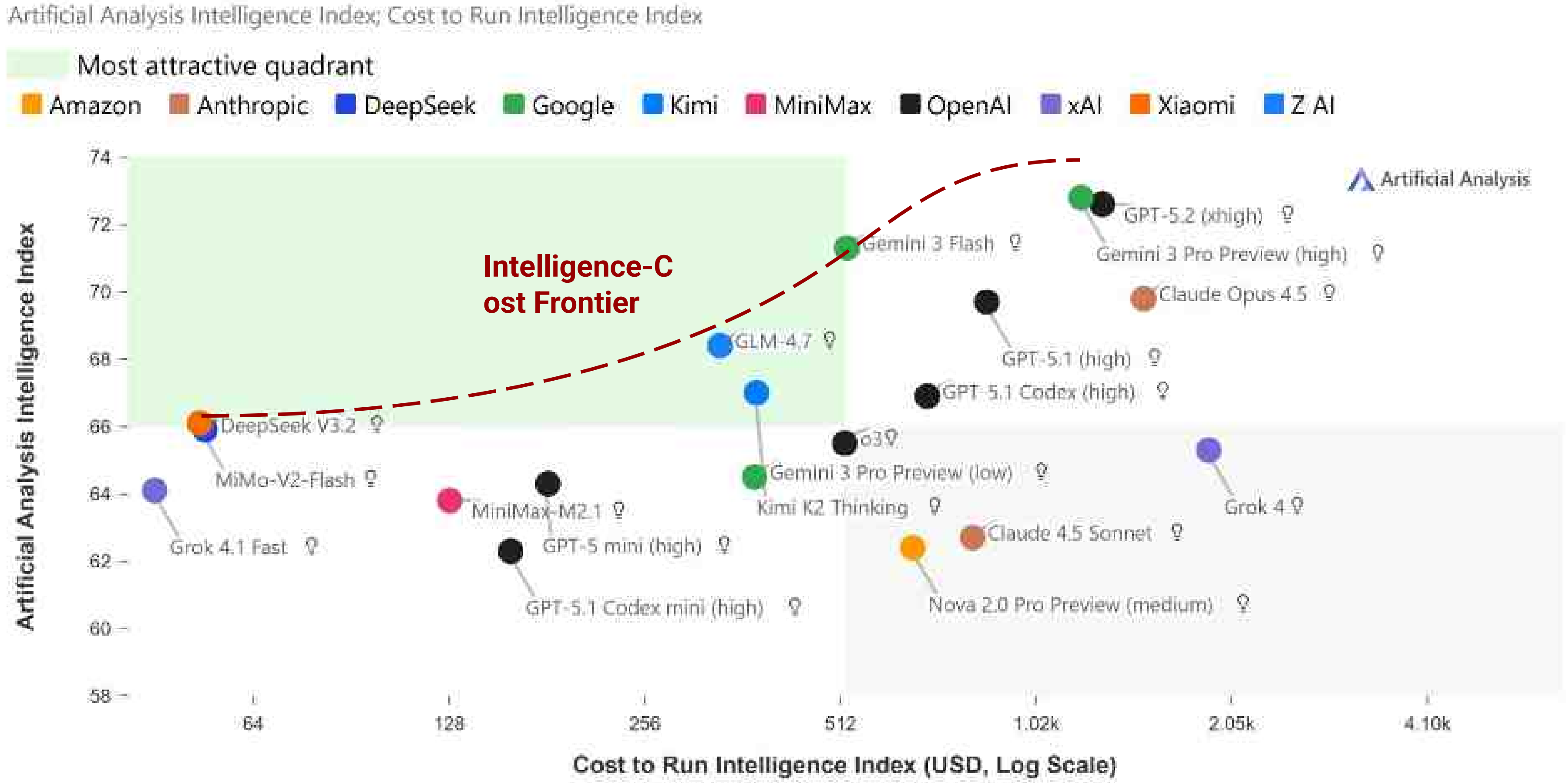


Source: Artificial Analysis

Not all intelligence cost equal

The 40x cost spread at similar intelligence levels shows architectural differences in how models perform reasoning. Grok 4 costs $1,900 to run benchmarks while DeepSeek costs $54 for comparable performance. Dense models activate all parameters for every token; sparse models like DeepSeek activate only a subset (37 billion of 671 billion parameters). The relationship between task complexity and compute consumption varies by architecture. A query might trigger thousands of reasoning tokens on one model and hundreds on another, even when both reach the same answer. This makes cost prediction difficult—you can't determine how much a model will spend on reasoning just by looking at the task. Developers should test multiple models on their specific workloads to find actual costs.

Intelligence vs Cost to Run Artificial Analysis Intelligence Index



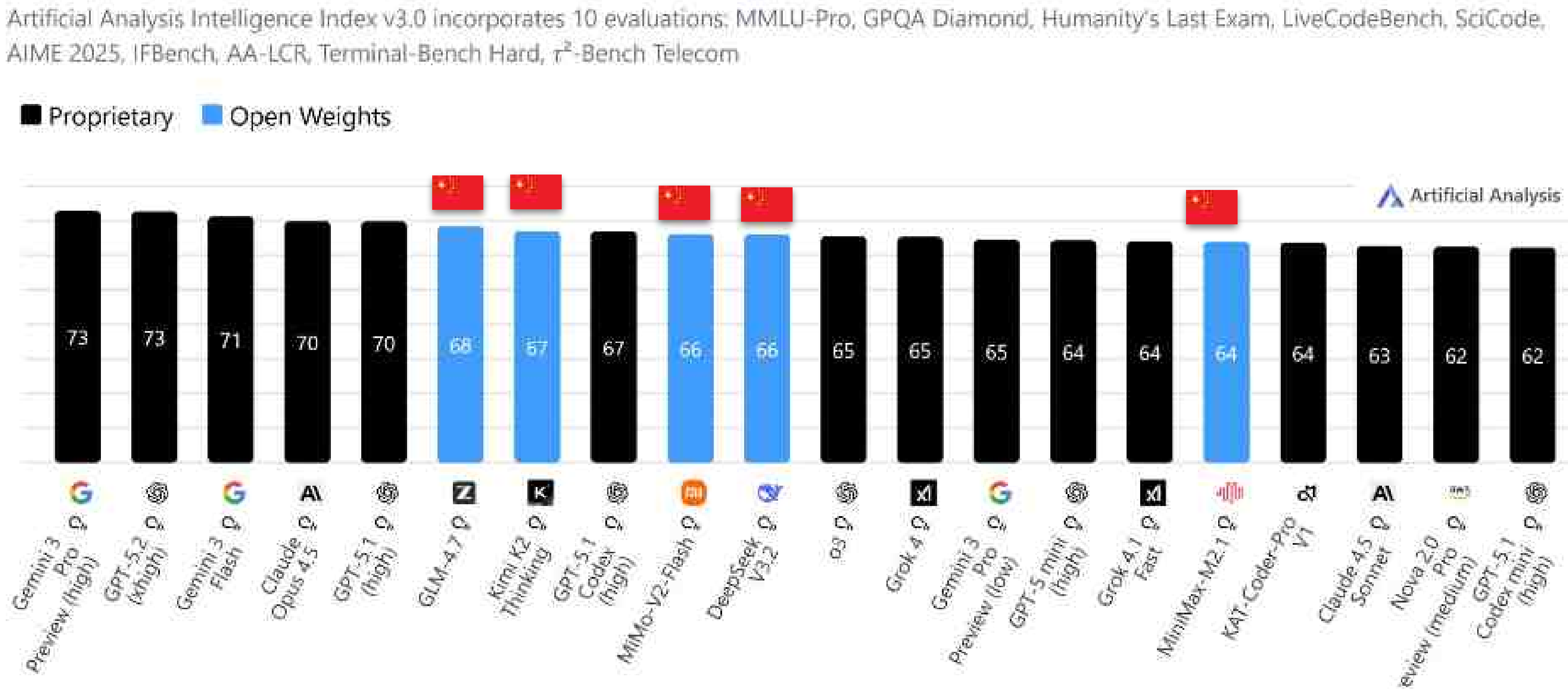
Source: Artificial Analysis

China is leading open source

All five of the top open-source models are Chinese: GLM 4.7, Kimi K2 Thinking, DeepSeek V3.2, Mimo-V2, and MiniMax M2.1. These models achieve frontier intelligence at the lowest cost—connecting to the previous slide's efficiency story. When open-source models reach Grok-4 performance and release under MIT license, they make frontier capability available at near-zero marginal cost, challenging the pricing models of proprietary alternatives.

China now leads global open-source AI model releases. In 2024, China published 23,695 AI papers, exceeding the combined output of the US, UK, and EU. China filed 35,423 AI patents, 13 times more than the US, UK, Canada, Japan, and South Korea combined. This reflects sustained investment in AI research infrastructure since 2017.

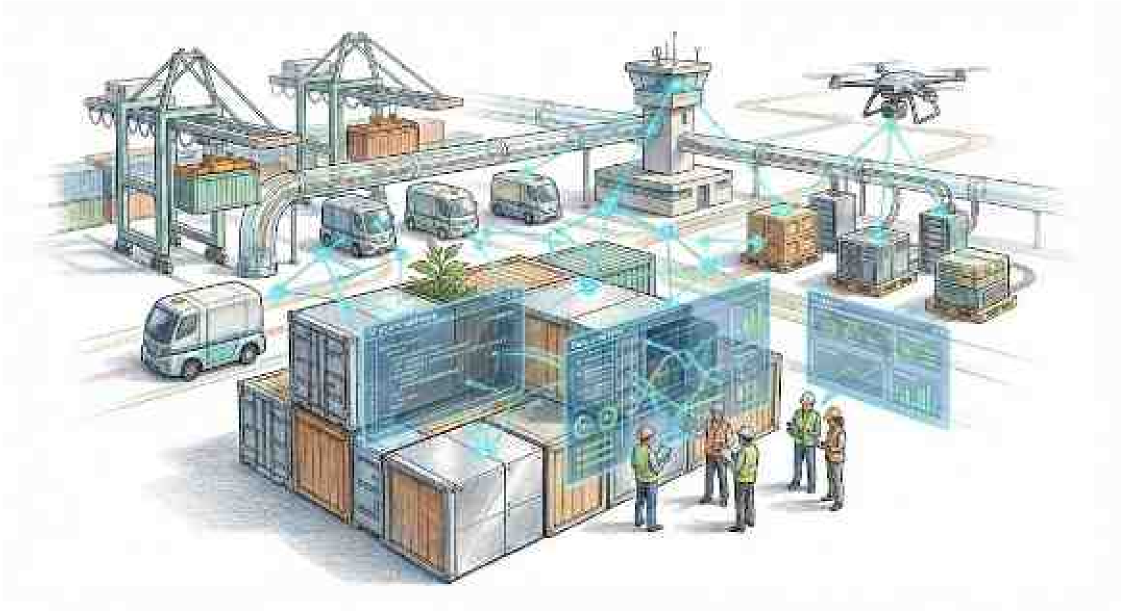
Artificial Intelligence Index by Open Weights vs Proprietary



Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom

Source: Artificial Analysis, Author analysis

Section 8: Products & Protocols



AI Crosses the Chasm in 2025

Generational

The three-year arc is clear. 2023 focused on enterprise experimentation with copilots and enterprise tools. 2024 made AI accessible to consumers through ChatGPT, Gemini, and Claude reaching mainstream audiences. 2025 was when AI crossed into production reliability. We entered the year with reasoning models barely a month old—o1 launched in September 2024. By year's end, AI had evolved from answering questions to autonomously completing 30+ hour tasks, writing production code at scale ($1B revenue in 6 months for Claude Code), and generating videos that topped app stores in 72 hours. Three companies raced through monthly releases: frontier models (GPT-5, Gemini 3, Claude 4.5 series), viral products (Sora 2, Nano Banana), and autonomous capabilities (computer use, agentic coding, browser control). ChatGPT reached 900M weekly users—evidence that AI had crossed from experimental tool to production infrastructure.

