OpenAI

[Visit Website]

OpenAI is an AI research laboratory.

#ai-models #ai

Details

HEADQUARTERS

San Francisco, CA

CEO

Sam Altman

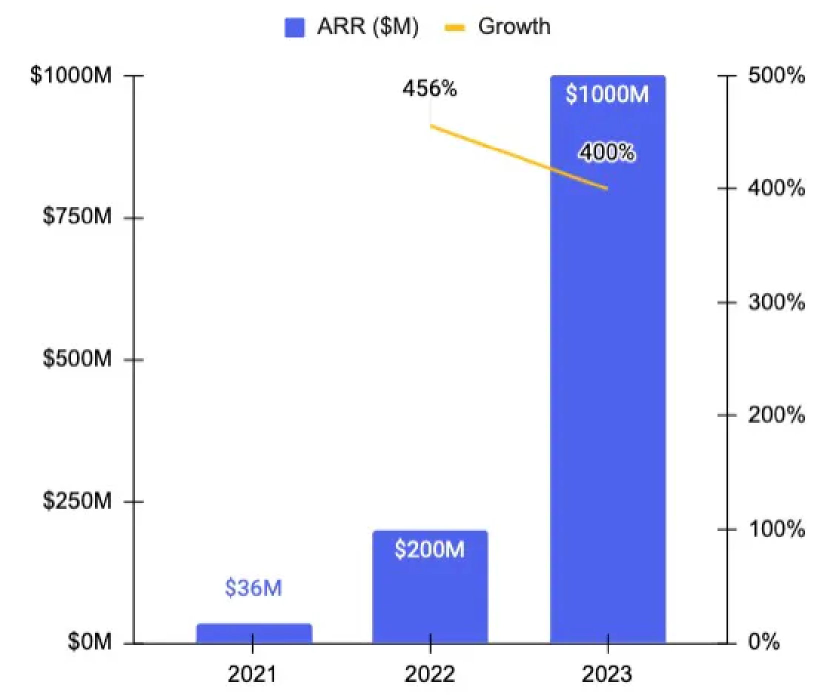
REVENUE	VALUATION	GROWTH RATE (Y/Y)
$1,000,000,000	$28,000,000,000	400%
2023	2023	2023

FUNDING

$11,300,000,000

2023

Revenue



is on track to ht $1B in annual recurring revenue by the end of 2023, up about 400% from an estimated $200M at the end of 2022.

OpenAI overall lost about $540M last year while developing ChatGPT, and those losses are expected to increase dramatically in 2023 with the growth in popularity of their consumer tools, with CEO Sam Altman remarking that OpenAI is likely to be “the most capital-intensive startup in Silicon Valley history.”

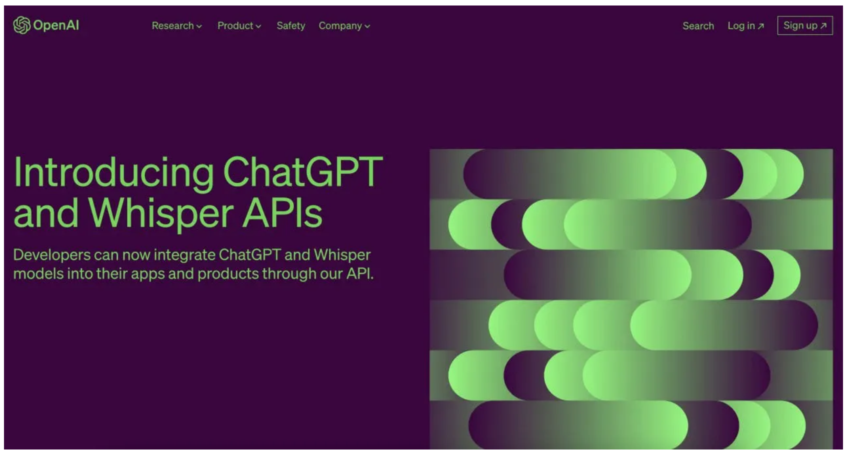
The reason for that is that operating ChatGPT is massively expensive. One analysis of ChatGPT put the running cost at about $700,000 per day taking into account the underlying costs of GPU hours and hardware. That amount—derived from the 175 billion parameter-large architecture of GPT-3—would be even higher with the 100 trillion parameters of GPT-4.

