SACRA

EQUITY RESEARCH

UPDATED
05/08/2025

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.

OpenAI

Visit Website

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

#ai-models #ai

REVENUE
$4,900,000,000
2024

VALUATION
$300,000,000,000
2025

GROWTH RATE (Y/Y)
206%
2025

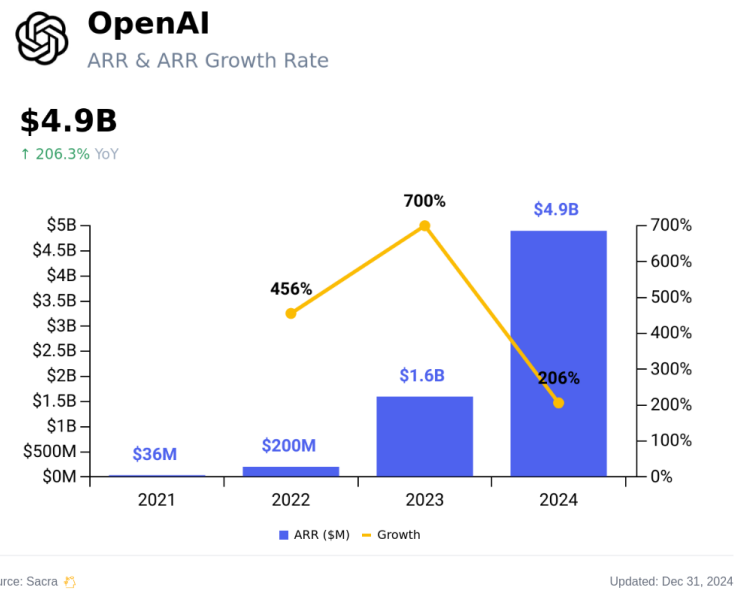
FUNDING
$16,600,000,000
2024

Details

HEADQUARTERS
San Francisco, CA

CEO
Sam Altman

Revenue



Source: Sacra

Updated: Dec 31, 2024

.

Sacra estimates that OpenAI hit $4.9B in annual recurring revenue (ARR) at the end of 2024, growing 206% year-over-year. The company generated about $3.7B in total revenue for the year of 2024, about 4x the previous year.

ChatGPT remains OpenAI's dominant revenue engine, reaching $2.9B in ARR by the end of 2024—up 2x year to date and well ahead of Anthropic's Claude ($120M ARR). OpenAI's consumer moat lies not just in model performance, but in distribution: with over 500 million weekly active users, its biggest competition is increasingly Google ($328B TTM revenue).

OpenAI currently operates at ~40% gross margins, far below the cloud software average of ~74%, but it expects margin expansion as inference efficiency improves. New infrastructure like prompt caching, plus architectural advances (e.g. GPT-4o vs. GPT-4), are already driving down per-token costs. The company projects margins to rise to nearly 70% by 2029.

Looking ahead, OpenAI has told investors it expects to reach $125B in revenue by 2029 and $174B by 2030. OpenAI expects to serve 3B monthly active users by 2030, with 900M DAUs. Monetization of free users, largely untapped today, is projected to generate $25B/year by 2029, possibly through affiliate revenue and shopping integrations rather than ads. The company has also floated other long-term monetization paths—including AI chips and robotics.

Valuation

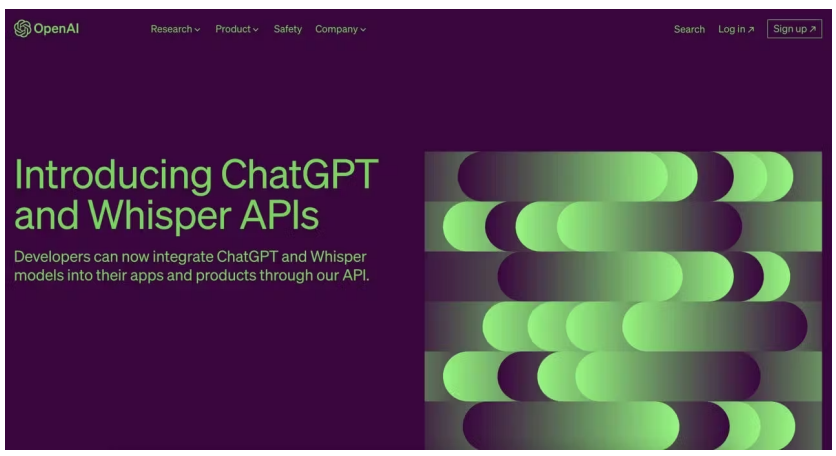
OpenAI is valued at $300 billion as of March 2025, following a $40 billion Series F led by SoftBank, with participation from Microsoft, Thrive Capital, Altimeter, and Coatue.