2025	Q1	Q2	Q3	Q4
OpenAI *ChatGPT has 900M WAU as of Dec*	GPT-4.5 Operator agent	o3 and o4-mini Codex	GPT-5 gpt-oss ChatGPT agent Sora app ChatGPT shopping ChatGPT Pulse	GPT-5.1 GPT-5.2 Group ChatGPT chats ChatGPT Atlas browser ChatGPT Images 1.5
Google *Gemini has 650M MAU as of Oct*	Gemini 2.5 Pro	Veo 2 Flow video editing	Gemini 2.5 Deep Think Genie 3 world model Nano Banana image	Gemini 3.0 Pro GenTabs Veo 3.1 video model Gemini Agent Gemini in Google Home and Android Auto
ANTHROP\C *Claude Code, GA'd in Apr and reached $1B run rate in Dec*	Claude 3.7 series Claude Code	Claude 4 series	Claude 4.5 series Claude for Chrome	

Frontier open weights model from the US

Viral video generation app

Viral image generation model

Personalized browser UX

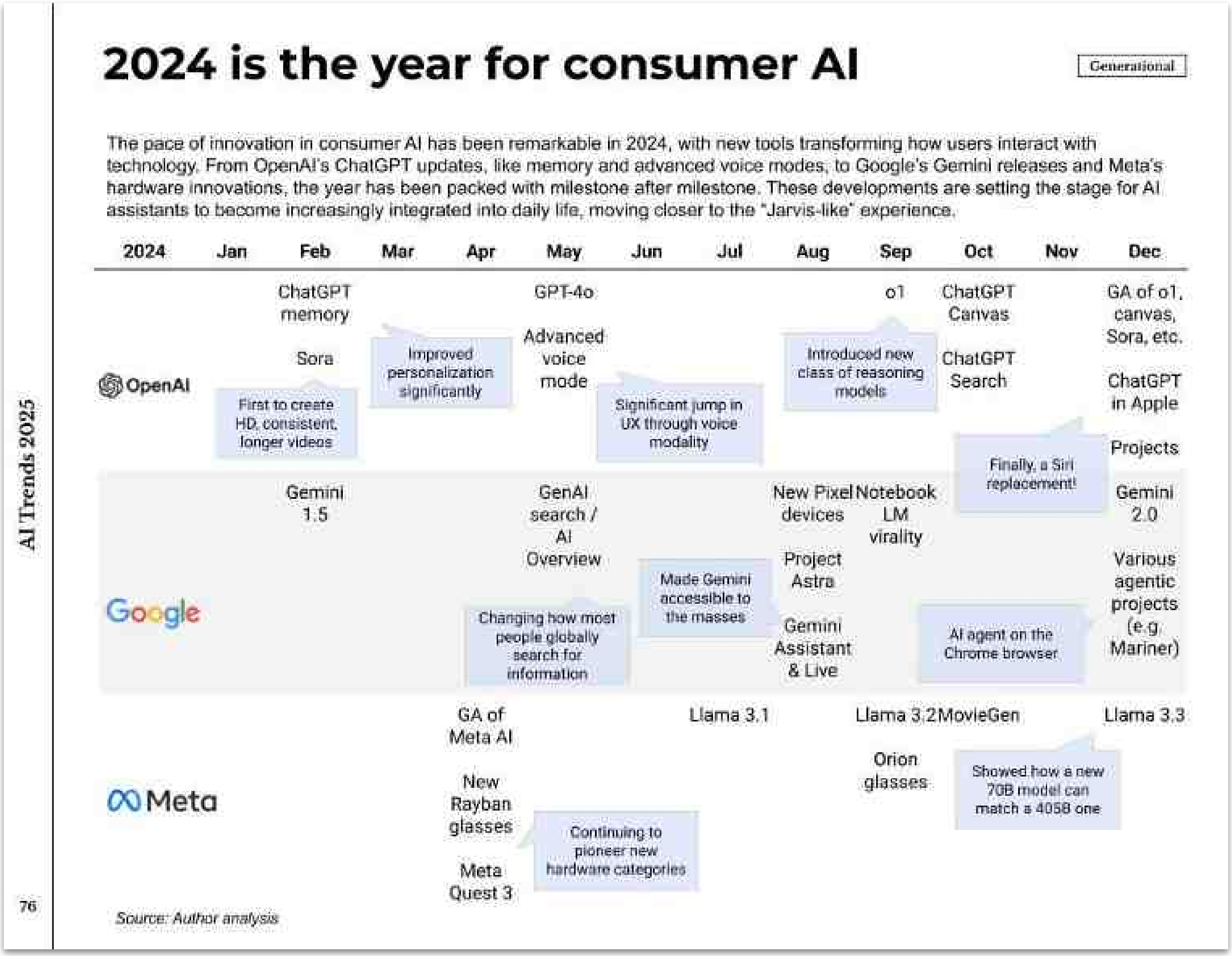
First truly autonomous agentic product

Source: Author analysis

2025: AI shifted from useful to reliable

Generational

AI in 2024 had the vibe of being useful and fun (see last year's report). In 2025, that perception shifted to reliability—companies began trusting AI to execute complete jobs and replace roles within 12-24 months. The timeline below shows only the most significant developments for readability, but 2025 actually saw roughly three times as many product releases as 2024. The acceleration is clearest in model iterations: companies that released 1-3 models at the 9-figure training cost in 2024 released 3-5 models in 2025 at the 10-figure level.



2024 is the year for consumer AI

Generational

The pace of innovation in consumer AI has been remarkable in 2024, with new tools transforming how users interact with technology. From OpenAI's ChatGPT updates, like memory and advanced voice modes, to Google's Gemini releases and Meta's hardware innovations, the year has been packed with milestone after milestone. These developments are setting the stage for AI assistants to become increasingly integrated into daily life, moving closer to the "Jarvis-like" experience.

2024	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
OpenAI		ChatGPT memory Sora			GPT-4o Advanced voice mode				o1	ChatGPT Canvas ChatGPT Search		GA of o1, canvas, Sora, etc. ChatGPT in Apple Projects
Google		Gemini 1.5			GenAI search / AI Overview			New Pixel devices Project Astra Gemini Assistant & Live	Notebook LM virality			Gemini 2.0 Various agentic projects (e.g. Mariner)
Meta				GA of Meta AI New Rayban glasses Meta Quest 3			Llama 3.1		Llama 3.2 Orion glasses	MovieGen		Llama 3.3

First to create HD, consistent, longer videos

Improved personalization significantly

Significant jump in UX through voice modality

Introduced new class of reasoning models

Finally, a Siri replacement!

Changing how most people globally search for information

Made Gemini accessible to the masses

AI agent on the Chrome browser

Continuing to pioneer new hardware categories

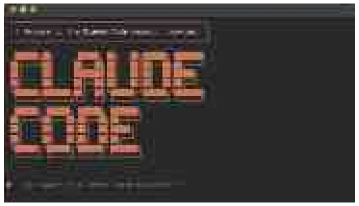
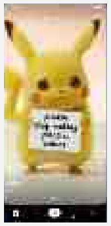
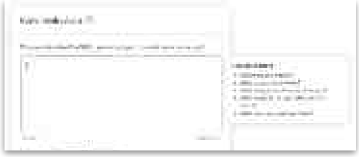
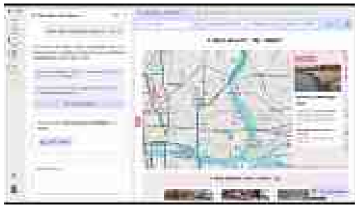
Showed how a new 70B model can match a 405B one

Source: Author analysis

Increasingly powerful user experience

Generational

As AI models improve, so do the ways we interact with them. Copilots have gone from simple chat interfaces in 2022 to handling actions and running programs by 2024. They now remember conversations, adapt to user preferences, and support multimodal inputs like text, visuals, and more. This shift highlights how personalization and real-time interaction are making AI feel more intuitive and integrated into daily workflows.

UX theme	2022	2023	2024	2025
Capabilities	Chat (Claude) 	Answer (Claude w/ Search) 	Action (Claude Artifacts) 	Autonomy (Claude Code) 
Modality	Text-to-text (ChatGPT) 	Text-to-image (ChatGPT Dall-E) 	2D multimodal (ChatGPT Advanced Voice mode) 	3D multimodal (Sora) 
Personalization	None (ChatGPT) 	Custom instructions (ChatGPT Instructions) 	Memories (ChatGPT Memories) 	Personalized UX (Google Gentabs) 

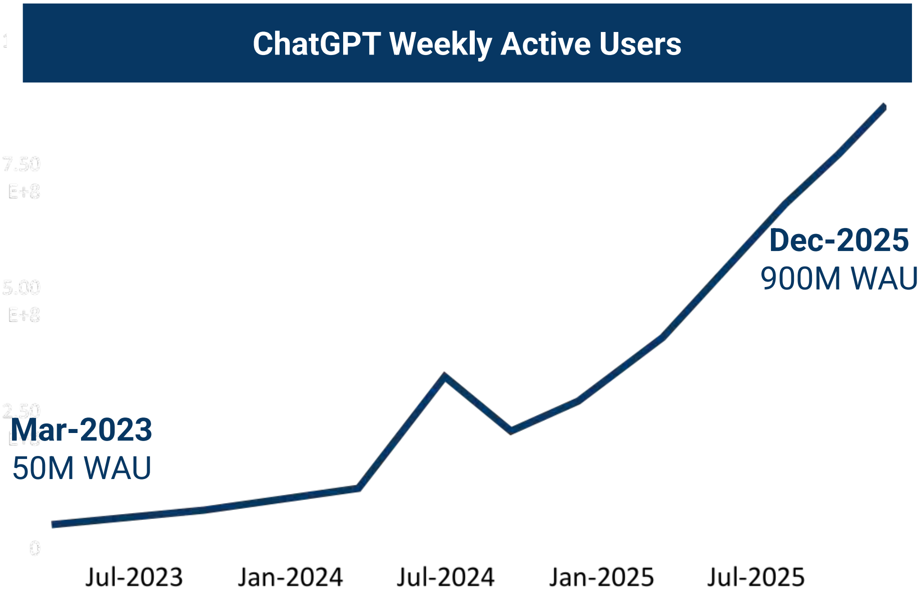
Source: Company websites, media, Author analysis

Consumer AI Product of 2025 – ChatGPT at >900M WAU and >$12B ARR

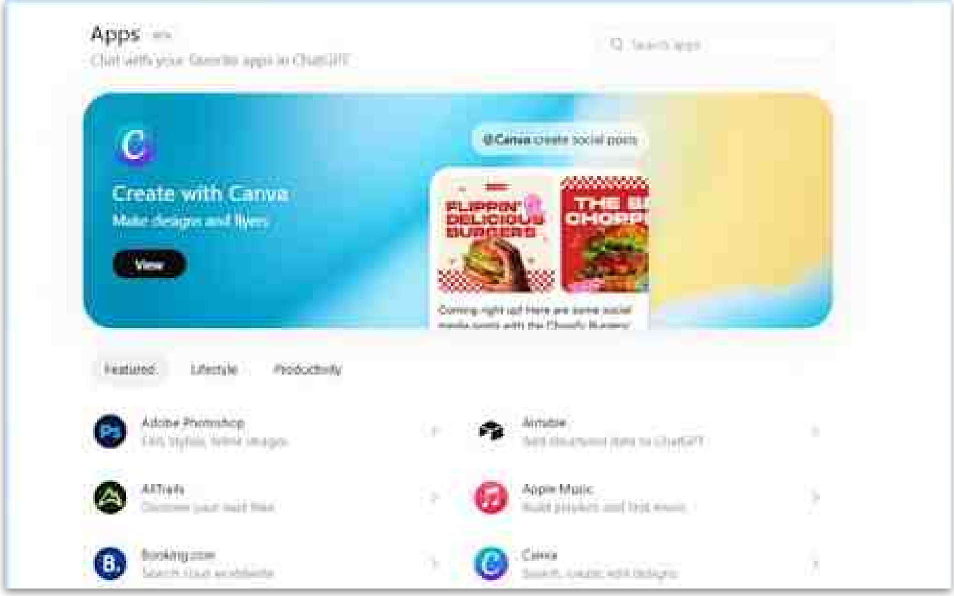
Generational

ChatGPT became the default brand for AI in 2025, reaching $12 billion ARR by July and 900 million weekly active users by December through relentless product velocity. As one Google employee commented at an internal strategy meeting, ChatGPT is becoming synonymous to AI the same way Google is to search.

GPT-5 launched August with unified multimodal capabilities, Codex coding agent achieved mass adoption among developers for autonomous software engineering, the app store arrived December integrating Spotify and DoorDash into conversations, and the viral "Year of ChatGPT" feature delivered personalized Spotify Wrapped-style recaps. The company maintained ~80% market share as fewer than 10% of users even tried competitors.



