

27 Apr 2026 16:37:38 ET | 61 pages

Artificial Intelligence

Anthropic: Driving the Acceleration of Enterprise AI

CITI'S TAKE

Anthropic is the leader in enterprise AI, a position it has cultivated through focus on commercial applications, particularly code development, safety/responsibility, and distribution across inference providers and developer platforms. Sustaining this lead will depend on maintaining technological and product strength while navigating challenges from increasingly enterprise-focused competitors, capacity constraints, and pricing (Figure 36) for both compute and models. With a \$4.2T+ TAM in AI models (Figure 3), we [continue to see](#) several large, durable enterprises emerge (Figure 33) from the current AI hyper-growth period, though the near-term will be defined by compute capacity constraints, capital requirements, enterprise adoption pace, and competition among AI labs, hyperscalers, and open-source providers. Having set the bar for performance and price with recent model releases (Figure 24) including Mythos, a new model class withheld from full release due to cybersecurity concerns, we expect Anthropic to continue gaining share through its singular focus on enterprise opportunities.

Product. Anthropic's strong product execution to date, most notably Claude Code, has compounded the share gains driven by its top-performing models. As agentic workflows diffuse across enterprises (Figure 2), the competitive landscape moves from benchmark performance to access to inference capacity, go-to-market resources, platform integration, and workflow lock-in. These dynamics are increasingly central to Anthropic's strategy.

Focus. Anthropic's early focus and continued prioritization of enterprise applications, particularly in code development, has enabled it to meaningfully capture high-value commercial AI use-cases, driving durable revenue growth at scale. While other major AI labs have recently shifted their focus to the enterprise, Anthropic's lead is considerable, based on its estimated 40% share of the enterprise LLM API market in 2025, per Anthropic investor Menlo Ventures ([Menlo Ventures](#), Dec. 9, 2025). The ability to maintain that lead will be determined by a competitive environment increasingly defined by capacity constraints, token economics, broadening user applications and access to capital.

Economics. The most important question in AI is the relationship between computational resources and performance. If Anthropic was able to train Mythos on "fairly mundane capacity, and a fairly mundane amount of it", as Nvidia CEO Jensen Huang said ([Dwarkesh Podcast](#), Apr. 15, 2026), this kind of infrastructure efficiency could redefine the economics of AI through algorithmic, rather than infrastructure, efficiency. This produces a structurally favorable setup for Anthropic, which has invested heavily in those dimensions.

Heath Terry^{AC}

Shelby Spencer

Ashley Kim

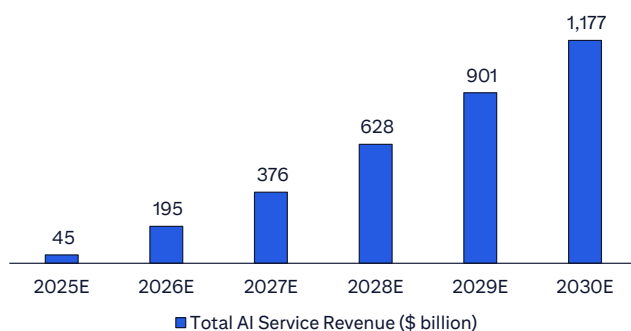
Janna Withrow

*See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations.

Portfolio Manager Summary

Anthropic is the leader in enterprise AI. As we wrote in August ([Artificial Intelligence: AI: The Information Era's Apex Technology](#)) we see the biggest opportunity in AI in the enterprise where we estimate a 90%+ CAGR over the next five years, growing into a \$1.2T market by 2030 (Figure 1). Whether purely by strategic focus, chance, or a combination of both, Anthropic's early focus on commercial applications, specifically in code development, has put it at the center of the accelerating adoption (Figure 2) of AI in the enterprise.

Figure 1. Citi AI Industry Service Revenue Forecasts Through 2030

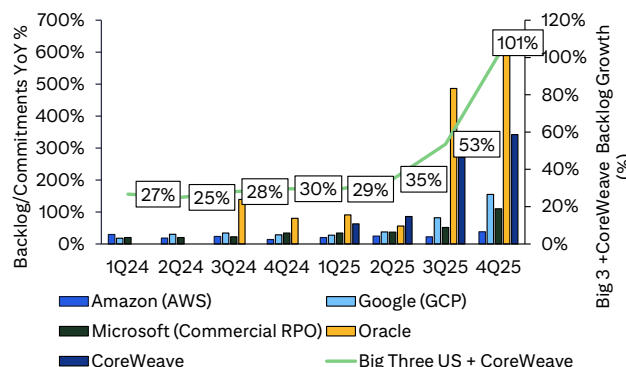


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: As of Mar. 26, 2026. We estimate AI service demand across the major foundation model providers based on a token production and token pricing methodology, leveraging public company commentary and reporting on revenue achievements, token processing volumes and growth. Our estimates intend to capture the totality of monetizable token production and does not claim to ascribe value beyond that frame. See additional details on methodology on the Cover page of [Artificial Intelligence: AI Industry Model Update](#).

Source: Citi Research

Figure 2. Hyperscaler Backlog Growth Shows Significant Demand Build-Up



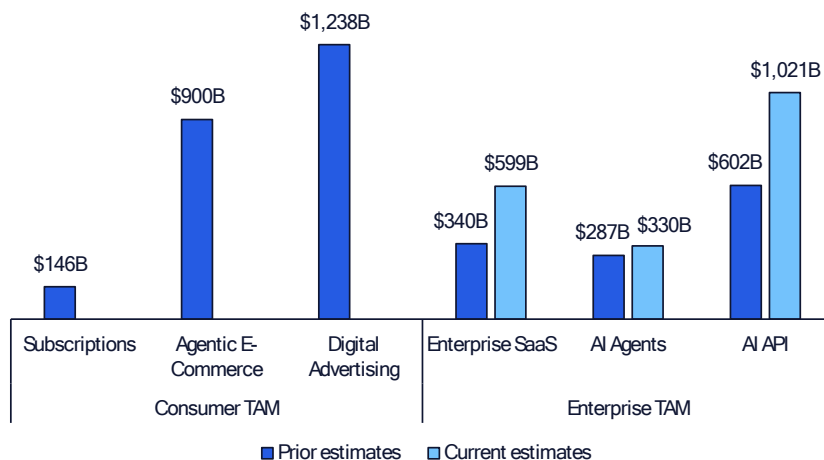
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Big Three are Amazon, Google and Microsoft

Source: Citi Research, Company Reports

While Anthropic's ability to maintain this lead as other AI labs reorient themselves to focus on the enterprise opportunity will depend on successfully maintaining its technological and product strength (Figure 3) while navigating challenges in go to market development, hyperscaler partnerships, capacity constraints, and pricing strategies, we believe that the \$4.2T+ addressable market in AI models will support multiple participants, not unlike the current cloud computing market. Our [earlier \\$3.5T+ AI TAM](#) for 2030 increases to \$4.2T+, driven by upward revisions to our Enterprise AI Subscriptions TAM (\$340B prior to new upper bound of ~\$600B), AI Agents TAM (\$287B prior to \$330B) and LLM APIs TAM (\$602B prior to new upper bound estimate of \$1.02T).

Figure 3. +4.2 Trillion AI TAM in 2030



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: See figures 17-22 for enterprise AI TAM calculations and methodologies. See [OpenAI: Architecting the Abstraction Layer for Everything](#) for consumer TAM calculations and methodologies.

Source: Citi Research, Bureau of Labor Statistics, ILO, Gartner, FactSet, NAICS Association, National Center for the Middle Market, Dun & Bradstreet

We believe that Anthropic's success will be determined by three primary factors:

Product. Despite being founded after other leading AI labs, Anthropic's models have shown to be competitive in objective benchmarks (Figure 24), particularly for highly commercial applications like coding, while its strong product execution to date, most notably Claude Code, has compounded the share gains driven by its top-performing models (Figure 23). The relevant competitive question is not which lab has the best model today, but how many months of lead any given lab holds at the capability tier that unlocks a new class of enterprise work. Mythos, still in private previews, is scoring significantly above competitors' recently released models, according to industry benchmark evaluations. That said, what matters more is Mythos's potential to move autonomous agentic work from "assisted" to "completed".

Model leadership has proven transient, with the most performant model typically being the most recently released. **In contrast, product leadership compounds. As agentic workflows diffuse across the enterprise, the competitive terrain will shift from benchmark performance to distribution, integration depth and workflow lock-in.**

Focus. We continue to see AI's primary value in the near term being realized at the enterprise level. The cost structure of AI, where inference costs scale with usage, means that monetization can't wait until companies have reached the billions of users that consumer "freemium" models have historically supported. **Anthropic's enterprise focus has proven to be an advantage, particularly in the highly economic field of code development which we estimate to be a \$800B+ enterprise AI opportunity.** This is a rough estimate derived from using Anthropic's traffic distribution associated with the Computer and Mathematical occupational category for 1P API (~60%) and Claude.ai (35%) (see Figure 42) as proxies for coding's share of our LLM API TAM estimate (Figure 18) and upper bound of our Enterprise AI Subscription TAM range (Figure 20). While other major AI labs have recently shifted their focus to the enterprise, Anthropic's lead is considerable,

based on market share. **The company's ability to maintain that focus amid a competitive environment increasingly defined by capacity constraints, token economics, broadening user applications and access to capital, will be an important competitive advantage.**

Economics. The massive gap between accelerating inference demand as reflected in hyperscaler backlogs (Figure 2) and the infrastructure needed to service it has made the ability to generate a [return on that investment](#) perhaps the primary determinant of competitive advantage among AI labs. While the price of tokens is declining as AI architecture continues to advance, frontier models are becoming increasingly token consumptive driving the price of output tokens higher (Figure 37).

Managing that spread, the main determinant of AI gross margins, will enable the AI labs without significant legacy advertising, software, or ecommerce businesses, to make the investments in infrastructure necessary to attract and scale customers. Anthropic leadership has spoken to the view that managing the pace of that investment and the risk associated with it will impact the race for AI leadership.

Remarkable Revenue Velocity, Tested Cost Structure

Anthropic's financial trajectory is one of the fastest in technology history. Anthropic's annualized revenue run-rate (ARR) has grown from \$1B in 2024 to \$30B by April 2026 (Figure 52), a 30x increase in ~15 months, driven by enterprise API adoption, the breakout success of Claude Code (\$2.5B ARR within nine months of launch), and broad distribution across all three major CSPs. This results in a ~13x EV/headline ARR at its February 2026 Series G valuation of \$380B (Figure 61).

The path to margin expansion and FCF generation will likely come under increasing focus as the company navigates the tension between explosive enterprise demand and the cost pressures of serving it at scale. Per [The Information](#) (Jan. 21, 2026), the company missed its 2025 gross margin target by 10pp (~40% realized vs. 50% projected), as inference costs ran materially above expectations against a backdrop of persistent supply chain constraints – HBM capacity, advanced packaging, power infrastructure – that keep spot prices elevated and readily-available compute scarce. This dynamic, coupled with a cash-flow break-even timeline that's been revised from 2027 to 2028 ([The Information](#), Jan. 27, 2026) underscores that the path to profitability is less a function of demand and more a function of whether infrastructure capacity can scale in line with it. Given the rapidly changing AI environment, we would caution against relying on any financial forecast from a specific point in time.

"We are a large consumer of Anthropic technology. [We] really admire their focus on security. Really admire their focus on safety."

– Jensen Huang, CEO, Nvidia, Mar. 19, 2026 on [All-In-Podcast](#)

Anthropic Background & Vision

Anthropic's Origins

To know Anthropic is to understand the central role safety and governance has played in the company's formation and strategic priorities. Founded in 2021 as a public benefit corporation by siblings Dario and Daniela Amodei, Anthropic emerged from growing concerns about AI alignment and safety, with the company's founders seeking to build safer, more interpretable and more steerable AI systems than existing approaches had prioritized ([Anthropic](#), Dec. 20, 2024).

Notably, the founding team includes several researchers and executives who worked on the development of GPT-2 and GPT-3 (released June 2020), whose release at the time validated the 'scaling laws' -- that scaling up model size and training data could lead to emergent abilities.

Seven person founding bench. While most start-ups have 1-3 founders, Anthropic's seven-person co-founder model appears deliberately designed to address the unique challenges of AI safety and development as each member of the team brings complementary expertise spanning research, policy, alignment and operations. The collective likely serves a governance function – creating built-in checks and balances that ensure multiple perspectives inform critical decisions related to model development and commercialization. Operationally, the company centers on the Amodeis who handle the brunt of the company's public facing engagements. While having seven co-founders could require more complex governance structures and potentially slow down decision-making processes that benefit from speed in early-stage companies, we do not see any meaningful evidence that Anthropic's choice to establish a deep founding bench has inhibited the company's ability to execute on its stated mission or gain market share in the enterprise.

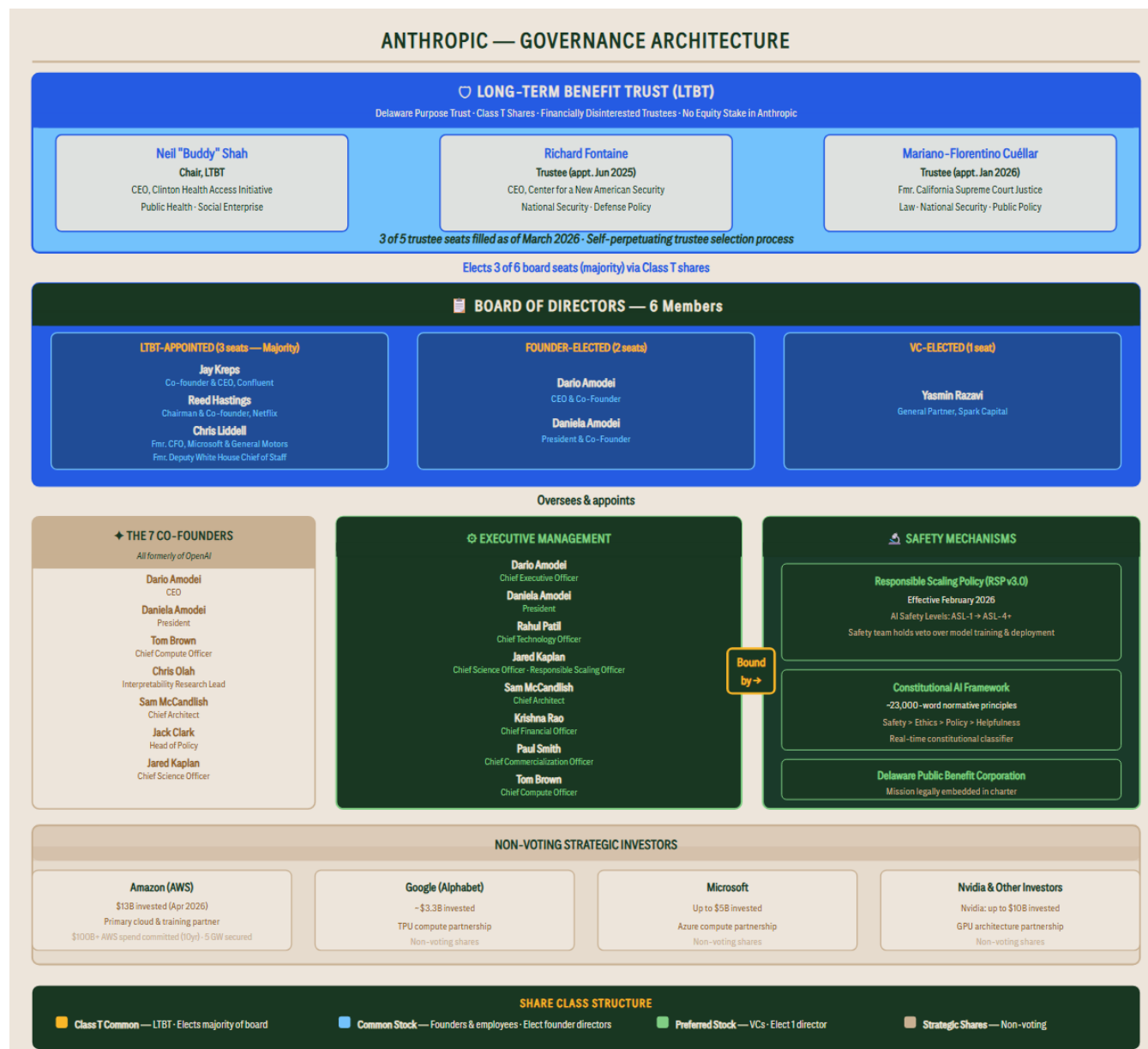
Governance

Like OpenAI, Anthropic operates as a public benefit corporation whose purpose is the responsible development and maintenance of advanced AI for the long-term benefit of humanity. This structure, along with the company's "experimental" [Long Term Benefit Trust](#) (LTBT), produces a distinctive structure which aims to align the company's governance with its mission (Anthropic, Sep. 19, 2023).

Anthropic's governance architecture reflects management's belief that the potential for extreme externalities resultant from AI requires dedicated handling with humanity's best interest in mind, rather than alternative priorities (Figure 4). Rather than concentrating control with founders through dual-class shares or ceding influence to lead investors proportional to capital deployed, Anthropic routes majority board authority through its LTBT, a financially disinterested independent body whose trustees hold zero equity in the company. In summary, LTBT trustees appoint majority board seats, with founders/VC accounting for the remainder. The \$6B fundraising threshold that triggered LTBT majority control was cleared in 2H24 – more than three years ahead of the original May 2027 deadline.

Figure 4. Anthropic's Bespoke Governance Architecture

as of April 2026



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Company Reports, The New York Times

We see three structurally distinctive features associated with this governance construction: (1) escalating control is tied to commercial milestones (since the LTBT's election rights increase as Anthropic raises capital, commercial success accelerates its authority rather than diluting it); (2) co-existence with conventional large scale venture investment despite non-voting participation; and (3) IPO survival design for Class T shares with no direct precedent in the public markets (though functionally analogous to a controlled company). Notably, the company has embedded a "failsafe" provision that allows for changes to the Trust and its powers WITHOUT consent of the Trustees if, "sufficiently large supermajorities of the stockholders agree."

“Our business model is straightforward: we generate revenue through enterprise contracts and paid subscriptions, and we reinvest that revenue into improving Claude for our users.”

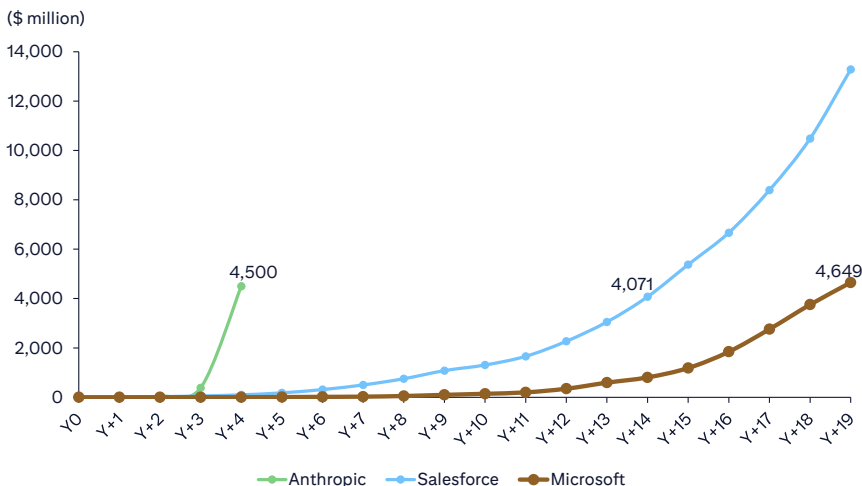
– Anthropic’s [website](#), Feb. 4, 2026

Rise in the Enterprise

Anthropic’s >10x annualized revenue run-rate growth in each of the past three years and continued acceleration through early April, with ARR crossing \$30B, has been supported by rapid adoption of its models by developers and enterprises ([Anthropic](#), Apr. 6, 2026). The company is estimated to have generated \$4.5B of revenue in 2025, four years after its founding, based on its internal projection ([The Information](#), Jan. 21, 2026). The exponential pace of Anthropic’s growth is notable, especially when compared to enterprise SaaS incumbents such as Microsoft, which took 20 years to reach this level of revenue, and Salesforce, which took 15 years. The slope of the curve speaks to the diffusion and distribution benefits made possible by the scaling of prior technologies such as cloud computing.

The company has two primary revenue drivers: API services and subscriptions. Roughly 86% of total revenue in 2025 was expected to come from its API business, according to the company ([The Information](#), Jan. 21, 2026). Additionally, CEO Amodei attributed roughly 80% of Anthropic’s business to enterprises and 20% to consumers, reflecting a focus on enterprises as a more predictable and stable source of income ([CNBC](#), Jan. 21, 2026).

Figure 5. Revenue Each Year (Y+N) Since Company Founding



© 2026 Citigroup Inc. No redistribution without Citigroup’s written permission.

Source: Citi Research, Company Reports, The Information

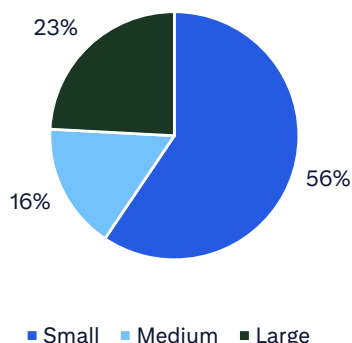
Customer Analysis

Our illustrative analysis of Anthropic’s published customer stories shows that **56% of customers are classified as “small” by the company and close to half of its sampled customer base operate in the software vertical.** Apart from its software concentration, Anthropic serves a diverse set of industries including Financial Services (9%), Education (6%), Professional Services (6%), Healthcare (5%), Cybersecurity (5%) and others. Geographically, ~70% of customers are in North America, 16% are in Europe and 9% are in Asia Pacific. About 40% of sampled customers fit into the “AI-native” bucket, similar in range to OpenAI’s 40% startup customer exposure and ~36% AI-native exposure based on our earlier analysis of [OpenAI’s customer composition](#). Heavy exposure to fast-scaling startups naturally carries risks to the upside (i.e. network effects) and downside (i.e. greater churn risks during macro slowdowns). This trade-off was clearly demonstrated in 2023,

impacting SaaS vendors with concentrated digital-native customer bases. Importantly, the above statistics represent the companies that Anthropic has chosen to highlight and thus isn't representative of any sort of average customer.

Figure 6. Anthropic Customer Mix by Company Size (%)

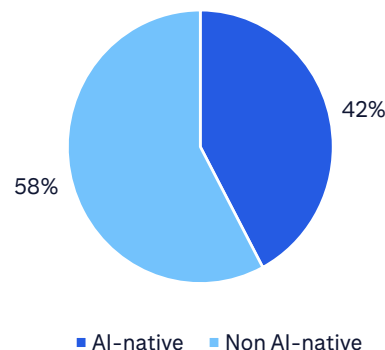
Based on 218 customer stories



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Company Reports

Figure 7. Anthropic Customer Mix of AI-Native vs. Non AI-Native (%)

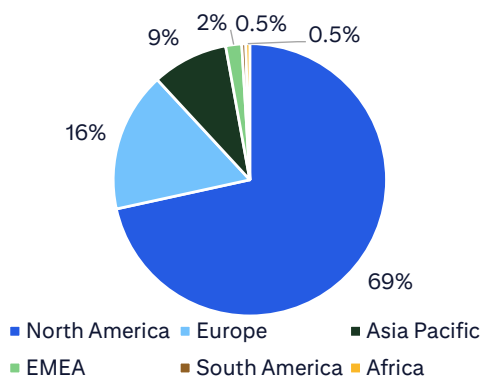
Based on 218 customer stories



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Company Reports

Figure 8. Anthropic Customer Mix by Geography (%)

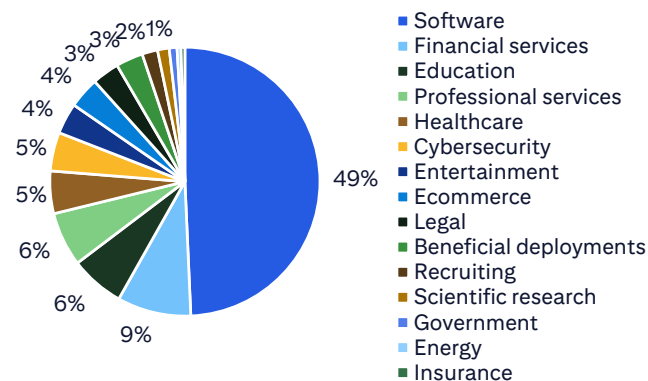
Based on 218 customer stories



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Company Reports

Figure 9. Anthropic Customer Mix by Vertical (%)

Based on 218 customer stories



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Company Reports

Compute Investment Strategy

Anthropic's approach to compute infrastructure is defined by three structural tensions: (1) a deliberate multi-cloud neutrality that preserves optionality at the cost of pricing power; (2) a self-imposed capital discipline that Amodei has publicly defended but which leaves the company with less locked-in capacity than its primary competitors; and (3) a cost structure that, even as per-token pricing for the same level of given intelligence falls industry-wide, remains exposed to capacity constraints that could slow enterprise adoption at the margin. While Anthropic elects for the first and second, many industry participants face pressure from the third.