Based on their September 2024 ARR of $4B, OpenAI was trading at a 75x revenue multiple at the time of the raise.

The company has raised about $64B in total primary funding to date, including strategic investments from Microsoft, Nvidia, and SoftBank. Other key backers include Thrive Capital, Khosla Ventures, and Abu Dhabi's MGX.

Product

OpenAI was founded in December 2015 as a non-profit dedicated to developing "safe" artificial intelligence. Its founding team included Sam Altman, Elon Musk, Greg Brockman, Jessica Livingston, and others.



OpenAI's first products were released in 2016—Gym, their reinforcement learning research platform, and Universe, their platform for measuring the intelligence of artificial agents engaged in playing videogames and performing other tasks.

In 2018, OpenAI published a paper introducing the Generative Pre-trained Transformer, or GPT.

Shortly following the development of the GPT, OpenAI did two things: 1) they launched GPT-2 without disclosing the model code and weights, reneging on their original conception of full transparency, and 2) they transitioned from a non-profit model to a “capped” for-profit model to raise VC and better attract potential employees.

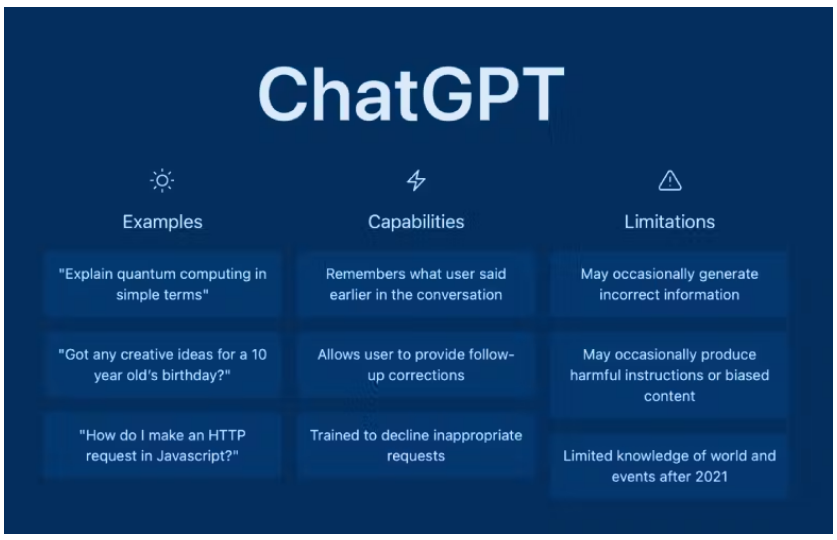
As of 2025, OpenAI offers a small set of core models and a vertically integrated product stack across text, audio, and image generation:

GPT-4o (May 2025): OpenAI’s flagship, natively multimodal model that handles text, images, and audio with high performance and low latency. It replaces GPT-4-turbo as the default model for ChatGPT (Free and Plus) and API use.

o4-mini & o4-mini-high (early 2025): Lightweight versions of GPT-4, optimized for latency and cost, used in enterprise deployments and ChatGPT Team/Enterprise tiers.

o3 (2023–2024): Earlier version of OpenAI’s GPT-4-turbo, now deprecated in most consumer-facing applications.

ChatGPT



In 2015, Magic found extreme product-market fit (and a Sequoia Series A term sheet) with 17,000 requests in 48 hours for its text-based assistant that could get you anything—but it failed to scale because it was completely human-powered behind the scenes.

Magic and similar products like Fin (started by Venmo founder Andrew Kortina and ex-VP of Product at Facebook Sam Lessin) pivoted into becoming workforce automation and analytics platforms for large human teams.

That pattern of product-market fit for an on-demand intelligent assistant re-emerged with OpenAI’s flagship consumer product ChatGPT, consumer-facing app that brought LLMs mainstream.

Today, 500 million people weekly use ChatGPT for tasks like code generation, research, Q&A, therapy, medical diagnoses, and creative writing.

With Voice Mode, launched alongside GPT-4o, ChatGPT allows real-time spoken conversations with the assistant, merging Siri-like UX with GPT-level intelligence.