Source: OpenAI, News

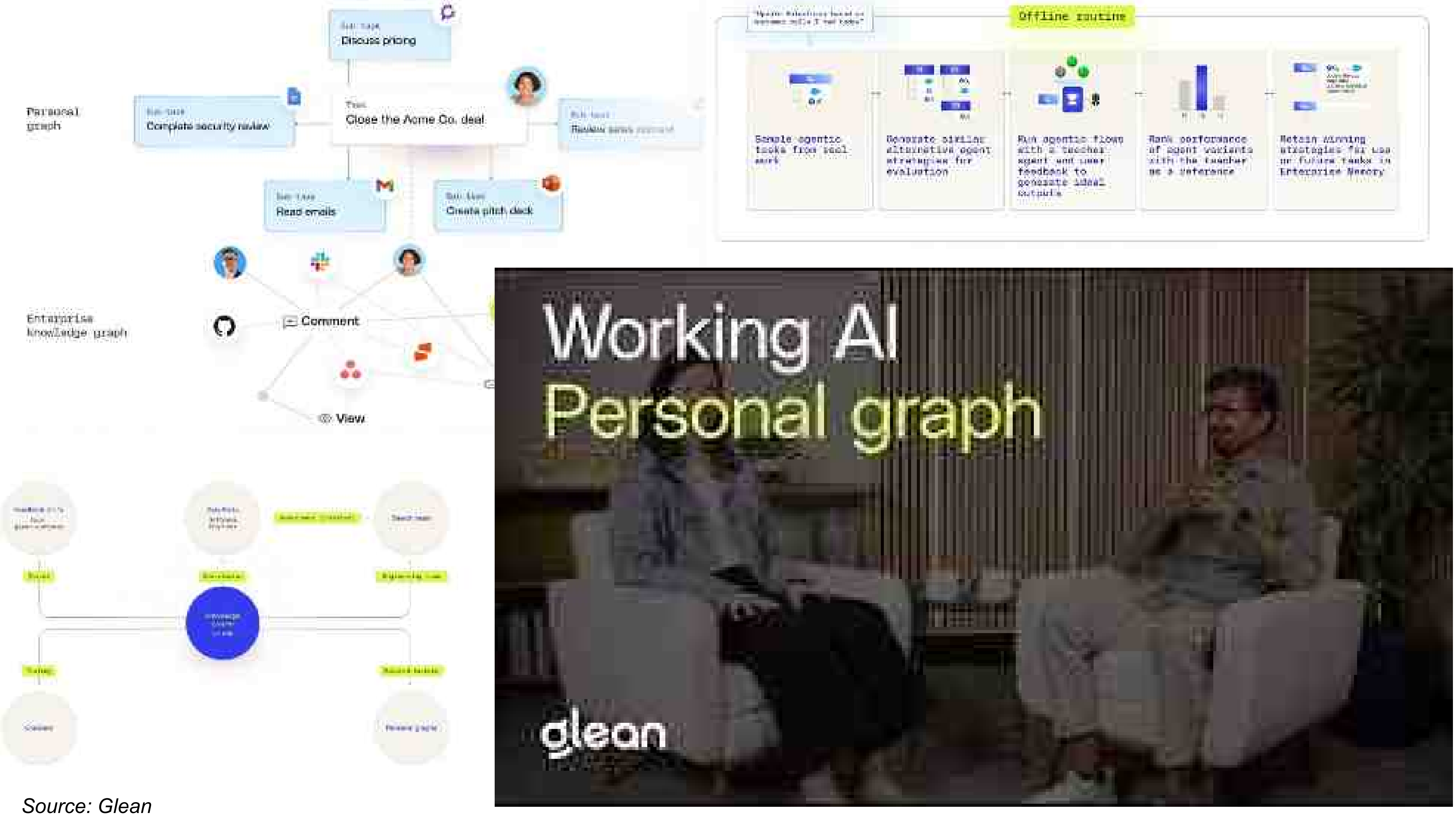


Top:ChatGPT App Store
Bottom: Fun "Year of ChatGPT" feature reflecting on your conversations

Enterprise AI Product of 2025 – Glean at $200M ARR engineering work context

Generational

Glean is the enterprise AI platform that reached $200M ARR in December 2025 by pioneering personalized AI through its dual-graph architecture. The platform combines an Enterprise Graph that understands company-wide systems, information, and workflows with a Personal Graph built for every employee that captures their projects, collaborators, and work style. This architecture enables AI that adapts to how individuals actually work rather than forcing uniform experiences. When a support agent receives a ticket, Glean doesn't just search documents—it understands which past resolutions are relevant to this specific person's role, permissions, and working patterns.


Personal graph
Sub-task
Discuss pricing
Complete security review
Task
Close the Acme Co. deal
Read emails
Create pitch deck
Enterprise knowledge graph
Comment
View
Offline routine
Sample agentic tasks from real work
Generate similar alternative agent strategies for evaluation
Run agentic flows with a teacher agent and user feedback to generate ideal outputs
Rank performance of agent variants with the teacher as a reference
Retain winning strategies for use on future tasks in Enterprise Memory
Working AI
Personal graph
glean

Source: Glean

Multimedia AI Products of 2025 – Sora & Nano Banana amplified creativity

Sora 2 and Nano Banana represent the multimedia breakthrough — video and image generation that went mainstream in 2025. Sora 2, OpenAI's video and audio generation model released in September, creates realistic clips with synchronized sound. The iOS app topped app stores within 72 hours of launch and operates like TikTok with a feed-based sharing model for AI-generated videos. Disney invested $1 billion in December to enable generation of 200+ copyrighted characters on the platform, signaling commercial viability. Nano Banana, Google's image generation model, creates professional visuals with accurate text rendering in multiple languages and resolutions up to 4K. Integrated across Google's product suite—Gemini app, Slides, Vids, NotebookLM—it became the default image generator for millions of daily users. Both products demonstrated that AI-generated multimedia moved from experimental novelty to consumer necessity, with viral adoption driving platform-level business models.

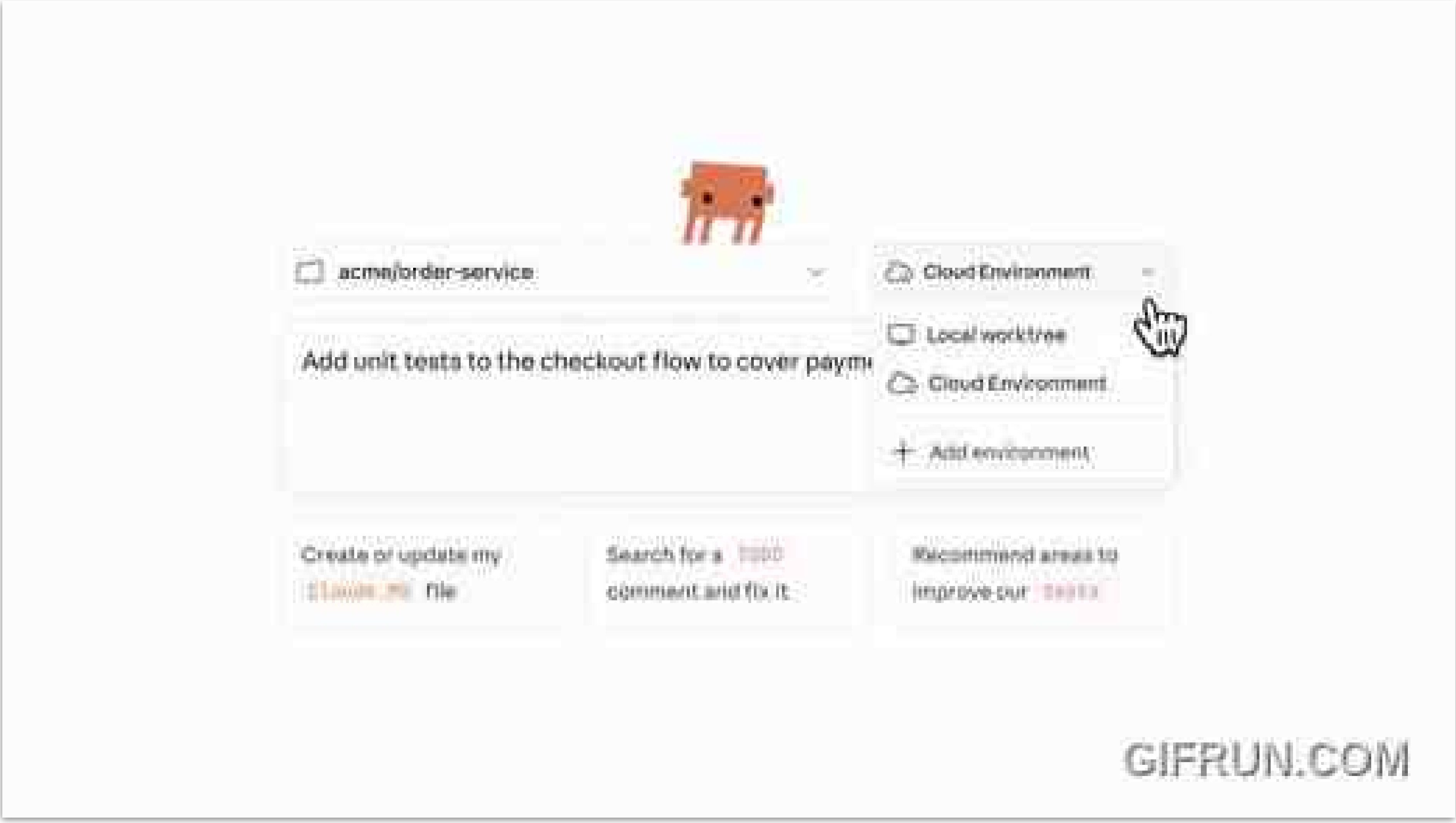
Top: Nano Banana precisely editing the same image
Bottom: various Sora generated videosinc. including Sam Altman being apprehended in Target (rightmost video)

Source: News

Agentic AI Product of 2025 – Claude Code reached $1 billion annualized rev

Generational

Claude Code operates from terminal and desktop, autonomously handling coding tasks—searching code, editing files, running tests, and committing to GitHub. The economics have shifted dramatically: features costing me $1,000 and 3 days contractor time earlier in 2025 now complete in under an hour on a $100/month plan. Individual developers handle workloads previously requiring 20-30% more headcount. The tool reached $1 billion run-rate revenue by November 2025, six months post-launch, capturing 54% of enterprise AI coding market share.



Add unit tests to the checkout flow to cover paym

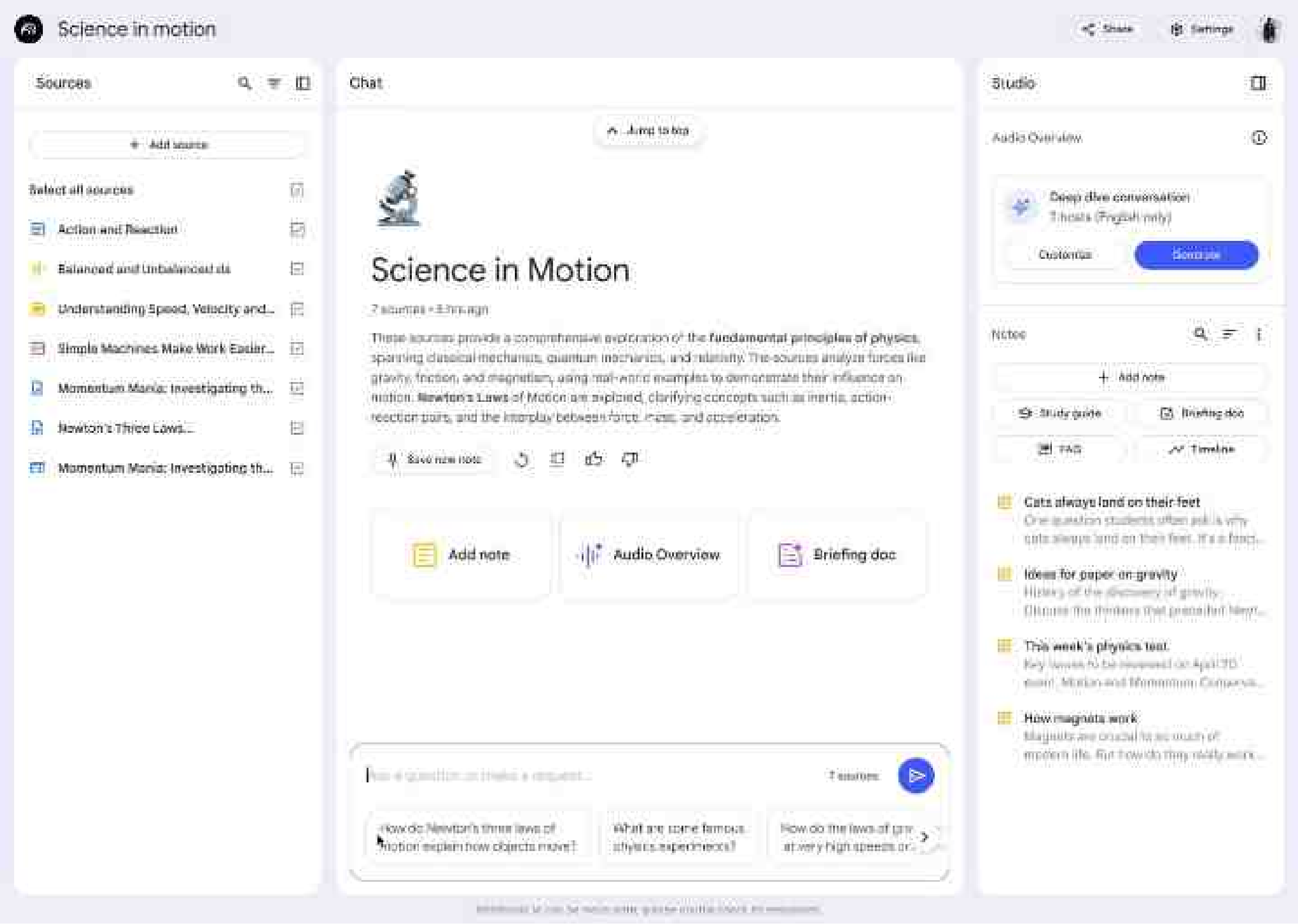
acme/order-service
Cloud Environment
Local worktree
Cloud Environment
Add environment
Create or update my CLAUDE.md file
Search for a TODO comment and fix it
Recommend areas to improve our tests
GIFRUN.COM

Source: Anthropic

Innovative UX AI Product of 2025 – NotebookLM has 30M monthly active users

Generational

NotebookLM is Google's AI research assistant that pioneered consumer AI audio generation with its viral Audio Overviews feature. Launched in September 2024, Audio Overviews converted documents into podcast-style conversations between two AI hosts with remarkably natural audio quality, sparking the AI-generated podcast trend. Spotify partnered with NotebookLM for its December 2024 Wrapped campaign, generating personalized AI podcasts for millions of users. By 2025, NotebookLM evolved into a comprehensive multimedia platform with eight output formats: Audio Overviews, Video Overviews, Slide Decks, Infographics, Mind Maps, Flashcards, Quizzes, and Data Tables. This format flexibility reached an estimated 30 million monthly active users.



