



Important Disclosures

Independent Third-Party Research Report

Important Disclosure

This research report was prepared by an independent third-party provider and is made available by StartEngine Crowdfunding, Inc. ("StartEngine") for informational purposes only. StartEngine and its affiliates, including StartEngine Primary, LLC, a registered broker-dealer, have not independently verified the information contained herein and do not guarantee its accuracy or completeness.

The views expressed in this report are those of the independent provider and do not necessarily reflect the views of StartEngine, its affiliates, or any of their employees.

This report does not constitute an offer to sell or a solicitation of an offer to buy any securities, nor does it constitute a recommendation to buy, sell, or hold any securities. Please carefully review the risks, conflicts of interest, and disclosures provided in the **Important Disclosures** section located at the end of this document.



Important Disclosures

Independent Third-Party Research

Important Disclosures

This research report was prepared by an independent third-party research provider without input, review, or influence by StartEngine Crowdfunding, Inc. ("StartEngine") or its affiliates and is distributed from StartEngine for informational purposes only. StartEngine has not independently verified the information contained herein and does not guarantee its accuracy or completeness. This report reflects the views of the independent provider and does not necessarily represent the views of StartEngine or its affiliates.

StartEngine does not endorse or make any representations regarding the conclusions or recommendations contained in this report. This report does not constitute a recommendation or solicitation by StartEngine to buy, sell, or hold any securities.

Investing in private companies involves a high degree of risk, including the loss of your entire investment. Past performance does not guarantee future returns.

This report is made available to all individuals on startengine.com and may be distributed to individuals who have expressed interest in specific companies or industries.

Please carefully review the company-specific risks, conflicts of interests, and disclosures below.

Company Specific Disclosures:

Conflict of Interest: StartEngine or its affiliates sell membership interests of:

- Series 32-1, a Series of StartEngine Private LLC, which owns fractional interest in an OpenAI Group PBC Convertible Interest Rights Agreement.
- Series 32-2, a Series of StartEngine Private LLC, which owns fractional interest in an OpenAI Group PBC Convertible Interest Rights Agreement.
- Series 11-1, a series of StartEngine Private Funds LLC, which indirectly owns A-2 Preferred Shares of OpenAI Group PBC via a special purpose vehicle.
- Series 11-2, a series of StartEngine Private Funds LLC, which indirectly owns A-2 Preferred Shares of OpenAI Group PBC via a special purpose vehicle.
- Series 32-3, a series of StartEngine Private Funds LLC, which indirectly owns Series C Preferred Stock of OpenAI Group PBC via a special purpose vehicle.
- Series 32-4, a series of StartEngine Private Funds LLC, which indirectly owns Series C Preferred Stock of OpenAI Group PBC via a special purpose vehicle.
- Series 12-1, a series of StartEngine Private Funds LLC, which indirectly owns Series C Preferred Stock of OpenAI Group PBC via a special purpose vehicle.

- Series 17-1, a series of StartEngine Private Funds LLC, which indirectly owns Series C Preferred Stock of OpenAI Group PBC via a special purpose vehicle.

Security Ownership: StartEngine or its affiliates owns fractional interest in a Convertible Interest Rights Agreement for OpenAI Group PBC, Series C Preferred Stock, and A-2 Preferred Shares of OpenAI Group PBC via a special purpose vehicle.

For further questions regarding this report or its distribution, please contact StartEngine at contact@startengine.com representative.



EQUITY RESEARCH

UPDATED

04/01/2026

OpenAI

TEAM

Jan-Erik Asplund
Co-Founder
jan@sacra.com

Marcelo Ballve
Head of Research
marcelo@sacra.com

DISCLAIMERS

This report is for information purposes only and is not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. Nothing in this report constitutes investment, legal, accounting or tax advice or a representation that any investment or strategy is suitable or appropriate to your individual circumstances or otherwise constitutes a personal trade recommendation to you.

This research report has been prepared solely by Sacra and should not be considered a product of any person or entity that makes such report available, if any.

Information and opinions presented in the sections of the report were obtained or derived from sources Sacra believes are reliable, but Sacra makes no representation as to their accuracy or completeness. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a determination at its original date of publication by Sacra and are subject to change without notice.

Sacra accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that liability arises under specific statutes or regulations applicable to Sacra. Sacra may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect different assumptions, views and analytical methods of the analysts who prepared them and Sacra is under no obligation to ensure that such other reports are brought to the attention of any recipient of this report.

All rights reserved. All material presented in this report, unless specifically indicated otherwise is under copyright to Sacra. Sacra reserves any and all intellectual property rights in the report. All trademarks, service marks and logos used in this report are trademarks or service marks or registered trademarks or service marks of Sacra. Any modification, copying, displaying, distributing, transmitting, publishing, licensing, creating derivative works from, or selling any report is strictly prohibited. None of the material, nor its content, nor any copy of it, may be altered in any way, transmitted to, copied or distributed to any other party, without the prior express written permission of Sacra. Any unauthorized duplication, redistribution or disclosure of this report will result in prosecution.