Multi-cloud neutrality a deliberate strategy. Anthropic has avoided granting any single hyperscaler exclusive compute rights. Rather, the company distributes its commitments across all three hyperscalers (the first to do so), neoclouds, and data center developers. As such, no single provider controls Anthropic's compute stack – a strategy that stands in contrast to OpenAI's early and exclusive commitment to Azure, which secured locked-in compute at below-market rates (though this commitment has since evolved to allow for alternative compute providers). Reports indicate that the company believes working with all three CSPs provides distribution advantages over exclusivity tied peers ([The Information](#), Feb. 17, 2026).

Figure 10. Anthropic Compute Infrastructure Partnerships Summary

Partner	Deal Type	Amount (\$B)	Capacity (GW)	Key Details	Announced
 AWS Amazon	Equity investment + enhanced partnership	\$13 committed + up to \$20 milestone-contingent (total potential: \$33B)	5 GW secured ~1GW online end 2026; remainder 2027-2028	Amazon's total committed investment reaches \$13B (\$8B prior + \$5B Apr 2026), with up to \$20B more tied to commercial milestones (total potential: \$33B). Anthropic commits to spending \$100B+ on AWS over 10 years covering Trainium2, Trainium3, Trainium4, and Graviton. AWS designated primary cloud and primary training partner. Full Claude Platform available natively within AWS. Claude Opus 4.7 available on Amazon Bedrock (Apr 16, 2026).	Sep 2023 Mar 2024 Nov 2024 Apr 2026
 AWS Project Rainier	AI compute cluster	\$11 Part of \$100B+ 10yr commitment	2.2 GW full build Dedicated training capacity; incremental to 5 GW total	\$11B Indiana campus (New Carlisle); 30 planned data centers. Deployed 500K+ Trainium2 chips in 2025; scaling to 1M+ chips. Phase 1 operational at 455 MW; full build 2.2 GW. Project Rainier's 2.2 GW is dedicated model training capacity — incremental to the broader 5 GW total AWS capacity secured under the Apr 2026 enhanced partnership. Trainium3 capacity coming online 2026; Trainium4 on roadmap.	Sep 2024 Oct 2025 Apr 2026
 GCP Google Cloud	TPU purchase	\$21+ \$10B initial order (Dec 2025) + \$11B subsequent order	5 GW§ +1.5 GW incremental to prior 3.5 GW	Up to 1 million Google TPU v7 "Ironwood" chips via Broadcom (\$21B). >1GW of inference capacity online in 2026. Multi-gigawatt expansion announced Apr 2026 (starting 2027). Google Cloud capacity updated to 5 GW per Bloomberg reporting; see §. Inference—primary workload.	Oct 2025 Apr 2026
 GCP Google	Equity investment	up to \$40 \$10B cash at signing + up to \$30B milestone-contingent	—	Google announced up to \$40B investment in Anthropic (Apr 24, 2026), structured as \$10B in immediate cash plus up to \$30B contingent on performance milestones. Executed as a strategic extension of Anthropic's Series G at a \$350B pre-money valuation. Google's total cumulative capital commitment to Anthropic reaches ~\$43B (inclusive of \$3B+ invested since 2023). Google held ~14% stake as of Mar 2025; current stake subject to dilution. Google Cloud Vertex AI remains primary inference distribution channel.	Oct 2023 Mar 2025 Apr 2026
 Fluid Fluidstack	Data center development	\$50	Multiple GW†	Custom-built data centers in Texas and New York, with additional sites planned. First facilities expected online throughout 2026.	Nov 2025
 Azure Microsoft	Compute capacity	\$30	1+	\$30B Azure compute purchase commitment. Claude available via Azure AI Foundry, Copilot Studio, GitHub Copilot, and M365 Copilot. Claude designated as Microsoft subprocessor (Jan 2026). MACC-eligible.	Nov 2025
 NVIDIA Nvidia + Microsoft	Joint investment	\$15	1+‡	Nvidia commits up to \$10B; Microsoft up to \$5B in Anthropic. Joint engineering to optimize Claude for NVIDIA Grace Blackwell and Vera Rubin architectures. Up to 1GW of NVIDIA compute capacity via Azure.	Nov 2025
 CW CoreWeave	AI cloud services	n/d	n/d	Multi-year agreement for dedicated NVIDIA GPU-powered cloud infrastructure to run Claude models. Variety of NVIDIA chip architectures. Initially focused on a phased infrastructure roll-out. Financial terms and capacity not publicly disclosed.	Apr 10, 2026
Total Committed		\$143B+	14.2+ GW*	* Excludes Fluidstack and CoreWeave (capacity not publicly disclosed) and Amazon milestone-contingent \$20B. † Azure and Nvidia+Microsoft GW may overlap; non-additive. Per The Information (Feb 2026), Anthropic has privately discussed securing at least 10 GW of capacity over the next several years. § Google Cloud capacity reflects incremental +1.5 GW added to prior 3.5 GW commitment.	

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Google Cloud capacity of 5 GW per Bloomberg reporting (Apr 2026); it is unclear whether this figure is incremental to or inclusive of the existing 3.5 GW commitment — the compute subtotal reflects only the incremental 1.5 GW. All capacity figures are targets subject to construction and grid timelines. List may not be exhaustive.

Source: Citi Research, The Information, Bloomberg, Thomson Reuters, The New York Times, CNBC

As summarized in Figure 10, Anthropic's compute infrastructure spans three different AI server chips: Amazon's Trainium, Google's TPU, and NVIDIA's GPU portfolio. This strategy has helped their models operate more efficiently as the company assigns tasks to different chips depending on where the company

determines performance is best ([The Information](#), Nov. 10, 2025). The company's \$50B initiative with Fluidstack to build custom-designed data centers marks Anthropic's first foray into longer-termed owned infrastructure ([Anthropic](#), Nov. 12, 2025). Anthropic's deal with CoreWeave likely reflects the company's need to access readily deployable dedicated GPU clusters. We expect that the company exceeded its 10GWs of capacity target as of April 2026 ([The Information](#), Feb. 9, 2026), following its \$100B 10-yr commitment to AWS to secure up to 5GW of new capacity for training and inference ([Anthropic](#), Apr. 21, 2026).

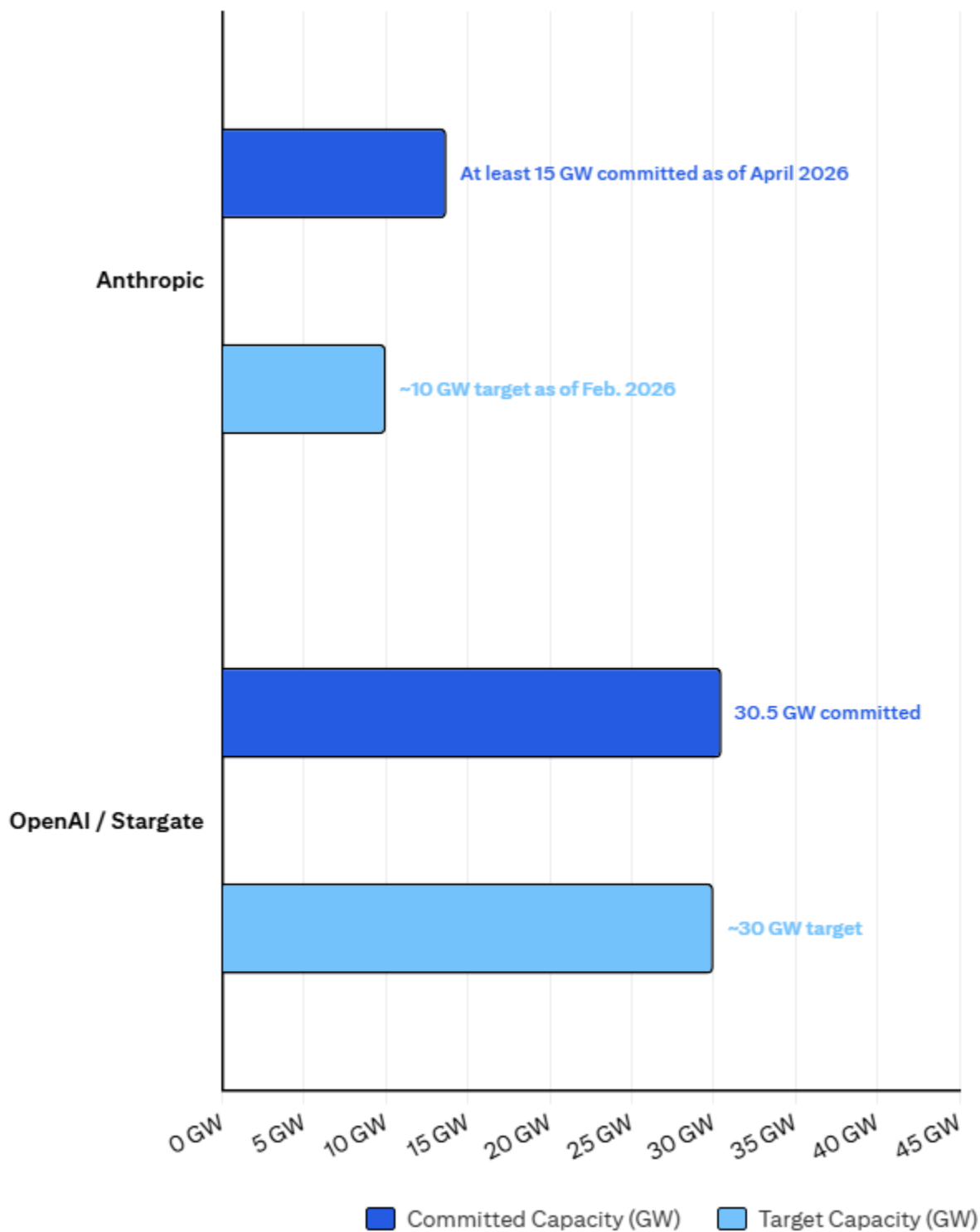
While the multi-architecture approach hedges against supply constraints in any single chip ecosystem and allows for broader enterprise distribution without exclusivity concessions, the company remains exposed to spot rates, which have been pushed higher due to supply chain tightness, resulting in even the oldest chips holding value ([Business Insider](#), Apr. 6, 2026). This dynamic likely contributed to the company lowering its 2025 gross margin expectations from 50% to 40% on higher inference costs – a pressure made more acute by the company's expected reseller distribution obligations ([The Information](#), Feb. 24, 2026).

“I could buy \$1T of compute that starts at the end of 2027. If my revenue is not \$1T, if it is even \$800B, there's no force on earth that could stop me from going bankrupt if I buy that much compute...I think it is true we're spending somewhat less than some of the other players.”

– Dario Amodei, CEO, Anthropic, Dwarkesh Podcast, Feb. 13, 2026

Capital Discipline versus Competitive Necessity. Dario Amodei has been candid about the existential arithmetic of the compute investment calculus facing frontier AI labs, warning that companies making massive infrastructure commitments risk insolvency if revenue timing slips by even a single year. This sentiment extends to a belief that large compute commitments made without rigorous revenue linkage signal financial recklessness rather than competitive ambition. Anthropic's ~15GW of committed compute as of April 2026, against OpenAI's 30+ GW commitment under Stargate, generally reflects this philosophy (Figure 11).

Figure 11. Compute Capacity Commitments for Anthropic and OpenAI



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

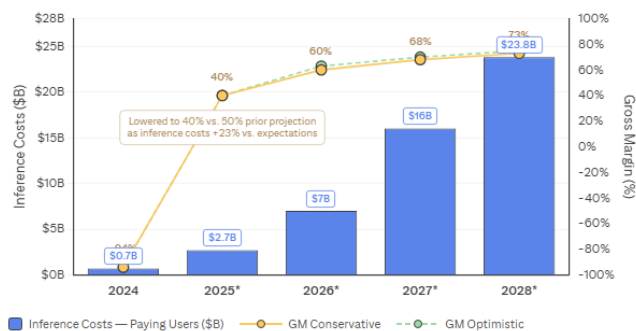
Includes CoreWeave (est. 1.0 GW) for Anthropic. Anthropic target of ~10 GW per The Information (Feb. 9, 2026); GW figures may overlap as vendors provide different components of the same data center infrastructure. Commitments may not be exhaustive.

Source: Citi Research, The Information, Company Reports

In a market where compute availability is itself a competitive determinative — for both model quality and inference capacity — a more conservative approach may constrain Anthropic's ability to serve demand at the pace its recent revenue trajectory implies. Anthropic's identifying capital discipline could prove prescient but the company's lowered 2025 gross margin expectations and ongoing compute capacity constraints suggest the balance between over investment and under investment remains difficult to calibrate in practice. Some Claude users have considered switching to Codex in response to Claude outages driven by constrained capacity (though both companies have faced similar challenges), a possibility made more probable by the release of GPT-5.5, which scores materially above Opus 4.7 and 3.1 Pro in Terminal-Bench 2.0 (Medium, Apr. 16, 2026).

Anthropic's recent postmortem may serve as a useful support for retention (Claude, Apr. 24, 2026). It traced recent Claude Code quality complaints to three changes affecting reasoning effort, cached thinking, and a verbosity-reducing system prompt. The detailed public accounting, the reset of usage limits for affected subscribers, and the commitments to broader eval coverage and gradual rollouts suggest that at least some of what users experienced may have been addressable engineering issues that Anthropic has now identified and corrected.

Figure 12. Anthropic Gross Margin & Inference Costs on Paying Users (\$B)

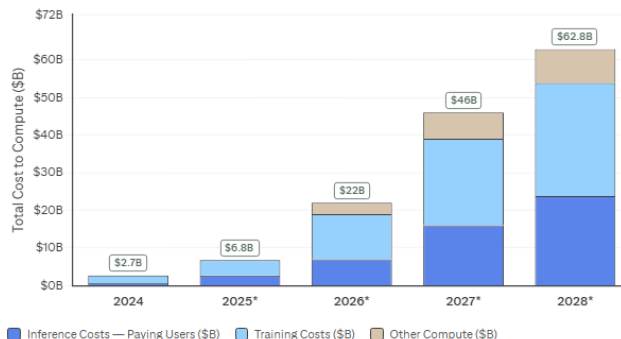


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

* Projected figures. Inference costs and gross margin projections per internal Anthropic estimates as reported by The Information (Jan. 2026). Gross margin defined as (Gross Revenue – Inference Costs on Paying Users – Other COGS) / Gross Revenue. Figures are not audited and should not be relied upon as confirmed financial data.

Source: Citi Research, The Information

Figure 13. Anthropic Total Cost to Compute (\$B)



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

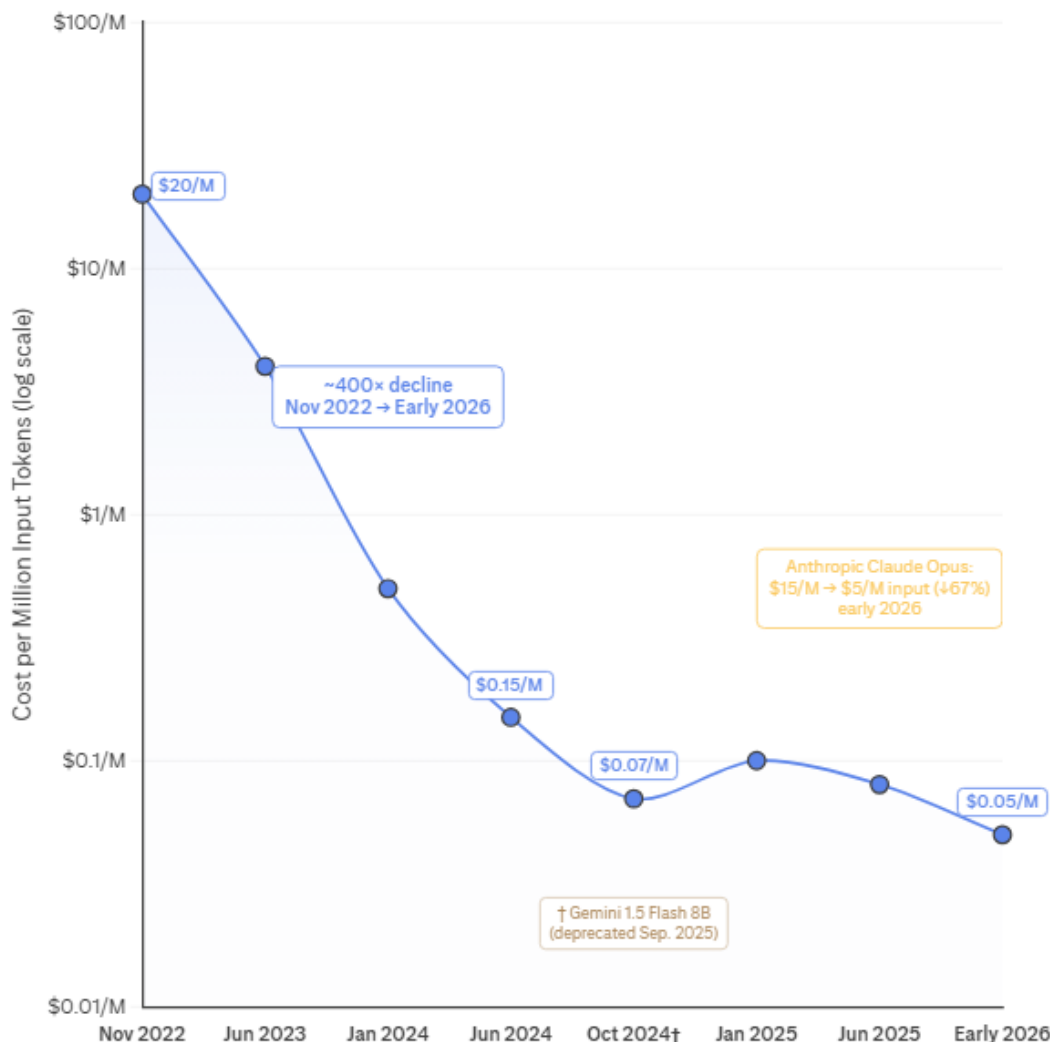
* Projected figures. Cost to compute data per internal Anthropic projections as reported by The Information (Jan. 2026). Training costs updated to reflect Dec-25 estimates. Other compute includes costs associated with buying chips. Figures are not audited and should not be relied upon as confirmed financial data.

Source: Citi Research, The Information

Pricing for a given level of intelligence is falling as capacity constraints pressure costs. The secular trend in AI inference pricing is unambiguously deflationary, with the cost per million tokens for GPT-3.5-equivalent performance declining approximately 280-fold from Nov. 2022 to Oct. 2024 (Stanford HAI, April 2025). On a like-for-like intelligence basis, AI is getting cheaper rapidly, and this trend is expected to continue as hardware efficiency improves and competition intensifies across the model tier spectrum.

Figure 14. Pricing for GPT-3.5- equivalent performance has declined ~400x since Chat GPT's debut in Nov. 2022

Cost per million input tokens — GPT-3.5-equivalent performance benchmark



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

GPT-3.5-equivalent benchmark tracks the lowest-cost model achieving 64.8% MMLU accuracy at each date. All prices are input token pricing only. Early 2026 reflects pricing for Gemini 2.5 Flash-Lite (\$0.05/M input, current).

Source: Citi Research, Epoch AI, Company Reports

That said, **this price descension is occurring against a backdrop of structural capacity constraints** — HBM memory sold out through 2026 ([CNBC](#), Jan. 10, 2026), [advanced packaging bottlenecks](#), onsite power infrastructure [lead times of 18-24 months](#) for natural gas reciprocating engines, among other constraints — that are keeping spot compute prices elevated which could limit Anthropic's ability to serve demand at the pace its revenue trajectory implies.

Industry Overview

A \$4.2T Overall AI TAM, \$1.9T Related to Enterprise AI

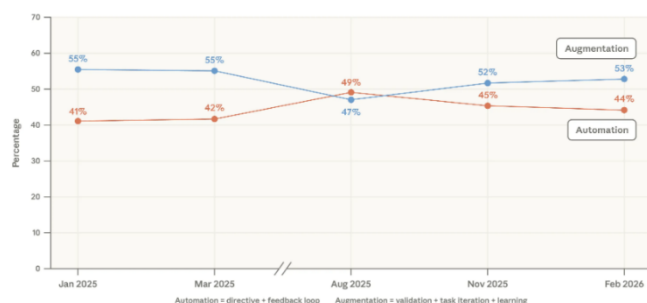
In the following sections, we revisit our earlier AI TAM analysis focusing on the markets for enterprise AI. Our [earlier \\$3.5T+ AI TAM](#) for 2030 increases to \$4.2T+, driven by upward revisions to our Enterprise AI Subscriptions TAM (\$340B prior to new upper bound of ~\$600B), AI Agents TAM (\$287B prior to \$330B) and LLM APIs TAM (\$602B prior to new upper bound estimate of \$1.02T). We view the overall AI industry opportunity through the lens of monetizable consumer and enterprise AI use-cases across enterprise software and services, infrastructure, advertising, commerce, hardware, and beyond. Steady breakthroughs in model [efficiencies](#) and [capabilities](#) are stimulating higher AI consumption across the global economy as consumer and enterprise adoption continues to accelerate. Within the enterprise domain where Anthropic is focused, companies are increasingly automating more work by developing and embedding agentic AI systems.

AI is becoming a coworker for coding and non-coding teams

The data from Anthropic's Economic Index studies over the past year reveals that users are gaining more comfort with autonomous AI behavior. "Directive" conversations with Claude jumped from 27% of interactions in January 2025 to 39% in August 2025, before moderating to 32% in November 2025, revealing a growing propensity to delegate tasks to AI tools ([Anthropic Economic Index Report](#), Jan. 15, 2026). Usage of Claude.ai is more evenly split between automation (e.g. full task delegation) vs. augmentation (e.g. "collaborative refinement and learning"), at 53% vs. 44% share of conversations while 1P API usage reveals a stronger preference for automation at 68% of transcripts vs. 17% for augmentation ([Anthropic Economic Index Report](#), Mar. 24, 2026). The contrast is even starker when it comes to economic tasks—97% of API usage was tied to automation vs. 47% for Claude.ai ([Anthropic Economic Index Report](#), Sep. 15, 2025).

Figure 15. Claude.ai automation vs. augmentation over time

Collaboration mode share over time by platform. Collaboration mode frequencies across Anthropic Economic Index Reports.



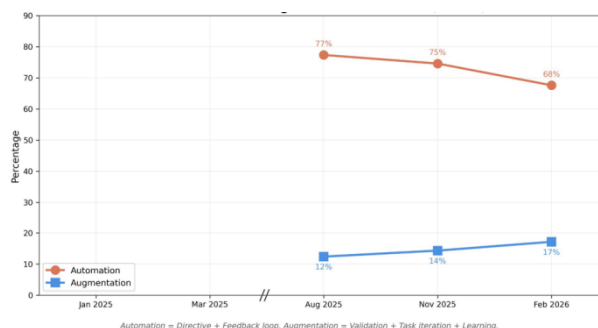
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Enter Footnote

Source: Citi Research, Anthropic

Figure 16. 1P API automation vs. augmentation over time

Collaboration mode share over time by platform. Collaboration mode frequencies across Anthropic Economic Index Reports.



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Enter Footnote

Source: Citi Research, Anthropic

LLM APIs a \$1T Market, Based on Top-Down Analysis

LLM APIs allow businesses to integrate frontier model capabilities and chatbots, like Claude and ChatGPT, with their own digital products/services and have become an important revenue source for model providers. During DevDay 2025, OpenAI handed out "Tokens of Appreciation" milestone award plaques to its most active developer customers that surpassed API token usage thresholds of 1T tokens

(Blue tier), 100B tokens (Black tier), and 10B tokens (Silver tier) ([OpenAI](#), Oct. 9, 2025). Microsoft and Google have referenced similar thresholds when providing token commentary. In their most recent earnings calls, Microsoft executives stated that over 250 customers were on track to process over 1T tokens on Foundry this year while Google executives shared that nearly 350 customers each processed over 100B tokens in December alone (implying a 1.2T token annual run-rate).