Science in motion

Share Settings

Sources

+ Add source

Select all sources

- Action and Reaction
- Balanced and Unbalanced.xls
- Understanding Speed, Velocity and...
- Simple Machines Make Work Easier...
- Momentum Mania: Investigating th...
- Newton's Three Laws...
- Momentum Mania: Investigating th...

Chat

Jump to top

Science in Motion

7 sources • 3 hrs ago

These sources provide a comprehensive exploration of the **fundamental principles of physics**, spanning classical mechanics, quantum mechanics, and relativity. The sources analyze forces like gravity, friction, and magnetism, using real-world examples to demonstrate their influence on motion. **Newton's Laws** of Motion are explored, clarifying concepts such as inertia, action-reaction pairs, and the interplay between force, mass, and acceleration.

Save new note

Add note | Audio Overview | Briefing doc

Ask a question or make a request... 7 sources

How do Newton's three laws of motion explain how objects move? | What are some famous physics experiments? | How do the laws of gra... at very high speeds or...

NotebookLM can be inaccurate, please double check its responses.

Studio

Audio Overview

Deep dive conversation
2 hosts (English only)

Customize | Generate

Notes

+ Add note

Study guide | Briefing doc
FAQ | Timeline

Cats always land on their feet
One question students often ask is why cats always land on their feet. It's a fasci...

Ideas for paper on gravity
History of the discovery of gravity. Discuss the thinkers that preceded Newt...

This week's physics test
Key issues to be reviewed on April 30 exam: Motion and Momentum, Compare...

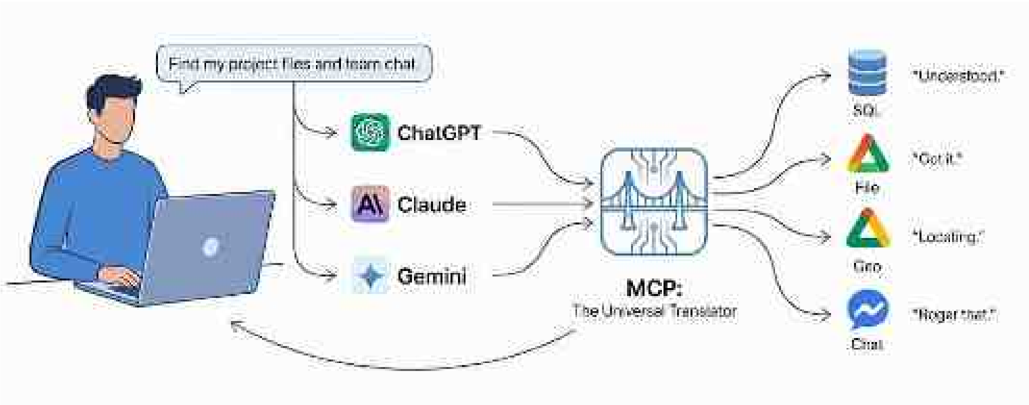
How magnets work
Magnets are crucial to so much of modern life. But how do they really work...

Source: Google

Open Source Product of 2025 – Model Context Protocol adopted by all AI firms

Generational

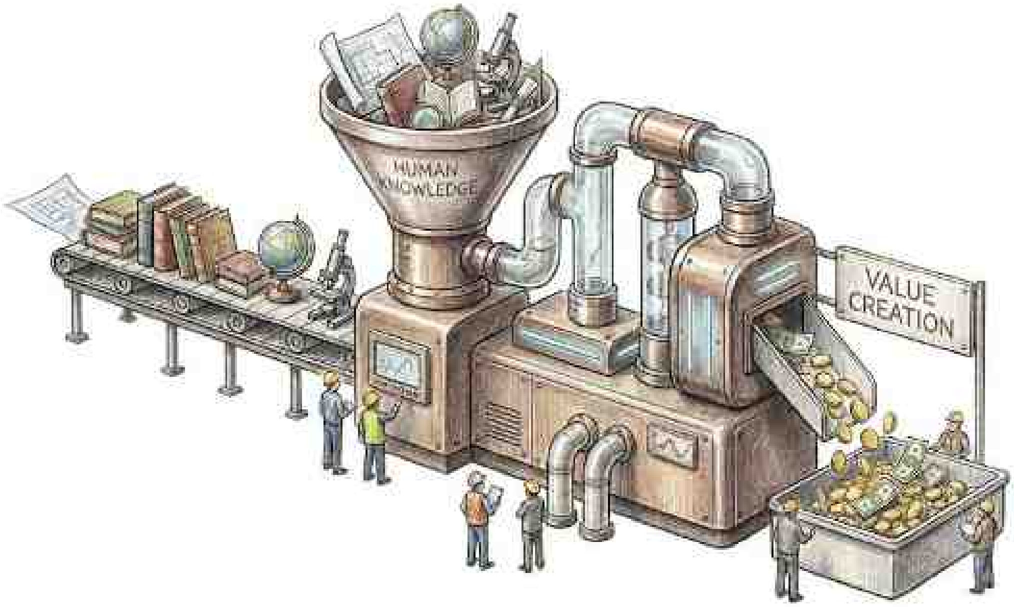
Model Context Protocol (MCP), launched by Anthropic in November 2024, became the industry standard for AI-to-tool integration within one year. The protocol solved the fragmentation problem of custom integrations by providing a universal interface for AI systems to access external data, comparable to USB-C for devices. Adopted by OpenAI, Google, and Microsoft, MCP reached 97 million monthly SDK downloads and over 5,800 servers by December 2025. The protocol was transferred to the Linux Foundation's Agentic AI Foundation (co-founded by Anthropic, Block, OpenAI, backed by Google, Microsoft, AWS) in December 2025, ensuring vendor-neutral governance.



	2024		2025											
	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Vetted Server Count (approx.)	400	600	800	1,200	2,700	3,600	4,000	4,700	5,200	5,600	5,900	6,100	6,500	7,000
Notable Companies Adopting	Block, Apollo, Replit, Windsurf, Sourcegraph	-	-	-	OpenAI	Google	Microsoft, GitHub	Salesforce	-	-	-	Cursor	-	AWS, Cloudflare, Bloomberg
Key Protocol Updates	Initial launch: basic connections	-	-	-	OAuth security, Streamable HTTP, audio support	-	-	Enhanced security controls	-	-	Server registry	-	Interactive apps, background tasks, extensions	-

Source: Model Context Protocol, PulseMCP

Section 9: AI Agent Opportunity



Spectrum of generative AI technology

Generative AI is a spectrum. Each step raises the ceiling on what can be automated. Automation capability rises as we move from base models to agents that can execute tasks with tools, memory, and reasoning.

Technology	Base Multimodal Language Model	Agents	Robotic Systems
Definition	Processes/generates text, images, audio; excels with digital information but lacks physical actions or human-like logic.	Builds on Copilot with autonomy, reasoning, and digital environment interaction. Manages software, web, and executes tasks independently.	Combines AI Agents with physical capabilities for interaction in the real world. Used in robots, autonomous vehicles, etc.
Example products available today	GPT-4o Llama 3 DeepSeek R1	Claude ChatGPT CURSOR	WAYMO FIGURE

Source: Author analysis

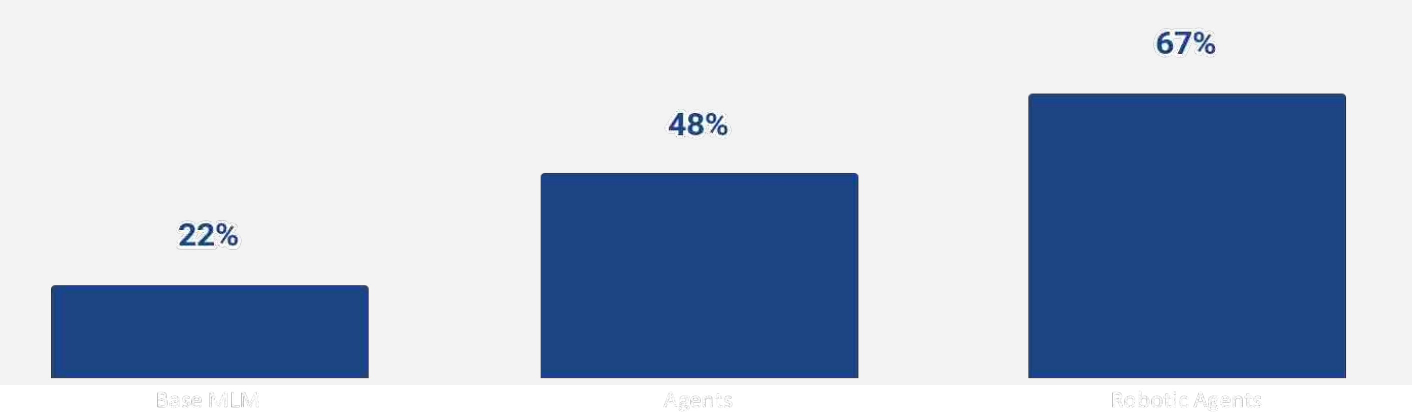
Agents are the automation frontier

Generational

A year ago I wrote that agents were early and would mainstream in one to three years. The last twelve months moved faster than expected: as teams shipped true agents—systems with tools, memory, and supervision—automation scores stepped up meaningfully beyond chat and copilots. I'm also shortening my robotics horizon from 5–10 years to 2–5 years, reflecting rapid progress in perception, planning, and low-cost hardware. The bars at the bottom summarize the average automation level across tasks for each tier.

Technology	Base Multimodal Language Model	Agents	Robotic Systems
Time to majority adoption	Now	1-2 years	2-5 years
Most automatable jobs	Interpreters & translators, Writers and Authors, Proofreaders and Copy Markers	Bookkeeping, Accounting, and Auditing Clerks, Insurance Claims Processing Clerks	Light Truck Drivers, Taxi Drivers
Least automatable jobs	Oil wellhead pumpers, nuclear power operators	Dishwashers, Roofers	Psychiatrists, Judges & magistrates

Average automation score across tasks



Source: Author analysis

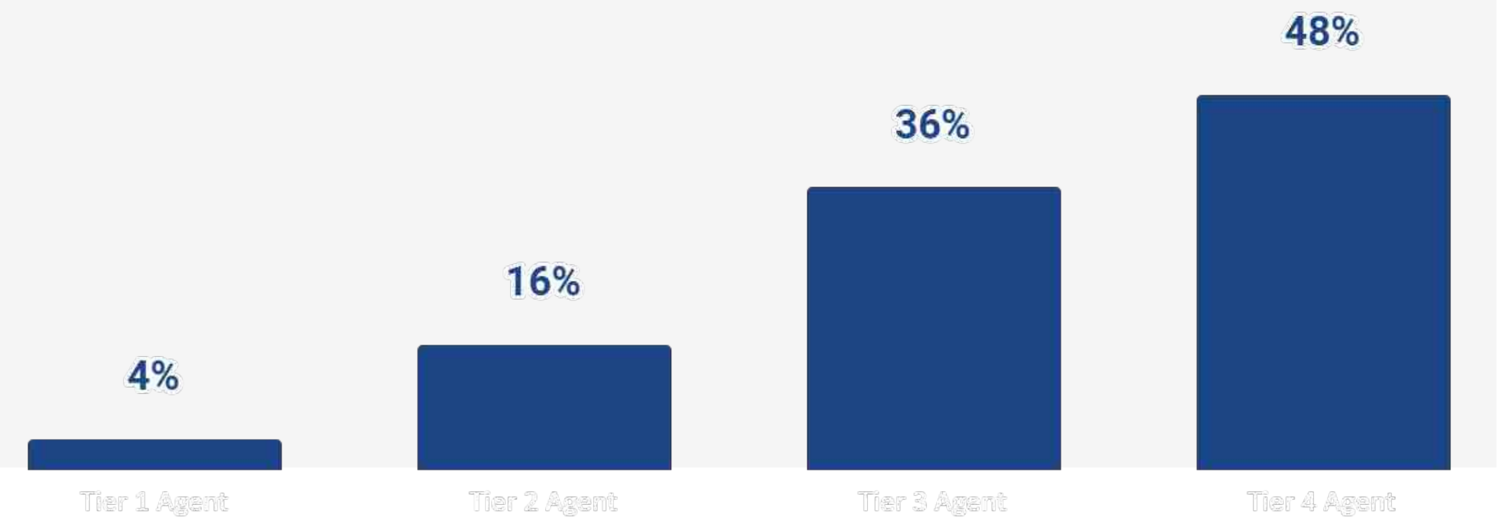
Agentic automation by model capability

Generational

The largest step-ups appear at Tier-3/Tier-4 capability combined with sound system design. The distribution also widens: more tasks cross the threshold from partially automatable to consistently automatable. In the pages that follow, we use Tier-4 agent scores as the baseline for mapping opportunities.