As of 2025, native image generation in ChatGPT allows state-of-the-art image generation built into GPT-4o.

API

OpenAI launched its API business in mid-2020, beginning with access to GPT-3. The move marked a pivotal shift from research lab to commercial platform, offering developers access to powerful language models via simple HTTP requests—pricing based on token usage.

Early use cases centered on natural language processing: summarization, classification, and basic question-answering. But as models improved with GPT-3.5 and GPT-4, usage expanded to code generation, customer support, product search, document analysis, and more.

The shift to GPT-4o in May 2025 dramatically improved speed and cost—reducing latency and inference costs by orders of magnitude. OpenAI also offers developer tools around:

Function calling: Structured outputs that let developers define API behavior via schema.

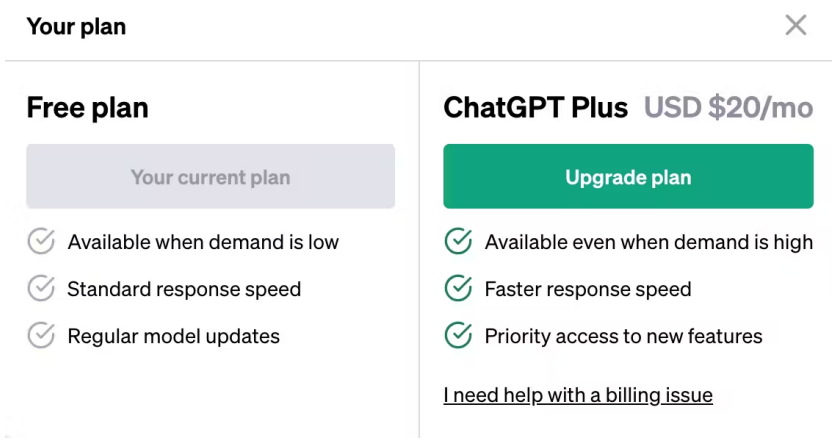
Assistants API: A higher-level framework for building persistent, tool-using AI agents.

File and retrieval tools: Vector storage and retrieval-augmented generation (RAG) for querying custom corpora.

Business Model

OpenAI makes money in a few different ways.

Subscriptions



OpenAI generates the majority of its revenue today through its suite of subscription products under the ChatGPT brand, with API access and licensing forming a secondary revenue stream. Together, these offerings put OpenAI on pace for more than $6.5B in annualized revenue as of mid-2025.

The consumer-facing ChatGPT Plus plan, launched in early 2023 at $20/month, remains the company’s most important single revenue line, with approximately 15.5 million active subscribers.

In addition to Plus, OpenAI has rolled out Pro ($200/month for power users), Team ($25–30/month per seat for SMBs), and Enterprise (custom contracts with published rates around $60/month per seat).

These newer tiers are growing quickly, especially on the business side: by early 2025, OpenAI had approximately 2 million paid business users, a number that includes educational institutions using a discounted $18/month "Edu" plan. Internal targets suggest this could double by year-end.

APIs

GPT-4

With broad general knowledge and domain expertise, GPT-4 can follow complex instructions in natural language and solve difficult problems with accuracy.

Learn about GPT-4

Model	Input	Output
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

GPT-3.5 Turbo

GPT-3.5 Turbo models are capable and cost-effective.

`gpt-3.5-turbo` is the flagship model of this family and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

Model	Input	Output
4K context	$0.0015 / 1K tokens	$0.002 / 1K tokens
16K context	$0.003 / 1K tokens	$0.004 / 1K tokens

OpenAI's second major line of business is its API platform, where developers pay on a usage basis to access models.

Pricing depends on both the model and the context window: for instance, GPT-4 with an 8K context costs $0.03 per 1,000 prompt tokens and $0.06 for completions, while GPT-3.5 is much cheaper at $0.002 per 1,000 tokens.

These APIs power third-party apps and SaaS tools, and are complemented by access to other OpenAI models like the DALL-E image generator, the Whisper audio transcription model, and tools for fine-tuning and embeddings. While this line of business brings in less than subscriptions—about 15–20% of total revenue—it plays a crucial strategic role by embedding OpenAI's models across the broader software ecosystem.

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.