Valuation

In April 2023, OpenAI raised its latest round of $300M at a roughly $29B valuation from Sequoia Capital, Andreessen Horowitz, Thrive and K2 Global.

Assuming OpenAI was at roughly $300M in ARR at the time, that would have given them a 96x forward revenue multiple.

Product



Introducing ChatGPT and Whisper APIs

Developers can now integrate ChatGPT and Whisper models into their apps and products through our API.

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.

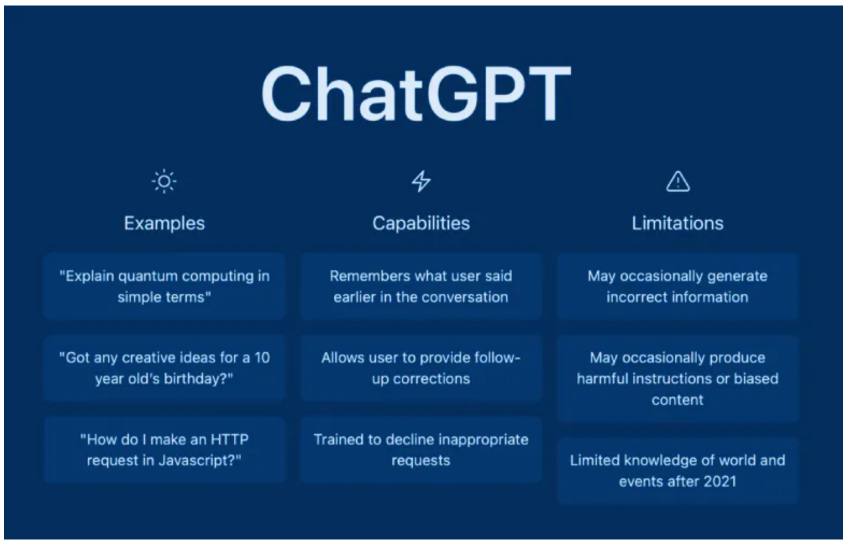
OpenAI’s first products were released in 2016—Gym, their reinforcement learning research platform, and Universe, their platform for measuring the intelligence of artificial agents engaged in playing videogames and performing other tasks.

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.

Today, OpenAI has about two dozen different products across AI-based text, images, and audio generation, including its GPT-3 and GPT-4 APIs, Whisper, DALL-E, and ChatGPT.

ChatGPT



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

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. Compare the $5/hr/person unit cost of overseas labor delivered via a chat interface vs less than $20/month for near-unlimited use of a large language model (LLM).

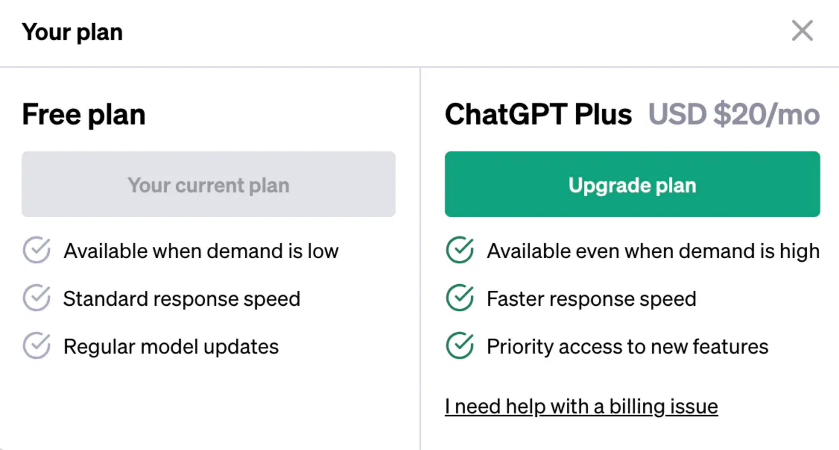
Today, millions of people converse with ChatGPT every day, using it for tasks like code generation, research, Q&A, therapy, medical diagnoses, and creative writing.