Model	Tier 1	Tier 2	Tier 3	Tier 4
Definition	Basic text generation with limited reasoning, tiny 4K context, no reliable tool use and frequent errors — error prone for most tasks	Improved instruction following for tasks, larger 64K context, basic tool use, brittle autonomy, still requires oversight	implicit reasoning, 128K context, versatile tool use, multimodal understanding, better self-correction	Human-level reasoning, 1M context, robust tool use, strong multimodality, self-correction
Approximate year	2022	2023	2024	2025
Representative Models	GPT-3.5	GPT-4 Turbo	GPT-4o Gemini 2.0 Claude 3.5	GPT-5 Gemini 3.0 Pro Claude 4.5

Average agent automation score across tasks



Source: Author analysis

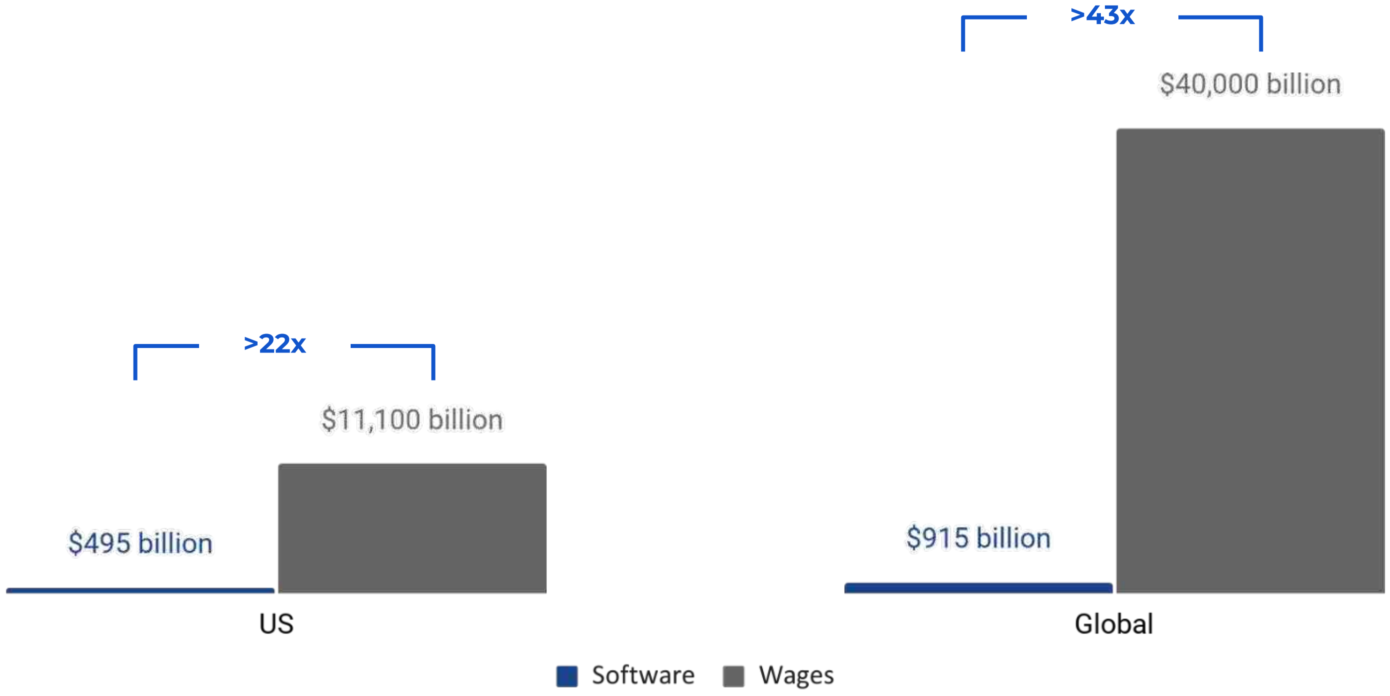
AI agent software opportunity

Generational

As we've seen models now beat professionals with over a decade in experience, the AI agent opportunity is even more imminent. AI agents go beyond traditional software, performing tasks that could gradually reshape roles across various industries, including the service sector. This shift suggests a potential realignment in the labor market, as AI may begin handling tasks typically associated with human expertise. In this context, wages become a proxy for a broader market opportunity, which is >20-40x bigger than the software market. Put into this context, the ROI question seems less apprehensive. This reframes ROI. We are selling outcomes against wage pools, not licenses.

United States

Global



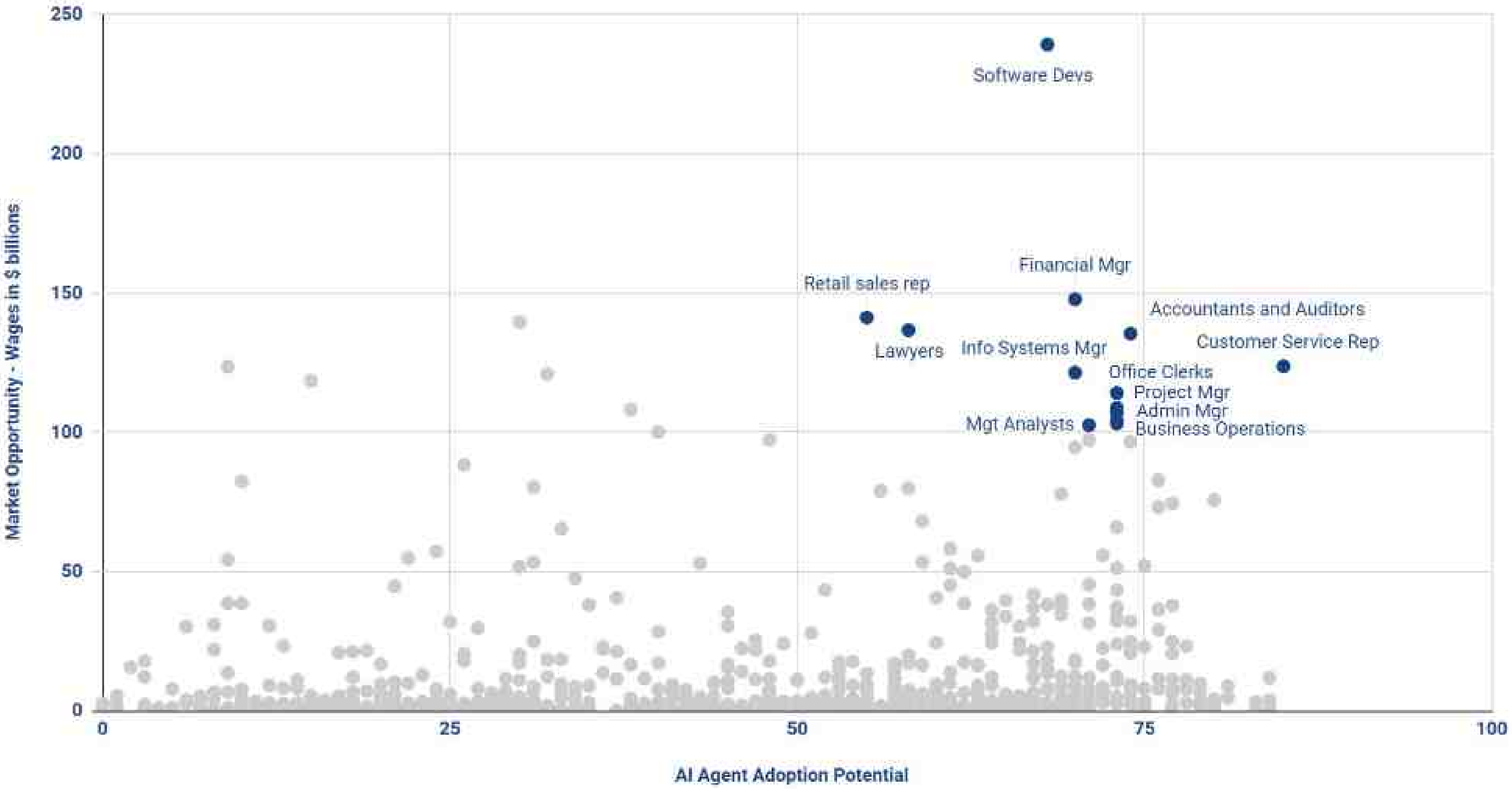
Source: Author analysis, US Bureau Labor of Statistics, FRED, IBIS World

Framework for agentic automation

Generational

To tactically frame the opportunity, I analyzed my study of 19,000 job tasks, this framework identifies jobs with high automation potential and significant market opportunity. Jobs with high automation scores are prime candidates for AI adoption, offering faster ROI and a clear path for investment. A useful proxy for market size is total wages, with most roles earning under $4 billion annually. However, outliers like software developers, with $230 billion in wages, represent the most attractive opportunities. The chart below maps jobs by automation potential and wage market size, highlighting roles at the forefront of agentic automation.

This analysis updates last year's study with 3,000 additional tasks & new wage data. The consistency of results validates the framework's usefulness: the jobs identified as high-potential & high-value attracted the most venture capital and grew the fastest.



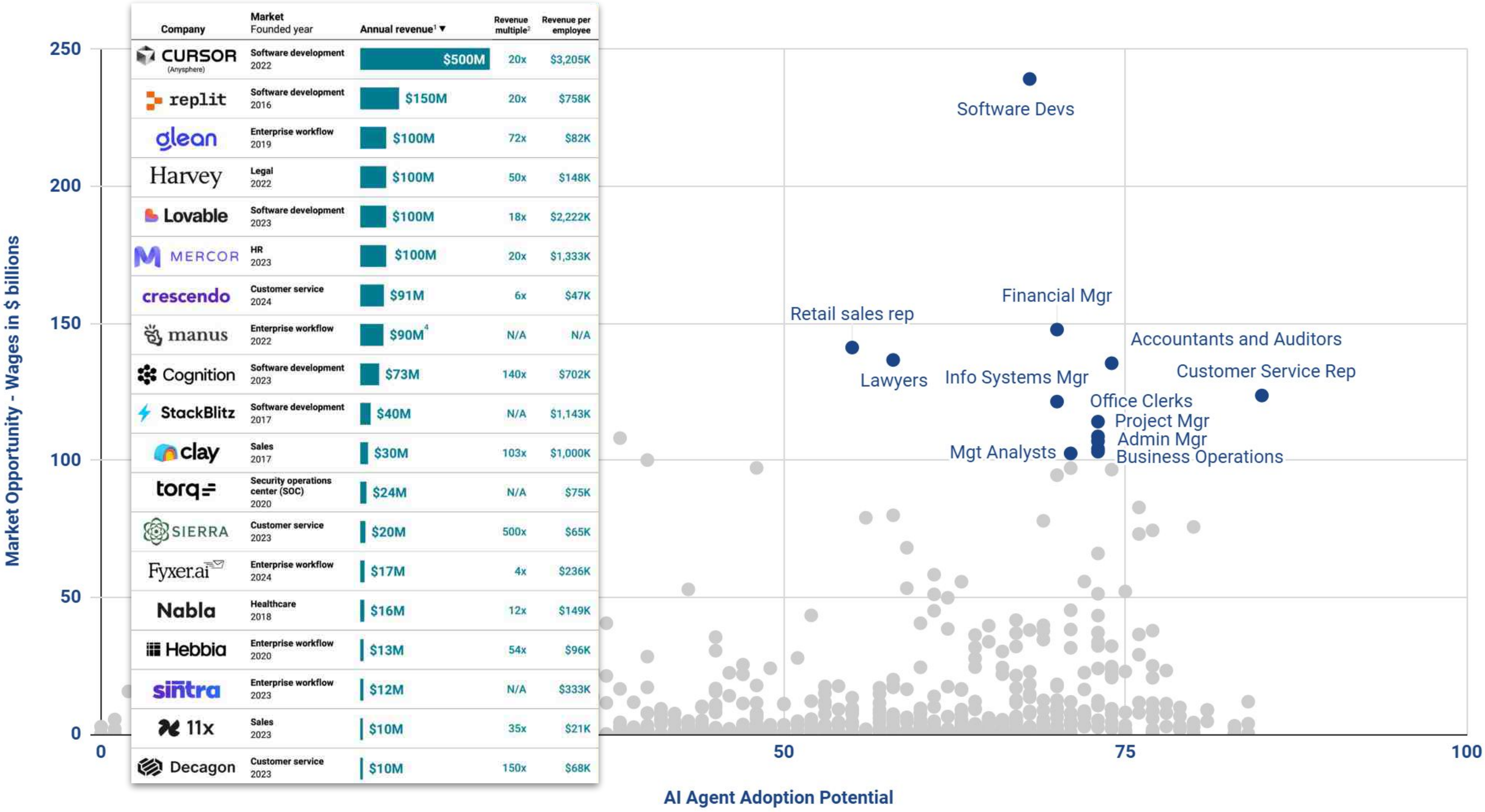
Source: Author analysis, US Bureau Labor of Statistics, CBInsights

Framework for agentic automation

Generational

To tactically frame the opportunity, I analyzed 19,000 job tasks across 830 jobs. This framework identifies jobs with high automation potential and significant market opportunity. Jobs with high automation scores are prime candidates for AI adoption, offering faster ROI and a clear path for investment. A useful proxy for market size is total wages, with most roles earning under $4 billion annually. Outliers like software developers, with $230 billion in wages, represent the most attractive opportunities. The chart below maps jobs by automation potential and wage market size, highlighting roles at the forefront of agentic automation.