OpenAI

AI research lab offering GPT models via API and ChatGPT for consumers

#ai-models #ai

[Visit Website](#)

Details

HEADQUARTERS
San Francisco, CA

CEO
Sam Altman



REVENUE

\$25,000,000,000

2026

VALUATION

\$852,000,000,000

2026

FUNDING

\$178,000,000,000

2026

Revenue

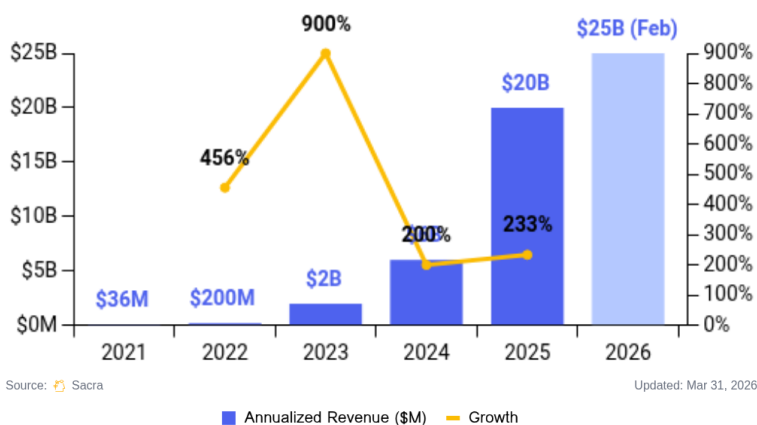


OpenAI

Annualized Revenue & Annualized Revenue Growth Rate

\$25.0B

↑ 233.3% YoY



Sacra estimates that OpenAI hit \$25B in annualized revenue in February 2026, up from \$20B at the end of 2025.

OpenAI CFO Sarah Friar confirmed the \$20B 2025 figure in January 2026, alongside \$6B in 2024 and \$2B in 2023, reflecting a 3x year-over-year growth rate. The acceleration tracks with rising adoption of ChatGPT products across consumers and enterprises, as weekly active users grew to 910M, up from 800M in October and 700M in July, and paying business users surpassed 9M as of February 2026, up from 5M in August.

In September 2025, OpenAI disclosed first-half results of approximately \$4.3B in revenue and \$2.5B cash burn, noting \$6.7B in R&D spend and nearly \$2.5B in stock compensation. The company held approximately \$17.5B in cash and securities at the end of H1 2025. OpenAI projected full-year 2025 revenue of approximately \$13B and cash burn of approximately \$8.5B, with compute and technical talent costs expected to consume approximately 75% of total revenue over the period.

OpenAI monetizes primarily through ChatGPT subscriptions, which account for the bulk of revenue, complemented by API usage and enterprise sales of tools like ChatGPT Deep Research. The company has added productivity features like spreadsheet and presentation editing to drive deeper use cases, while offering bundling discounts on enterprise contracts.

OpenAI posted a 33% gross margin, constrained by inference costs that reached \$8.4B in 2025 and are projected to rise to \$14.1B in 2026, with paying users accounting for approximately 66% of inference spend. OpenAI projects reaching \$85B in revenue by 2030.

The annualized revenue figures exclude Microsoft licensing revenue and large one-time deals, focusing on core subscription and API business. Under a renegotiated partnership in October 2025, OpenAI agreed to pay Microsoft 20% of total revenue through 2032, extended from the original 2030 timeline. OpenAI projected paying more than \$13B in total revenue share—mostly to Microsoft—across 2026 and 2027.

The revised terms allow OpenAI to defer a portion of revenue-sharing payments to later years, with deferred payouts more heavily weighted toward the years leading up to 2032. OpenAI projected cash burn of approximately \$9B in 2025 and \$17B in 2026, not turning cash-flow positive until 2030.

Valuation & Funding

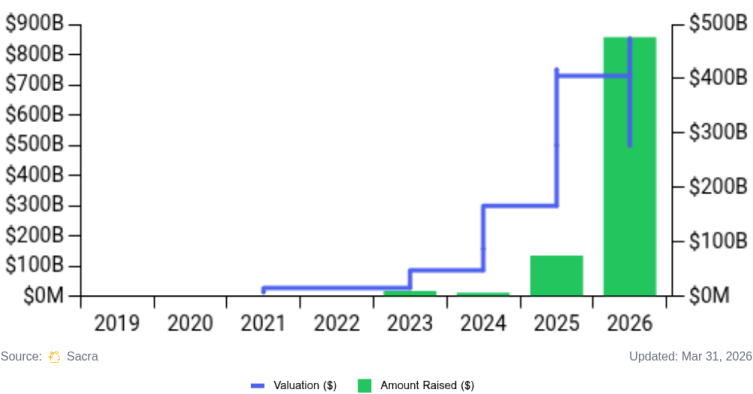


OpenAI

Valuation & Amount Raised

\$852.0B

↑ \$568.4B Total Raised



In April 2026, OpenAI announced \$122B in new investment at a \$852B valuation in a round co-led by SoftBank alongside Andreessen Horowitz, D.E. Shaw Ventures, MPX, and TPG, alongside existing investors.