With ChatGPT Plugins—OpenAI's 3rd-party app store for ChatGPT—that intelligent assistant can take action for you inside of the apps you use, including Zapier and ~1000 others. With Zapier, for example, you can type "find my last email from Jeff" in ChatGPT and your assistant will use your Zapier plugin to pull the latest email from Gmail and display it in chat.

Business Model

OpenAI makes money in a few different ways.

Subscriptions



Your plan

Free plan

Your current plan

✓ Available when demand is low

✓ Standard response speed

✓ Regular model updates

ChatGPT Plus USD $20/mo

Upgrade plan

✓ Available even when demand is high

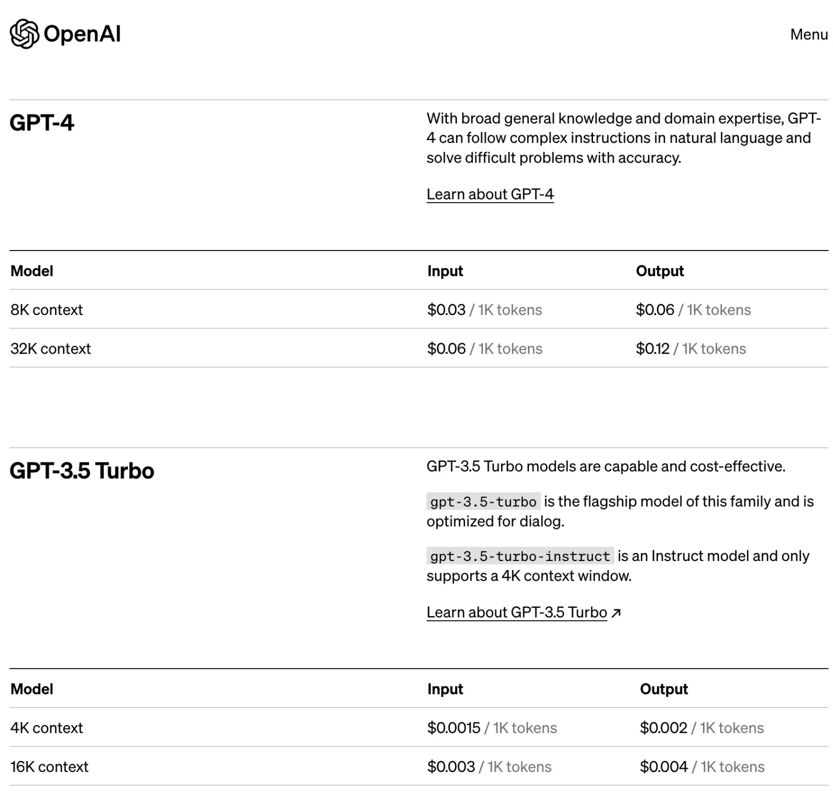
✓ Faster response speed

✓ Priority access to new features

I need help with a billing issue

The main way that OpenAI makes money today is via subscriptions to its ChatGPT Plus product. Pricing is on the flat rate of $20/month. Total sales hit $100M annualized shortly after the launch of ChatGPT's paid tier.

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

The other main way OpenAI makes money is via the usage-based pricing for their APIs offered to businesses building on top of models like GPT-3 and GPT-4. Pricing differs based on the size of the context window and the particular model—for example, GPT-4 with a smaller 8K context window costs $0.03 per 1K tokens for prompts and $0.06 for completions, while the less powerful gpt-3.5-turbo model optimized for chat dialogue costs just $0.002 per 1K tokens.