Using OpenAI's 1T/100B/10B token volume thresholds as a framework, we conduct a top-down analysis to estimate the global LLM API market opportunity in 2030. Our analysis is based on assumptions around the total volume of tokens processed via direct API calls and associated token pricing. We bifurcate the addressable market by large enterprise, mid-market and SMB customers, assuming 1T Tokens/Blue tier represents large enterprises, 100B Tokens/Black tier represents mid-market firms, and 10B Tokens/Silver tier represents startups/SMBs. We acknowledge there are exceptions to the mapping as fast scaling startups could be some of the largest token burning entities. For token volumes, we assume usage growth of 200% per year through 2030. On token pricing, we start with average blended token costs/1M tokens of \$2 for large enterprises, \$1.5 for mid-market firms, and \$0.5 for startups/SMBs in 2025 and forecast a 50% annual decline in token pricing through 2030. The outcome of this PxQ model gets us to ~\$1T in 2030, compared to our bottoms-up AI API forecast of ~\$600B from our [Artificial Intelligence - OpenAI: Architecting the Abstraction Layer for Everything](#) report.

Figure 17. Framework for Getting to Average Token Consumption per Firm

OpenAI Token Tiers for "Tokens of Appreciation" Milestone Awards

	Cumulative tokens used / firm (B)	Billion tokens used / firm / day	Billion tokens used / firm / year
1T Tokens (Blue)	1,000	1.143	417
100B Tokens (Black)	100	0.114	42
10B Tokens (Silver)	10	0.011	4

OpenAI API release date	3/1/2023
Cut-off date for token awards	7/23/2025
Days in between	875

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Methodology: For token assumptions, average daily token usage per firm is calculated by dividing the cumulative tokens consumed by the number of days between OpenAI's ChatGPT API release date (Mar. 1, 2023) and milestone awards cutoff date (Jul. 23, 2025). We then annualize the daily token consumption rate to arrive at annual token consumption estimates of ~400B for large enterprises, ~40B for mid-market firms, and ~4B for startups/SMBs.

Source: Citi Research, OpenAI

Figure 18. AI API TAM Analysis

Global addressable firm base	Est. # global registered businesses	% mix	Est. # that would use LLM APIs:	% of registered that would use APIs	Mapped Token Tiers
Large enterprise (1,000 employees+)	69,000	0.03%	62,100	90%	1T Tokens (Blue)
Mid-market (100-999 employees)	3,450,000	2%	1,035,000	30%	100B Tokens (Black)
SMB (1-99 employees)	226,481,000	98%	11,324,050	5%	10B Tokens (Silver)
Total	230,000,000	100%	12,421,150	5%	-

Token Tier	Firm Size	Annual token usage run-rate in 2025 (B)	Assumed token growth CAGR from 2025-30	Blended token price assumption in 2025	Token price reduction assumed by 2030
1T Tokens (Blue)	Large enterprise	400	200%	\$2.00	-97%
100B Tokens (Black)	Mid-market	40	200%	\$1.50	-97%
10B Tokens (Silver)	Startup/SMBs	4	200%	\$0.50	-97%

Token Tier	Firm Size	Token usage potential / firm in 2030 (B)	Addressable firms #	Blended token price (in 2030E)	TAM (\$B)
1T Tokens (Blue)	Large enterprise	97,200	62,100	\$0.06	\$377
100B Tokens (Black)	Mid-market	9,720	1,035,000	\$0.05	\$472
10B Tokens (Silver)	Startup/SMBs	972	11,324,050	\$0.02	\$172
Total:					\$1,021

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: the \$1T TAM is our upper bound estimate for the LLM API TAM, assuming rapid diffusion and growth of token intensive workloads. The annual token usage volume could vary significantly per company.

Source: Citi Research, NAICS Association, National Center for the Middle Market, Dun & Bradstreet

Based on 1P API traffic patterns for Claude, close to half of the LLM API market could be related to code-related tasks. Anthropic's Economic Index report showed that over 60% of 1P API traffic in February 2026 was concentrated in the Computer and Mathematical occupational category, an increase from ~55% in August (see Figure 42; [Anthropic](#), Mar. 24, 2026). Coding continues to represent the largest use-case for Claude.ai as well, representing 35% of conversations in February 2026, though this is down from 40% in March 2025 ([Anthropic](#), Mar. 24, 2026).

Enterprise AI Subscriptions TAM Potential of Up to ~\$600B

"There are two specific problems I am worried about: labor market displacement, and concentration of economic power. Let's start with the first one. This is a topic that I warned about very publicly in 2025, where I predicted that AI could displace half of all entry-level white collar jobs in the next 1-5 years, even as it accelerates economic growth and scientific progress."

– Dario Amodei, CEO, Anthropic in his "The Adolescence of Technology" essay, January 15, 2026

We refine our prior TAM analysis to consider different scenarios and outcomes in which 15-30% of global knowledge work is handled by AI by 2030. A new variable added to the sensitivity analysis is the subscription price per seat for Business and Enterprise tiers. We believe an argument can be made in either direction regarding the trajectory of AI subscription costs. On the one hand, the costs of AI chatbot services like Claude or ChatGPT could decline as more efficient models yield more abundant and cheaper units of intelligence. On the other hand, we believe it is possible that AI infrastructure and application providers would raise subscription prices over time as model capabilities improve and automate more white-collar work. This would follow how traditional SaaS vendors have justified past price increases by the added value associated with incremental bundled features.

Our analysis reveals a range of TAM outcomes between \$184B to \$599B in 2030. Under a scenario in which 20-25% of knowledge work is handled by AI and the business tier/enterprise tier subscription costs \$30/\$60 per seat per month, we expect a narrower TAM range of \$365B-\$389B. Our TAM range is based on the view that companies will choose one primary AI partner to provide employees with an everyday AI assistant like Claude or ChatGPT, though it may understate the opportunity as certain companies could opt for a multi-vendor strategy in

enterprise IT procurement. We believe the prosumer market, not captured below, could also represent a sizeable near-term opportunity. Employees who don't yet have access to their preferred AI chat interface may choose to pay for a consumer subscription in the meantime and utilize it for personal and work purposes. Tellingly, Anthropic's Mar. 2026 AI Economic Index report revealed that Claude.ai usage comprises 45% work, 12% coursework, and 42% personal use-cases.

Additionally, based on the Bureau of Labor's 2024 employment statistics, we estimate that close to 60% of the US TAM opportunity could be tied to industries with a high percentage of white-collar work such as Healthcare, Government, Professional services, Finance/Insurance and Real Estate, where we are seeing early AI adoption and concerted GTM efforts by AI players.

Figure 19. Enterprise AI Subscriptions TAM and Industry Breakdown

	US	Global
Knowledge workers estimate (mn)	100	1,000
% access to productivity suite	95%	70%
Addressable knowledge workers (mn)	95	700
Business Tier subscription pricing (\$/mo./user)	\$30	\$30
SMB mix %	45%	65%
Enterprise Tier subscription pricing (\$/mo./user)	\$60	\$60
Enterprise mix %	55%	35%
Enterprise AI subscriptions TAM (\$ mn)	\$53,010	\$340,200

	US TAM (\$ mn)	Employment (2024, mn)
Healthcare and social assistance	\$13,220	23.7
Government, excluding state/local education, hospitals	\$6,399	11.5
Professional, scientific, and technical services	\$6,041	10.8
Finance and insurance	\$3,740	6.7
Real estate and rental and leasing	\$1,376	2.5
5 industries total	\$30,776	55.2
% of total	58%	55%

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Methodology: Our enterprise AI subscriptions TAM is a function of the addressable knowledge worker base, SMB/enterprise mix and business/enterprise tier monthly pricing plans. The top-down projection starts with Gartner and ILO informed estimates for US and global knowledge worker bases and applying assumptions for digital work adoption rates. Business tier pricing is informed by published pricing of leading AI subscription vendors incl. Microsoft, OpenAI, Google, and Anthropic, while enterprise tier pricing is informed by what's been reported by TechCrunch. Enterprise pricing is typically custom quoted and depends on company size, usage, context window sizes, security requirements, support SLAs, and other negotiation terms therefore is often not publicly disclosed as a single fixed number. Many companies may bundle AI capabilities within broader productivity suites, which means the AI cost is part of a larger service contract. We make assumptions around the SMB vs. enterprise mix for the US and globally. The US TAM breakdown for specific verticals relies on 2024 employment estimates from the Bureau of Labor Statistics.

Source: Citi Research, Bureau of Labor Statistics, ILO, Gartner

Figure 20. Enterprise AI Subscriptions TAM Under Different AI Labor Impact Scenarios

Addressable knowledge worker pool in 2030 under different scenarios (millions)								
		% of Knowledge Work Handled by AI in 2030E						
		10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%
SMBs		585	553	520	488	455	423	390
Enterprises		315	298	280	263	245	228	210
Total		900	850	800	750	700	650	600

SMB / Business Tier AI Subscription TAM (\$B)								
		% of Knowledge Work Handled by AI in 2030E						
		10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%
Business Tier pricing (\$/mo./user)	\$15	\$105	\$99	\$94	\$88	\$82	\$76	\$70
	\$18	\$123	\$116	\$109	\$102	\$96	\$89	\$82
	\$20	\$140	\$133	\$125	\$117	\$109	\$101	\$94
	\$23	\$158	\$149	\$140	\$132	\$123	\$114	\$105
	\$25	\$176	\$166	\$156	\$146	\$137	\$127	\$117
	\$28	\$193	\$182	\$172	\$161	\$150	\$139	\$129
	\$30	\$211	\$199	\$187	\$176	\$164	\$152	\$140
	\$33	\$228	\$215	\$203	\$190	\$177	\$165	\$152
	\$35	\$246	\$232	\$218	\$205	\$191	\$177	\$164
	\$38	\$263	\$249	\$234	\$219	\$205	\$190	\$176
	\$40	\$281	\$265	\$250	\$234	\$218	\$203	\$187
	\$43	\$298	\$282	\$265	\$249	\$232	\$215	\$199
	\$45	\$316	\$298	\$281	\$263	\$246	\$228	\$211

Enterprise Tier AI Subscription TAM (\$B)								
		% of Knowledge Work Handled by AI in 2030E						
		10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%
Enterprise Tier pricing (\$/mo./user)	\$45	\$170	\$161	\$151	\$142	\$132	\$123	\$113
	\$48	\$180	\$170	\$160	\$150	\$140	\$130	\$120
	\$50	\$189	\$179	\$168	\$158	\$147	\$137	\$126
	\$53	\$198	\$187	\$176	\$165	\$154	\$143	\$132
	\$55	\$208	\$196	\$185	\$173	\$162	\$150	\$139
	\$58	\$217	\$205	\$193	\$181	\$169	\$157	\$145
	\$60	\$227	\$214	\$202	\$189	\$176	\$164	\$151
	\$63	\$236	\$223	\$210	\$197	\$184	\$171	\$158
	\$65	\$246	\$232	\$218	\$205	\$191	\$177	\$164
	\$68	\$255	\$241	\$227	\$213	\$198	\$184	\$170
	\$70	\$265	\$250	\$235	\$221	\$206	\$191	\$176
	\$73	\$274	\$259	\$244	\$228	\$213	\$198	\$183
	\$75	\$284	\$268	\$252	\$236	\$221	\$205	\$189

Total AI Subscription TAM (\$B)								
		% of Knowledge Work Handled by AI Agents in 2030E						
		10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%
Blended pricing (\$/mo./user)	\$26	\$275	\$260	\$245	\$230	\$214	\$199	\$184
	\$28	\$302	\$286	\$269	\$252	\$235	\$218	\$202
	\$31	\$329	\$311	\$293	\$275	\$256	\$238	\$220
	\$33	\$356	\$337	\$317	\$297	\$277	\$257	\$238
	\$36	\$383	\$362	\$341	\$320	\$298	\$277	\$256
	\$38	\$410	\$388	\$365	\$342	\$319	\$296	\$274
	\$41	\$437	\$413	\$389	\$365	\$340	\$316	\$292
	\$43	\$464	\$439	\$413	\$387	\$361	\$335	\$310
	\$46	\$491	\$464	\$437	\$410	\$382	\$355	\$328
	\$48	\$518	\$490	\$461	\$432	\$403	\$374	\$346
	\$51	\$545	\$515	\$485	\$455	\$424	\$394	\$364
	\$53	\$572	\$541	\$509	\$477	\$445	\$413	\$382
	\$56	\$599	\$566	\$533	\$500	\$466	\$433	\$400

Methodology: The above TAM tables leverage assumptions from the prior table around total global knowledge workers and SMB vs. enterprise mix percentages. SMB/Business Tier AI Subscription TAM formula is $((\text{global knowledge worker count} \times \text{SMB mix \%}) \times (1 - \% \text{ of knowledge work handled by AI in 2030E})) \times (\text{monthly business tier subscription pricing} \times 12) / 1000$. Enterprise Tier AI Subscription TAM formula is $((\text{global knowledge worker count} \times \text{enterprise mix \%}) \times (1 - \% \text{ of knowledge work handled by AI in 2030E})) \times (\text{monthly enterprise tier subscription pricing} \times 12) / 1000$.

Source: Citi Research

“A ‘country of geniuses in a data center’ could represent the largest change to the global labor market in human history...”

– Dario Amodei, Anthropic CEO, on Dwarkesh Podcast, Feb. 13, 2026

“Country of Geniuses in a Data Center”

CEO Dario Amodei believes that there is a 50:50 chance that we are 1–3 years away from AGI (artificial general intelligence) and is 90% confident that it will be achieved within 10 years. He expects this milestone to be marked by AI capabilities reaching a level of a “country of geniuses in a data center,” which he defines as “a set of AI agents that are more capable than most humans and most things and can coordinate at superhuman speed.”

~\$330B Agents TAM Reflects a More Nascent Market

Growing demand for autonomous AI agents may drive another step-function increase in enterprise AI adoption. Though the market is still nascent, LangChain’s year-end survey of 1,300+ professionals across different company sizes and industries revealed that 57% of respondents already had agents running in production environments in 2025 (up vs. 51% in 2024) while 30% were actively developing agents with plans for production ([LangChain State of Agent Engineering Report](#), Dec. 16, 2025). As tools and agentic frameworks mature, and the unit of automation evolves from tasks to entire roles, agentic systems are expected to fundamentally reshape how businesses leverage AI. Model Context Protocol (MCP), introduced by Anthropic in Nov. 2024, and Agent-to-Agent (A2A) protocol, introduced by Google in Apr. 2025, are examples of critical frameworks enabling the development of multi-agent systems for both enterprise and consumer applications. MCP standardizes tool access (e.g. APIs, databases, SaaS systems) for LLMs/agents, while A2A standardizes agent-to-agent communication and coordination involving delegation, negotiation, shared states and multi-step workflows.

In [Artificial Intelligence – OpenAI: Architecting the Abstraction Layer for Everything](#) we estimated **\$290B** in potential AI agent spend in 2030 in a 5% GDP growth CAGR and 15% “moderate GenAI impact on labor” scenario. We refine our Total Addressable Market (TAM) analysis, projecting what agentic system spending could look like across diverse department functions and firm sizes. This involves key assumptions regarding department mix, FTE reduction, and agent costs as a percentage of salaries. Our bottom-up analysis gets us to a \$330B TAM.

Figure 21. AI Agent TAM – Key Assumptions

Department function	Distribution	FTE Reduction %	Average annual salary	Agent Cost % of FTE	Roles
Finance	5%	40%	\$125,000	5%	Close cycles, reconciliations, etc. become agent-first
Support	15%	50%	\$80,000	5%	L1/L2 handled by agents; humans handle exceptions
HR Ops	3%	30%	\$105,000	5%	Onboarding, payroll prep, policy routing
IT/Ops	5%	15%	\$130,000	5%	Humans shift to governance, policy, agent ops
Sales/BD	20%	15%	\$150,000	5%	Agents handle research prep, follow-up; humans handle relationships
Legal/Compliance	2%	20%	\$160,000	5%	Agents do monitoring, drafting, checks; humans handle judgment related work

# Employees	Midpoint size	Addressable # firms
51-200	125	170,245
201-500	350	95,290
501-1,000	750	41,966
1,001-5,000	3,000	47,428
5,001-10,000	7,500	8,244
10,000+	15,000	8,836

© 2026 Citigroup Inc. No redistribution without Citigroup’s written permission.

Methodology: Distribution column shows headcount mix assumptions by function. Our analysis does not include every business function. FTE reduction % shows assumptions around the reduction in full-time employees for each function by 2030 due to potential agentic AI-related impacts. Annual salary assumptions are taken from Indeed.com. We assume an agent in 2030 will cost 5% of a FTE salary today. The addressable firm numbers are taken from The Org's company counts based on employee ranges.

Source: Citi Research, The Org, Indeed.com

Figure 22. AI Agent TAM - Sizing per Function and Firm Size

# Full-Time Employees (FTE)	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	6	18	38	150	375	750
Support	19	53	113	450	1,125	2,250
HR Ops	4	11	23	90	225	450
IT/Ops	6	18	38	150	375	750
Sales/BD	25	70	150	600	1,500	3,000
Legal/Compliance	3	7	15	60	150	300

# FTE Reduction	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	2	7	15	60	150	300
Support	10	27	57	225	563	1,125
HR Ops	1	3	7	27	68	135
IT/Ops	1	3	6	23	56	113
Sales/BD	4	11	23	90	225	450
Legal/Compliance	1	1	3	12	30	60

# FTE in 2030	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	4	11	23	90	225	450
Support	10	27	57	225	563	1,125
HR Ops	3	8	16	63	158	315
IT/Ops	5	15	32	128	319	638
Sales/BD	21	60	128	510	1,275	2,550
Legal/Compliance	2	6	12	48	120	240

Labor Savings (\$)	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	\$300,000	\$900,000	\$1,900,000	\$7,500,000	\$18,750,000	\$37,500,000
Support	\$760,000	\$2,120,000	\$4,520,000	\$18,000,000	\$45,000,000	\$90,000,000
HR Ops	\$126,000	\$346,500	\$724,500	\$2,835,000	\$7,087,500	\$14,175,000
IT/Ops	\$117,000	\$351,000	\$741,000	\$2,925,000	\$7,312,500	\$14,625,000
Sales/BD	\$562,500	\$1,575,000	\$3,375,000	\$13,500,000	\$33,750,000	\$67,500,000
Legal/Compliance	\$96,000	\$224,000	\$480,000	\$1,920,000	\$4,800,000	\$9,600,000

Agent System Costs (\$)	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	\$15,000	\$45,000	\$95,000	\$375,000	\$937,500	\$1,875,000
Support	\$38,000	\$106,000	\$226,000	\$900,000	\$2,250,000	\$4,500,000
HR Ops	\$6,300	\$17,325	\$36,225	\$141,750	\$354,375	\$708,750
IT/Ops	\$5,850	\$17,550	\$37,050	\$146,250	\$365,625	\$731,250
Sales/BD	\$28,125	\$78,750	\$168,750	\$675,000	\$1,687,500	\$3,375,000
Legal/Compliance	\$4,800	\$11,200	\$24,000	\$96,000	\$240,000	\$480,000
Total	\$98,075	\$275,825	\$587,025	\$2,334,000	\$5,835,000	\$11,670,000

Agent System TAM (\$B)	51-200	201-500	501-1,000	1,001-5,000	5,001-10,000	10,000+
Finance	\$3	\$4	\$4	\$18	\$8	\$17
Support	\$6	\$10	\$9	\$43	\$19	\$40
HR Ops	\$1	\$2	\$2	\$7	\$3	\$6
IT/Ops	\$1	\$2	\$2	\$7	\$3	\$6
Sales/BD	\$5	\$8	\$7	\$32	\$14	\$30
Legal/Compliance	\$1	\$1	\$1	\$5	\$2	\$4
Total TAM by firm size category (\$B)	\$17	\$26	\$25	\$111	\$48	\$103
Total TAM (\$B)	\$330					

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Methodology: # FTEs are calculated by taking the midpoint employee count for different firm sizes (e.g. 125 for 51-200 employee range, 350 for 201-500, etc.) multiplied by the distribution percentages for each function. # FTE Reduction takes the # FTEs multiplied by the FTE reduction percentage assumptions from the prior table, with # FTEs in 2030 showing the difference of the two. Labor Savings (\$) is a function of FTE reduction * average annual salary assumptions in the prior table. Agent system cost is a function of Labor Savings*Agent Cost % of FTE. Agent System TAM is a function of agent system costs*addressable firms per function and firm size.

Source: Citi Research, FactSet

The Two Exponentials

In a recent interview, CEO Dario Amodei described two simultaneous exponentials: 1) capability exponential, and 2) diffusion exponential ([Dwarkesh Podcast](#), Feb. 13, 2026). Capability exponential describes model improvements that align with scaling laws: increased compute directly translates to gains in intelligence. Diffusion exponential refers to the rate at which AI capabilities are being integrated with real economic activity. While AI diffusion outpaces prior technology cycles, its adoption still lags its true capability. This gap is driven by significant constraints: organizational change management, complex security permissions, stringent legal and compliance approvals, and the substantial time and training required for new workflows. The growing "[capability overhang](#)"—a term coined by OpenAI executives—describes the widening gap between AI's potential and its practical deployment (OpenAI, Jan. 18, 2026). This should diminish as companies collaborate with forward-deployed engineers and solution architects from model providers and Global System Integrators (GSIs) to establish robust data, security, and governance foundations crucial for agentic workflows.

Anthropic Positioning in a Competitive Landscape

The AI industry has entered a phase where the basis of competition is shifting from generalized model intelligence to productization and go to market strategy. This is reflected in the increasing focus among AI lab leadership on providing enterprise tools (specifically around code development), forward deployable engineers, and the inference capacity necessary to deploy tools and agents at scale.

Through 2024 and much of 2025, the defining question was which lab could produce the most capable foundation model. That question has not become irrelevant, but it has become insufficient. Benchmark convergence across frontier providers and the rapid closing of the open-source capability gap has compressed the window in which raw model quality alone can sustain a durable competitive advantage (Figure 23; [Artificial Intelligence: OpenAI: Architecting the Abstraction Layer for Everything](#)).

Figure 23. Frontier Model Leaderboard

Model Provider	Leading Model	Artificial Analysis Intelligence Index	Speed (Output Tokens/Second)	Blended Price (USD/1M Tokens)	Context Window (1K Words)
OpenAI	GPT-5.5 (xhigh)	60	74	\$11.25	922
Google	Gemini 3.1 Pro Preview	57	116	\$4.50	1000
Anthropic	Claude Opus 4.7 (max)	57	48	\$10.00	1000
Moonshot AI	Kimi K2.6	54	136	\$1.71	256
Xiaomi	MiMo-V2.5-Pro	54	62	\$1.50	1000
DeepSeek	DeepSeek V4 Pro (Max)	52	37	\$2.17	1000
Alibaba	Qwen3.6 Max Preview	52	34	\$2.92	256
Z.ai	GLM-5.1	51	51	\$2.15	200
MiniMax	MiniMax-M2.7	50	50	\$0.53	205
NVIDIA	NVIDIA Nemotron 3 Super	36	155	\$0.41	1000
Amazon	Nova 2.0 Pro Preview (medium)	36	139	\$3.44	256
Inception	Mercury 2	33	702	\$0.38	128
Arcee AI	Trinity Large Thinking	32	117	\$0.40	512
Mistral	Mistral Small 4	28	160	\$0.26	256
InclusionAI	Ling 2.6 Flash	26	205	\$0.15	262
Meta	Llama 4 Maverick	18	110	\$0.47	1000
Baidu	ERNIE 4.5 300B A47B	15	23	\$0.48	131
Cohere	Command A	13	44	\$4.38	256
IBM	Granite 4.0 H Small	11	313	\$0.11	128
AI21 Labs	Jamba 1.7 Large	11	60	\$3.50	256

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Artificial Analysis has not yet released output speeds for GPT-5.5.

Source: Citi Research, Artificial Analysis (as of Apr. 27, 2026)

We believe three interrelated dynamics will define in the competitive landscape for AI Labs through 2026.