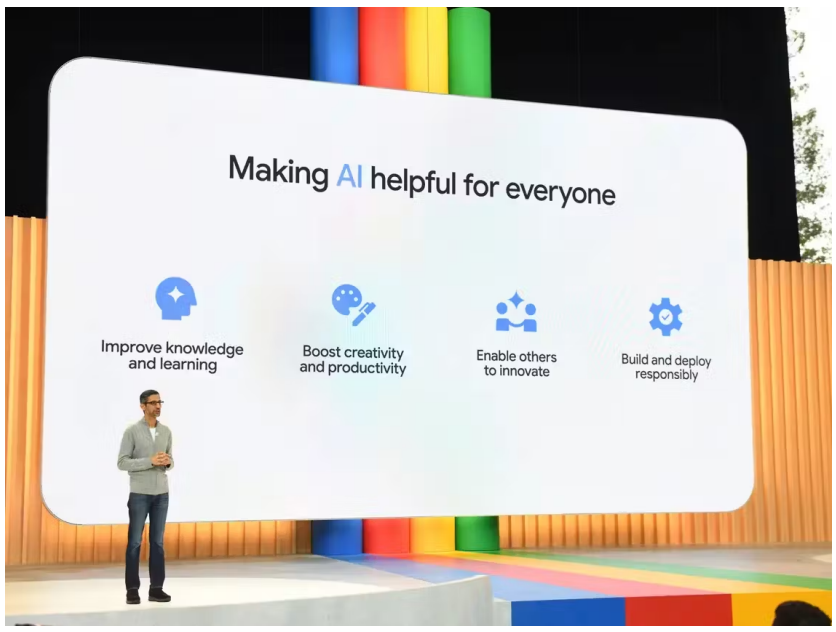
After Microsoft has recovered its $13B, the split flips: Microsoft receives 50% of profits until it reaches a total return of $92B—at which point it too hits its cap. Once that happens, OpenAI reverts fully back to nonprofit control and retains 100% of future profits.

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.

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



In 2023, Google merged its DeepMind and Google Brain AI divisions in order to develop a multi-modal AI model to go after OpenAI and compete directly with GPT-4 and ChatGPT. The model is currently expected to be released toward the end of 2023.

Gemini is expected to have the capacity to ingest and output both images and text, giving it the ability to generate more complex end-products than a text-alone interface like ChatGPT.

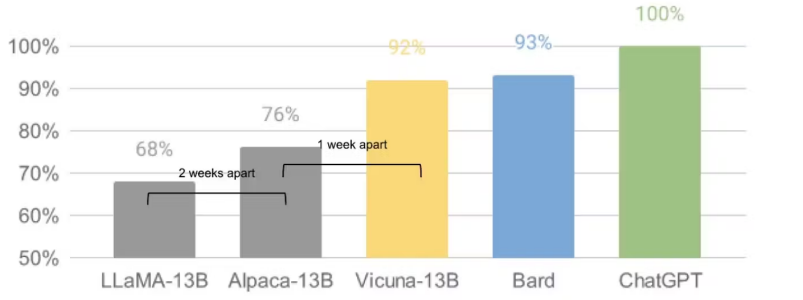
One advantage of Google's Gemini is that it can be trained on a massive dataset of consumer data from Google's various products like Gmail, Google Sheets, and Google Calendar—data that OpenAI cannot access because it is not in the public domain.

Another massive advantage enjoyed by Google here will be their vast access to the most scarce resource in AI development—compute.

No company has Google's access to compute, and their mastery of this resource means that according to estimates, they will be able to grow their pre-training FLOPs (floating point operations per second) to 5x that of GPT-4 by the end of 2023 and 20x by the end of 2024.

Meta

Meta has been a top player in the world of AI for years despite not having the outward reputation of a Google or OpenAI—software developed at Meta like Pytorch, Cicero, Segment Anything and RecD have become standard-issue in the field.



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/

When Meta's foundation model LLaMA leaked to the public in March, it immediately caused a stir in the AI development community—where previously models trained on so many tokens (1.4T in the case of LLaMa) had been the proprietary property of companies like OpenAI and Google, in this case, the model became "open source" for anyone to use and train themselves.

When it comes to advantages, Meta—similar to Google—has the benefit of compute resources that they can use both for developing their LLMs and for recruiting the best talent. Meta have the 2nd most H100 GPUs in the world, behind Google.