This analysis updates last year's study with 3,000 additional tasks & new wage data. The consistency of results validates the framework's usefulness: the jobs identified as high-potential & high-value attracted the most venture capital and grew the fastest.

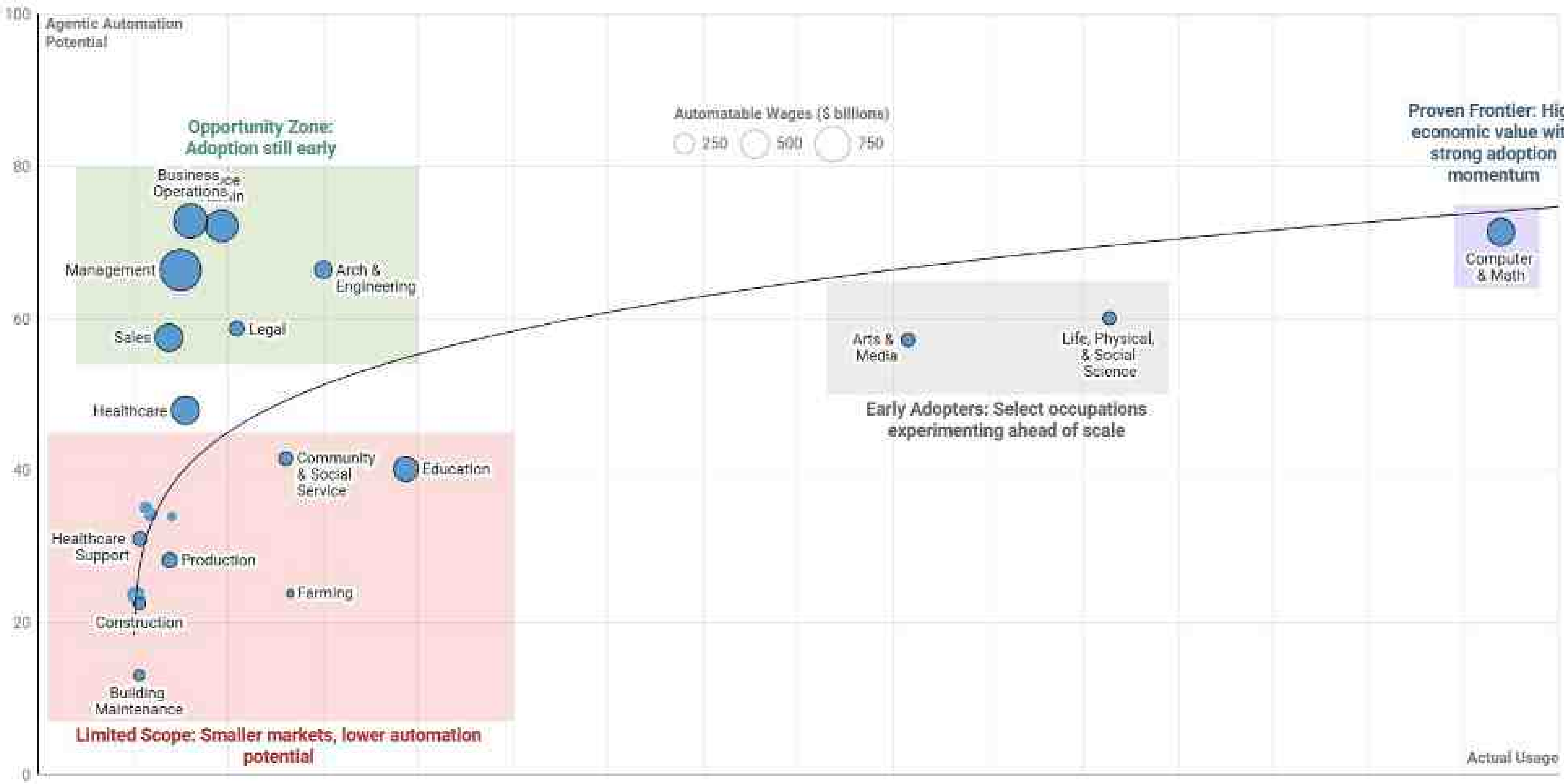


Company	Market Founded year	Annual revenue[1] ▼	Revenue multiple[2]	Revenue per employee
CURSOR (Anysphere)	Software development 2022	$500M	20x	$3,205K
replit	Software development 2016	$150M	20x	$758K
glean	Enterprise workflow 2019	$100M	72x	$82K
Harvey	Legal 2022	$100M	50x	$148K
Lovable	Software development 2023	$100M	18x	$2,222K
MERCOR	HR 2023	$100M	20x	$1,333K
crescendo	Customer service 2024	$91M	6x	$47K
manus	Enterprise workflow 2022	$90M[4]	N/A	N/A
Cognition	Software development 2023	$73M	140x	$702K
StackBlitz	Software development 2017	$40M	N/A	$1,143K
clay	Sales 2017	$30M	103x	$1,000K
torq	Security operations center (SOC) 2020	$24M	N/A	$75K
SIERRA	Customer service 2023	$20M	500x	$65K
Fyxer.ai	Enterprise workflow 2024	$17M	4x	$236K
Nabla	Healthcare 2018	$16M	12x	$149K
Hebbia	Enterprise workflow 2020	$13M	54x	$96K
sintra	Enterprise workflow 2023	$12M	N/A	$333K
11x	Sales 2023	$10M	35x	$21K
Decagon	Customer service 2023	$10M	150x	$68K

Source: Author analysis, US Bureau Labor of Statistics, CBInsights

Framework for agentic automation

Thinking about automation potential together with actual AI usage creates a practical framework for deciding where to build and invest hext. Jobs in the Proven Frontier—notably Computer & Mathematics—combine high potential with strong usage, indicating mature adoption and near-term impact. The Opportunity Zone—including Management, Business Operations, and Office/Admin—shows high potential but lower current usage, signaling white space for new products and company formation. This slide uses usage observed at the 22 major occupation-group level. The previous slide drills down from these high-level groupings to ~830 specific occupations, but is missing usage data.



Source: Author analysis, US Bureau Labor of Statistics, Anthropic

Example: Lawyers

Generational

Occupation description: Represent clients in criminal and civil litigation and other legal proceedings, draw up legal documents, or manage or advise clients on legal transactions. May specialize in a single area or may practice broadly in many areas of law.
Number employed: 731,340
Average annual wage: $176,470

Technology	MLM	Agent	Robot
Example tasks			
Interpret laws, rulings and regulations for individuals and businesses. `	Low	Med	Med
Analyze the probable outcomes of cases, using knowledge of legal precedents.	Low	High	High
Gather evidence to formulate defense or to initiate legal actions	Low	Med	Med
Represent clients in court or before government agencies.	N/A	Low	Low
Evaluate findings and develop strategies and arguments in preparation for presentation	Low	Med	Med
Advise clients concerning business transactions, claim liability, advisability of suits.	Low	Med	Med
Examine legal data to determine advisability of defending or prosecuting lawsuit.	Low	Med	Med
Prepare, draft, and review legal documents, such as wills, deeds, patent applications	Med	High	High
Study Constitution, statutes, decisions, and ordinances of quasi-judicial bodies	Low	High	High
Negotiate settlements of civil disputes.	Low	Med	Med
Supervise legal assistants.	N/A	Low	Low
Negotiate contractual agreements.	Low	Med	Med
Confer with colleagues with specialties in appropriate areas of legal issue	Low	Med	Med
Search for and examine public and other legal records to write opinions	Low	High	High
Perform administrative and management functions related to the practice of law.	Low	High	High

Source: Author analysis, US Bureau of Labor Statistics

Spellbook – Legal agent

Generational

Spellbook uses AI to review and draft contracts and they have always been at the forefront of adopting the latest models to legal work, such as launching the first agentic product for lawyers, Spellbook Associate. Spellbook is used by over 2,600 law firms, professional services, & in-house teams including the likes of Addleshaw Goddard (Global Law 200), Nestle (Fortune 100), and BDO (top 5 auditing firm). Spellbook recently raised a $20 million series A from Inovia Capital and strategic investor Thomson Reuters.

For a deep dive into how PolyAI builds products, read Building enterprise AI products with Spellbook. Here's a preview:

- **Embrace the "Skepticism Window":** Capitalize on the period when a new AI technology faces skepticism but shows promise. Scott's experience with agentic AI mirrors the early days of GPT models, suggesting that this skepticism often precedes rapid adoption and improvement.
- **Focus on Hard Sub-Problems:** In complex AI systems like Spellbook Associate, solving difficult sub-problems (e.g., manipulating a 100-page legal document) is crucial before tackling higher-level tasks. This approach ensures a solid foundation for more advanced features.
- **Leverage Existing Workflows:** Spellbook's success partly comes from integrating with lawyers' familiar processes, like track changes. This reduces adoption friction and addresses potential concerns about AI errors.
- **Sequence Feature Expansion:** Start with core workflows and gradually broaden functionality. This strategy allows for mastering specific use cases before expanding, as seen in Spellbook's evolution from single to multi-document workflows.
- **Ride the Foundation Model Wave:** Build on top of rapidly improving foundation models to benefit from their advancements. Scott emphasizes the importance of capturing upside from new model releases like GPT-4 and o1.
- **Value-Based Pricing in High-Stakes Industries:** In fields like law where professional time is extremely valuable, flat-fee pricing can be more appealing than usage-based models. This approach simplifies billing and emphasizes the product's value proposition.
- **Balance AI Aggressiveness with Cost Control:** Encourage aggressive AI usage in development while monitoring for excessive resource consumption. Scott mentions allowing developers to use AI freely but intervening when necessary to prevent runaway costs.

Example: Customer service reps

Generational

Occupation description: Interact with customers to provide basic or scripted information in response to routine inquiries about products and services. May handle and resolve general complaints. Excludes individuals whose duties are primarily installation, sales, repair, and technical support.
Number employed: 2,858,710
Average annual wage: $43,520

Technology	MLM	Agent	Robot
Example tasks			
Confer with customers by telephone or in person to provide information about products	Med	High	High
Keep records of customer interactions or transactions, recording details of inquiries	Low	High	High
Check to ensure that appropriate changes were made to resolve customers' problems.	Low	High	High
Contact customers to respond to inquiries or to notify them of claim investigation results	Low	High	High
Determine charges for services requested, collect deposits or payments	Low	High	High
Complete contract forms, prepare change of address records, or issue service request	Low	High	High
Refer unresolved customer grievances to departments for further investigation.	Low	High	High
Resolve customers' service or billing complaints by performing activities	Low	Med	Med
Review insurance policy terms to determine whether a particular loss is covered	Low	High	High
Solicit sales of new or additional services or products.	Low	Med	Med
Compare disputed merchandise with original requisitions and information from invoices	Low	High	High
Obtain and examine all relevant information to assess validity of complaints	Low	High	High
Recommend improvements in products, packaging, shipping, service, or billing methods	Low	Med	Med

Source: Author analysis, US Bureau of Labor Statistics

PolyAI – Customer service agent

Generational

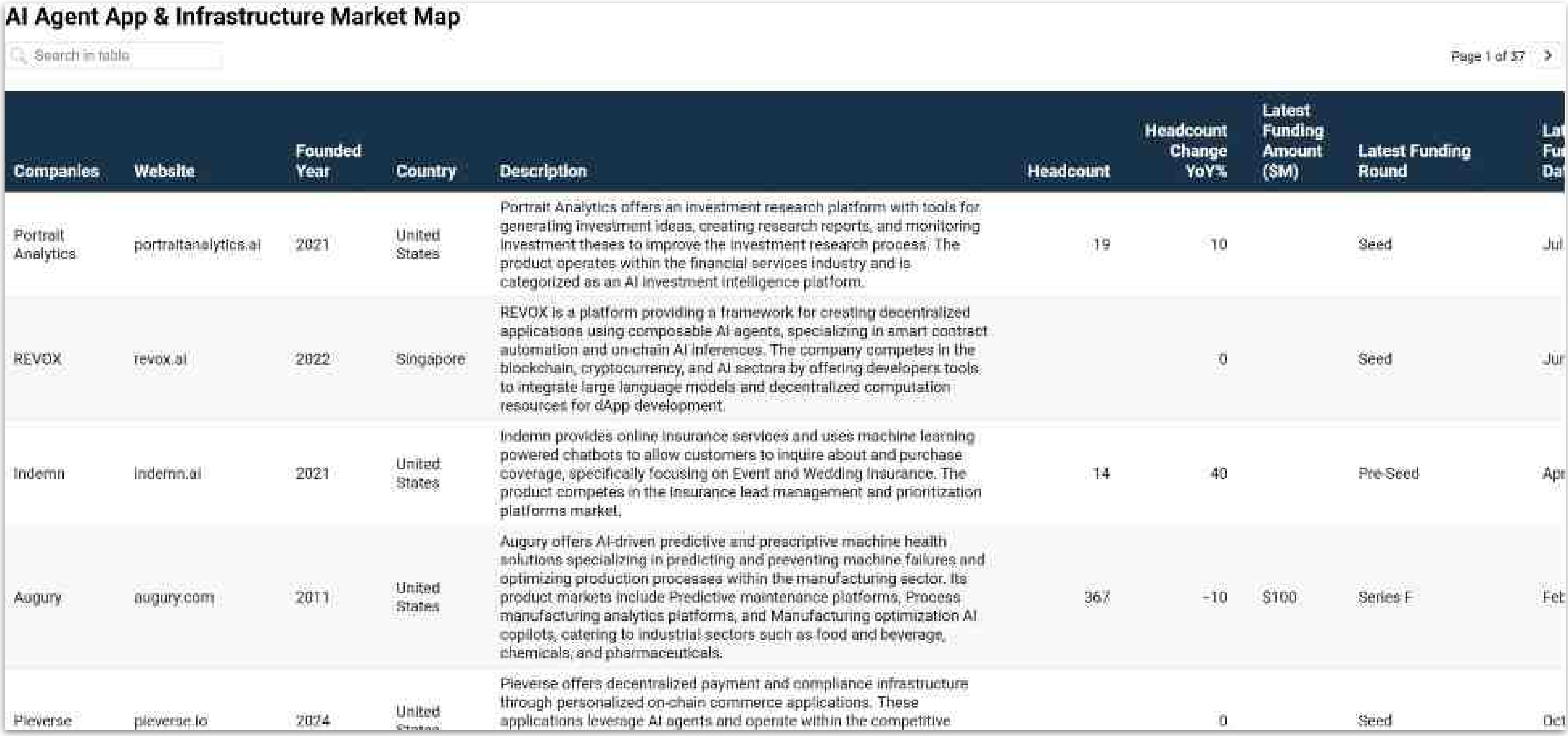
PolyAI develops enterprise conversational assistants that engage in natural conversations with customers to resolve their issues. These assistants understand customers regardless of their phrasing or manner of speaking. While voice interaction has gained recent popularity due to GPT-4o, PolyAI has been at the forefront of this technology since 2017. The company recently raised a $50M Series C round led by Hedosophia and NVentures (NVIDIA's venture arm), bringing their total funding to $120M from prominent investors including Khosla Ventures, Georgian, Point72 Ventures, and others.