1. **The migration of value from standalone models to integrated workflows, as exemplified by Anthropic's progression from API to Claude Code to Cowork.** As base model quality converges across providers, costs are shifting up the stack from inference endpoints to the orchestration layers, product surfaces, and workflow integrations that sit between the model and the end user. We expect providers that own the workflow surface area (e.g., the IDE, the desktop agent, the enterprise connector, etc.) will capture disproportionate value. This strategy seems to underpin Anthropic's product expansion from API to Claude Code to Cowork. Each layer adds switching costs and deepens the relationship with the customer's actual work.
2. **The shift from copilots to agents as a product paradigm.** 2025 was the year agentic AI moved from research demos to early pilot deployments. 2026 is shaping up to be the year of enterprises standardizing and building guardrails on agentic platforms for real workloads like coding, editing, customer service,

and knowledge work. The distinction lies in how copilots suggest while agents execute. That is, copilots augment individual tasks; agents take ownership of multi-step workflows, coordinate across systems, and operate with bounded autonomy over extended time horizons. This transition fundamentally changes the competitive calculus. Building a useful copilot requires a good model. Building a reliable agent requires a good model plus tool integration, memory management, error recovery, safety guarantees, deep understanding of domain-specific workflows, and so on. Beyond simply meeting the next benchmark, the companies solving these systems-level challenges in partnership with enterprises are defining the next era of enterprise AI. The January 2026 launch of Claude Cowork as well as Claude Code's \$2.5B in run-rate revenue in February 2026 represent early evidence that the agentic product thesis can generate revenue at scale ([Anthropic](#), Jan. 26, 2026; [Anthropic](#), Feb. 12, 2026). As of April 2026, the run-rate revenue of the entire company exceeds \$30B, up from \$9B in December 2025 ([Anthropic](#), Apr. 6, 2026).

3. **The emergence of coding as the first vertical to be structurally transformed by Claude may serve as a template for what follows.** Software development is the leading indicator for AI's broader impact on knowledge work. It is the vertical where AI capabilities are most measurable, where testing and verification are most natively accessible, where ROI is most immediately quantifiable, and where adoption has moved fastest from experimentation to production. In the six months after its launch, Claude Code achieved \$1B in run-rate revenue within six months, a pace on par with even ChatGPT's early consumer ramp of \$2B ARR in 2023, \$6B in 2024, and \$20B+ in 2025 since the launch of its ChatGPT app in 2022 ([Anthropic](#), Dec. 3, 2025; [OpenAI](#), Jan. 18, 2026). Claude Code's trajectory demonstrates that when AI can demonstrably do the work, users and enterprises look to pay for it at scale. 2026 and 2027 may see this pattern replicate across adjacent verticals like legal, finance, healthcare, security and design, each following a similar arc from pilot to production as agentic tools mature for domain-specific workflows. The competitive question for each AI provider, including Anthropic, will be whether they can extend their coding foothold into these new verticals before incumbents or new entrants establish their own footholds.

The past year has been marked by an acceleration in deployment as corporations continue to leverage AI to do more with less. For Anthropic, that adoption has meant a 3x increase in signed 8-9 figure deals and 7x growth in customers with ARR >\$100K in 2025 versus 2024 ([Venture Beat](#), Aug. 8, 2025; [Anthropic](#), Feb. 12, 2026). The number of customers spending over \$1M has also expanded from a dozen two years ago to 500 in mid-February, which then doubled to more than 1,000 in early April ([Anthropic](#), Feb. 12, 2026; [Anthropic](#), Apr. 6, 2026).

Figure 24. Frontier Model Performance Across Key Benchmarks

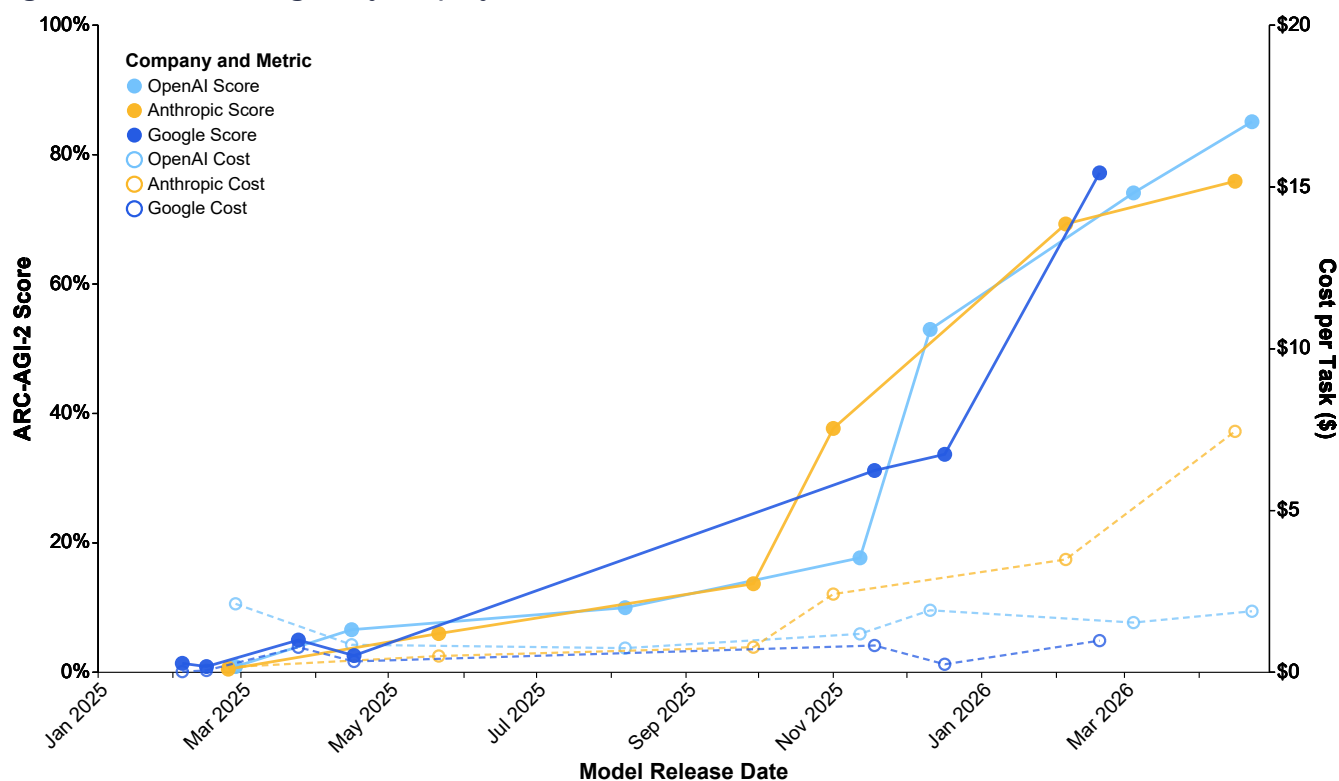
Release Date		4/23/2026	4/16/2026	4/7/2026	3/5/2026	2/19/2026	2/5/2026	12/18/2025	12/11/2025	11/24/2025
Benchmark Category	Benchmark	GPT-5.5 Thinking	Opus 4.7	Mythos Preview	GPT-5.4 Thinking	Gemini 3.1 Pro	Opus 4.6	Gemini 3 Pro	GPT-5.2 Thinking	Opus 4.5
Reasoning and Knowledge	GPQA Diamond (No Tools)	93.60%	94.20%	94.60%	92.80%	94.30%	91.30%	91.90%	92.40%	87.00%
Reasoning and Knowledge	Humanity's Last Exam (No Tools)	41.40%	46.90%	56.80%	39.80%	44.40%	40.00%	37.50%	34.50%	30.80%
Reasoning and Knowledge	Humanity's Last Exam (with Search and Python)	52.20%	54.70%	64.70%	52.10%	51.40%	53.10%	45.80%	45.50%	43.20%
Software Engineering	Terminal-Bench 2.0	82.70%	69.40%	82.00%	75.10%	68.50%	65.40%	56.90%	62.20%	59.80%
Agentic Tasks	r ² -Bench (Telecom)	98.00%	-	-	98.90%	99.30%	99.30%	98.00%	98.70%	98.20%
Agentic Tasks	MCP Atlas	75.30%	77.30%	-	68.10%	73.90%	75.80%	54.10%	60.60%	62.30%
Abstract Reasoning	ARC-AGI-2 (Verified)	85.00%	75.83%	-	73.30%	77.10%	68.80%	31.10%	52.90%	37.60%
Multimodality	MMMU Pro (with Tools)	83.20%	-	-	82.10%	80.50%	77.30%	81.00%	80.40%	73.90%
Long Context	MRCR v2 (128K-256K 8-Needles)	87.50%	59.20%	-	79.30%	84.90%	91.90%	77.00%	70.00%	74.00%
Long Context	BrowseComp Long Context	84.40%	79.30%	86.90%	82.70%	85.90%	84.00%	59.20%	89.80%	72.89%

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Anthropic, OpenAI, Google

The data reinforces how quickly the competitive dynamics are materializing. Although newer models tend to improve on older models, GPT-5.5 Thinking, Gemini 3.1 Pro, and Opus 4.7 are now often separated by single-digit differences in percentage scores across reasoning, coding, agentic tasks, and multimodality (Figure 24). This degree of parity would have been shocking even twelve months ago. For instance, all three labs vaulted from near-zero scores on ARC-AGI-2 to the 76%-85% range with marked progress clustered within weeks (Figure 25). Advantages on raw reasoning benchmarks are now measured in months instead of quarters. Leads can be erased within a single model release cycle as providers grow increasingly familiar with benchmarks and continue to refine their training architectures.

Figure 25. ARC-AGI-2 Progress by Company

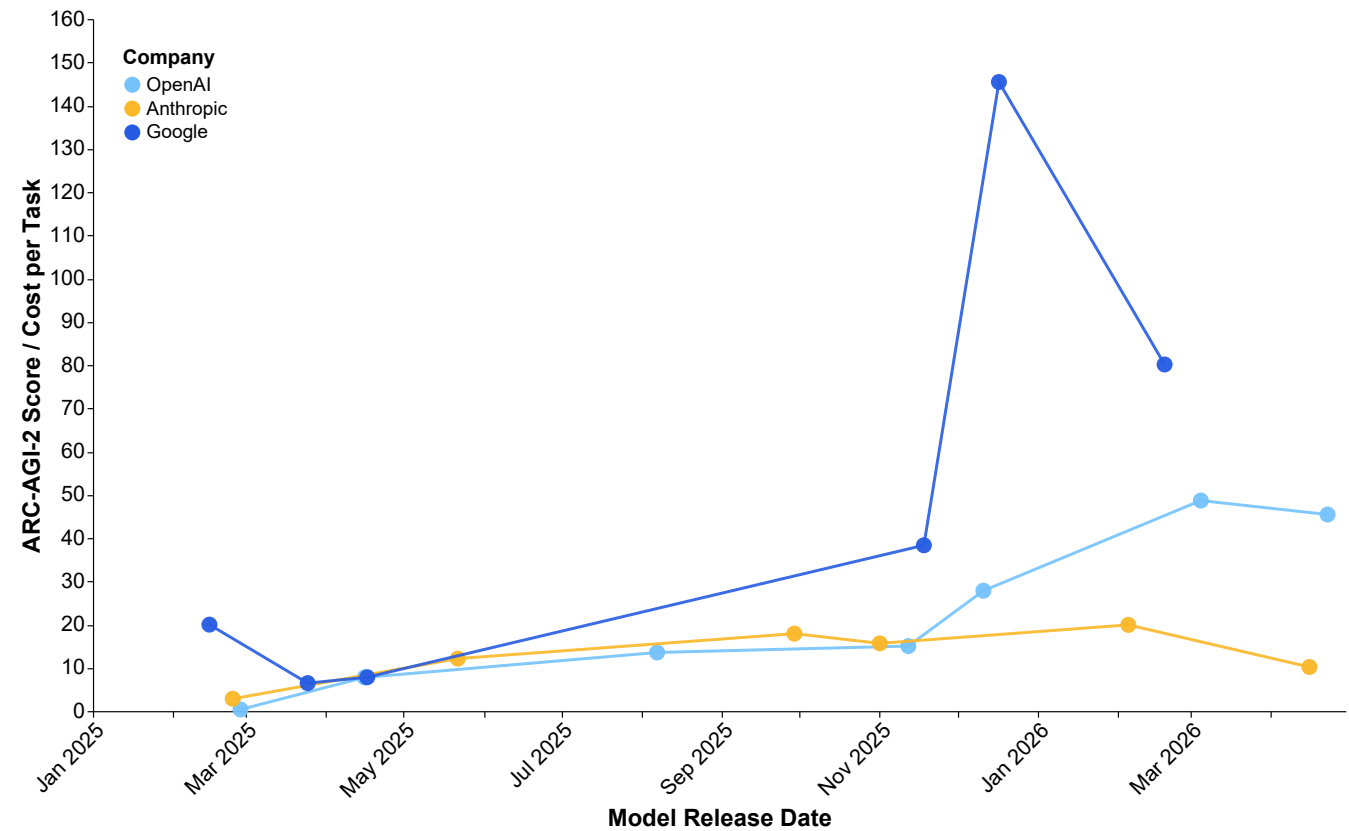


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, ARC Prize

While all three labs have achieved similar ARC-AGI-2 scores, the cost of harnessing that progress differs (Figure 26). OpenAI, Anthropic, and Google saw improvements in their intelligence ratio over successive model releases. However, each lab's most recent release reversed this trend, suggesting that the latest capability gains came at a disproportionately higher cost. **The implication is that convergence in benchmark scores may mask a widening gap in cost-efficiency, a dimension that will increasingly matter as enterprises evaluate which provider delivers the most capability per dollar.**

Figure 26. ARC-AGI-2 Performance per Dollar

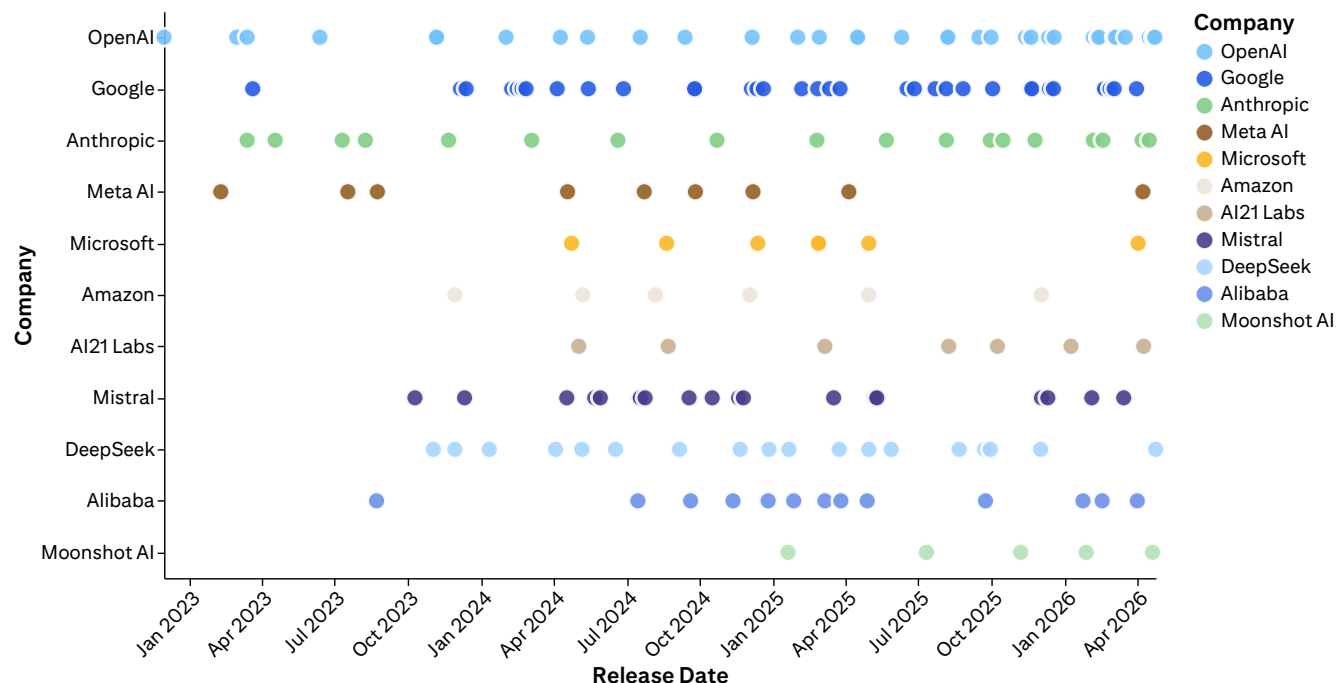


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, ARC Prize

Meanwhile, the sheer volume and accelerating pace of model releases across the industry has grown from a handful of labs shipping sporadically in 2023 to top providers releasing updates for several model types per quarter by early 2026, as they try to stay ahead of the competition and take advantage of the rapidly evolving technology (Figure 27).

Figure 27. AI Model Release Timeline



*Historical Events Through April 24, 2026

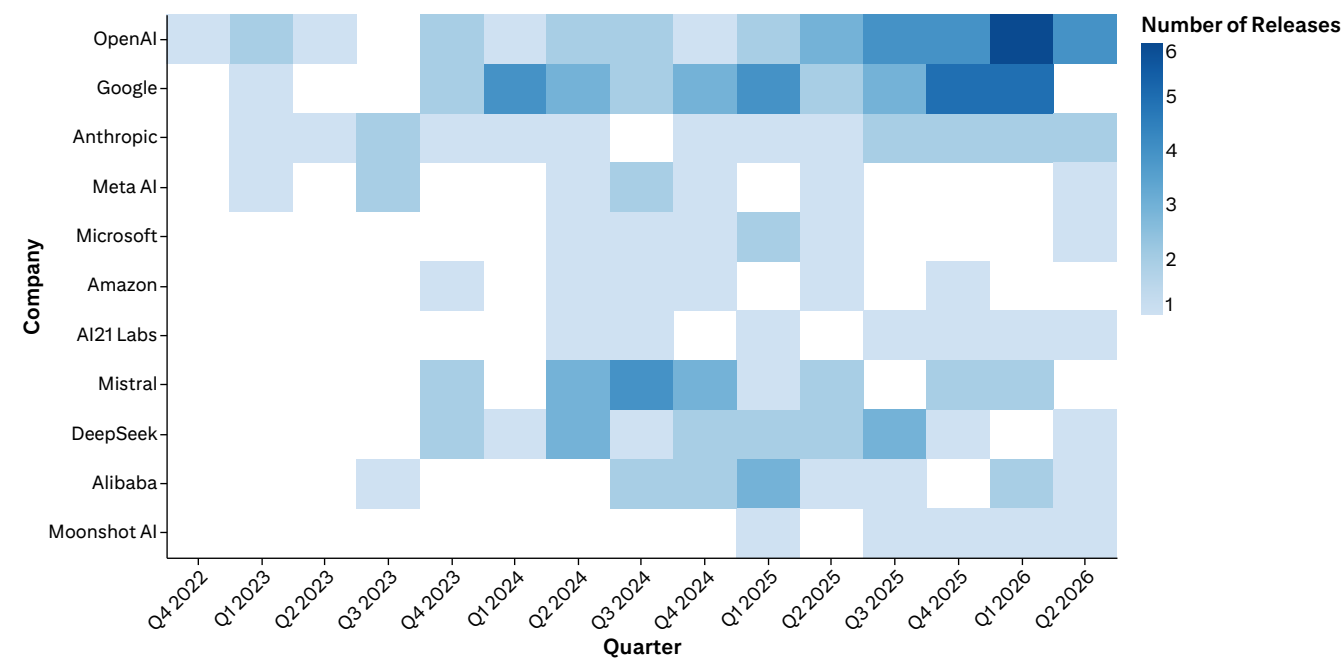
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, OpenAI, Google, Anthropic, Meta AI, Microsoft, Amazon, DeepSeek, Alibaba, Moonshot AI

Anthropic maintains major model updates roughly quarterly with the Claude 3 family in Q1 2024, Claude 3.5 in Q2-Q4 2024, Claude 3.7 in Q1 2025, Claude 4 in Q2 2025, Claude 4.1 in Q3 2025, Claude 4.5 in Q4 2025, Claude 4.6 in Q1 2026, and Claude 4.7 in Q2 2026 (Figure 27). They sometimes group product innovation around major model launches like Artifacts with Sonnet 3.5 and Claude Code with the 4.0 family.

However, their innovation timelines for models and products are comparatively less in lockstep when contrasted against the patterns followed by OpenAI and Google, whose lockstep strategies support platform building by unlocking differentiated application layers with each foundational improvement in their model families. Anthropic's recent acceleration, with Opus 4.1, 4.5, 4.6, and 4.7 all released since August 2025, may signal a response to competitive pressure and a maturation of Anthropic's research pipeline. DeepSeek, Alibaba, Mistral, and other have emerged as credible model providers, and a frequency heatmap for model releases shows the cadence per quarter intensifying (Figure 28).

Figure 28. AI Model Release Frequency



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

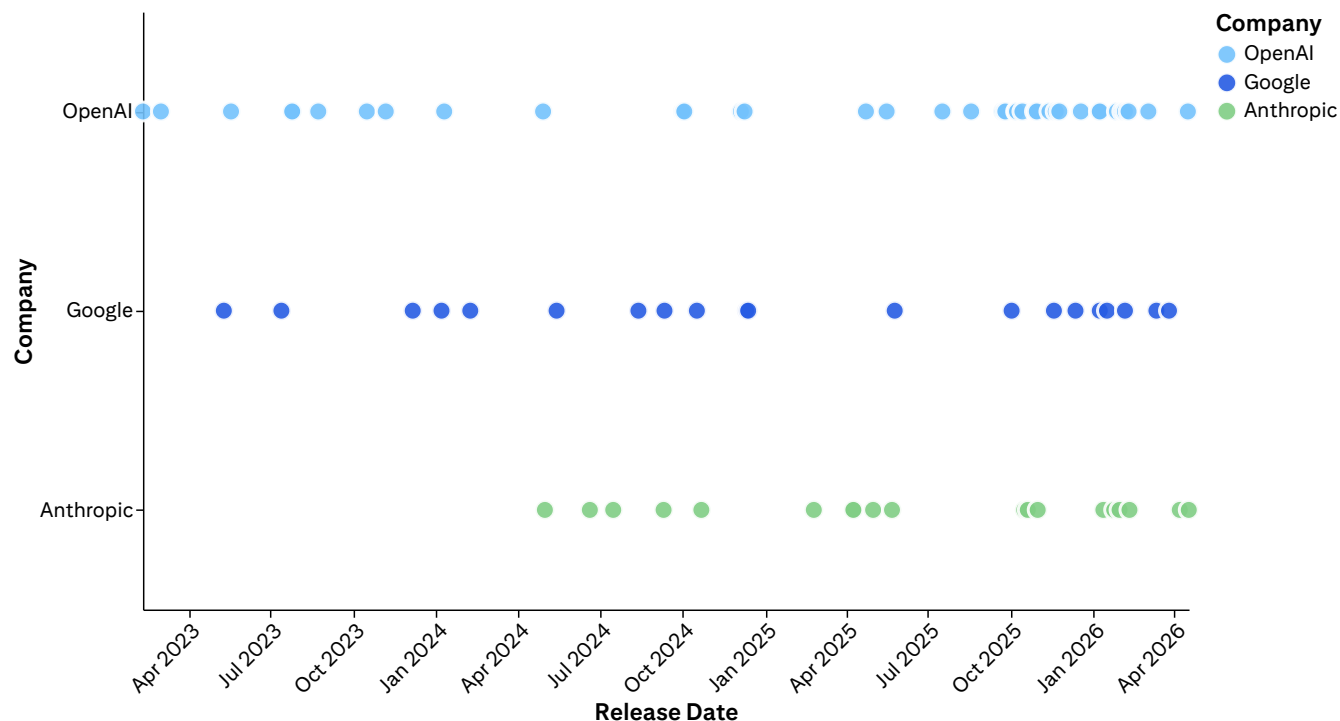
See the AI Model Release Timeline above for details on the models shown in this heatmap.

Source: Citi Research, OpenAI, Google, Anthropic, Meta AI, Microsoft, Amazon, DeepSeek, Alibaba, Moonshot AI

A broader timeline of products and features underscores a related point. Competition is no longer confined to model releases alone (Figure 29). OpenAI, Google, Anthropic, and other companies have each shipped a steady drumbeat of product-level launches including new interfaces, enterprise features, developer tools, and platform integrations, with the pace accelerating sharply in late 2025. The race to build products and workflows around the model is now matching the intensity of the race to advance the model itself.



Figure 29. Products and Features Timeline



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, OpenAI, Google, Anthropic

This product release sequence reveals Anthropic's pursuit to further sell its agentic capabilities. With offerings such as their Team plan (May 2024), Enterprise plan (Sep. 2024), tiered Max plans (\$100/\$200 monthly in Apr. 2025), and Claude Cowork (Jan. 2026), enterprises now have access to a wider range of self-serve and customizable options, enabling them to choose the product or combination of products that best meets their needs. More tellingly, the integration timeline demonstrates systematic productization of agentic workflows with, for instance, Claude Code moving from research preview (Feb. 2025) to GA with MCP/Files API (May 2025) to web availability (Oct. 2025) to Cowork plugins (Jan. 2026). This development and branching contrasts with OpenAI's broader consumer push (e.g., ChatGPT Go geographic expansion, Shopping Research, Health verticals, etc.) and Google's search-centric strategy.