OpenAI also makes money from renting access to their DALL-E image model, Whisper audio model, and fine-tuning and embedding. Enterprise customers spending upwards of $45K/month get discount pricing.

Hybrid structure

The sky-high of training and deploying models like GPT-4 helps explain the unusual terms of Microsoft's roughly $13B of investment in OpenAI over the last few years.

OpenAI's hybrid corporate structure, with both a for-profit business wing and a non-profit research lab wing, determines how investors in the company will eventually be paid out. A wrinkle in the for-profit wing of the business is that profits are capped: OpenAI's earliest investors and employees are limited to making 100x their initial investment. The combined organization is run by OpenAI's non-profit arm.

Once the for-profit business arm begins to return profits, the first people to get paid out will be the very earliest investors in the company, who will get their principal paid back.

After those early investors are paid out their principal, 25% of OpenAI's profits will go to employees and to pay early investors (until they hit their profit cap), while 75% will go to Microsoft until it recoups its $13B principal.

After Microsoft recoups its $13B, it will get 50% of all OpenAI profits until it gets to $92B (at which point they'll hit the profit cap), while 49% will go to early investors and employees and 2% will go to OpenAI's non-profit arm.

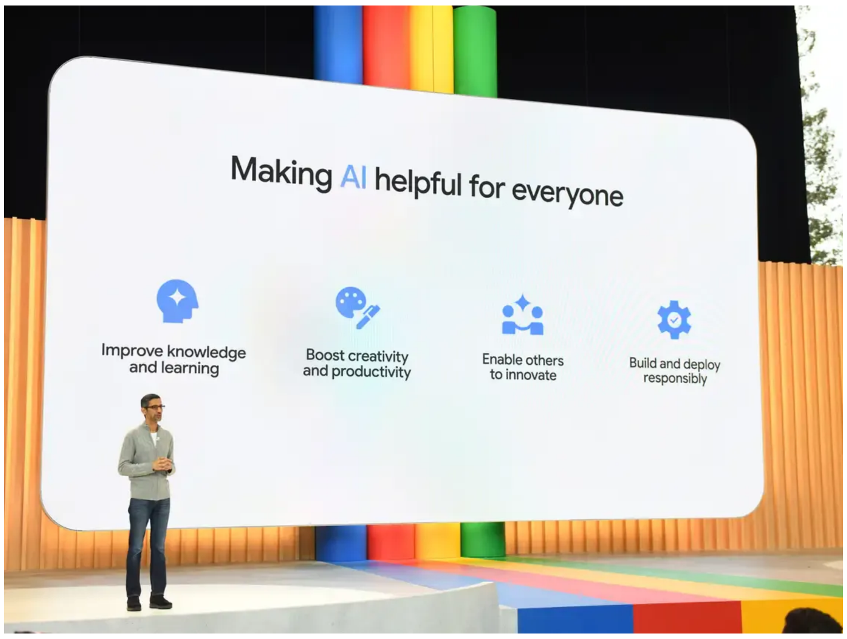
Once $92B in profit is generated and paid to Microsoft—along with that $13B in principal—all equity reverts back to OpenAI, along with 100% of future profits.

This structure operates like a hedge—for OpenAI, it is a way to make sure that the company has the capital and institutional backing that it needs to survive in the short-term given that the profit-making capacity of the company is still unproven, with a large long-term reward in the event that they're successful in making it work.

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



Making AI helpful for everyone
Improve knowledge and learning
Boost creativity and productivity
Enable others to innovate
Build and deploy responsibly

Earlier this year, 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.