In October 2025, OpenAI completed a recapitalization converting the for-profit subsidiary into OpenAI Group PBC while keeping the nonprofit—renamed the OpenAI Foundation—in control. The Foundation's equity stake was valued at approximately \$130B; Microsoft's stake in the PBC was valued at approximately \$135B, representing roughly 27% on an as-converted diluted basis. Prior to that, in October 2025, current and former staff sold approximately \$6.6B of stock in a secondary transaction that valued OpenAI at \$500B, with buyers including SoftBank, Thrive Capital, Dragoneer, MGX, and T. Rowe Price.

In March 2025, OpenAI raised \$40B at a \$300B post-money valuation.

OpenAI is a developer of large-scale generative AI models and consumer-facing products, including ChatGPT, DALL-E, Whisper, and developer APIs.

OpenAI is actively laying the groundwork for a 2026 IPO, having hired Cynthia Gaylor, former CFO of DocuSign, as its first head of investor relations to sharpen investor messaging and governance ahead of the listing. Internal targets discussed include a filing in H2 2026 and a 2027 listing, with the company potentially valued at up to \$1 trillion.

Product

ChatGPT

OpenAI was founded in 2015 as a non-profit AI research lab and later restructured into a capped-profit company to attract funding. Over time, it has developed a suite of AI models and consumer products across text, image, and audio generation. Ahead of its planned IPO, OpenAI is refocusing resources on coding and enterprise offerings and cutting side projects to accelerate monetization and sharpen its investor story.

OpenAI's breakthrough consumer product is the ChatGPT assistant, which in 2022–2023 brought large language models into the mainstream. ChatGPT has since evolved into a real-time, multimodal, voice-enabled agent used by more than 900 million people weekly.

The core model powering ChatGPT is GPT-5.4, a frontier model designed for professional work and long-running agents, offering stronger performance on complex, multi-step tasks. GPT-5.4 follows GPT-5.2 (December 2025) and targets further improvements in agentic AI capabilities; Microsoft has already integrated GPT-5.2 into Microsoft 365 Copilot, leveraging its advanced "Thinking" capabilities for enterprise users.

ChatGPT supports real-time spoken conversation, native image understanding and generation, and tool-based actions ranging from web browsing and code execution to manipulating files and integrating with third-party services. Built-in productivity features allow users to work on spreadsheets, presentations, and data analysis directly in ChatGPT, blurring the boundary between AI assistant and cloud productivity suite. ChatGPT for Excel extends this further, enabling AI-assisted analysis inside Microsoft Excel with integrations for financial data providers including FactSet, Dow Jones Factiva, LSEG, S&P Global, and Moody's — available to Business, Enterprise, Edu, Pro, and Plus users in the US, Canada, and Australia.

ChatGPT also offers a dedicated shopping research mode powered by a GPT-5 mini model, designed to improve product information accuracy. Users select the mode, which browses for several minutes and returns a buyer's guide with pros and cons; routine price checks remain faster in regular ChatGPT.

OpenAI has expanded into healthcare with ChatGPT Health, a dedicated space for health conversations with integrations for Apple Health, Function, and MyFitnessPal. The product isolates health chats from general conversations and does not use them to train models, addressing privacy concerns while capitalizing on heavy health usage patterns OpenAI observed in ChatGPT. To lead this effort, OpenAI hired a vice president from Facebook/Instagram, positioning ChatGPT to provide advanced medical insights in the growing AI health market.

OpenAI's first detailed public study of ChatGPT usage found that 73% of chats are non-work related and nearly half of conversations come from users aged 18 to 25, with women making up the majority of the user base.

API

Beyond ChatGPT, OpenAI also provides developer APIs for direct model access. The OpenAI API (launched 2020) lets developers embed GPT capabilities into their own applications on a pay-per-use basis.

Model access through the API has continually improved — from the original GPT-3 to the latest GPT-5 — with major boosts in speed and cost-efficiency in 2025. OpenAI has introduced features like function calling (letting the model return structured data that can trigger programmatic functions) and an Assistants API for building persistent AI agents that can use tools and remember context.

The API also encompasses other model endpoints like the DALL-E 3 image generator and Whisper speech-to-text, reflecting OpenAI's broader AI offerings beyond text. All together, OpenAI's product ecosystem — spanning ChatGPT (consumer and enterprise), developer APIs, and specialized tools — positions the company as a leading AI platform in 2025.

Codex

Codex is OpenAI's dedicated coding product: a developer-focused CLI and native Mac application built for real-time, AI-assisted software development.

Rather than being positioned as a standalone model, Codex functions as an integrated coding environment powered by OpenAI's frontier models. It is designed for interactive programming workflows — writing new code, editing existing files, refactoring large codebases, running tests, and iterating rapidly inside a local development environment.

The Codex CLI embeds directly into the terminal, allowing developers to prompt, patch, and modify code without leaving their workflow. It supports multi-file context, long-running edits, and iterative refinement, making it suitable for professional engineering use rather than one-off code snippets. The native Mac app provides a dedicated desktop coding experience optimized for speed and low-latency interaction, with deep integration into local filesystems and developer tooling, positioning Codex as a direct competitor to tools like Anthropic's Claude Code.