For a deep dive into how PolyAI builds products, read Building enterprise AI products with PolyAI. Here's a preview:

- **Enterprise AI adoption requires balancing multiple stakeholders**: While customer experience heads traditionally were the main decision-makers, generative AI brings security, IT, branding, and legal teams into the conversation. This complicates the sales process but also creates opportunities for education and addressing diverse concerns.
- **Enterprises often hold AI to higher standards than humans**: As Devidas noted, AI assistants may be expected to stay rigidly on-topic in ways that human agents are not. This creates challenges in making AI seem natural while still meeting strict enterprise requirements.
- **Customization and control are critical**: Some enterprises have extremely specific restrictions, like forbidding an AI from stating basic facts unrelated to their business. AI systems need to be highly configurable to meet these idiosyncratic needs.
- **Practical solutions trump theoretical ideals**: Shawn emphasized the importance of being practical rather than trying to build the perfect technical solution. Time-to-market, packaging, marketing strategy, and sales execution are as important as the underlying technology.
- **Layered safeguards are necessary**: PolyAI uses multiple layers of protection, from general content filters to project-specific customizations. This allows tailoring the AI's behavior to different levels of risk tolerance.
- **Iterative testing with clients is crucial**: Finding edge cases and potential issues requires extensive testing, both internally and with clients. This process accumulates knowledge over time that can be applied to future projects.
- **Transparency about limitations builds trust**: Being open about the current state of generative AI technology and its limitations, while showing a clear roadmap for addressing concerns, helps enterprises feel more comfortable adopting these solutions.
- **Self-serve capabilities are becoming important**: With generative AI, some enterprises want more control in maintaining or even building their own assistants. Providing tools for this can be a differentiator.

AI Agent Market Map: 1,800 companies and 180 tha matter the most

The AI agent market exploded in 2025. I compiled 1,800 companies across the app and infrastructure layer and identified the top 10%—the ones with traction, differentiated positioning, or solving high-value problems. The full table is available in the Generational Substack article about this report or through this link here.



AI Agent App & Infrastructure Market Map

Companies	Website	Founded Year	Country	Description	Headcount	Headcount Change YoY%	Latest Funding Amount ($M)	Latest Funding Round	Lat Fu Da
Portrait Analytics	portraitanalytics.ai	2021	United States	Portrait Analytics offers an investment research platform with tools for generating investment ideas, creating research reports, and monitoring investment theses to improve the investment research process. The product operates within the financial services industry and is categorized as an AI investment intelligence platform.	19	10		Seed	Jul
REVOX	revox.ai	2022	Singapore	REVOX is a platform providing a framework for creating decentralized applications using composable AI agents, specializing in smart contract automation and on-chain AI inferences. The company competes in the blockchain, cryptocurrency, and AI sectors by offering developers tools to integrate large language models and decentralized computation resources for dApp development.		0		Seed	Jun
Indemn	indemn.ai	2021	United States	Indemn provides online insurance services and uses machine learning powered chatbots to allow customers to inquire about and purchase coverage, specifically focusing on Event and Wedding Insurance. The product competes in the insurance lead management and prioritization platforms market.	14	40		Pre-Seed	Apr
Augury	augury.com	2011	United States	Augury offers AI-driven predictive and prescriptive machine health solutions specializing in predicting and preventing machine failures and optimizing production processes within the manufacturing sector. Its product markets include Predictive maintenance platforms, Process manufacturing analytics platforms, and Manufacturing optimization AI copilots, catering to industrial sectors such as food and beverage, chemicals, and pharmaceuticals.	367	-10	$100	Series F	Feb
Pieverse	pieverse.io	2024	United States	Pieverse offers decentralized payment and compliance infrastructure through personalized on-chain commerce applications. These applications leverage AI agents and operate within the competitive		0		Seed	Oct

Source: Various company databases

Section 10: 2026 Outlook

2025 Scorecard

How did last year's outlook hold up? Four of five predictions are on track—the exception being interactable virtual worlds, which lagged behind text and video generation.

Outlook	Score	Assessment
Smartphones become the default GenAI device	Green	This was expected – all phones released this year marketed themselves as AI phones. No other consumer device has the right combination of compute, battery, portability and ubiquity.
Glasses will be the next GenAI device	Green	AI glasses have proliferated. Since last year's outlook, Meta has sold millions of glasses. While the only major consumer product in the US has been Meta's, there are lots of vendors in Asia. Google will also be releasing their first post-GenAI erra glasses in Q1 2026 and Apple is also pivoting their AR experience to glasses.
Robotaxis will become the norm	Green	Waymo has made rapid progress and we will continue to see this trend across US and China. So far this year, Waymo has completed over 14 million paid trips, three times more than in 2024
We'll create our own virtual worlds	Red	My expectations here were specific to models creating interactable worlds. Models have not developed as fast as their textual counterparts. There also hasn't been a viral use case. That said, there are adjacent viral use case in video generation this year with OpenAI's Sora.
Everyone will have their own agents	Green	This is true because mainstream AI apps like ChatGPT and Google Gemini have agentic capabilities by default. While more advanced agentic capabilities might be expensive, they are still generally available.

5M AI glasses will be sold in US in 2026, >3x in 2025

Generational

Smartphones are versatile but limited by their reliance on touch interfaces, which require users to actively engage by pulling out the device. For AI-powered experiences to feel truly seamless, they need to integrate naturally into our environment. Smart glasses offer the most immediate solution.

Meta has proven the market exists. Ray-Ban Meta glasses have sold over 2 million pairs, with sales tripling in 2025—driving the overall smart glasses market to 110% YoY growth. Meta plans to produce 10 million pairs annually by 2026. Google and Samsung are responding with Android XR glasses powered by Project Astra (Gemini AI), scheduled for Q1 2026, with Warby Parker as a retail partner.

The two main US players driving innovation remain Meta and the Google/Samsung/Qualcomm coalition—but Meta has a 2-year head start in market learning.

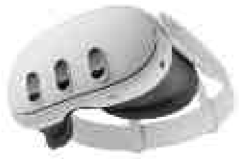
Experience with Meta Ray-Ban Display glasses

Google’s Project Astra demo. Click to play video

Source: Meta

Smart glasses primer

Generational

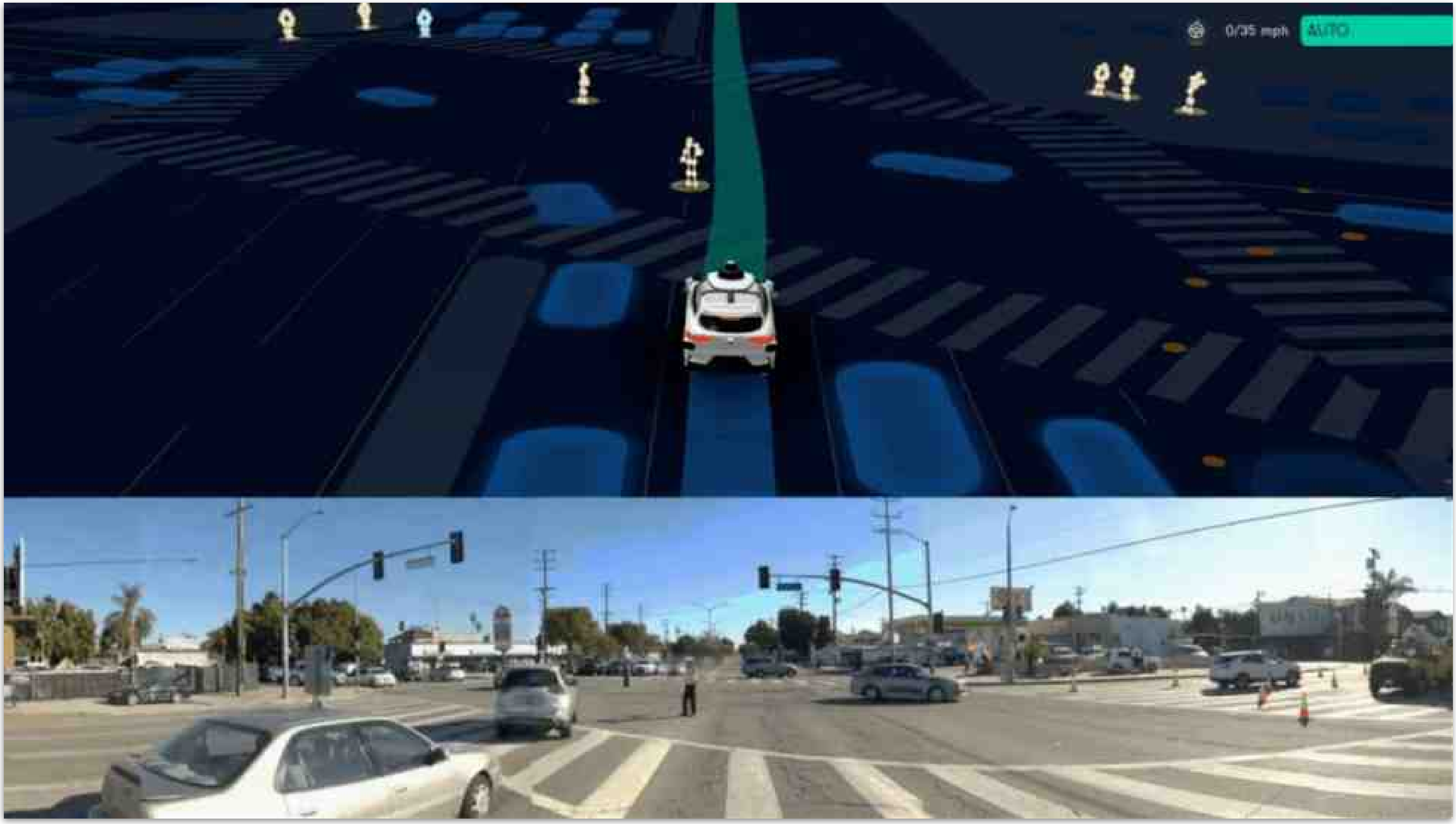
Technology	Level 1 AV Glasses	Level 2 AV Glasses w/ HUD	Level 3 Companion AR	Level 4 True AR	Level 5 Immersive MR
Definition	Smart eyewear with no display, focusing on hands-free audio features like voice assistants and music.	Glasses with a small HUD for simple static overlays, displaying information but not interacting with the environment.	Glasses that extend your screen, offering a virtual display for media and productivity but lacking true AR anchoring.	True AR glasses that overlay interactive 3D content onto the real world, with spatial awareness and real-time interaction.	Mixed reality headsets capable of both AR and VR, blending immersive environments with real-world elements.
Visual Approach	No display, audio (built-in speakers / mic) and camera	Small corner display (heads-up), static overlay	Larger virtual screen, often tethered to phone/PC	See-through display with environment mapping (SLAM)	Fully enclosed visor or high-res video passthrough
XR Capabilities	None	Basic AR (text/graphics only)	2D/3D overlays (not world-locked)	Full AR (world-locked 3D)	AR + VR (mixed reality)
Typical Use Cases	Music, calls, discreet notifications	Hands-free checklists, directions, remote video feed	Private big-screen media, multi-monitor productivity	Immersive training, design visualization, 3D overlays	High-end simulation, VR/AR gaming, virtual collaboration
Example products	 Meta Rayban	 Meta Rayban Display	 XReal Air2 Pro	 Magic Leap 2	 Meta Quest 3

Source: Author analysis

US robotaxis will drive 50M trips in 2026, >3x from 2025

Generational

Autonomous vehicles have crossed the threshold from pilot to scale. Waymo completed 14 million trips in 2025—3x the prior year—and now serves 450,000 weekly paid rides across 5 US cities with a fleet of 2,500 robotaxis. Tesla is also rapidly testing its robotaxis across 10 cities. Both are expanding aggressively: Waymo plans 20+ new cities in 2026 including Tokyo and London. Tesla is aiming to produce its first Cybercab car and have AV services in 30 cities in 2026. Morgan Stanley expects 30% of all rideshare miles driven will be done through AV cars.