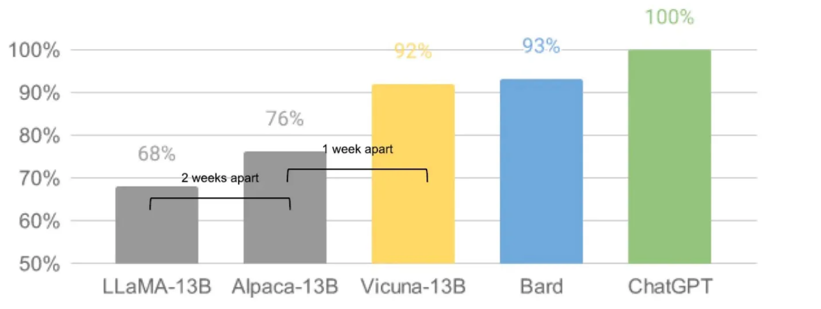
One advantage of Google's Gemini is that it can be trained on a massive dataset of consumer data from Google's various products like Gmail, Google Sheets, and Google Calendar—data that OpenAI cannot access because it is not in the public domain.

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

A next-generation AI assistant for your tasks, no matter the scale

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.

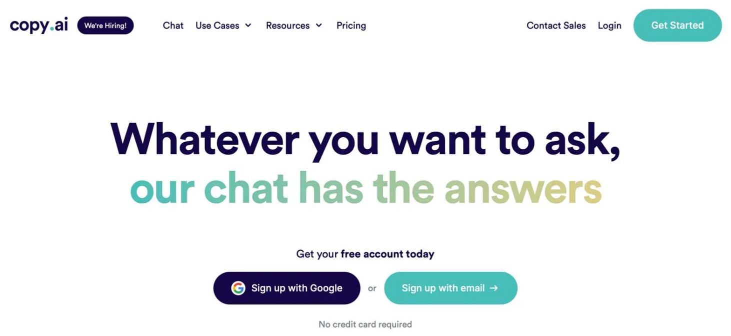
Claude's 100K token context window means that it is specifically a better fit than ChatGPT for many use cases across the enterprise—something we've already seen play out across Notion, DuckDuckGo, and Quora, as well as companies like Robin AI (a legal tech business using Claude to suggest alternative 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.



Whatever you want to ask, our chat has the answers

Get your free account today

No credit card required

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.