Codex is designed around interruption-friendly, real-time collaboration between developer and model. Engineers can steer changes mid-generation, request targeted diffs instead of full rewrites, and use it for structured tasks such as debugging, migrations, and test generation.

Codex has expanded into cybersecurity with Codex Security (research preview), an AI security agent that builds project-specific threat models, scans repository history and new code, attempts to reproduce suspected flaws in an isolated environment, and proposes patches for human review. Formerly known as Aardvark, the tool connects to GitHub repositories and — across a 30-day private beta — scanned more than 1.2 million commits, identifying 792 critical findings and 10,561 high-severity findings, and has reported critical vulnerabilities to open-source projects including OpenSSH, GnuTLS, and Chromium, resulting in 14 CVEs assigned.

Sora

OpenAI originally introduced Sora in February 2024 as a text-to-video model that brought object permanence and early world-simulation behaviors to AI video. The product has since evolved into Sora 2, a video-and-audio generation model with improved physical realism and controllability, alongside an invite-only iOS social app rolling out in the U.S. and Canada.

The app centers on "cameos" — a one-time video-and-audio verification that lets users insert their likeness (and permitted friends' likenesses) into Sora-generated scenes — plus a TikTok-style feed for sharing and remixing. The app is free at launch, with monetization limited to paying for extra generations during periods of high demand. ChatGPT Pro users can access the experimental Sora 2 Pro model via sora.com and, soon, in-app. Safety and controls include visible/invisible provenance signals, parental controls via ChatGPT, default limits for teens, and granular permissions to revoke cameo access or remove videos at any time.

Sora's entertainment footprint has expanded through a landmark agreement with The Walt Disney Company, enabling Disney character content within the app and broadening the product's use cases across the entertainment industry.

Business Model

OpenAI monetizes its technology via a combination of subscription services and usage-based fees, with an emphasis on turning its popular ChatGPT into a revenue engine.

Subscriptions

Paid subscriptions to ChatGPT account for the majority of OpenAI's revenue. The flagship offering is ChatGPT Plus for consumers, which provides faster responses and early access to new features at \$20/month. This plan has amassed roughly 15 million active subscribers as of mid-2025, making it the single largest revenue source.

Building on this, OpenAI rolled out higher-priced tiers: ChatGPT Pro at \$200/month for power users (with expanded usage limits and priority access), ChatGPT Team (around \$25–30 per user/month for small businesses), and ChatGPT Enterprise (custom-priced, roughly \$60 per seat at list price) for large organizations. These higher tiers have quickly grown the business user base, which stood at more than 9M paying business users as of February 2026. As of December 2025, more than one million organizations use OpenAI's technology.

OpenAI incentivizes enterprise adoption by offering bundle discounts (on the order of 10–20% off) for large deployments. It also continues to add value to subscriptions — for example, the inclusion of new spreadsheet and presentation tools in ChatGPT Plus/Enterprise directly challenges Microsoft and Google's productivity suites.

OpenAI has also introduced geographic-specific pricing to accelerate adoption in key markets. ChatGPT Go, an India-only plan priced at \$5 per month, provides roughly 10× more messages and image generations compared with prior consumer tiers, intended to accelerate user growth and strengthen OpenAI's competitive position in India.

APIs

The second major revenue stream is the API and licensing business. Developers pay usage-based fees to access OpenAI's models via cloud API, which typically costs on the order of \$0.03 per 1K tokens for GPT-4 and \$0.002 per 1K tokens for GPT-3.5, among other pricing tiers.

This model-as-a-service business contributes an estimated 15–20% of OpenAI's total revenue. While smaller in share, it is strategic: by powering hundreds of third-party applications and enterprise software (for instance, powering features in apps like Notion, Salesforce, or Bing), the API extends OpenAI's reach and cements its models as a de facto platform.

OpenAI also earns some licensing income from partnerships — for example, the company's deal with Microsoft integrates GPT-4 into Bing and Azure services (Microsoft effectively pays OpenAI for API usage, entangled with their broader investment deal).

Advertising

OpenAI has added advertising as a third revenue stream, running a limited test of ads inside ChatGPT for logged-in adults in the US on the Free and Go tiers, with Pro, Business, and Enterprise subscriptions remaining ad-free. Ads appear at the bottom of answers when a relevant sponsored product or service is tied to the user's current conversation; they are clearly labeled and separated from organic responses, and include user controls for dismissal and personalization preferences. OpenAI blocks ads from appearing near sensitive topics such as health, mental health, and politics, and does not show ads to users believed to be under 18. The company has sought minimum advertiser commitments of \$200,000 in its tightly controlled beta. COO Brad Lightcap has described the rollout as "iterative," emphasizing that OpenAI is focused on maintaining user trust and getting privacy right as it scales the channel.

Hybrid structure

OpenAI's unusual hybrid structure — combining a capped-profit, for-profit subsidiary with a controlling nonprofit parent — shapes how the company's investors and employees are ultimately compensated. This structure was designed to allow the organization to raise significant outside capital while preserving a mission-aligned governance framework.