The releases of Computer Use (Oct. 2024) and Agent Skills (Oct. 2025) position Anthropic for the enterprise automation market, where their longstanding focus on risk mitigation and safety may help command premium pricing ([Anthropic](#), Mar. 3, 2023). This direction could pair well with brand-driven differentiation if "Claude" becomes synonymous with AI that prioritizes reliability, similar enterprise buyers paying premiums for IBM or AWS despite comparable technical alternatives.

Anthropic's private preview release of Claude Mythos Preview, a model that scored 83.1% on CyberGym (versus 73.1% for Opus 4.7) and autonomously discovered thousands of zero-day vulnerabilities across major operating systems and browsers, further sharpens this enterprise positioning ([Anthropic](#), Apr. 7, 2026; [AISI](#), Apr. 13, 2026).

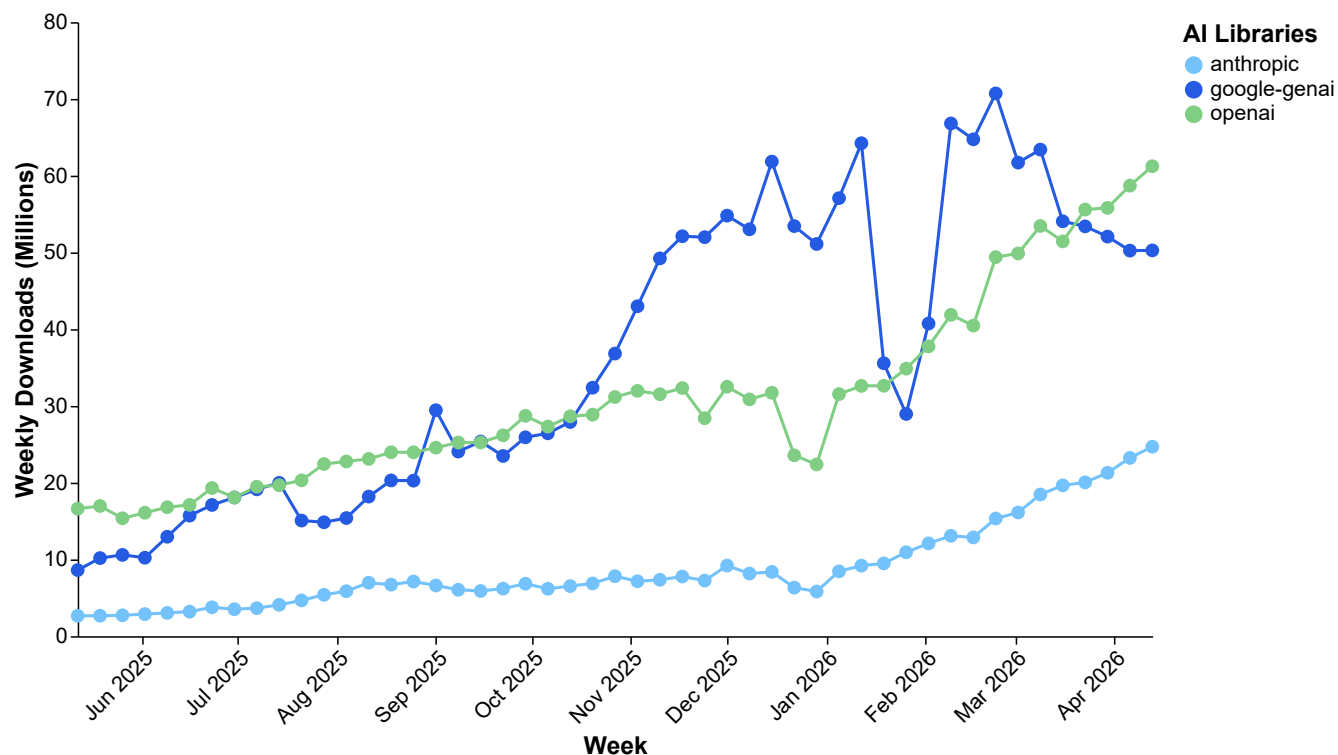
“[W]hile we do not plan to make Claude Mythos Preview generally available, our eventual goal is to enable our users to safely deploy Mythos-class models at scale [...]. To do so, that also means we need to make progress in developing cybersecurity (and other) safeguards that detect and block the model’s most dangerous outputs. We plan to launch new safeguards with an upcoming Claude Opus model, allowing us to improve and refine them with a model that does not pose the same level of risk as Mythos Preview.”

– [Anthropic’s Frontier Red Team, Apr. 7, 2026](#)

Due to concerns over Mythos being too powerful to release broadly, Anthropic has withheld it from general availability. It is instead channeling Mythos through Project Glasswing to over fifty partners, which include Palo Alto Networks, AWS, and Microsoft ([Anthropic](#), Apr. 7, 2026). OpenAI now offers a rival model, GPT-5.4-Cyber, through its Trusted Access for Cyber program ([OpenAI](#), Apr. 14, 2026). While Mythos compresses what previously required years of specialized exploit-development expertise into an API call, CrowdStrike has expressed that enterprise security teams may at times favor portfolios of smaller, purpose-built models for deterministic detection and response over monolithic models like Mythos and GPT-5.4-Cyber, which suggests the near-term commercial opportunity may accrue more to platform vendors that operationalize these capabilities within existing workflows for security operations ([CrowdStrike Holdings Inc \(CRWD.O\) - Takeaways from Conversation with SVP Counter Adversary Ops](#)).

The developer ecosystem illustrates expanding competition among AI providers. Weekly public downloads of AI libraries, while not representative of enterprise adoption and best viewed as a rough indicator of momentum among individual developers integrating AI APIs, show the google-genai package rising from 10M in mid-2025 to over 70M by early 2026 (Figure 30). This adoption overtook OpenAI’s library, which now averages 50M–60M downloads per week and has since regained the top position. Anthropic’s library, though smaller in absolute volume, has seen a notable rise in downloads since January 2026. This trend aligns with a surge in Claude Code adoption and suggests a market where developer mindshare is becoming increasingly distributed among multiple providers rather than consolidating around a single leader (Figure 30).

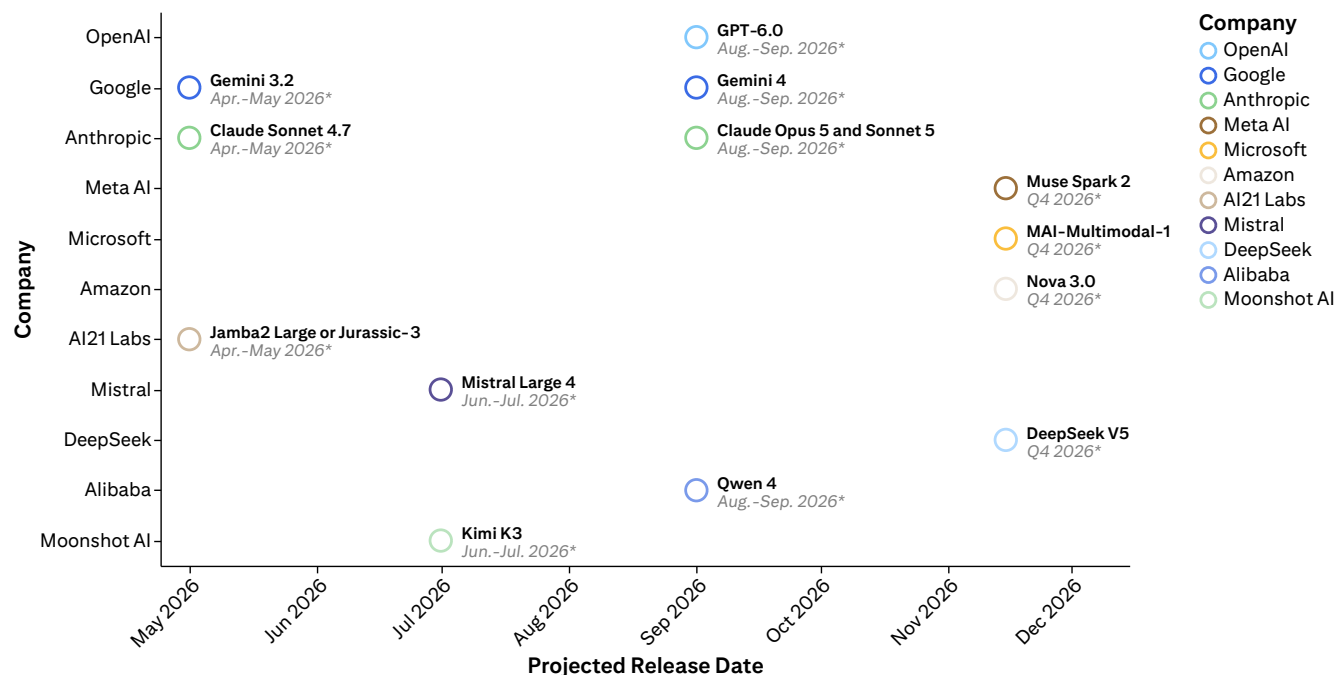
Figure 30. Weekly Public Downloads of AI Libraries



© 2026 Citigroup Inc. No redistribution without Citigroup’s written permission.
Source: Citi Research, PyPI Stats

Looking ahead, release cadence and the public suggest to us that every major lab will ship a new model by year end, meaning the benchmark parity we observe today is likely to persist or tighten further as each provider refreshes its lineup using the latest hardware on clustering release timelines (Figure 31).

Figure 31. All Companies Are Expected to Release a New Model by Year End



**Projected Events Beyond April 24, 2026*

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research

In the next couple of quarters, Anthropic and other AI Labs are likely to face coincident releases from all major competitors. The anticipated Claude Opus 5 and Sonnet 5 launch in mid-to-late 2026 is probable to again coincide with major releases by OpenAI and Google, as seen in 2024 and 2025. When major releases happened in quick succession late last year, Anthropic dropped Claude 4.5 first, followed by Gemini 3 and then GPT-5.2. Shortly thereafter, Anthropic released the more capable Claude 4.6 models to challenge those competitive offerings which were followed by Gemini 3.1 and GPT 5.4. After the release of Claude 4.7 and GPT-5.5 in April 2026, it is likely that a new Gemini model will soon follow, with each iteration aiming to meet or surpass the latest scores on benchmarks.

This sequencing pattern exposes a fundamental tension. Being first to market telegraphs capability, allowing followers to dynamically adjust inference-time compute allocation or model selection thresholds to marginally exceed published metrics. The industry could approach a Nash equilibrium where no player wants to move first, creating potential release delays or strategic ambiguity around launch dates. A potential delay in DeepSeek V4 (which was initially expected around Lunar New Year from Feb. 17 to Mar. 4 but was ultimately released on Apr. 24) could be a recent example of this hesitation ([The Information](#), Jan. 9, 2026). It followed the release of GPT-5.4 on Mar. 5 and GPT-5.5 on Apr. 23.

The capital fueling this competitive landscape is substantial. According to PitchBook, OpenAI and Anthropic were valued at \$840B and \$380B respectively as of February 2026, with round-over-round step-ups of 1.7x and 2.1x (Figure 32). Companies across the AI stack from Databricks to Anysphere (i.e., Cursor) command valuations that suggest investor conviction that the market is still in its early stages.

Figure 32. An Overview of the VC Landscape in AI

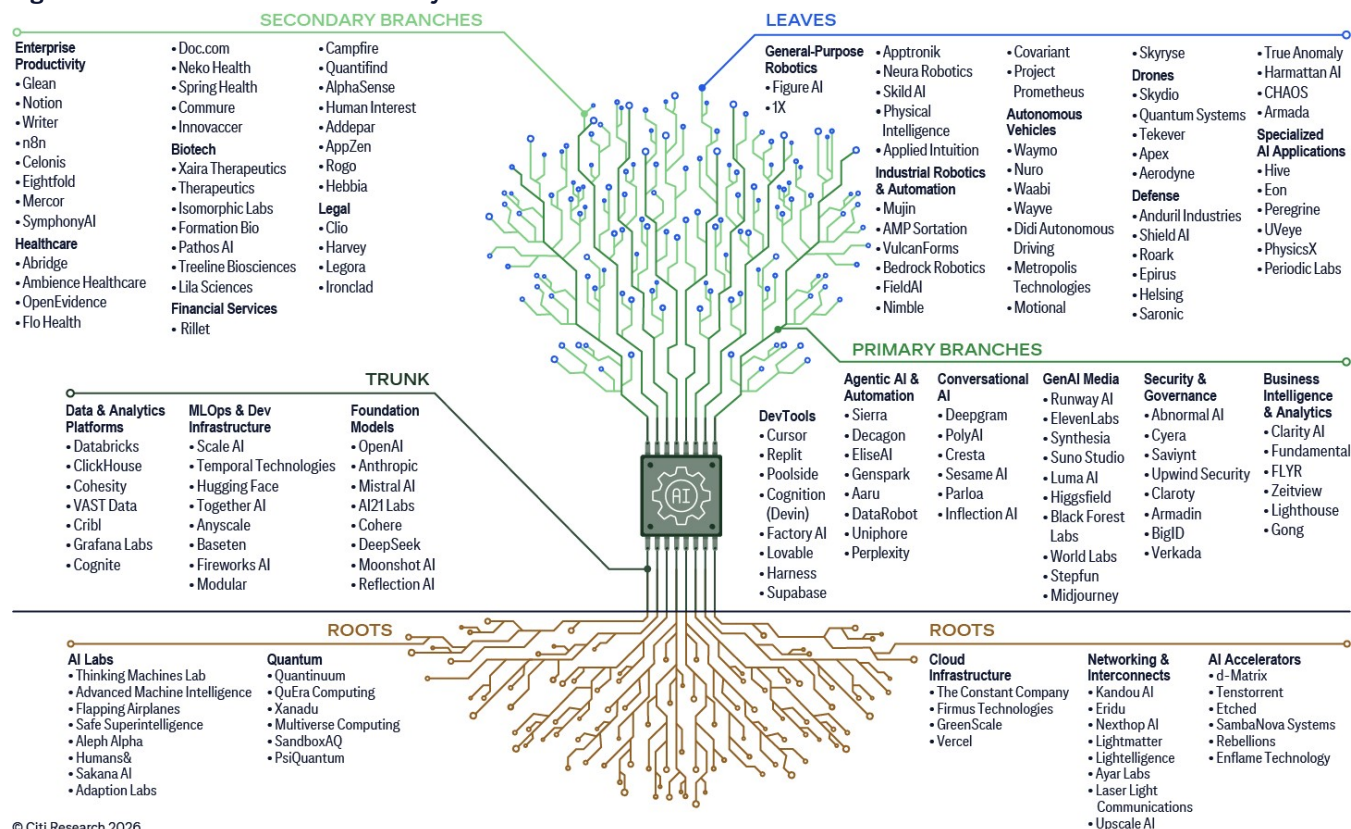
Company	Last Financing Valuation (\$ mn)	Last Financing Date	Prior Financing Valuation (\$ mn)	Prior Financing Date	Round/Round step-up
OpenAI	\$840,000	Feb-26	\$500,000	Oct-25	1.7x
Anthropic	\$380,000	Feb-26	\$183,000	Sep-25	2.1x
Databricks	\$134,000	Feb-26	\$100,000	Sep-25	1.3x
Scale AI	\$29,000	Jun-25	\$14,600	Dec-23	2.0x
Figure AI	\$39,500	May-25	\$2,600	Feb-24	15.2x
Ramp	\$32,000	Nov-25	\$22,500	Jul-25	1.4x
Safe Superintelligence	\$32,000	Apr-25	\$5,000	Sep-24	6.4x
Anduril Industries	\$30,500	Jun-25	\$14,000	Sep-24	2.2x
VAST Data	\$30,000	Oct-25	\$9,100	Nov-23	3.3x
Anysphere (aka Cursor)	\$29,300	Nov-25	\$9,900	Jun-25	3.0x
Perplexity	\$20,000	Sep-25	\$18,000	Jul-25	1.1x
Deel	\$17,300	Oct-25	\$12,000	Feb-25	1.4x
ClickHouse	\$15,000	Jan-26	\$6,665	Oct-25	2.3x
Skild AI	\$14,000	Jan-26	\$4,700	May-25	3.0x
Mistral AI	\$13,664	Sep-25	\$6,506	Jun-24	2.1x
Helsing	\$13,659	Jun-25	\$5,321	Jul-24	2.6x
OpenEvidence	\$12,000	Jan-26	\$6,000	Oct-25	2.0x
ElevenLabs	\$11,000	Feb-26	\$6,600	Jan-26	1.7x
Notion	\$11,000	Jan-26	\$10,000	Sep-21	1.1x
Cognition	\$10,200	Sep-25	\$4,000	Mar-25	2.6x
1X	\$10,000	Sep-25	\$210	Jan-23	47.6x
Crusoe	\$10,000	Oct-25	\$2,800	Dec-24	3.6x
Mercor	\$10,000	Feb-26	\$2,000	Feb-25	5.0x
Sierra	\$10,000	Sep-25	\$4,500	Nov-24	2.2x
Thinking Machines Lab	\$10,000	Jun-25	\$9,000	Feb-25	1.1x

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, PitchBook

The breadth of capital deployment across these companies reflects the full complexity of the AI stack (Figure 33). The VC ecosystem for AI has already expanded dramatically over the past eight months, as the field of competitors at every layer continues to widen ([Artificial Intelligence – AI: The Information Era's Apex Technology](#)).

Figure 33. An Overview of the VC Ecosystem for AI



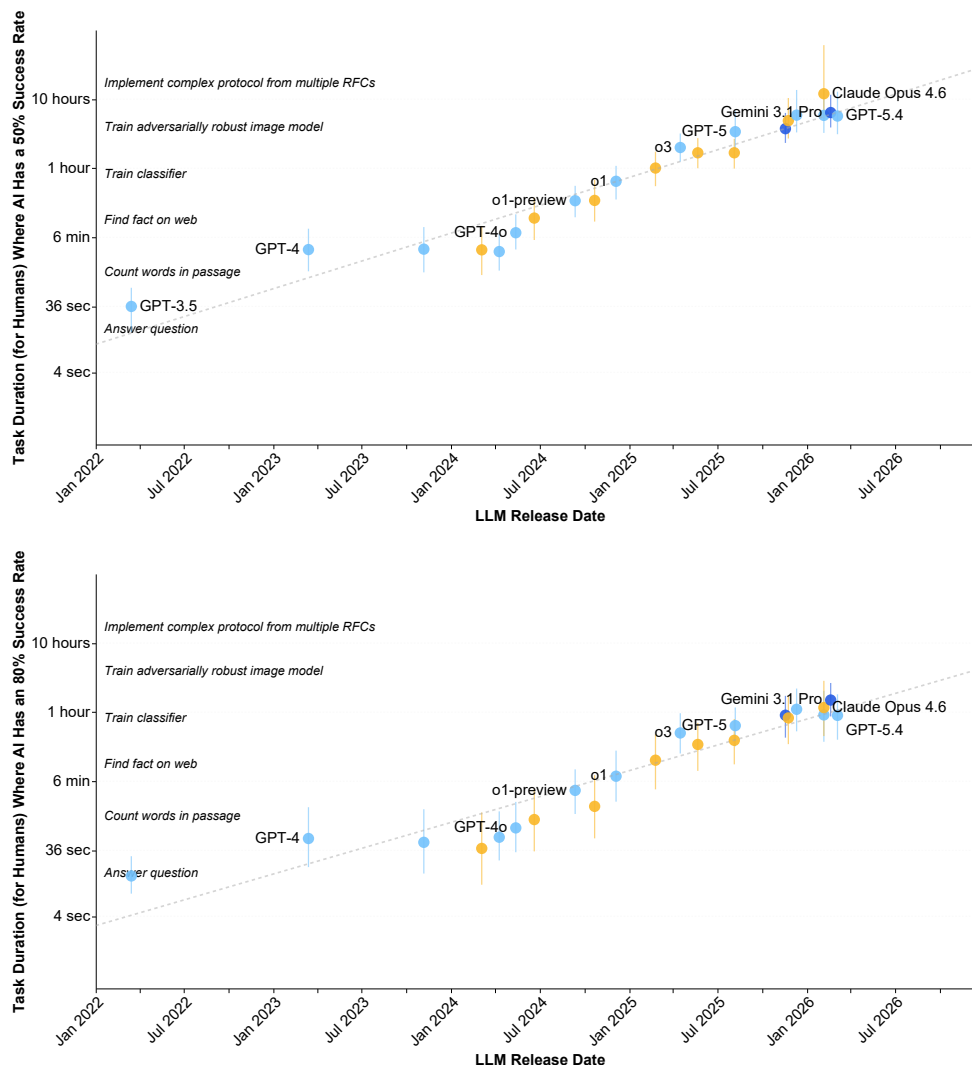
© Citi Research 2026

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, PitchBook

Yet time horizons for task completions tell perhaps the most consequential story about today's inflection point (Figure 34). It plots the task-completion time horizons at which AI models can operate reliably with GPT-3.5 answering simple questions in seconds to today's frontier models sustaining multi-hour autonomous workflows. Opus 4.6, GPT-5.4, and Gemini 3.1 Pro were tightly clustered at the frontier in March of 2026, with all of them capable of sustained operation on tasks that would take a human the better part of a workday. This achievement is the technical foundation of the copilot-to-agent transition. Models can now stay on task long enough to do real work, not just answer questions. When three labs can each clear nearly 70% on ARC-AGI-2 and sustain multi-hour autonomous execution, the challenge becomes who builds the best products on top of that capability. They are competing for enterprise customers that are looking for who translates hours of reliable autonomous operation into workflows that these multifaceted businesses can trust, adopt, and pay for at scale.

Figure 34. Task-Completion Time Horizons of AI Models



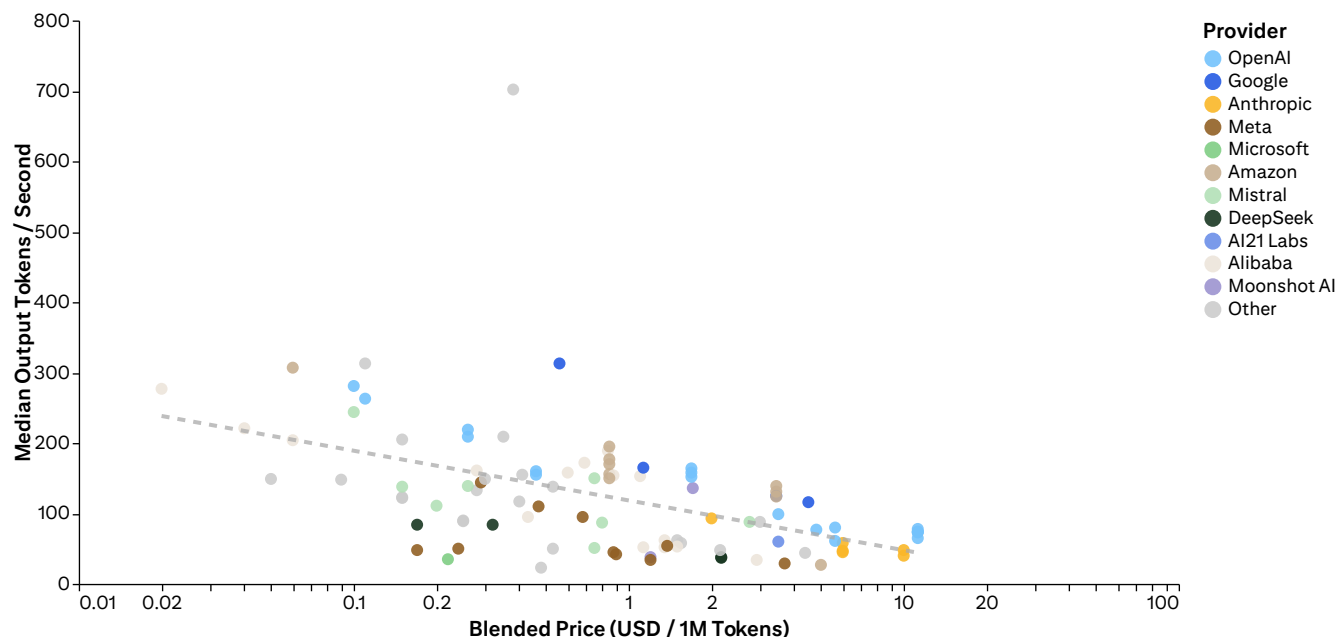
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

METR last updated its measurements on Mar. 19, 2026.