Waymo recognizing the traffic cop

Source: Waymo

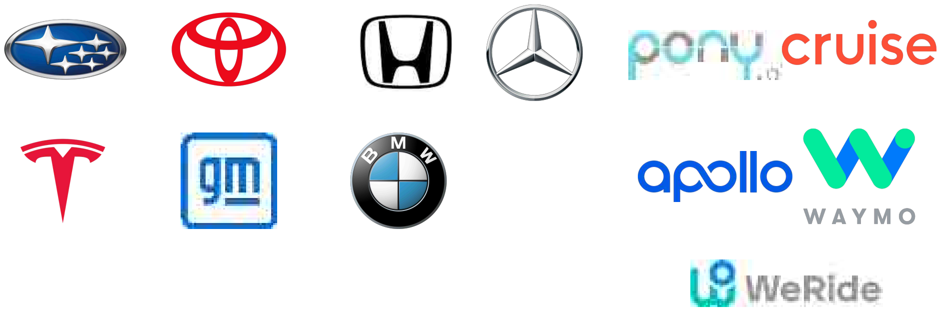
Autonomous vehicles primer

Generational

The Society of Automotive Engineers (SAE) levels of automation provide a framework for classifying vehicle automation from no automation (Level 0) to full automation (Level 5). These levels are defined based on the extent of driver involvement and the vehicle's ability to handle dynamic driving tasks under specific conditions. This framework has been guiding automakers, regulators, and consumers in understanding and deploying autonomous technologies responsibly. Most personal cars today are in level 1-2 and with commercial autonomous vehicles in L4.

Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
No automation: the driver is in complete control of the vehicle at all times.	**Driver assistance**: the vehicle can assist the driver or take control of either the vehicle's speed, through cruise control, or its lane position, through lane guidance.	**Occasional self-driving**: the vehicle can take control of both the vehicle's speed and lane position in some situations, for example on limited-access freeways.	**Limited self-driving**: the vehicle is in full control in some situations, monitors the road and traffic, and will inform the driver when he or she must take control.	**Full self-driving under certain conditions**: the vehicle is in full control for the entire trip in these conditions, such as urban ride-sharing.	**Full self-driving under all conditions**: the vehicle can operate without a human driver or occupants.

Car Brands by Highest Achieved Autonomous Vehicle Level

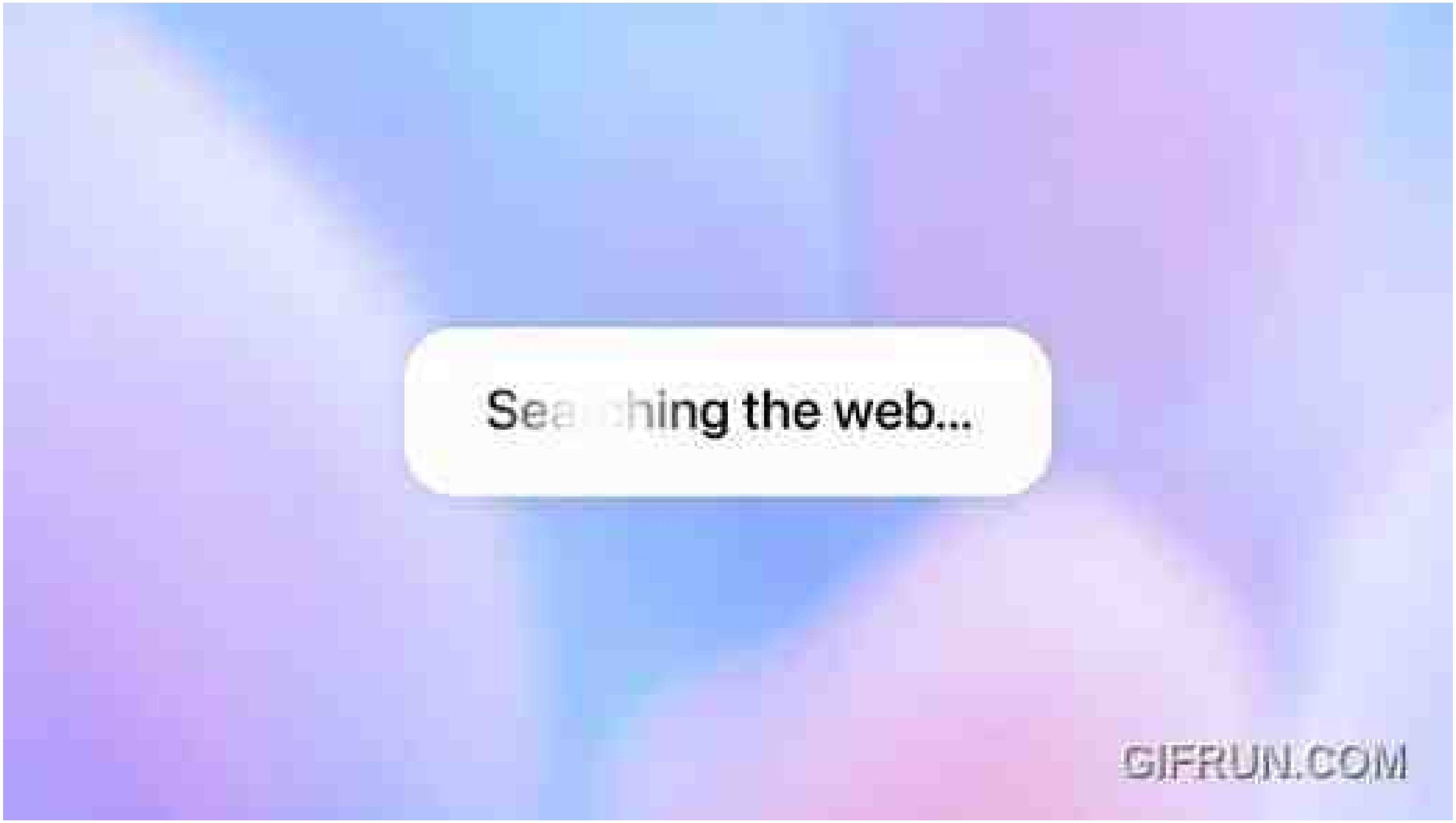


Source: SAE

Agents will process $1B of e-commerce sales in 2026

Generational

Consumer habits are already shifting toward AI-mediated discovery. Half of all internet users now engage AI when searching online, and 44% of those users say AI is their primary and preferred source—not a supplement, but a replacement for traditional search. The audience is already there. What's missing is the transaction layer—the ability to move from "find this" to "buy this" without human handoff. That infrastructure arrived in late 2025.

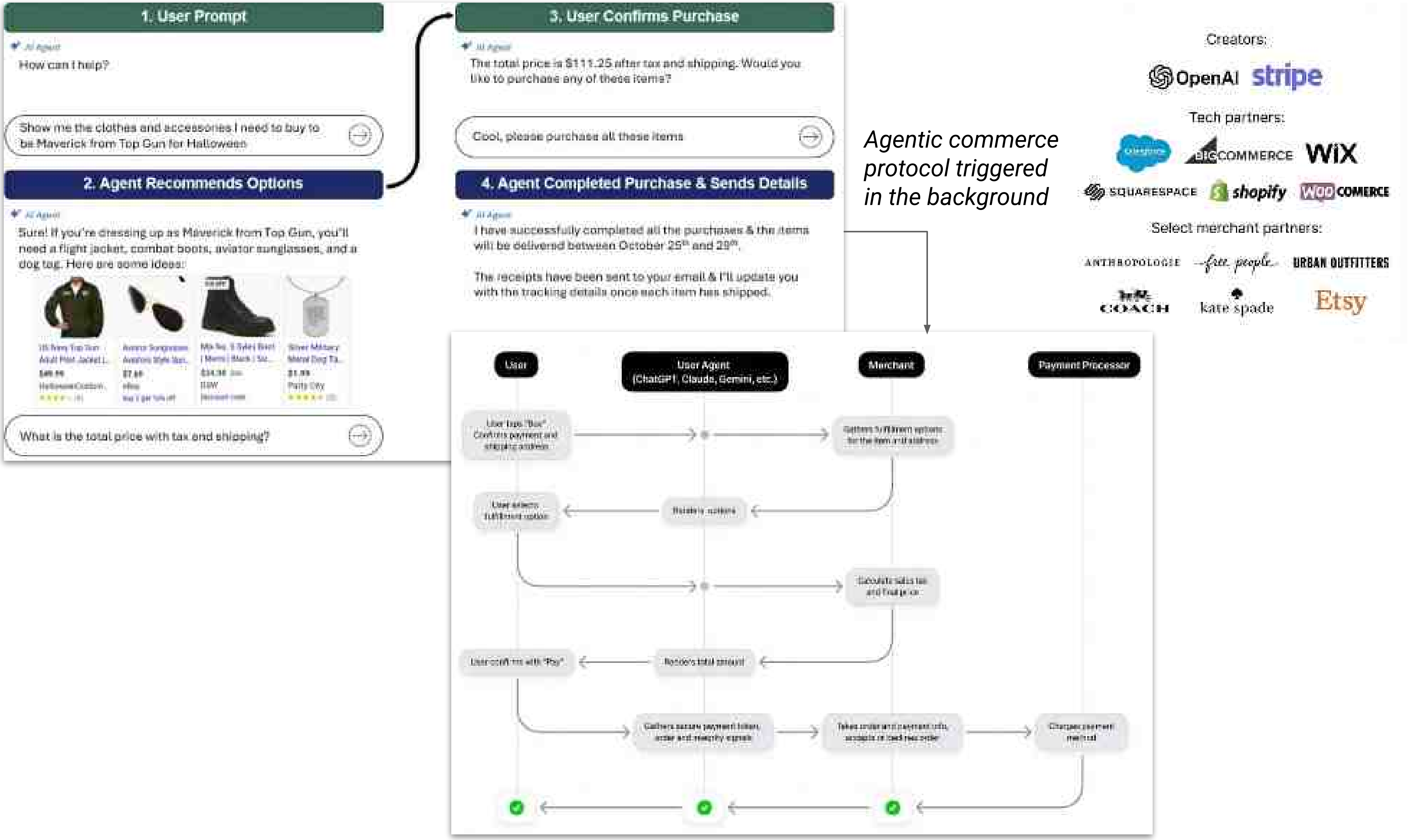


Source: OpenAI

Agentic commerce primer

Generational

Agentic commerce is when AI agents shop, negotiate, and transact on behalf of users—transforming shopping from discrete steps (search → browse → compare → buy) into a continuous, intent-driven flow. The infrastructure is already live. OpenAI and Stripe launched the Agentic Commerce Protocol (ACP)—an open standard enabling any AI agent to transact with any merchant. There are also competing protocols like Agent Payments Protocol (AP2) by Google and Trusted Agent Protocol by Visa.



1. User Prompt

AI Agent
How can I help?

Show me the clothes and accessories I need to buy to be Maverick from Top Gun for Halloween

2. Agent Recommends Options

AI Agent
Sure! If you're dressing up as Maverick from Top Gun, you'll need a flight jacket, combat boots, aviator sunglasses, and a dog tag. Here are some ideas:

US Navy Top Gun Adult Pilot Jacket L...
$49.99
HalloweenCostum...
Aviator Sunglasses Aviators Style Sun...
$7.69
eBay
Buy 1 get 10% off
Mix No. 6 Sylvia Boot | Men's | Black | Si...
$34.98
DSW
[illegible]
Silver Military Metal Dog Ta...
$1.99
Party City

What is the total price with tax and shipping?

3. User Confirms Purchase

AI Agent
The total price is $111.25 after tax and shipping. Would you like to purchase any of these items?

Cool, please purchase all these items

4. Agent Completed Purchase & Sends Details

AI Agent
I have successfully completed all the purchases & the items will be delivered between October 25th and 29th.

The receipts have been sent to your email & I'll update you with the tracking details once each item has shipped.

Agentic commerce protocol triggered in the background

User | User Agent (ChatGPT, Claude, Gemini, etc.) | Merchant | Payment Processor

User taps "Buy" Confirms payment and shipping address → Gathers fulfillment options for the item and address → Relays options → User selects fulfillment option → Calculates sales tax and final price → Renders total amount → User confirms with "Pay" → Gathers secure payment token, order and identity signals → Takes order and payment info, accepts or declines order → Charges payment method

Creators:

OpenAI stripe

Tech partners:

salesforce BIGCOMMERCE WiX
SQUARESPACE shopify WOO COMMERCE

Select merchant partners:

ANTHROPOLOGIE free people URBAN OUTFITTERS
COACH kate spade Etsy

Source: Morgan Stanley, OpenAI