Microsoft's \$13B investment in OpenAI over the past few years reflects both the company's capital intensity and this hybrid incentive structure. Microsoft does not hold equity in OpenAI LP; instead, it receives a share of profits. Early investors and employees are entitled to returns capped at 100× their principal. Once OpenAI becomes profitable, those earliest investors get paid back first. Then, 25% of all profits go to early investors and employees (until they hit their cap), while 75% go to Microsoft until it recoups its \$13B in principal.

After Microsoft has recovered its \$13B, the split flips: Microsoft receives 50% of profits until it reaches a total return of \$92B — at which point it too hits its cap. Once that happens, OpenAI reverts fully back to nonprofit control and retains 100% of future profits.

This structure functions like a hedge: it allows OpenAI to raise the capital it needs to survive in a compute-intensive, uncertain market, while preserving a long-term mission-focused structure if the company succeeds. It also helps explain why OpenAI has been so aggressive in monetizing ChatGPT so early — it's not just about product-market fit, but also about proving that the capped-profit structure can sustain a cutting-edge AI company at scale.

OpenAI completed a major recapitalization (October 2025), converting the for-profit into OpenAI Group PBC while keeping the nonprofit — now the OpenAI Foundation — in control. The Foundation now holds equity in the for-profit valued at approximately \$130B and gains additional ownership at future valuation milestones. Microsoft's stake is aligned with the new PBC structure under a restructured commercial agreement that includes a \$250B commitment to purchase Azure cloud services.

Data centers & infrastructure

OpenAI's business model increasingly depends on direct investment and partnerships in large-scale data center infrastructure. The Stargate initiative — a \$500B plan (over four years) to build 10GW of AI data center capacity in the U.S. — represents the company's most ambitious infrastructure effort. Key partners include SoftBank (financial lead), Oracle (operations and hosting), Microsoft (cloud services), NVIDIA (hardware), and Crusoe (Abilene site buildout). The initiative launched in 2025 with \$100B deployed immediately.

The flagship Stargate I site in Abilene, Texas is co-owned by Crusoe and Oracle, with Oracle delivering racks of Nvidia GB200 GPUs. As of mid-2025, the site is partially operational, supporting early GPT-5 training and inference. Five additional US Stargate data center sites have since been announced, bringing the initiative to nearly 7 GW of capacity and over \$400B in investment over approximately three years, with a target of 10 GW and \$500B overall. The expansion is expected to create approximately 25,000 onsite jobs.

OpenAI has committed to over \$500B in disclosed cloud capacity across multiple providers as of early 2026, marking a deliberate strategy to diversify compute sources beyond Microsoft. Under revised partnership terms (October 2025) that removed Microsoft's right of first refusal on new OpenAI cloud workloads, OpenAI contracted to purchase \$250B of Azure cloud services from Microsoft. OpenAI has also expanded its AWS arrangement under a deal in which Amazon will invest \$50B and OpenAI will consume approximately 2 gigawatts of Trainium capacity over 8 years. Separately, OpenAI has committed to a massive cloud deal with Oracle beginning in 2027, widely reported as totaling about \$300B over roughly five years, or about \$30B of annual revenue for Oracle. The cumulative cloud commitments — \$250B Microsoft Azure, expanded AWS, ~\$300B Oracle — represent some of the largest infrastructure deals in technology history.

OpenAI is also pursuing international expansion (e.g., Stargate UAE, Stargate Norway, planned India data center), often in partnership with local governments and infrastructure providers, shifting away from exclusive reliance on Microsoft Azure, adding Oracle, CoreWeave, and considering dedicated storage data centers to control costs and latency.

OpenAI's chip strategy spans partnerships with Nvidia, AMD, Broadcom, Cerebras, and potentially Arm that will power 26GW of data center capacity, implying \$1T+ in build-out costs. The Broadcom collaboration — running for approximately 18 months according to Greg Brockman — focuses on a co-designed inference chip manufactured by TSMC, targeting 10GW between 2026–2029. In parallel, OpenAI is in discussions with Arm on an Arm-designed CPU to pair with its AI server chip; Arm shares rose 11% following the Broadcom announcement. OpenAI has also partnered with Cerebras Systems to add 750 megawatts of ultra-low latency AI compute to its platform, with capacity coming online in multiple tranches through 2028. The multi-year Cerebras arrangement is valued at more than \$10 billion, and OpenAI has already put its first model into production on Cerebras chips: GPT-5.3-Codex-Spark runs on Cerebras' Wafer Scale Engine 3 as part of a "latency-first serving tier" integrated into OpenAI's production stack. OpenAI has emphasized that GPUs remain foundational across training and inference pipelines, with Cerebras complementing that base for workflows requiring extremely low latency.

OpenAI's infra push is both a moat and a bet: control over compute is essential to model development, product delivery, and future margin expansion in the face of rising competition and hardware costs.

Competition

OpenAI's biggest competitors to date are Google, who have their own decade-plus long research in AI now coming to fruition, Meta, whose LLaMa language model competes with GPT-4 from an open source direction, and competing private AI research laboratory Anthropic.