Anthropic

Anthropic is an AI research company started in 2021 by Dario Amodei (former VP of research at OpenAI), Daniela Amodei (former VP of Safety and Policy at OpenAI) and nine other former OpenAI employees, including the lead engineer on GPT-3, Tom Brown. Their early business customers include Notion, DuckDuckGo, and Quora.

Notion uses Anthropic to power Notion AI, which can summarize documents, edit existing writing, and generate first drafts of memos and blog posts.

DuckDuckGo uses Anthropic to provide “Instant Answers”—auto-generated answers to user queries.

Quora uses Anthropic for their Poe chatbot because it is more conversational and better at holding conversation than ChatGPT.

In March 2023, Anthropic launched its first product available to the public—the chatbot Claude, competitive with ChatGPT. Claude’s 100K token context window vs. the roughly 4K context window of ChatGPT makes it potentially useful for many use cases across the enterprise.

Despite the advanced state of OpenAI, there are a few different avenues for Anthropic to become a major player in the AI space.
1. Anthropic gives companies optionality across their LLMs

DuckDuckGo’s AI-based search uses both Anthropic and OpenAI under the hood. Scale uses OpenAI, Cohere, Adept, CarperAI, and Stability AI. Quora’s chatbot Poe allows users to choose which model they get an answer from, between options from OpenAI and Anthropic.

Across all of these examples, what we’re seeing is that companies don’t want to be dependent on any single LLM provider.

One reason is that using different LLMs from different providers on the back-end gives companies more bargaining power when it comes to negotiating terms and prices with LLM providers.

Working with multiple LLM companies also means that in the event of an short-term outage or a long-term strategic shift, companies aren’t dependent on just that one provider and have a greater chance of keeping their product going in an uninterrupted manner.

This means that even if OpenAI were to be the leader in AI, Anthropic would still have a great opportunity as a #2—as the Google Cloud to their AWS in a world of multi-cloud, and as a vital option for companies to use to diversify their AI bill.
2. Anthropic is focused on B2B use cases, OpenAI on B2C

Different AI chatbots have different strengths and weaknesses. For example, Anthropic’s Claude chatbot is more verbose than ChatGPT, more natural conversationally, and a better fit for many B2B use cases. On the other hand, ChatGPT is better at tasks like generating code or thinking about code, and for many B2C use cases.

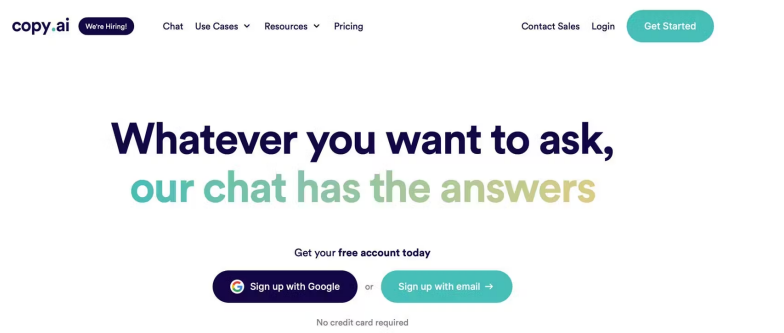
Claude’s 100K token context window means that it is specifically a better fit than ChatGPT for many use cases across the enterprise—something we’ve already seen play out across Notion, DuckDuckGo, and Quora, as well as companies like Robin AI (a legal tech business using Claude to suggest alternative language in briefs) and AssemblyAI (a speech AI company using Claude to summarize and drive Q&A across long audio files). Legal doc review, medical doc review, financial doc review—Claude has applications across industries where large amounts of text and information need to be processed.

Another aspect of Claude that makes it potentially more useful than ChatGPT for professional use cases is the fact that it has been trained specifically to be “more steerable” and produce predictably non-harmful results.

Claude’s more prescriptive approach means it can be relied on to provide more consistent answers with less hallucinations—a tradeoff that might make it less useful than ChatGPT for all-purpose consumer applications, for exploring novel information, or for generating new information like code, but which makes it more useful for e.g. basic service and support tasks that involve retrieving information from a knowledge base and synthesizing it for customers.
3. Anthropic serves businesses rather than competing against them

Anthropic’s focus on the business/enterprise use case of building AI chatbots could be powerful not just in terms of what it allows customers to build but in how it allows them to avoid hitching their wagon to OpenAI.

OpenAI’s hit product is the consumer chatbot ChatGPT, which therefore makes OpenAI potentially competitive with any product building an AI product for consumers.