Source: Citi Research, METR

The economic dimension of this competition is also important. Language models with a lower price per million tokens tend to generate output at a higher speed (Figure 35). Lower-cost models often employ fewer parameters, simpler architectures, less reasoning, or more aggressive quantization and distillation techniques. While these approaches sometimes affect output quality, they also reduce the computational overhead per token, which can lead to less floating-point operations, less memory bandwidth consumption, and faster inference. Anthropic's models land in the middle ground, sitting tightly along the trendline with a relatively consistent price-to-speed ratio across the lineup.

Figure 35. Output Speed Versus Blended Price



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Artificial Analysis

With the release of GPT-5.5, the pricing trends for OpenAI and Anthropic have become more closely aligned. Regarding frontier model pricing per million tokens, GPT-5.5 Thinking, Opus 4.7, and Gemini 3.1 Pro now sit within a narrow band on input pricing of \$2–\$5 per million tokens (Figure 36). Output pricing ranges from \$12–\$30 per million tokens. Cached input token pricing, a feature all three providers now offer, has compressed to \$0.20–\$0.50, further reducing the marginal cost of agentic workflows that rely on repeated context.

Figure 36. Frontier Model Pricing Per Million Tokens

Model	Release Date	Input Tokens	Cached Input Tokens	Output Tokens
GPT-5.5 Thinking	4/23/2026	\$5.00	\$0.50	\$30.00
Opus 4.7	4/16/2026	\$5.00	\$0.50	\$25.00
GPT-5.4 Thinking	3/5/2026	\$2.50	\$0.25	\$15.00
Gemini 3.1 Pro	2/19/2026	\$2.00	\$0.20	\$12.00
Opus 4.6	2/5/2026	\$5.00	\$0.50	\$25.00
Gemini 3 Pro	12/18/2025	\$2.00	\$0.20	\$12.00
GPT-5.2 Thinking	12/11/2025	\$1.75	\$0.18	\$14.00
Opus 4.5	11/24/2025	\$5.00	\$0.50	\$25.00

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, OpenAI, Anthropic, Google

It is against this backdrop that we evaluate Anthropic's competitive positioning across its three core product lines of the API, Claude Code, and Cowork.

API Business: Claude as Premium AI Infrastructure

As referenced earlier, **86% or ~\$3.9B of Anthropic's internally projected \$4.5B in 2025 revenue was from its API business.** The API offering positions Anthropic as

an essential infrastructure provider for intelligence, similar to how other software companies have successfully built their businesses around API-driven platforms such as Stripe for payments or Twilio for communications. By offering Claude via API, Anthropic allows developers to easily integrate cutting-edge AI into their products, with revenue directly tied to the usage of those AI models through token consumption. As partners and customers extend, embed, and automate functionality through APIs, Anthropic drives broader adoption and ecosystem growth for its distinct brand of AI, Claude.

One factor behind Anthropic's strong API revenue performance may be its ability to consistently command a premium for Claude models. Even as input/output token pricing for Opus 4.5 (released Nov. 2025) declined sharply by 67% vs. the prior Opus 4.1 model (Figure 37), pricing remained higher than the comparable GPT-5.1/GPT-5.2 and Gemini 3 Pro models. Opus 4.6 and Sonnet 4.6 maintained a higher price point than OpenAI's GPT-5.4 model, with the premium persisting despite GPT-5.4's 17% token price increase over GPT-5.2, which itself was 40% more expensive than GPT-5.1. Though GPT-5.5's output token pricing is now \$5 higher than Opus 4.7, the \$25/\$125 per million input/output token pricing for partners involved in Anthropic's Claude Mythos Preview, a limited research preview tied to its Project Glasswing cybersecurity initiative, suggests there could be further model pricing/tiering upside with future model releases ([Anthropic](#), Apr. 7, 2026). The company does not plan to make Claude Mythos Preview generally available but will continue developing safeguards that would eventually allow such capable models to be safely deployed at scale. Mythos is the most capable [model Anthropic has] built to date" and represents "a step change" in AI performance, according to an Anthropic spokesperson ([The Information](#), Mar. 26, 2026; [Fortune](#), Mar. 31, 2026).

Figure 37. Claude Model Pricing History

	Base Input (\$/1M tokens)	Base Output (\$/1M tokens)	Blended (\$/1M tokens)	Chg. vs. Prior generation %
Claude Mythos *	\$25.0	\$125.0	\$50.0	400%
Claude Opus 4.7	\$5.0	\$25.0	\$10.0	0%
Claude Opus 4.6	\$5.0	\$25.0	\$10.0	0%
Claude Sonnet 4.6	\$3.0	\$15.0	\$6.0	0%
Claude Opus 4.5	\$5.0	\$25.0	\$10.0	-67%
Claude Sonnet 4.5	\$3.0	\$15.0	\$6.0	0%
Claude Haiku 4.5	\$1.0	\$5.0	\$2.0	25%
Claude Opus 4.1	\$15.0	\$75.0	\$30.0	-
Claude Sonnet 4	\$3.0	\$15.0	\$6.0	-
Claude Haiku 3.5	\$0.8	\$4.0	\$1.6	-

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

* Mythos Preview pricing that will be available to Project Glasswing participants. Anthropic does not plan to make Claude Mythos Preview generally available. API pricing can be viewed on <https://platform.claude.com/docs/en/about-claude/pricing>.

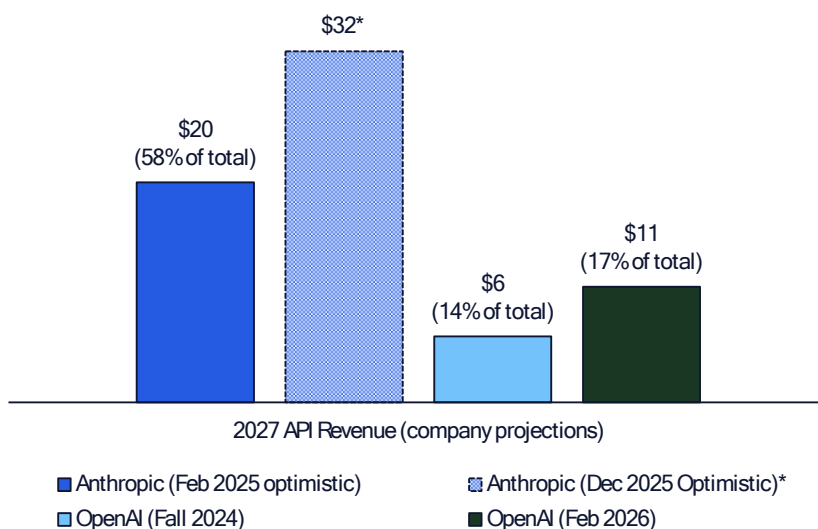
Source: Citi Research, Anthropic

Anthropic's last API revenue projection from Feb. 2025 called for \$20B in 2027 or 58% of its \$34.5B optimistic total revenue projection at the time ([The Information](#), Feb. 12, 2025). Since then, the company's total revenue projection (optimistic scenario) has been revised up twice, by 13% during the summer and another 41% as of Dec. 2025 ([The Information](#), Nov. 4, 2025; [The Information](#), Jan. 27, 2026). **Applying the same 58% projected API mix to the most recent total revenue**

projection would imply \$32B in API revenue in 2027 in an optimistic scenario. This is about 3x higher than OpenAI's projected API revenues, which had been capped by its strategic partnership agreement with Microsoft as the exclusive cloud provider for OpenAI's stateless APIs until the [updated partnership agreement](#) announced on Apr. 27, 2026. That said, given the recent acceleration in growth in enterprise adoption and related AI Lab revenues, any prior forecasts are likely to be outdated at best.

Figure 38. Anthropic vs. OpenAI Internal API Revenue Projections

(\$ billions)



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: the \$32B API revenue projection for Anthropic is derived by multiplying the company's optimistic total revenue projection for 2027 of \$55B (as of December 2025) by its earlier 58% projected API revenue mix (as of February 2025).

Source: Citi Research, The Information

Advancing AI capabilities toward AGI could diminish model differentiation and intensify pricing pressure on today's premium models and API margins, even though custom LLM applications/agents will retain some switching costs. While the API business will remain a major growth driver for Anthropic, we believe the firm will pursue stronger moats in subscriptions and agentic systems this year, as already evidenced by offerings like Claude Code and Claude Cowork.

Building Stickiness via Subscriptions

Anthropic's subscription plan evolution began with the September 2023 launch of Claude Pro, followed by the release of additional tiers for Claude Team, Claude Enterprise and Claude Max over the next 1.5 years (Figure 39). Since the start of 2026, Anthropic has demonstrated a more aggressive GTM strategy to gain application market share by lowering the price of Claude Team seats by \$5/seat in January and launching a self-serve Enterprise option at \$20/month per seat in February (Figure 37). Our checks also suggest greater flexibility around Enterprise seat minimums for negotiated contracts. The pricing trajectory reflects a trend to expand the addressable market at every tier while preserving margin on high-usage power users, a playbook that mirrors the SaaS industry's proven freemium-to-enterprise ladder.

Figure 39. Subscription Timeline for Anthropic

Date	Subscription Plan Release Timeline
Sep-23	Introduced Claude Pro plan
May-24	Introduced Claude Team plan and iOS app: Team plan is \$25 per month billed annually, or \$30 per user per month, with a minimum of 5 seats.
Sep-24	Claude Enterprise plan introduced
Apr-25	Introduced Claude Max Plan \$100 per month for expanded usage (5x higher rate limits) \$200 per month for max flexibility (20x higher rate limits)
Jan-26	Lowered pricing for Claude Team (for teams of 5 to 150): Standard seats are now \$20 per month with annual discount, or \$25 per month billed monthly Premium seats are now \$100 per month with a new annual discount, or \$125 per month billed monthly
Feb-26	Claude Enterprise self-serve option launched: Claude Enterprise uses a simple seat-plus-usage model. Usage is billed at API rates, so your organization only pays for what it uses. Administrators can set spend caps at the organization and individual user levels to reliably manage costs. Self-serve seats are \$20/seat.

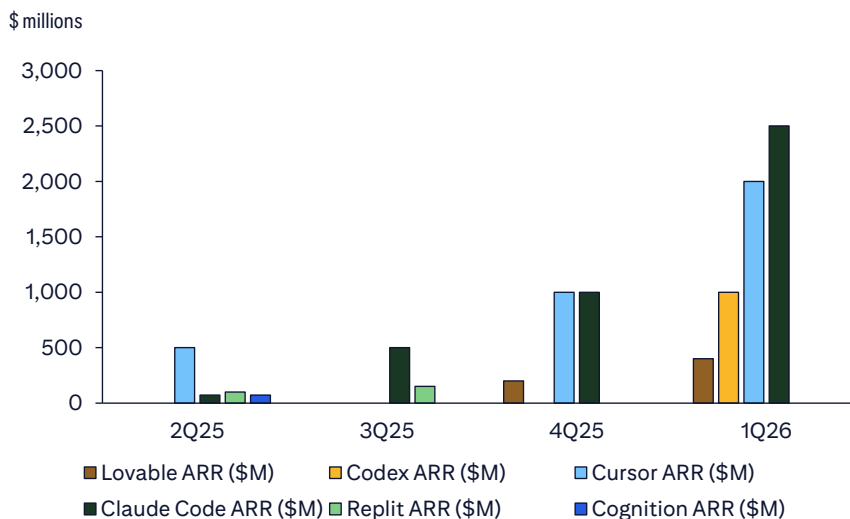
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Anthropic

The 3C's in AI Coding: Claude Code, Cursor, Codex

Claude Code was released as a research preview in February 2025 and became generally available on May 22, 2025. Within six months of launching to the public, Claude Code reached [\\$1B in run-rate revenue](#) in November, the same annualized run-rate as Cursor, which also crossed the [\\$1B milestone](#) around the same time (Anthropic, Dec. 3, 2025; Cursor, Nov. 13, 2025). Three months later, Claude Code [ARR crossed \\$2.5B](#) as of mid-February, surpassing that of Cursor which reached \$2B in February (Anthropic, Feb. 12, 2026; [Bloomberg](#), Mar. 2, 2026). **We see the productization of Claude's specialized coding capabilities as key to Anthropic capturing more subscription revenue while also driving higher consumption API revenue.**

Figure 40. ARR Milestones for AI Coding Tools



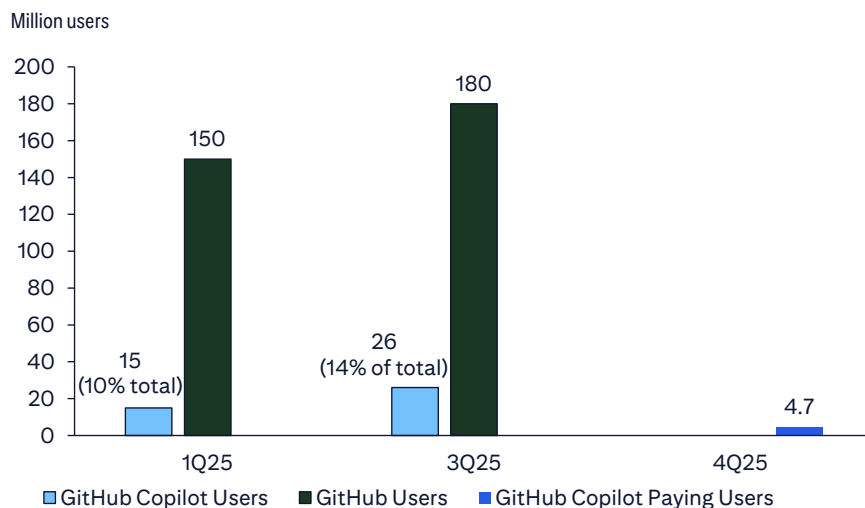
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: available ARR metrics sourced from press releases and news outlets (Bloomberg, VentureBeat, TechCrunch, X.com). Lovable crossed \$400M ARR in Feb. 2026, Codex crossed \$1B ARR by the end of Jan. 2026, Claude Code ARR was \$2.5B as of mid-Feb. 2026, Cursor ARR crossed \$2B in February and \$2.7B in March.

Source: Citi Research, Company Reports

Despite the impressive ARR trajectories, adoption of AI coding assistants remains underpenetrated as reflected in Microsoft's developer userbase statistics. GitHub Copilot penetration expanded from 10% to 14% from CQ1 to CQ3 2025, while the percentage of paying GitHub Copilot users was just 2.6% (when comparing CQ4's ~5 million paid seats against the last disclosed CQ3 GitHub user count of 180 million developers), underscoring the significant runway ahead. While the only datapoint that we have on Claude Code is that within roughly four months post launch, it had purportedly already reached 115K developers ([PPC Land](#), Jul. 6, 2025), **we think the number of paying Claude Code developers today may be anywhere between 2-4 million users, based on the \$2.5B ARR disclosure and an average spend assumption range of \$50-\$100 per developer.**

Figure 41. GitHub User Metrics Suggest AI Coding Assistants Market Remains Underpenetrated



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: available user metrics from Microsoft's quarterly earnings calls. Calendar quarter basis.

Source: Citi Research, Company Reports

A big question on investors' minds given the unrelenting pace at which the AI coding space is evolving is which tools will ultimately prevail. While the lines between code completers (e.g. GitHub Copilot, Gemini Code Assist, Tabnine), app builders (e.g. Replit, Bolt.new, V0 (Vercel), Lovable, Cursor), and autonomous agents (e.g. Claude Code, Codex, Factory, Cognition) are blurring as each company expands its product reach, **we don't expect a winner-take-all market, but rather a stratified playing field with multiple leaders that become the default primary or secondary tool depending on the software developer's particular role.**

For instance, IDE-centric tools like Cursor and AI-centric tools like Claude Code and Codex currently occupy distinct niches. Comparing Cursor's model-agnostic, editor-first approach against Claude Code's design (taken here as representative of the emerging class of model-specific coding tools that includes Codex) illuminates a deeper structural shift in how AI could be integrated into the development workflow.

Specifically, Cursor excels in IDE-native development workflows and codebase mapping with a VS Code feel, while Claude Code benefits high-level operations and complex reasoning tasks. However, their architectural trajectories suggest convergence toward a unified AI-first development paradigm. **A current hybrid approach, where some developers may leverage Claude Code's premium AI capabilities for substantial changes while using Cursor's superior contextual awareness for incremental work, represents a transitional phase exposing fundamental tensions between traditional IDE structures and AI-native workflows.**

Key differentiators include Claude Code's token consumption for file context reading versus Cursor's efficient default codebase mapping, as well as Cursor's file visibility and model-agnostic advantages against Claude Code's more versatile and Anthropic-native capabilities where thoughtful claude.md files can yield 5-11% better code quality for even premier Claude agents ([Arize AI](#), Nov. 20, 2025). While

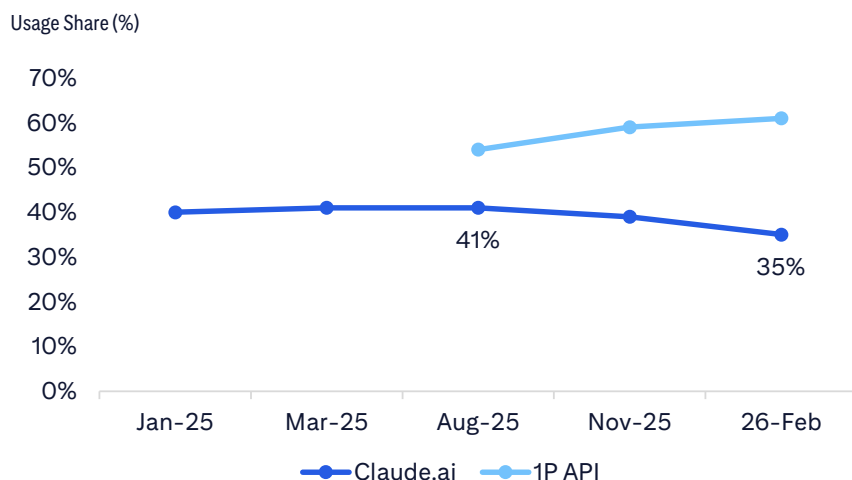
the free Cursor experience paired with a \$20/month Claude subscription proves surprisingly effective for mixed workflows (e.g., using Cursor's free models for patches and Claude Code's desktop app for substantial changes), this fragmented approach will likely consolidate under demand for superior efficiency and effectiveness.

As both platforms recognize that optimal productivity requires seamless integration of deep contextual understanding with sophisticated development tooling, Claude Code may adopt IDE-like features with better codebase mapping while Cursor integrates more powerful AI reasoning capabilities at lower or differently packaged prices. Currently, even with an existing Claude subscription, users must pay additional fees to access premium Claude models or APIs within Cursor. The platform offering the best capability-to-price ratio across variables like versatility, workflow type, model performance, and UI preferences will likely pull ahead, though user taste may ultimately drive platform selection alongside technical superiority.

This convergence is particularly timely given the broader shift toward AI agents that function as persistent assistants across local folder structures, suggesting that tomorrow's development environment could be less about choosing between coding tools and more about selecting the AI operating system that best understands and manipulates a user's entire digital workspace ([Cohorte](#), Mar. 14, 2026).

For Claude.ai, coding-related tasks continues to represent the largest use case at 35%, though is down from a peak of 41% in August 2025 ([Anthropic Economic Index Report](#), Mar. 24, 2026). Given software development's significant share in AI API traffic, we are not surprised at OpenAI's more aggressive focus on the opportunity over the last several months with Codex.

Figure 42. Task Usage Share Trends by Occupation Group: Computer & Mathematical



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Anthropic Economic Index Report

Differentiated Design Edge: Customizing Agentic Behavior

As evidenced by the accelerating ARR of Claude Code, Claude gained rapid adoption partly on account of its coding platform and model capabilities (Figure

40). As of December 2025, Anthropic is estimated to hold 54% of the coding market, followed by OpenAI at 21%, Google at 11%, and other providers collectively at 14% ([Menlo Ventures](#), Dec. 9, 2025). Many companies and users embrace its polished code style and design for backend architectures, data engineering, systems programming, terminal commands, and other core domains ([Claude](#), Mar. 20, 2026; [Alibaba](#), Mar. 12, 2026). Its system prompts, training, and internal design are likely very influential in its distinguished execution of best practices ([Claude](#), Feb. 17, 2026). Claude further supports style adaptation through auto-generated `claude.md` files, which users can edit to persist project-specific instructions and coding standards ([Claude](#), Nov. 25, 2025). These files can boost in-repo performance by over 10% for Claude Sonnet 4.5 after prompt learning ([Arize AI](#), Nov. 20, 2025). While OpenAI and Google models can obtain similar workflow context via `agents.md`, these files are not created by default and lack the same seamless integration at present ([Agents.md](#), Mar. 20, 2026). However, companies may build out this replicable capability as agentic coding gains further popularity.

In addition, Anthropic integrates Chat, Cowork, and Code within Claude Desktop, providing an easily navigated platform that helps to streamline workflows and reduce friction. OpenAI has now announced plans to similarly offer a centralized desktop app featuring ChatGPT, Codex, and Atlas browser ([CNBC](#), Mar. 19, 2026). Compared to Cursor and Codex, junior engineers at Notion, for example, reportedly prefer Claude Code for executing simpler tasks, as well as skillfully anticipating user intent with limited context and imprecise instructions ([The Information](#), Mar. 23, 2026). Meanwhile, Cursor is known for its IDE interface, while Codex is often believed to excel at accomplishing complex, long-running tasks when guided with precise prompting. Personality also plays a role in the adoption of these tools as assistants ([Andrej Karpathy](#), Mar. 20, 2026).

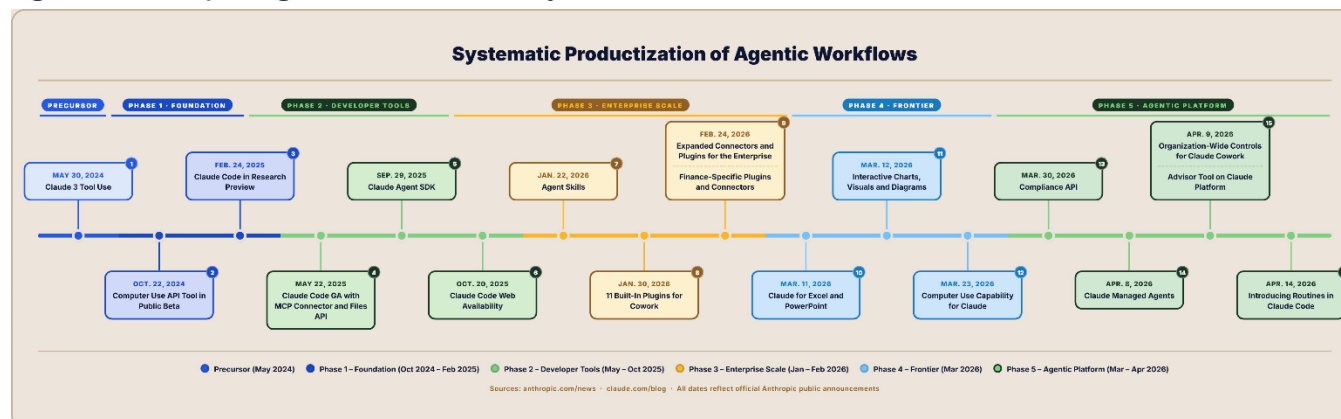
The model's internal context for and alignment with coding best practices, as well as the customization enabled by features like `claude.md` files, position Anthropic well for adapting to user preferences as agentic workflows proliferate ([Claude](#), Mar. 20, 2026). Since both models and developers have distinct coding styles, companies that favor Claude's approach and code design develop stickiness and could incur switching costs, particularly if those companies invest in agentic architectures and further customization. In a recent survey, 57% of respondents mentioned either Claude Code or Claude models as the tool they are most attached to ([The Pragmatic Engineer](#), Mar. 3, 2026). Claude's capacity to align with evolving user preferences and tastes as agentic capabilities expand may be central to securing premium pricing as well as long-term loyalty from both individual and enterprise customers.