Google

Google has long been a leader in AI research and now directly competes with OpenAI in large language models. In 2023, Google combined its Brain and DeepMind units to accelerate development of its next-generation AI, a multimodal model known as Gemini. Gemini handles text, images, and other modalities in an integrated way, aiming to match or surpass GPT-4's capabilities.

Google's key advantages are data and compute: it can train models on enormous troves of user data from Google Search, Gmail, YouTube, Android, etc., which are proprietary assets unavailable to OpenAI. Moreover, Google effectively owns the world's largest AI computing infrastructure, from custom TPUs to vast data centers. Estimates suggested Google could afford to train models with 5× the computing power of GPT-4 by end of 2023 and 20× by end of 2024.

This scale might yield more advanced models or cheaper inference. Google has integrated AI into many products (e.g. Bard, an LLM chat for search; AI-assisted features in Google Workspace) and has brought Gemini to external users across its apps and products. The latest iteration, Gemini 3, has surpassed rival models on industry benchmarks and reached 650M MAUs, with strong reception for its coding speed and automated design capabilities; tools like Antigravity and generative UI round out the release. Alphabet's stock rose on the positive reception, and Google has priced Gemini 3 Pro at competitive token rates — tightening pressure on ChatGPT as OpenAI acknowledged rough near-term vibes and Google's recent pretraining gains.

Google also benefits from ecosystem control — it can distribute AI features to billions of users via Chrome or Android updates — and from a cash-rich core business (search advertising) that can subsidize free AI offerings. OpenAI, lacking such distribution channels and ad revenue, must rely on partnerships (like the default Bing integration in ChatGPT) to reach end-users.

Meta

Meta (Facebook) has emerged as a major competitor by open-sourcing powerful language models. In early 2023, Meta's LLaMA model (65B parameters) was leaked and then intentionally released as LLaMA 2 under a permissive license by mid-2023. This marked a turning point — for the first time, a top-tier model rivaling GPT-3/GPT-4 was available to the public and researchers.

Developers worldwide have since built upon LLaMA variants, fine-tuning them for specific tasks and even running them on local hardware. Meta's strategy is to drive AI advancement through openness, which in turn pressures OpenAI's proprietary approach.

LLaMA 3 is anticipated by late 2024 or 2025, potentially extending Meta's leap in making cutting-edge models widely available. In addition to model quality, Meta enjoys a massive infrastructure advantage: it reportedly possesses the second-largest installation of NVIDIA H100 GPUs (after Google), giving it ample capacity to train and deploy AI models at scale.

Meta's long history in AI research (from PyTorch to advanced projects like CICERO and Segment Anything) means it has deep expertise. While Meta doesn't directly monetize its models (they are released for free use), its open-source releases threaten to commoditize the core technology. If anyone can download a model nearly as good as GPT-4 and run it cheaply, OpenAI could lose pricing power or see customers opt to fine-tune local models for cost or privacy reasons.

OpenAI is already facing competition from startups that deploy open models at lower costs. Meta's open approach also garners goodwill and a community of developers, which could indirectly benefit Meta's own products and reputation.

Anthropic

Anthropic is a San Francisco-based AI lab founded in 2021 by former OpenAI researchers (including Dario and Daniela Amodei) as a more safety-focused, enterprise-oriented rival. Anthropic's flagship model is Claude, an AI assistant similar to ChatGPT.

From the start, Claude differentiated itself with an ultralarge context window (up to 100,000 tokens), allowing it to digest very long documents or even book-length texts in one prompt.

This made Claude attractive for corporate use cases like analyzing lengthy financial reports or legal documents, where ChatGPT's earlier context limit (~4K–32K tokens) was insufficient. Claude is also tuned for a more cautious, "helpful and harmless" style, which businesses appreciate for reliability in things like customer service chatbots.

Anthropic's focus on B2B use cases and being a "model provider" (rather than building consumer apps) has paid off. By early 2026, Anthropic reached \$19B in annualized revenue, up from approximately \$4B in mid-2025. This surge is driven by large contracts with cloud providers and enterprises (Anthropic has partnerships with Google Cloud and Amazon AWS, and powers AI features in products like Notion and Quora). Investor enthusiasm is high: Anthropic is reportedly closing a new funding round of ~\$5B led by Iconiq at a \$170B valuation, up sharply from a \$20B valuation in early 2024. Such backing gives Anthropic resources approaching OpenAI's.

In competition with OpenAI, Anthropic positions Claude as the safer, enterprise-friendly alternative to ChatGPT — essentially "OpenAI for companies that don't want to rely on OpenAI." Many organizations adopt Claude to diversify their AI stack or to avoid dependency on a single vendor. If OpenAI ever stumbles (in uptime, pricing, or PR), Anthropic stands to benefit as the primary second source. Anthropic's close ties with major cloud platforms (Google invested in 2022; Amazon in 2023) also ensure Claude is well-distributed (e.g. offered as a service on AWS) and integrated into other enterprise tools.