Since the launch of the GPT-3 API, there’s been a wave of companies building text-based AI products—see AI writing assistants like Jasper and Copy.ai. Jasper and Copy.ai built their businesses reselling OpenAI’s GPT-3 output at ~60% gross margin. Then OpenAI released ChatGPT, with which users can upload a batch of text and have it edited via a chat interface just as they could have within Jasper or Copy.ai.

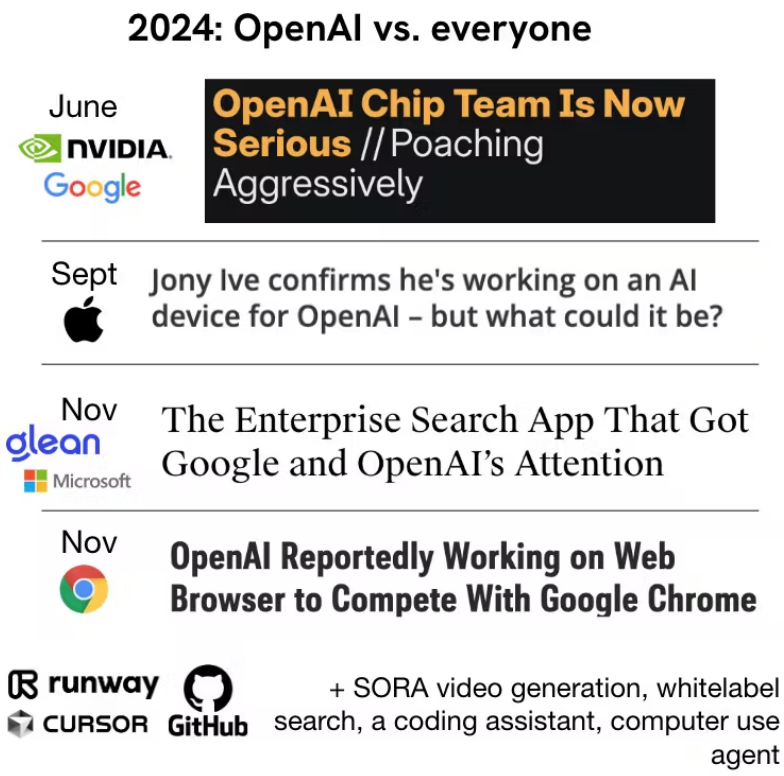
OpenAI’s hit consumer product ChatGPT, while a big success for OpenAI, therefore works at cross purposes to their ability to sell access to their APIs into businesses.

Anthropic, by not having a consumer-facing product like ChatGPT, avoids this issue.

Instead, they can fully focus on developing a product specifically responsive to the needs of businesses, which might mean higher customization, better integration capabilities, a stronger focus on scalability and reliability, white-labeling, or better data privacy controls.

TAM Expansion

OpenAI’s long-term goal is to develop the most capable artificial intelligence that is safe and aligned with human values. But its short-to-medium term growth will come from a much broader set of vectors—each one expanding its total addressable market far beyond the original bounds of chat and API access.



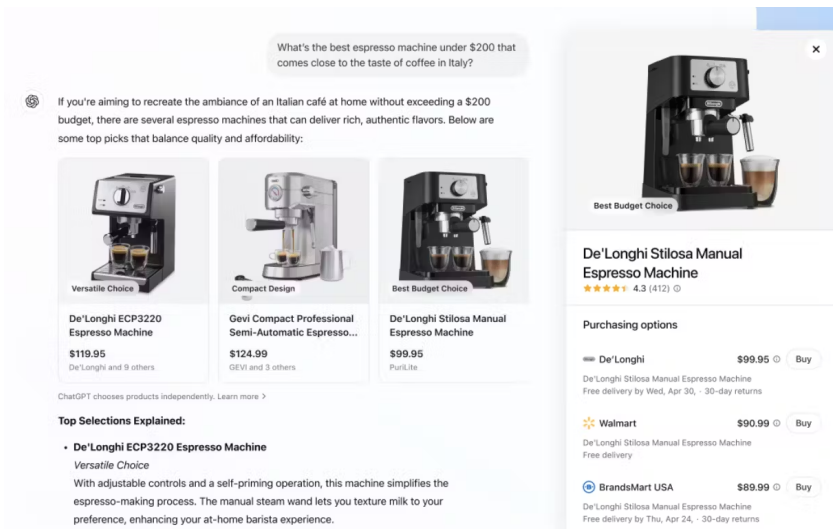
From agents that execute software tasks, to embedded shopping flows, to a full-stack AI operating layer that reaches chips and devices, OpenAI is building toward becoming the default intelligence layer across both consumer and enterprise computing.