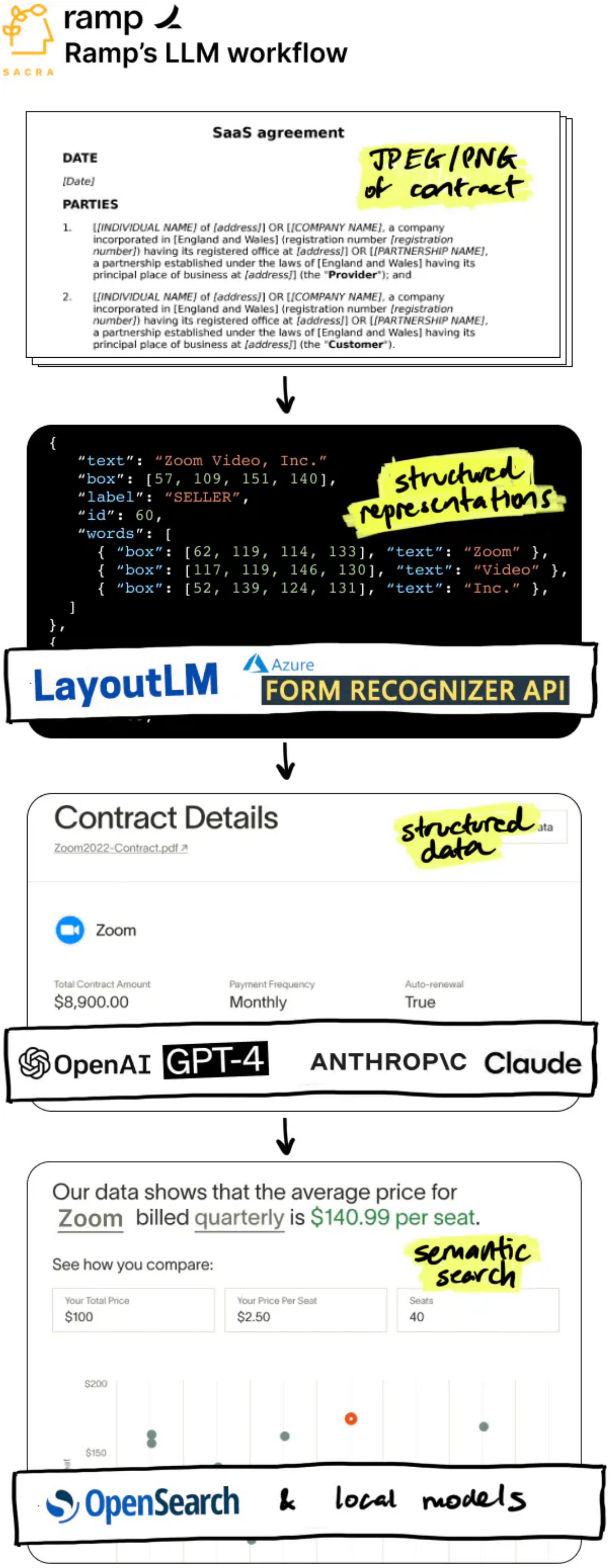
TAM Expansion

OpenAI's long-term goal is to develop the most capable artificial intelligence that is safe and aligned with human values.

While it's challenging to predict the long-term potential and value of an emerging technology like artificial intelligence, we can already make some predictions about how OpenAI might be valuable in the future based on changes that are happening now.

For example, OpenAI has triggered and is benefiting from shifts in how work is done across the B2B world—and as more and more spend moves towards AI-enabled software, OpenAI stands to become an AWS-like tax on the entire ecosystem.

Where card issuing enabled new, digital workflows that form the basis of Ramp and Brex's disruption of American Express, the next wave of disruption built on LLMs will automate away the human labor and judgment built into those workflows.



ramp

Ramp's LLM workflow

SaaS agreement

DATE

[Date]

PARTIES

1. [*[INDIVIDUAL NAME]* of *[address]*] OR [*[COMPANY NAME]*, a company incorporated in [England and Wales] (registration number *[registration number]*) having its registered office at *[address]*] OR [*[PARTNERSHIP NAME]*, a partnership established under the laws of [England and Wales] having its principal place of business at *[address]*] (the "**Provider**"); and

2. [*[INDIVIDUAL NAME]* of *[address]*] OR [*[COMPANY NAME]*, a company incorporated in [England and Wales] (registration number *[registration number]*) having its registered office at *[address]*] OR [*[PARTNERSHIP NAME]*, a partnership established under the laws of [England and Wales] having its principal place of business at *[address]*] (the "**Customer**").

```
{
   "text": "Zoom Video, Inc."
   "box": [57, 109, 151, 140],
   "label": "SELLER",
   "id": 60,
   "words": [
     { "box": [62, 119, 114, 133], "text": "Zoom" },
     { "box": [117, 119, 146, 130], "text": "Video" },
     { "box": [52, 139, 124, 131], "text": "Inc." },
   ]
},
{
```

Contract Details

Zoom2022-Contract.pdf ↗

Zoom

Total Contract Amount	Payment Frequency	Auto-renewal
$8,900.00	Monthly	True

Our data shows that the average price for Zoom billed quarterly is $140.99 per seat.

See how you compare:

Your Total Price	Your Price Per Seat	Seats
$100	$2.50	40

Rather than humans uploading receipts and filling out forms, AI will pre-classify transactions once they're made and merely prompt a human to confirm—changing the finance manager's role from pilot to auto-pilot monitor.

Finance work moving from humans to autonomous AI agents portends a shift from seats in payroll in Rippling towards spend on tokens and SaaS in tools like Ramp and Brex—ultimately benefiting OpenAI who can take a tax on the entire ecosystem.

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.

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 Oct 01st, 2023