The two companies have also diverged sharply on government and defense access. OpenAI signed a deal to deploy its models inside classified Pentagon environments, with technical safeguards enforcing three "red lines": no mass domestic surveillance, no use to direct autonomous weapons systems, and no use for high-stakes automated decisions such as social-credit scoring. Anthropic refused to accept a broad "any lawful use" standard without explicit carve-outs for surveillance and autonomous weapons; War Secretary Pete Hegseth subsequently directed the Pentagon to designate Anthropic a "supply chain risk" — an unprecedented designation for a US company that Anthropic said it would challenge in court. OpenAI publicly opposed the designation and urged the department to offer Anthropic the same terms.

TAM Expansion

OpenAI's initial market was simply providing AI text generation via chat and API. By 2025, however, the company is aggressively expanding its TAM by pushing into new domains and deeper into the tech stack.

OpenAI's long-term vision is to become an "intelligence layer" in both consumer and enterprise settings. In practice, this means moving beyond just answering questions to executing tasks, facilitating commerce, and integrating with user devices and infrastructure. Key vectors of TAM expansion include:

Enterprise agents

OpenAI is evolving from assisting humans to autonomously performing work on their behalf. With the introduction of GPT-4's function calling and the "Assistants API," developers and enterprises can create AI agents that carry out multi-step operations, not just single responses.

OpenAI has enabled these agents to use tools, browse the web, call APIs, and even control software via a desktop UI. For example, instead of a human filling out forms or clicking through enterprise SaaS menus, a GPT-4o-powered agent could handle tasks like filing expenses, scheduling meetings, updating CRM entries, or processing invoices based on a simple instruction.

Early versions of this are seen in ChatGPT's ability to act as a coding assistant that executes code or as a plugin-based agent that can order groceries or book travel. In companies, such AI agents could cut down the need for junior administrative work or customer support roles, effectively shifting spending from human labor to AI services. OpenAI's models would take a small fee for every task completed, which at scale across millions of tasks becomes a significant new revenue stream – a kind of meter on productive knowledge work.

If this trend takes off, OpenAI could tap into budgets currently spent on business software licenses or even labor outsourcing for routine white-collar work. OpenAI has made significant investments in training models for highly specialized professional tasks, including employing 100+ ex-investment bankers at \$150/hour to train AI that builds transaction-grade financial models for IPOs and restructurings. This effort aims to replace junior banker grunt work and expand agentic task execution across consulting, finance, and legal – sectors representing billions in current labor spend.

The competitive advantage here will be having the most reliable and capable agents, which OpenAI hopes to secure with its head start in model capabilities and its controlled ecosystem for tool use (plugins vetted for ChatGPT, etc.). Enterprise agents also deepen lock-in: once an organization builds an AI workflow around OpenAI's models, it may become as indispensable as an operating system. OpenAI's personal agent strategy has been further reinforced by the hire of Peter Steinberger, creator of the widely-adopted open-source agent framework OpenClaw. OpenClaw has grown rapidly among developers for enabling AI agents that handle multi-step workflows like scheduling and messaging on personal computers; the project will transition to a foundation structure that OpenAI will continue to support while keeping it open source and independent, giving OpenAI influence over a broadly-used agent ecosystem while Steinberger drives the company's next generation of personal agents.

From search to transactions

Another expansion vector is turning conversational AI into a commerce platform. OpenAI (and others like Microsoft) began experimenting with integrating shopping and services directly into chat interactions, fundamentally changing how users discover and purchase.

Rather than referring a user out to a search engine or e-commerce site, ChatGPT can act on a query like "I need a birthday gift for my 5-year-old niece" by showing one perfect product suggestion and a "buy" button right there. This fuses the traditional roles of search (finding information) and e-commerce (transaction) into one step.

OpenAI's monetization strategy extends beyond transaction commissions to include advertising. The company is building out an ads business — having recruited a head of ads to oversee all monetization efforts, including bringing advertising to ChatGPT — as part of a broader push to expand revenue streams beyond subscriptions and API usage. OpenAI stands to capture value through affiliate fees, lead generation payments from merchants, and potentially advertising placements – a new form of AI-age ad placement where brands might pay to have their product recommended for certain queries.

Users benefit from convenience (no need to wade through lists of links or reviews – the AI uses its judgment to present an optimal choice). OpenAI has partnered with platforms like Stripe for payments, and with retailers and aggregators to source real-time product info via plugins. If ChatGPT becomes a trusted purchase assistant, it could take a cut of a huge volume of online sales.

This positions OpenAI not just as a software provider, but as a participant in the retail transaction value chain. It's a high-margin opportunity (commissions can far exceed API token prices) and vastly expands TAM into sectors like shopping, travel booking, food ordering, and other consumer services. In effect, ChatGPT could become an AI concierge for everything – capturing a share of e-commerce without holding inventory or logistics (similar to how Google takes ad fees for steering customers to businesses).

AI operating system

OpenAI is also moving closer to the operating system layer, aiming to become a pervasive interface for users across devices.

ChatGPT was initially accessed mainly via a web browser or mobile app, but OpenAI has since introduced native desktop applications for Mac and Windows. The ChatGPT desktop app lets users call up the AI with a keystroke, use drag-and-drop (e.g. dropping a screenshot or document for analysis), and maintain conversation context across sessions. Moreover, OpenAI has implemented a form of long-term memory in ChatGPT: the AI can remember past interactions or user preferences over time, rather than each session starting fresh.