Enterprise agents

OpenAI is moving from assisting humans in SaaS tools to actively replacing them in executing software-based workflows. With the rollout of agents that can use tools, browse the web, and control a desktop UI, OpenAI is positioned to automate many of the rote, repetitive tasks that make up white-collar work today. What began as a chatbot helping with drafts or summaries is quickly evolving into an agent that can file expenses, schedule meetings, fill out forms, and book travel—shifting the role of the human from primary executor to monitor or approver.

As a result, spend that once went toward human seats in tools like Rippling, Expensify, or Greenhouse will begin to shift toward AI agent usage—measured not in users but in tasks completed. And because these agents run on OpenAI's models, the company stands to benefit from every marginal task that's automated, collecting a kind of metered tax across the ecosystem.

From search to transactions



With the launch of native shopping flows inside ChatGPT—including product recommendations, shoppable cards, and embedded links—OpenAI is beginning to turn intent into monetizable action. Where Google monetized search through advertising and Amazon through fulfillment, OpenAI is fusing the two by embedding the purchase journey directly inside the answer to a user's question. There's no list of links to scroll through—just the one product that best fits the query, ready to buy.

This new surface area for commerce turns ChatGPT into a native acquisition channel for brands and merchants, and into a potential destination for high-intent queries across categories like fashion, beauty, electronics, and home. It also opens up new high-margin revenue lines—affiliate, sponsorship, even eventual advertising—that extend far beyond API or subscription fees.

AI operating system

ChatGPT has evolved from a web-based chatbot into something closer to an AI-first operating layer. The new desktop app for Mac and Windows allows users to call up GPT instantly, pipe in screenshots or files, and carry context forward across sessions. The addition of long-term memory means the assistant knows who you are, what you care about, and how to help—shifting from a stateless tool to a persistent collaborator.

System-level integrations, like the new Apple Intelligence handoff in iOS, take this further by giving GPT access to the operating system itself. OpenAI doesn't need to own the platform—it simply needs to be available everywhere, ready to process and execute on user intent. The more ubiquitous it becomes, the more it becomes the de facto interface layer between humans and machines.

Vertical stack

OpenAI is rapidly building down and out from the model layer. It is designing its own chips to reduce dependency on Nvidia, potentially lowering inference costs and allowing for tighter control over model-hardware integration. It is reportedly exploring a consumer hardware device with Jony Ive, aiming to build a screen-less, always-on AI product that brings its capabilities into the physical world. At the same time, it is layering on services like shopping, memory, and file management to give users more reasons to stay inside the ChatGPT environment.

Each new layer adds defensibility and value capture. Where the early monetization path depended on tokens and subscriptions, the future looks more diversified—revenue from transactions, from devices, from vertical integrations, and from enabling entire AI-native workflows. OpenAI is positioning itself not just as a single product or model provider, but as a full-stack intelligence platform that can operate across interfaces, industries, and operating systems.

Risks

Compute constraints: OpenAI's growth and model development remain heavily dependent on access to scarce computing resources, particularly advanced GPUs from Nvidia. While the company is reportedly developing its own chips, any supply chain disruptions or pricing changes in computing infrastructure directly impact OpenAI's ability to train new models and scale existing ones. The $5B in losses during 2024 highlights the capital intensity of the business model.

Structural profitability: Unlike other tech giants that had clear monopoly-like profit engines to fund growth, OpenAI currently lacks a similar self-sustaining economic moat. With losses expected to increase to $14B by 2026, the company faces pressure to develop sustainable unit economics before capital markets tighten. The unusual capped-profit structure may also create misaligned incentives between early investors seeking their 100x returns and the need for long-term reinvestment.

Competitive displacement: OpenAI faces a multi-front competitive battle against tech giants with more resources (Google), open-source alternatives that could commoditize base capabilities (Meta's LLaMa), and specialized providers optimized for enterprise needs (Anthropic). With a 75x revenue multiple, investor expectations remain extremely high, requiring OpenAI to maintain both its technical leadership and business execution. Any significant advancement by competitors or shift toward open-source adoption could rapidly erode OpenAI's market position.

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.

Published on May 08th, 2025