Cowork: An Early Prototype of Knowledge Work Agents

As we wrote in our [industry initiation](#) and [OpenAI initiation](#), AI is introducing a new level of abstraction to work, expected to drive significant changes to the \$6T IT market and \$60T labor market inclusive of ~100 million US knowledge workers/~1 billion global knowledge workers. Cowork, launched as a research preview on Jan. 12 and made generally available on Apr. 9, marks Anthropic's first agentic product — and its clearest bid yet to capture the large, underpenetrated market for AI-driven knowledge work automation ([Anthropic](#)). As a clear example of AI-squared, Cowork was developed in the span of 1.5 weeks with Claude writing most of the code, and inspired by users creatively applying Claude Code to non-coding tasks such as building slide decks, organizing files and emails, cancelling subscriptions, among many other use-cases ([Boris Cherny tweet on X.com](#), Jan. 12, 2026). Rapid

follow-on announcements around Cowork's extensibility and customization underscored Anthropic's intent to bring Claude Code's autonomous, multi-step execution model to the enterprise.

Figure 43. Anthropic's Agentic AI Product Journey

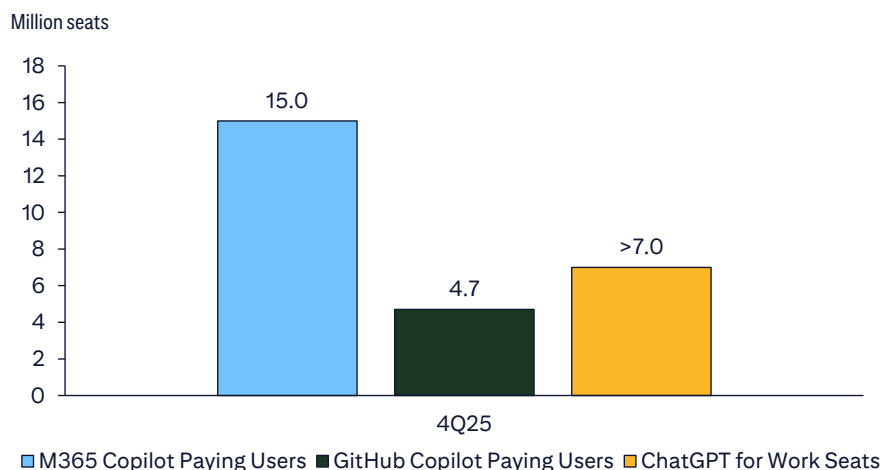


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Anthropic

The rationale behind Anthropic's Cowork push is clear: AI platform companies' recent disclosures on paid seats suggest **the relative opportunity for knowledge work automation may be at least three times as large as code development**, at the current pace of adoption. This is evident in Microsoft's most recent FQ2 FY26 disclosure of over 15 million paid M365 Copilot seats (and "multiples more enterprise chat users") which is ~3x higher than the company's 4.7 million paid GitHub Copilot seats. OpenAI has also already surpassed 7 million seats for ChatGPT for Work (OpenAI, Nov. 5, 2025). **An estimate of roughly 2-4 million paying Claude Code developers today and an assumption of a 3:1 knowledge worker-to-developer employee ratio suggests the potential for 6-12 million Cowork users in the near-term.**

Figure 44. Addressable knowledge work seats could be at least >3x higher than coding market



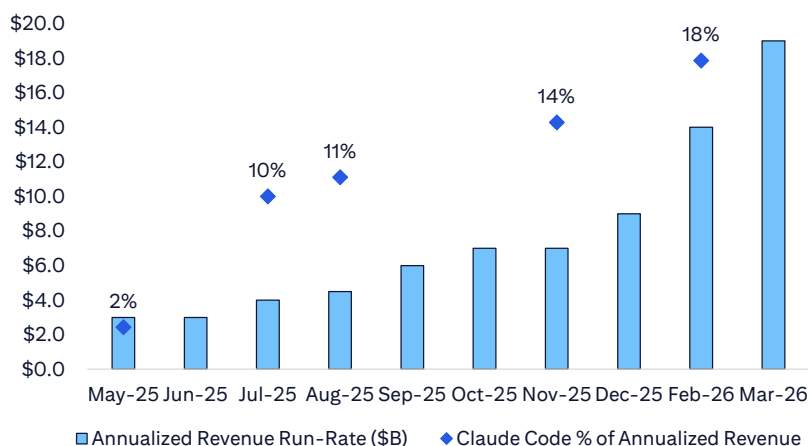
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Company Reports

Cowork's monetization logic mirrors Claude Code's — more capable agents drive more Claude usage, across both subscriptions and the API. Just as Claude Code's proliferation fueled rapid ARR growth, Cowork advances should similarly lift both revenue streams (Figure 45): seat additions from broader enterprise deployments and rising token consumption as agentic workloads grow in complexity and scale. Anthropic's early April ARR of \$30+ billion was already approaching its December 2025 projection of \$32 billion in annualized revenue exiting 2026 ([The Information](#), Apr. 7, 2026).

Enterprise uptake may be more gradual for Cowork than Claude Code. Claude Code penetrated the enterprise bottom-up as individual developer adoption led to spend scaling up organically. Cowork, aimed at non-developers, requires IT and management buy-in from the start, which suggests a fundamentally slower adoption curve, but a larger TAM upside. While the GA launch addressed the access control layer with role-based access controls, group spend limits, and per-tool connector controls ([Anthropic](#), Apr. 9, 2026), the product's broader scope carries higher risk, and adoption remains gated by security, governance, and auditability gaps, particularly for regulated workloads (HIPAA, FedRAMP, FSI). In particular, Cowork activity is not captured in Audit Logs, Compliance API, or Data Exports, which is critical for regulated workloads ([Anthropic](#)). Early movers like Zapier and [Thomson Reuters](#) are already piloting, but broader enterprise traction likely awaits clearer compliance frameworks. Near-term adoption may be prosumer-led with Pro and Max subscribers experimenting at the margins.

Figure 45. Claude Code's Contribution as a % of Anthropic's Annualized Revenue Run-Rate Has Steadily Risen



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Anthropic, The Information, VentureBeat

In addition to stronger governance and audit controls needed on the product side to accelerate adoption, the rate of agentic revenue growth will be influenced by how effective enterprise deployments are. **Anthropic's revenue capture from knowledge work automation will be fueled by multiple go-to-market motions:** 1) organic, user-led demand continues to drive adoption, 2) strategic technology integrations and GSI partnerships expand its reach, and 3) a growing salesforce champions land-and-expand deals. GSIs [partnerships](#) account for some of Anthropic's largest enterprise AI deals signed to date e.g. Claude rollout to Cognizant's 350,000 employees ([WSJ](#), Nov. 3, 2025), Deloitte's 470,000 employees ([Anthropic](#), Oct. 6, 2025), and 30,000 professionals through the

Accenture Anthropic Business Group ([Accenture](#), Dec. 9, 2025), and are aimed proving out ROI, ensuring successful implementations, and continuing to generate demand for Anthropic's offerings. Anthropic's broader suite of enterprise connectors relative to OpenAI is another meaningful commercial differentiator. By lowering the integration barrier for enterprises looking to connect Claude to existing SaaS and enterprise systems, it helps smooth the on-ramp to adoption at scale (Figure 46).

Figure 46. OpenAI and Anthropic Partnerships for Non-Lifestyle Apps/Connectors

Category	Common Apps/Connectors	OpenAI Exclusive Apps	Anthropic Exclusive Connectors
Productivity and Collaboration	Airtable, Asana, Atlassian Rovo, Base44, Box, Circleback, Clay, ClickUp, Egnyte, Granola, Jotform, Linear, Mem, Monday.com, Notion, Pylon, Slack	Dropbox, Dropbox Dash, GoDaddy, Habitify, Jimdo, Lucid, MarcoPolo, Network Solutions, Outlook Calendar, Outlook Email, Readwise, SharePoint, Teams, Upwork, Zoom	Calendly, Campfire, Clockwise, Close, Craft, Docusign, DocuSeal, Fellow.ai, Fireflies, Glean, Gmail, Google Calendar, Google Drive, Guru, Harmonic, Intercom, Krisp, Lumin, Make, Microsoft 365, Oracle NetSuite, Process Street, SignNow, Smartsheet, Zapier, Zoho Projects
Developer Tools	Hugging Face, Jam, Netlify, Replit, Vercel	Ace Knowledge Graph, Ace Quiz Maker, Alpic, Botlab, Botpress, Datadog, GitHub, Neon Postgres, Quicknode, Ranked AI, Retell AI, SendGrid, Sticklight	AirOps, AWS Marketplace, Benchling, Clerk, Cloudflare, Databricks, DevRev, GraphOS MCP Tools, Jentic, n8n, PagerDuty, pg-aiguide, PlanetScale, PlayMCP, Port IO, Postman, Sentry, Stytych, Supabase
Finance and Business Operations	Brex, CB Insights, Intuit QuickBooks, Moody's Analytics, Morningstar, MT Newswires, PitchBook, Quartr, Ramp, S&P Global	Bankrate, Carta CRM, Cube, Dow Jones Factiva, EMPLOYERS, Intuit Credit Karma, Intuit Mailchimp, Intuit TurboTax, Kraken, Neptune Flood, Simply Business, SubscriptionFlow, Third Bridge, TickerSage, Tinman AI	Airwallex, Aiwyn Tax, Blackbaud, Chronograph, Clarity AI, FactSet AI-Ready Data, LSEG, Mercury, MSCI, Pigment, Plaid, Yardi VirtuOSO, Zoho Books
Sales and Marketing	Common Room, HubSpot, Klaviyo, Vibe Prospecting, ZoomInfo	Brand24, Dex, Docket, HighLevel, Lona, Lovable, Outliner, Semrush, Serpstat, Ubersuggest, Wednesday.app	ActiveCampaign, Ahrefs, Apollo.io, Attio, Braze, Clarify, Customer.io, Day AI, G2, Gainsight (Staircase AI), LunarCrush, MailerLite, Outreach, Similarweb, Sprouts Data Intelligence, Zoho CRM
Content Creation and Design	Canva, Cloudinary, Figma, Gamma, Wix	Adobe Acrobat, Adobe Express, Adobe Photoshop, AI Voice Generator, AnyPDF, Apixel, Awesome Slides, Beautiful.ai, BluTranslate, Enhancv, Formbyte, Genspark AI Slides, HeyGen, Invideo, Manus, Shutterstock, Sider Recorder & Transcriber, SlidesGPT, Smallpdf, Speechify, Strive PDF Generator, VEED Fabric	Excalidraw, Magic Patterns, Mermaid Chart, Miro, Sanity, Webflow, WordPress.com
Data Analytics and BI	Amplitude, Daloopa, Hex, Mixpanel, MotherDuck	Statsig, ThoughtSpot Spotter	Bigdata.com, CData, Coupler.io, Crossbeam, DataGrail, Dremio Cloud, Google Cloud BigQuery, Google Compute Engine, Honeycomb, Omni Analytics, PostHog, Qlik, Snowflake, Starburst, Supermetrics, Synapse.org, Visier,
Specialized Industry Solutions	Aiera, Consensus, LegalZoom	Dewa, Domotz (Preview), Ethos, Fotocasa, Fyxr, GovTribe, Imovirtual, McAfee, Medeloop Grants, Norton, Otomoto, Pazy, Scite, SciSpace, Tuio	10x Genomics Cloud, Aura, Benevity, BioRender, bioRxiv, Candid, ChEMBL, Clinical Trials, CMS Coverage, Context7, Cortellis Regulatory Intelligence, Granted, Harvey, Intapp, lastminute.com, LILT, Medidata, Midpage Legal Research, NPI Registry, Open Targets, Owkin, PubMed, Scholar Gateway, Wyndham Hotels and Resorts
HR, Recruitment, and Learning	Gusto, Udemy	Coursera, DataCamp, edX, Indeed, Internshala, italki language learning, Khan Academy, LeapScholar, Quizlet, Scrimba, Skillshare, StraighterLine Degree Planner, Wayground, ZipRecruiter	Dice, Learning Commons Knowledge Graph, Metaview, Microsoft Learn
E-commerce and Payments	PayPal, Stripe	Helium 10, Remitly	Airwallex, Square
General Utilities and Integrations		Linkly URL Shortener, Linktree, Namecheap, Sider Scholar, Spaceship, Streak	Bitly, Tavily, Tool Universe

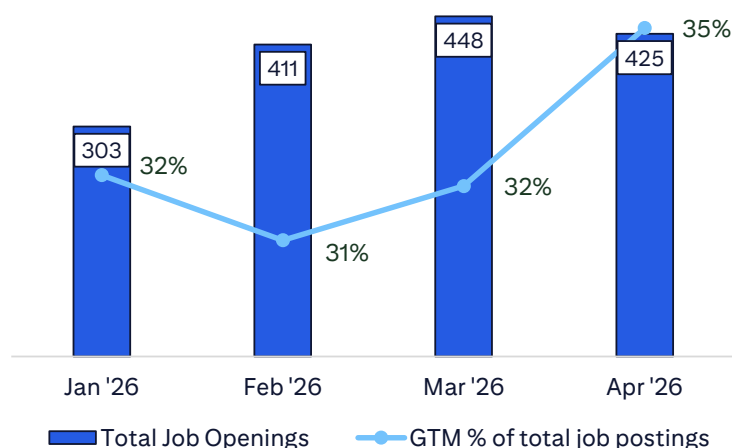
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

To narrow the focus to productivity, the table does not capture lifestyle apps for ChatGPT or health & wellness connectors for Claude.

Source: Citi Research, OpenAI (as of Apr. 23, 2026), Anthropic (as of Apr. 23, 2026)

Alongside important ecosystem partnerships, Anthropic's hiring data suggests it is building out a sophisticated and scalable PLG-to-enterprise salesforce, to drive broader Claude adoption across different customer segments (e.g. strategic enterprises, mid-market, SMBs, startups) and verticals. The GTM strategy is aligned with its deliberate product sequencing: Claude Code lands as the developer wedge, Claude for Enterprise converts shadow usage into formal contracts, and Cowork extends agentic AI to non-technical knowledge workers via MCP-connected data sources, private plugin marketplaces, and sandboxed local environments. Cowork is the surface area that converts a technical API relationship into an org-wide productivity platform.

Figure 47. Anthropic Ramps Up GTM Hiring



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: job postings taken as of Jan. 9, Feb. 8, Mar. 9, and Apr. 8, 2026

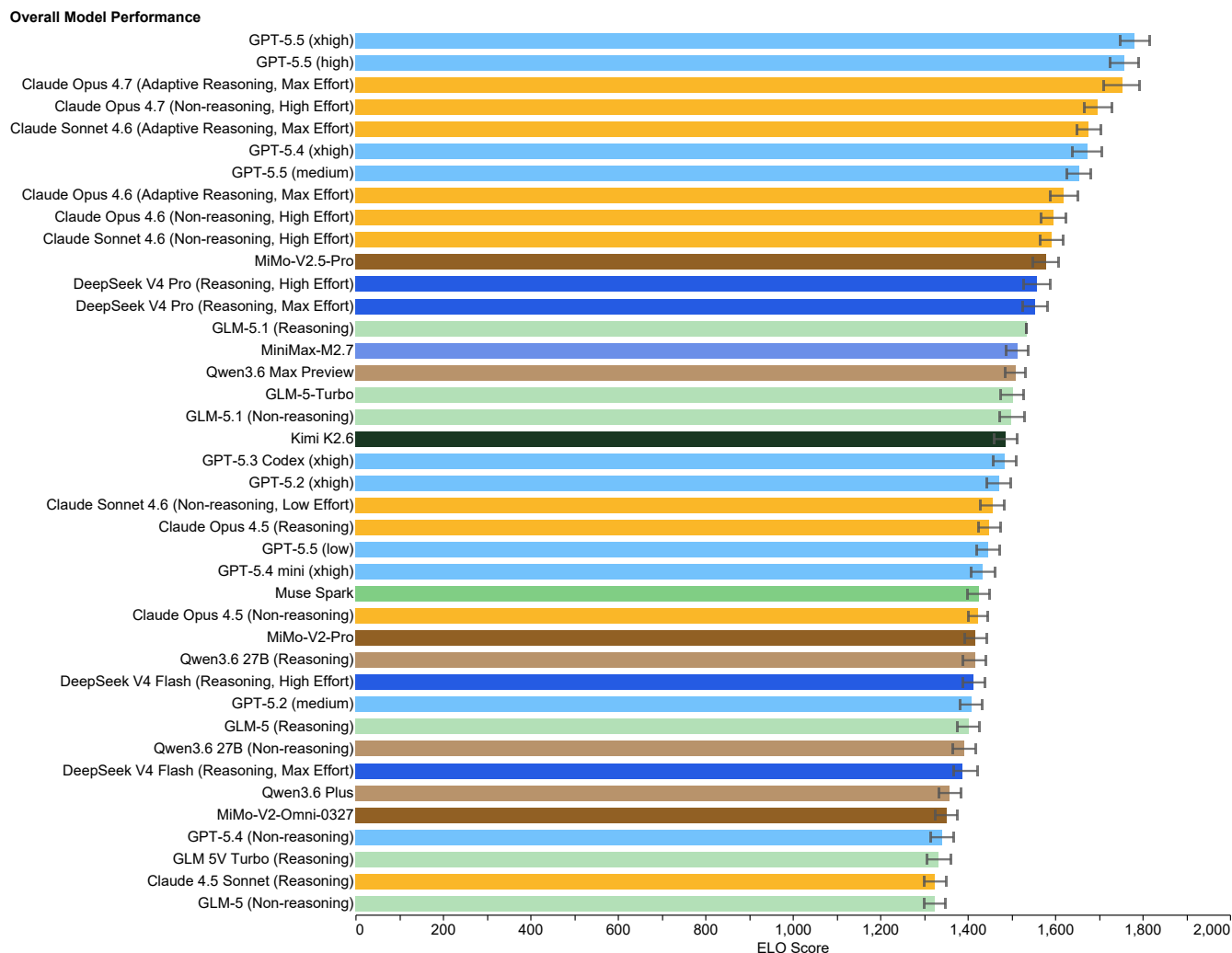
Source: Citi Research, Anthropic

Vertical Expansion

While Anthropic currently bundles Claude Code and Claude Cowork into the core Claude subscription, this strategy does not preclude future standalone, off-the-shelf agents. We expect Anthropic will broaden its agent portfolio, particularly as Claude is further tailored for specific verticals like Financial Services, Health and Life Sciences, Cybersecurity, Public Sector, and distinct job functions such as sales, legal, finance, and design. Before model providers can justify charging premium prices for advanced, standalone AI agents, they must first gain deep domain expertise through real-world production use cases. The practical lessons and challenges uncovered while working with startups and enterprises to build custom agents with Claude and ChatGPT models are essential for shaping successful, high-cost offerings. For example, it has been speculated on [in the press](#) that OpenAI has plans to sell agents at different tiers in the future, including high-end "PhD-level research agents" floated at a potential cost of \$20,000 per month. Investments of this scale require a validated product-market fit ([The Information](#), Mar. 5, 2025).

GDPval is a benchmark designed by OpenAI to track model performance across different occupational roles ([OpenAI](#), Sep. 25, 2025). Since OpenAI's leaderboard was last updated in December 2025, we look first to Artificial Analysis as a proxy for the current performance of frontier models ([OpenAI](#), Nov. 20, 2025; [Artificial Analysis](#), Mar. 18, 2026). The data from Artificial Analysis is less granular; it provides only one overall score for each model instead of breakdowns per occupation. Even so, several patterns emerge around Anthropic's positioning and the broader AI competitive landscape (Figure 48).

Figure 48. GDPval Leaderboard of Overall Model Performance Based on Data from Artificial Analysis



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

The ELO rating system is derived from blind pairwise comparisons and named after Arpad Elo.

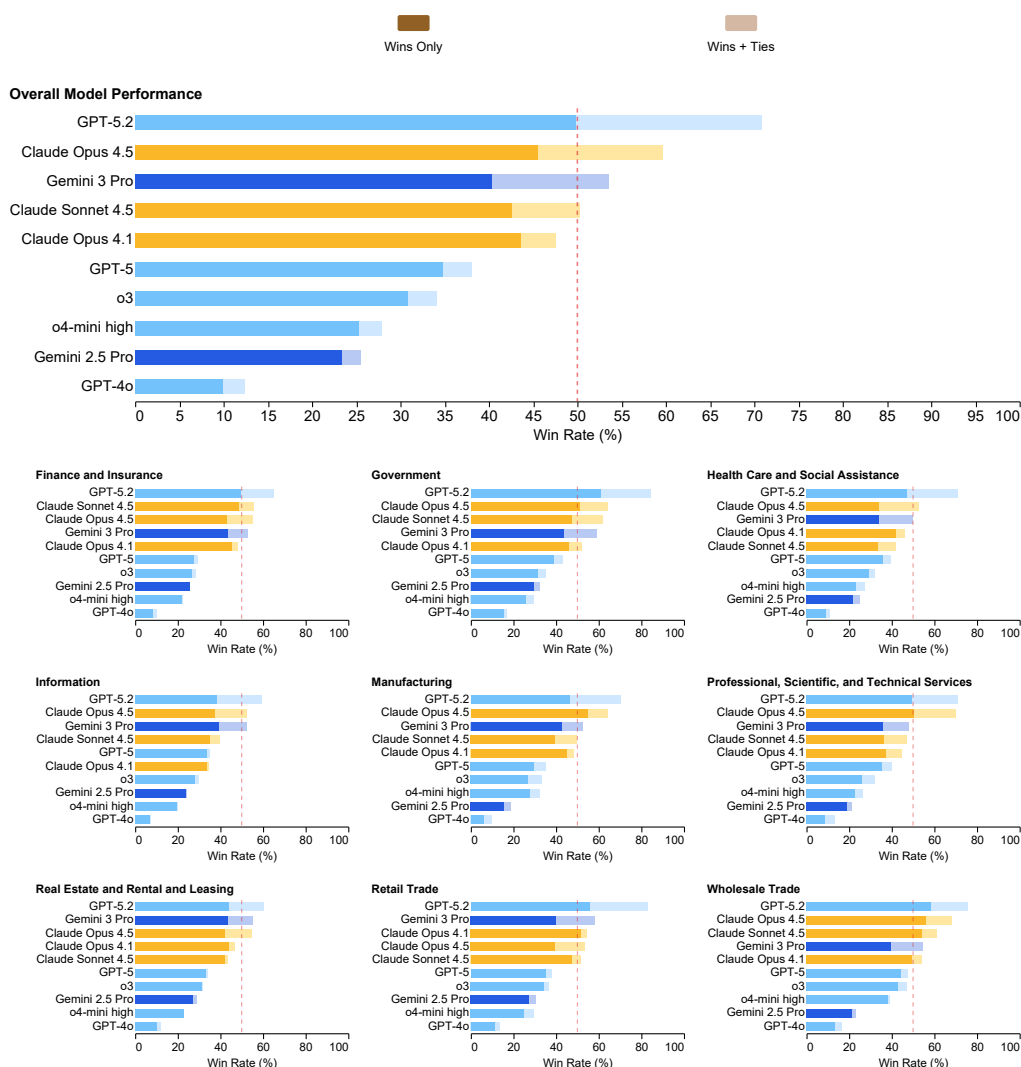
Source: Citi Research, Artificial Analysis (as of Apr. 27, 2026)

The GPT-5.5, Claude 4.7, GPT-5.4, and Claude 4.6 models represent steady improvements. OpenAI currently tops the chart with GPT-5.5 at 1785, and Anthropic follows with Claude Opus 4.7 at 1753 ELO. Additionally, six of the top ten models are Claude variants, while four of the top ten models are GPT variants. Sonnet 4.6 with Adaptive Reasoning (1633 ELO) and Opus 4.6 with Adaptive Reasoning (1606 ELO) markedly outperform Opus 4.5 with Reasoning (1400 ELO). Google and several APAC-domiciled players also rank in the top forty.

Turning to the data below from OpenAI's leaderboard, Opus 4.5, Sonnet 4.5, and Opus 4.1 consistently rank in the top five when compared to ten competitive offerings from Anthropic, OpenAI, and Google (Figure 49). This predominance reflects a strong focus on occupational workloads since at least the training of Opus 4.1, which was released to the public on August 5, 2025. To help put the rankings below into perspective, GPT 5.2 (December 11), Opus 4.5 (November 24), and Gemini 3 Pro (November 18) launched within months of each other in 2025.

Generally, Sonnet 4.5 and Gemini 3 Flash are viewed as default models, with Opus 4.5 and Gemini 3 Pro reserved as premium offerings. Even so, the data shows Sonnet 4.5 outperforming Gemini 3 Pro in finance, government, and wholesale trade (Figure 49). These sectors typically require synthesis and coordination over pure technical depth. Gemini 3 Pro shows slight advantages over Opus 4.6 in real estate and retail trade. GPT 5.2's blanket outperformance compared to Opus 4.6 and Gemini 3 Pro likely reflects OpenAI's strategic response to Anthropic's workplace specialization. The creation of this benchmark and OpenAI's improved results further evidence a broader shift in competitive focus from raw intelligence to workflow fit with application-specific optimization.