This means ChatGPT begins to act more like a persistent personal assistant that knows you – for example, it might recall your family members' names mentioned in past chats, or your ongoing projects at work, to better assist you continuously. On mobile, there are deep integrations as well; notably, Apple's iOS 19 introduced an "Intelligence Handoff" that allows Siri to pass complex user requests to ChatGPT behind the scenes, embedding GPT's capabilities at the system level rather than requiring OpenAI to build its own phone OS.

Rather than building its own phone OS, OpenAI is collaborating with platform providers to embed GPT's capabilities at the system level. The ultimate goal is for GPT-based assistance to be ubiquitous – available in every app, every context, whenever a user needs to solve a problem or automate a task.

If OpenAI's AI becomes as fundamental as an operating system, the TAM extends to potentially every digital interaction a person has. This could open up usage-based monetization analogous to an OS license or app store cut, and it greatly increases user switching costs (if your life/work is interwoven with an AI that knows all your context, you wouldn't easily switch to a competitor).

However, it also puts OpenAI in more direct competition with platform owners (like Apple, Google, Microsoft) who have their own assistant offerings – highlighting the importance of OpenAI's strategy to partner (as with Microsoft on Windows, and with Apple's shortcuts/Siri, etc.) to secure distribution.

Government & defense

OpenAI has opened a government and national-security vertical through an agreement with the Pentagon to deploy its AI models inside classified Defense Department environments. The arrangement is cloud-only, with OpenAI retaining control of its "safety stack" and placing cleared, forward-deployed engineers alongside safety researchers embedded with the government. The deal permits the Department to use its AI for all lawful purposes within existing DoD policy, and includes technical safeguards enforcing three "red lines": no mass domestic surveillance, no use to direct autonomous weapons systems, and no use for high-stakes automated decisions such as social-credit scoring. Financial terms and specific models were not disclosed.

OpenAI has framed the Pentagon deal as a template for the broader US AI industry, urging the government to offer the same terms to other labs — an implicit positioning of the arrangement as a reference contract for further federal and allied-government work. The government vertical represents a meaningful TAM expansion into procurement budgets that have historically flowed to defense contractors and specialized government IT vendors, and it gives OpenAI a foothold in classified environments where commercial AI has rarely operated.

Risks

Compute constraints: OpenAI's progress is tightly bound to access of expensive AI compute (specialized GPUs or future AI chips). The company's model development and operations require tens of thousands of GPU cards running in parallel, a resource that is in limited supply globally. Any supply crunch or cost spike in compute could slow OpenAI's model improvements or make its services uneconomical.

Structural profitability: OpenAI currently does not have a clear path to profitability, given its astronomical spending on R&D and infrastructure. Unlike early internet or software companies that enjoyed high margins, OpenAI's gross margins (~40%) are constrained by variable compute costs. While revenue is skyrocketing, expenses are rising just as fast — the company expects to burn \$8B in cash in 2025 on compute and other costs. Cumulative losses will continue to mount (projected \$14B in total losses by 2026 at current rate).

Regulatory & government exposure: OpenAI's Pentagon deal — which permits the Department to use its AI "for all lawful purposes" within existing DoD policy — exposes the company to legal, reputational, and geopolitical risk if its models are implicated in controversial military or surveillance applications. The deal also places OpenAI at the center of a politically charged dispute between the Trump administration and rival AI labs: when Anthropic refused to accept the same terms without explicit carve-outs for surveillance and autonomous weapons use, the Pentagon designated Anthropic a "supply chain risk" — a designation OpenAI publicly opposed. Any future administration could rescind or renegotiate the Pentagon terms, or expand restrictions that undermine OpenAI's government revenue channel.

DISCLAIMERS

This report is for information purposes only and is not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. Nothing in this report constitutes investment, legal, accounting or tax advice or a representation that any investment or strategy is suitable or appropriate to your individual circumstances or otherwise constitutes a personal trade recommendation to you.

This research report has been prepared solely by Sacra and should not be considered a product of any person or entity that makes such report available, if any.

Information and opinions presented in the sections of the report were obtained or derived from sources Sacra believes are reliable, but Sacra makes no representation as to their accuracy or completeness. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a determination at its original date of publication by Sacra and are subject to change without notice.

Sacra accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that liability arises under specific statutes or regulations applicable to Sacra. Sacra may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect different assumptions, views and analytical methods of the analysts who prepared them and Sacra is under no obligation to ensure that such other reports are brought to the attention of any recipient of this report.

All rights reserved. All material presented in this report, unless specifically indicated otherwise is under copyright to Sacra. Sacra reserves any and all intellectual property rights in the report. All trademarks, service marks and logos used in this report are trademarks or service marks or registered trademarks or service marks of Sacra. Any modification, copying, displaying, distributing, transmitting, publishing, licensing, creating derivative works from, or selling any report is strictly prohibited. None of the material, nor its content, nor any copy of it, may be altered in any way, transmitted to, copied or distributed to any other party, without the prior express written permission of Sacra. Any unauthorized duplication, redistribution or disclosure of this report will result in prosecution.

Published on Apr 01st, 2026