Figure 49. GDPval Leaderboard of Overall Model and Sector Performance



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

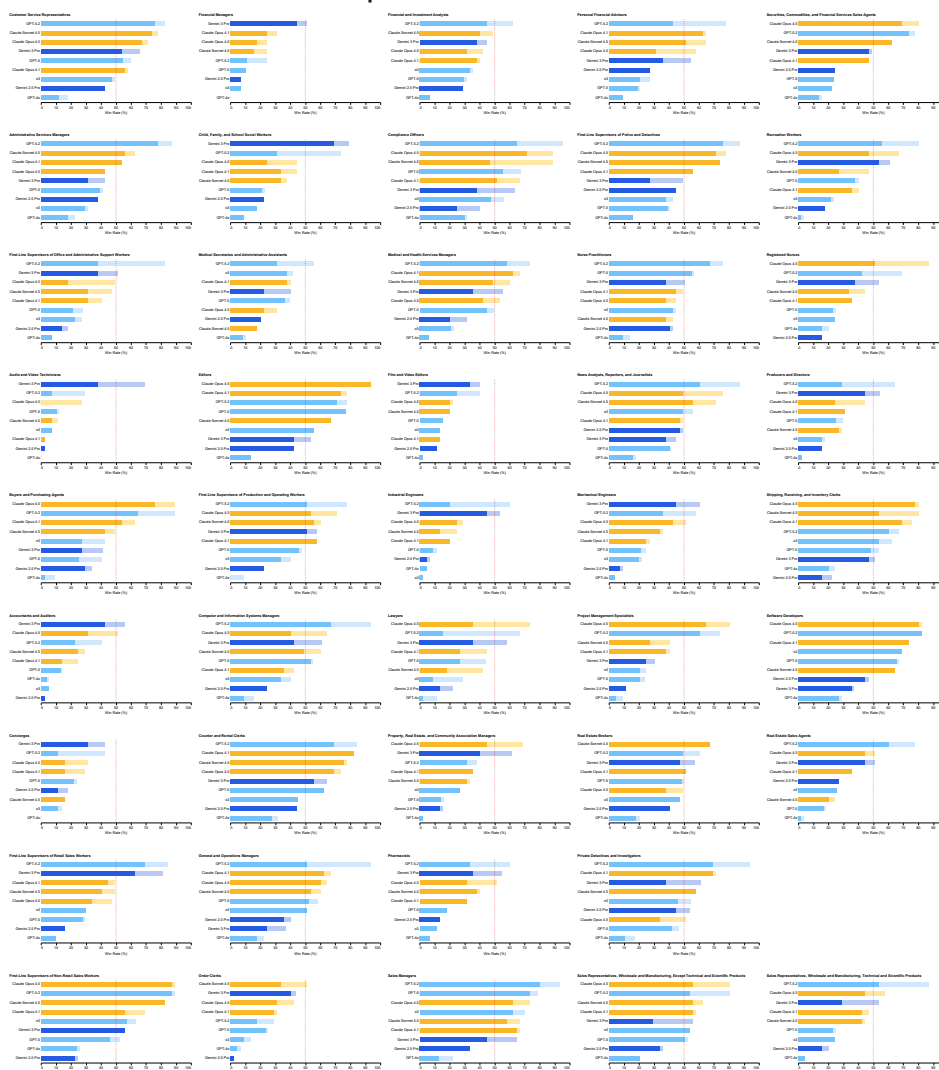
OpenAI last updated its leaderboard on Dec. 11, 2025.

Source: Citi Research, OpenAI

Trends in the occupation-specific data are more variable (Figure 50). Anthropic's models demonstrate strengths in occupations that require communication, coordination, and workflow management. Roles such as editors, project management specialists, software developers, and sales agents consistently show high win rates, indicating Claude's aptitude for synthesizing information, supporting collaborative tasks, and facilitating customer interactions.

Across all models, this data also reveals challenges in jobs requiring advanced technical expertise or visual skills. Opus 4.5 shows lower performance in roles such as financial managers, industrial engineers, and medical secretaries, where deep subject expertise and structured decision-making are essential. Even so, this data does not capture today’s most capable models. Anthropic is actively laying inroads in these specialized domains, as demonstrated by the rollout of Claude plugins for financial services and bio research (Anthropic, Mar. 17, 2026; Anthropic, Mar. 17, 2026).

Figure 50. GDPval Leaderboard of Sector and Occupation Performance



OpenAI last updated its leaderboard on Dec. 11, 2025.

Source: Citi Research, OpenAI

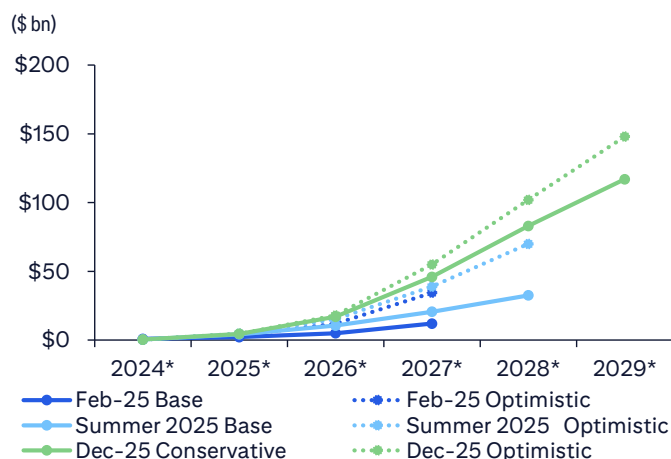
Financial Overview

Company Projected Financials

Like OpenAI, much of the financial information that Anthropic shares with its investors has made its way into the public domain through media outlets. While these figures are not audited and should not be relied upon as confirmed financial data, we highlight our key takeaways.

Revenue growth has consistently exceeded internal forecasts. Anthropic's revenue has rapidly scaled at an ~10x rate for the past three years from \$100M in 2023 to over \$30B by early April 2026 (Figure 52) driven by enterprise adoption, Claude Code success, and burgeoning CSP distribution partnerships. For example, the company exited 2025 at an \$9B ARR -- more than double its "optimistic" expectations as of early 2025 (Figure 51).

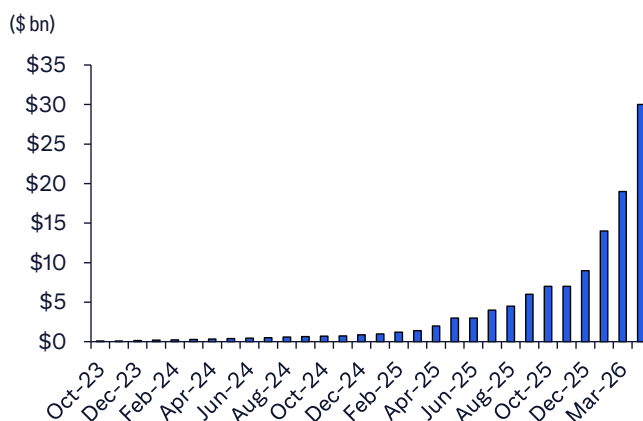
Figure 51. Anthropic Internal Revenue Projections History



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, The Information

Figure 52. Anthropic Annualized Revenue Run-Rate

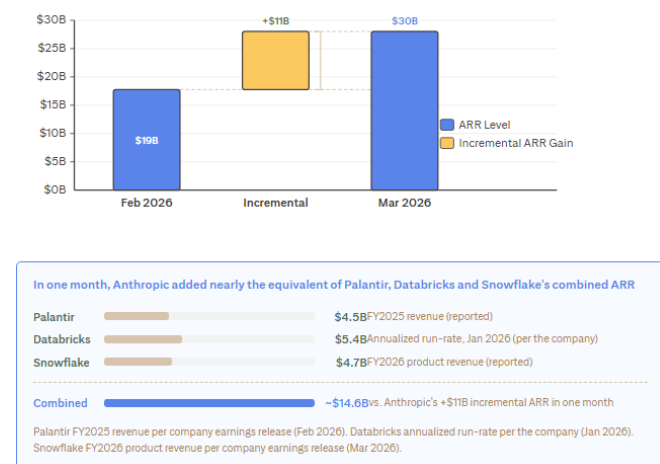


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, The Information, Company Reports

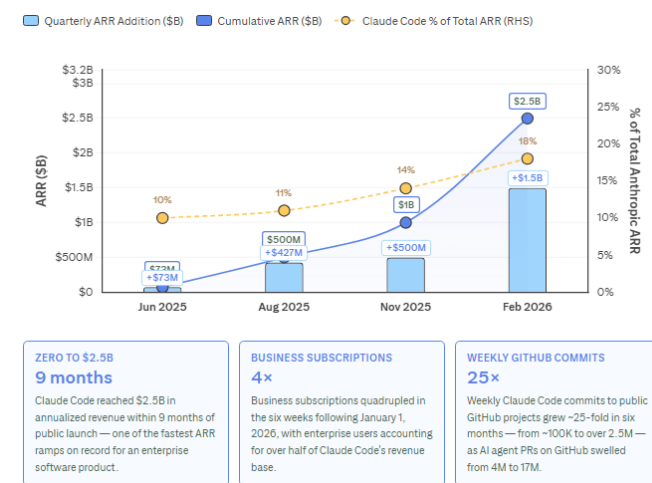
In March 2026 alone the company added an incremental ~\$10B in ARR. For context, \$10B is nearly the combined ARR's of Palantir, Databricks and Snowflake (Figure 53). As of April 2026, the company is running 8 months ahead of its 2026 ARR target of \$32B, based on the Information reporting, suggesting its full-year figure will exceed its December 2025 projection of \$18B. **Claude Code has been a significant product-level revenue driver** (Figure 54). In the nine months since its launch, Claude Code reached a \$2.5B ARR, growing from 10% of total ARR to nearly 20%.

Figure 53. In one month, Anthropic added \$11B in ARR



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Figures are not audited and should not be relied upon as confirmed financial data.
Source: Citi Research, Company Reports

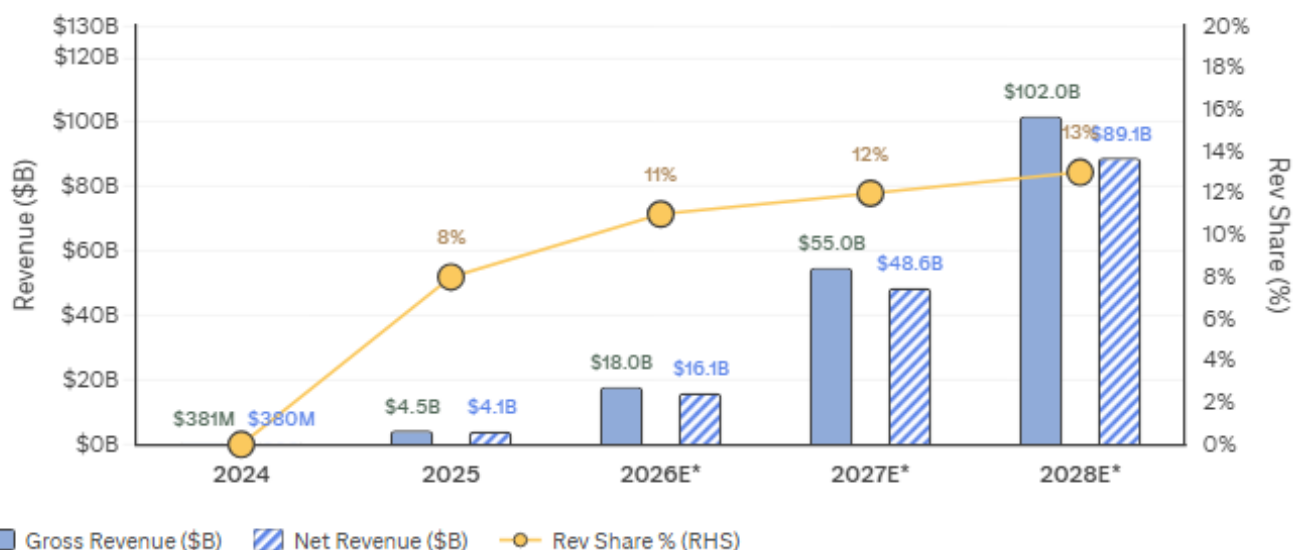
Figure 54. Claude Code has been a significant revenue driver



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Figures are not audited and should not be relied upon as confirmed financial data.
Source: Citi Research, Company Reports, GitHub, The Information

Gross-to-net revenue distinction matters for comparability. Press reports ([Reuters](#), Apr. 8, 2026) indicate that Anthropic's revenue is reported on a gross basis, including the resell revenue retained by AWS, GSP and Azure. Adjusting for estimated resell distributions -- \$2B in 2026, growing to \$13B in 2028 as of the company's Dec. 2025 projections ([The Information](#), Feb. 17, 2026) -- Anthropic's net revenue is ~\$20B less than gross revenue through 2028 (Figure 55). As resell distributions scale, the divergence between gross and net revenue could widen.

Figure 55. Anthropic Revenue Share with Partners (\$B)

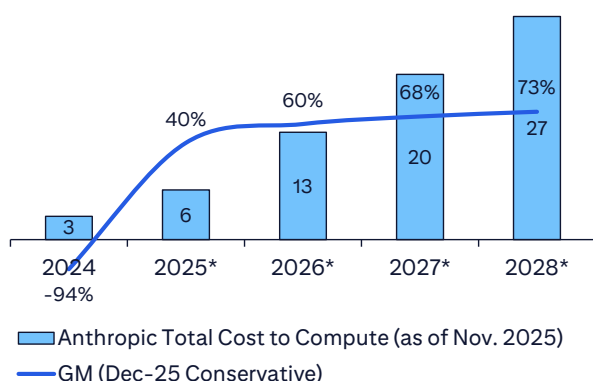


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
*Projected figures. Rev Share % = Partner Rev Share / Gross Revenue. Data per internal Anthropic projections as reported by The Information (Jan. 2026).
Figures are not audited and should not be relied upon as confirmed financial data.

Source: Citi Research, The Information

Figure 56. Anthropic Internal Projections for Cost to Compute and Gross Margins

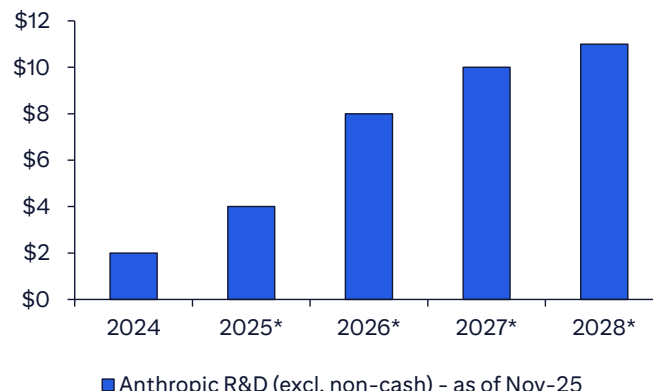
(\$ bn, %)



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, The Information (Nov. 10, 2026)

Figure 57. Anthropic Internal Projections for R&D

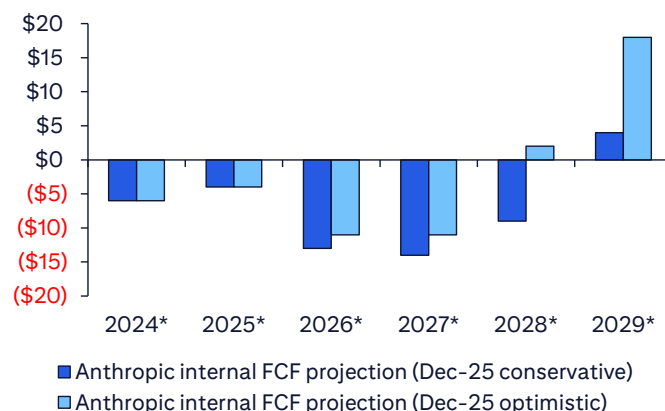
(\$ bn)



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, The Information (Nov. 10, 2026)

Figure 58. Anthropic Internal Projections for Free Cash Flow

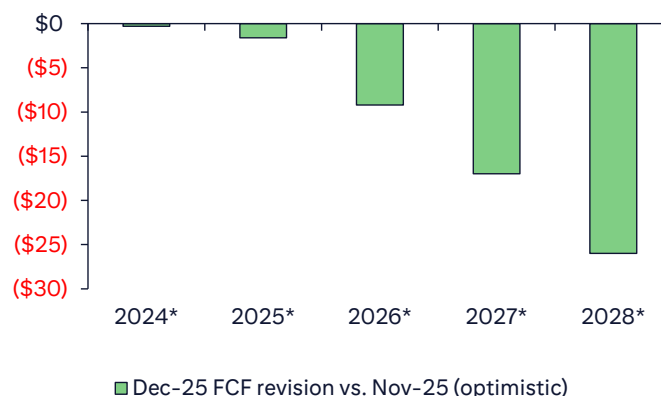
(\$ bn)



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, The Information (Nov. 10, 2026)

Figure 59. Anthropic Dec-25 Free Cash Flow Revision vs. Nov-25

(\$ bn)



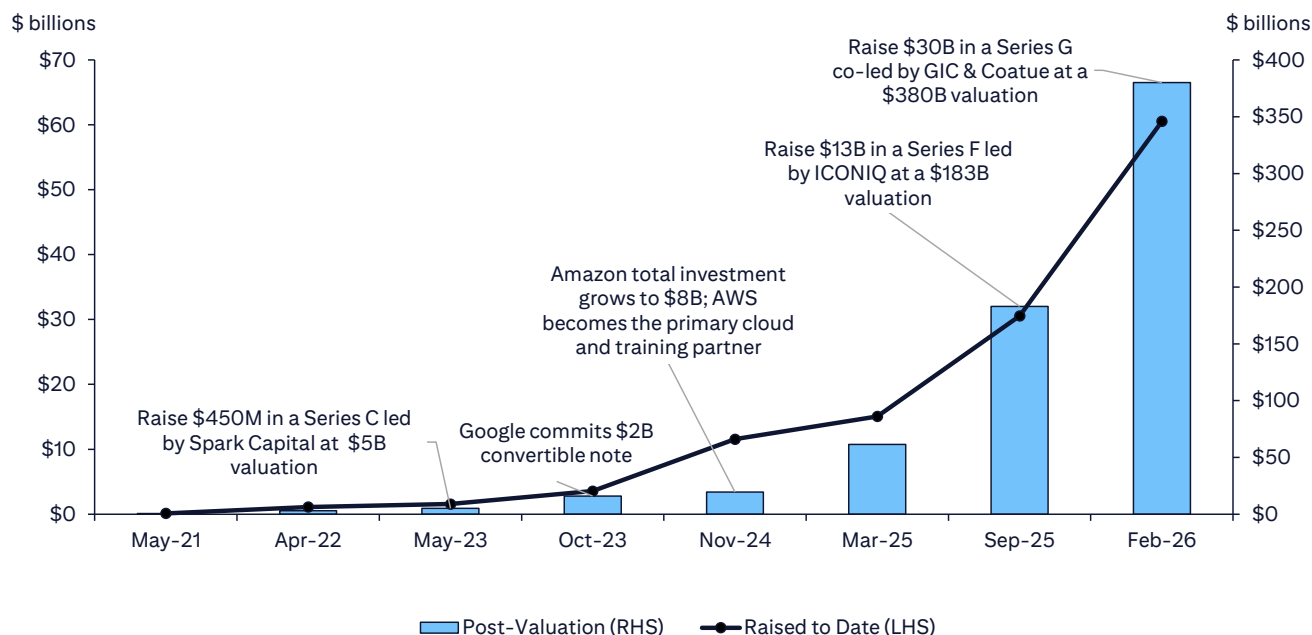
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, The Information (Nov. 10, 2026)

Funding History and Private Valuation

Anthropic has raised ~\$60B across ten funding rounds since its initial Series A in May 2021, culminating in a Series G \$380B post-money valuation as of February 2026 (Figure 60), according to [the Company](#) (Feb. 12, 2026). The company's Series A-C raised a combined \$1.6B at valuations ranging from \$461M to \$5.1B, led by traditional venture investors including Spark Capital. The company's funding profile shifted in late 2023 following Google's \$2B convertible investment and Amazon's two consecutive \$4B investments which solidified AWS as the company's primary cloud and training partner. More recently, the company has

raised \$46.5B in under 12 months via its Series E-G, the latter of which was led by GIC and Coatue. Recent press reports indicate the company has been courted with offers valuing the company at +\$800B, however it's unclear whether the company will raise at these levels ([Bloomberg](#), April 14, 2026). Additionally, Anthropic secured a \$2.5B revolver in May 2025 ([CNBC](#), May 16, 2025), thus diversifying its capital structure.

Figure 60. Anthropic Valuation History



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

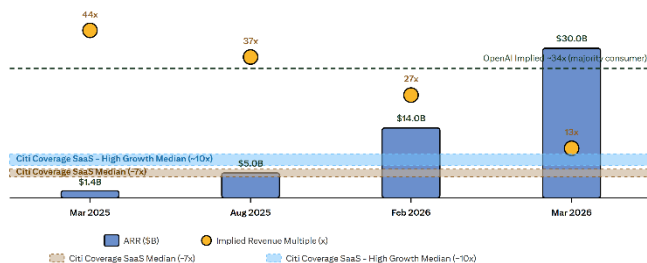
Source: Citi Research, Company Reports, Pitchbook

Anthropic's implied revenue multiple has compressed from 44x in March 2025 to 13x as of March 2026, reflecting the rapid scaling of Anthropic's revenue base over the past year and indicating that valuation support is increasingly underpinned by revenue scale rather than purely by forward expectations ([The Information](#), Mar. 11, 2025). When compared to high-growth public SaaS peers across Citi Software coverage ($\geq 20\%$ CY26 revenue growth, market cap $> \$10B$), Anthropic's most recent implied multiple remains at a slight premium to a ~10x median.

While measurement timing limits the comparability between the two, **Anthropic's discount to OpenAI's 34x** (based on its March 2026 \$852B valuation and Feb. 2026 \$25B ARR) **reflects OpenAI's scaled consumer revenue mix and generalized platform optionality** -- though consumer focused platform businesses typically attract lower revenue multiples than enterprise SaaS. Given that Anthropic reports its enterprise resell revenue on a gross basis, we adjust its share of enterprise revenue to reflect reports that the company expects to payout \$1.9B to its resale partners according to its winter 2025 forecasts ([The Information](#), Feb. 17, 2026). On this basis, Anthropic's implied adjusted net revenue multiple of 17x (Figure 61) remains at a significant premium to public SaaS averages, though Anthropic's resale payouts could grow as a share of gross revenue if the company's revenue mix shifts more towards CSP platform distribution.

Figure 61. Anthropic's March 2026 13x implied revenue multiple sits at a slight premium to high-growth public SaaS peers but at discount to OpenAI's 34x

Anthropic Implied Revenue Multiple



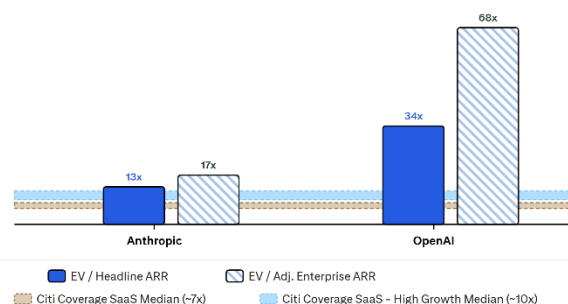
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
OpenAI's 34x based on its \$852B valuation and \$25B ARR

Source: Citi Research, The Information, Company Reports, Thomson Reuters

Figure 62. The gap between Anthropic and OpenAI's implied revenue multiples as of March 2026 is likely due to the latter's consumer mix and perceived platform optionality

Anthropic and OpenAI Implied Revenue Multiples, adjusted for revenue treatments

Metric	Anthropic	OpenAI
Headline EV	\$380B	\$852B
Headline ARR	\$30B	\$25B
% Consumer	20%	50%
Enterprise ARR	\$24B	\$12.5B
Net Adjustment	(\$2B)	—
Net Adj. ARR	\$22B	\$12.5B
EV / Headline ARR	13x	34x
EV / Adj. Enterprise ARR	17x	68x



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Per the Information and other outlets, Anthropic books revenue on a gross basis whereas OpenAI books on a net basis. Consumer/Enterprise split based on the Information reporting and management commentary. Anthropic's \$1.9B net adjustment based on the Information reporting.

Source: Citi Research, The Information, Company Reports