BATHAEE DUNNE LLP
Yavar Bathaee (CA 282388)
yavar@bathaeedunne.com
Andrew C. Wolinsky (CA 345965)
awolinsky@bathaeedunne.com
445 Park Avenue, 9th Floor
New York, NY 10022
Tel.: (332) 322-8835

Brian J. Dunne (CA 275689)
bdunne@bathaeedunne.com
Edward M. Grauman (*p.h.v. to be sought*)
egrauman@bathaeedunne.com
901 South MoPac Expressway
Barton Oaks Plaza I, Suite 300
Austin, TX 78746
Tel.: (213) 462-2772

Allison Watson (CA 328596)
awatson@bathaeedunne.com
3420 Bristol Street, Suite 600
Costa Mesa, CA 92626
Tel: (213) 462-2772

Attorneys for Plaintiffs and the Proposed Classes

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

SAMUEL BRYANT, DOMINIQUE CAVALIER, VICTORIA DONOVAN, ALICE EIDSON, ALEXANDER HALLORAN, TRENTON MARSOLEK, SAMIR OUIJDANI, ZACHARY PAYNE, JAKE WOLFSON, CARA ZAJAC, and JASON ZHANG, individually and on behalf of all others similarly situated,

Plaintiffs,

v.

MICROSOFT CORPORATION, a Delaware corporation,

Defendant.

Case No. 3:25-cv-8733

CLASS ACTION COMPLAINT

DEMAND FOR JURY TRIAL

TABLE OF CONTENTS

INTRODUCTION

1. This lawsuit arises from an anticompetitive restraint placed upon OpenAI and its customers—specifically, ChatGPT subscribers—by OpenAI's horizontal competitor Microsoft. As a result of this restraint, ChatGPT prices were inflated since the service's inception, with price levels reaching an eye-popping 100 to 200 times competitors' prices on a per-token basis amidst a February 2025 price war.

2. A secretive agreement struck between OpenAI and Microsoft early in OpenAI's development allowed Microsoft to control the supply of compute to its horizontal competitor's products, and for years after OpenAI essentially created the Consumer Generative AI Market ("CGAI") in late 2022 with the launch of ChatGPT, that is exactly what Microsoft did. It used an exclusivity clause to restrict OpenAI's product output, and to impose a price (or, conversely, output and quality) floor on its competitor OpenAI's ChatGPT products.

3. Microsoft harmed Consumer Generative AI purchasers—most notably, ChatGPT subscribers—for years by mercilessly choking OpenAI's compute supply, thereby supracompetitively inflating the then-market leader's prices while Microsoft hurried to ready its own competing CGAI products, including Copilot. Consumers across the CGAI were injured by this anticompetitive conduct, and the agreement that enabled it, none more so than ChatGPT subscribers, who paid prices multiples too high for the economic benefit they received because of the Microsoft compute restraint.

4. This restraint was recently relaxed, immediately reducing OpenAI token prices by as much as 80% and immediately leading to substantially improved product output, quality, selection, and speed for ChatGPT subscribers. But the restraint—and Microsoft's control over it—still remains, lingering as a sword of Damocles over OpenAI wielded by one of its principal competitors. Further, ChatGPT subscribers who overpaid substantially for a subpar product due to a Microsoft-imposed output restraint have not been compensated for their overcharge.

5. This lawsuit seeks damages and appropriate injunctive relief to remedy these wrongs visited upon OpenAI customers, including ChatGPT subscribers, by OpenAI's horizontal competitor Microsoft. Plaintiffs paid for price-inflated and quality-degraded ChatGPT subscriptions during a period in which Microsoft leveraged its anticompetitive agreement and OpenAI's then-market power to restrict

supply and extract supracompetitive prices. Even now, they face the threat of future injury so long as Microsoft continues to have the ability to constrain OpenAI's compute supply—and thereby the output of a horizontal competitor.

* * *

6. On November 30, 2022, OpenAI released a generative AI chatbot to the public, which it called ChatGPT.[1] Within two months, ChatGPT had become the fastest-growing consumer software application in history, with 100 million users. Shortly thereafter, in February 2023, OpenAI launched a premium service, ChatGPT Plus, that charged users a monthly fee for certain ChatGPT services, including access to the newest models, unlimited (or less-constrained) access to ChatGPT features, and purportedly faster response speeds and less downtime. OpenAI also sold priority access to ChatGPT to developers through API tokens.

7. The launch of ChatGPT Plus created a new product market in the United States—the Consumer Generative AI Market. Limited (increasingly limited, over time) ChatGPT features were available for free, while the latest models and priority access cost everyday consumers a flat fee per month (a familiar "freemium" model for software services). To developers, who use ChatGPT programmatically within their own applications and services, access was obtained through API tokens, which bore a publicly-accessible market price. Although the Generative AI Market was new, its economic model was not: Open AI was monetizing its new "killer app," ChatGPT, in a familiar "freemium" software pricing model. Users paid a monthly fee, and OpenAI would provide them ChatGPT features on-demand through an OpenAI-created user interface, through the web or through a native app.

8. But behind the scenes, a longstanding tech giant—and a recidivist violator of the United States antitrust laws—was about to put the squeeze on OpenAI and its customers. Central to OpenAI's Consumer Generative AI products, and necessary to their operation, is computation: when a user queries ChatGPT, OpenAI's models are run through powerful arrays of GPUs and other high-end processing units in order to generate a response. Without the powerful computing arrays, no response at all—and

[1] "GPT" stands for generative pre-trained transformer, a type of large language model (LLM).

certainly not a meaningful, competitive one—will be provided. This is the case whether a ChatGPT user seeks to answer a question, to generate or revise source code, to create an image, or just to chat.

9. Compute is a necessary input (and indeed, consumer-facing feature) of a Consumer Generative AI product like ChatGPT. Yet controlling the supply of compute to OpenAI's generative AI products, including ChatGPT, was Defendant Microsoft—OpenAI's horizontal competitor in the Consumer Generative AI Market.

10. Through a secretive agreement struck early in OpenAI's corporate development, Microsoft imposed a term on OpenAI that required the OpenAI's generative AI products to *exclusively* use Microsoft (specifically, its Azure cloud service) for compute services. This meant that Microsoft (supplier of its own Consumer Generated AI products, including Copilot) had the contractual ability to restrict the output of its horizontal competitor OpenAI, then the far-and-away market leader in the Consumer Generated AI market.

11. And that is exactly what Microsoft did. As demand for OpenAI's ChatGPT product skyrocketed, Microsoft exercised its control over compute supply to restrict output to ChatGPT customers, causing massive price inflation for OpenAI tokens (the price charged API users to access the OpenAI models that drive ChatGPT) and corresponding degradations in quality, speed, sophistication, and product choice within OpenAI's subscription ChatGPT products.

12. The price inflation from Microsoft's compute restraint on OpenAI's generative AI products was so substantial that, when DeepSeek AI entered the CGAI market in early 2025 and sparked a price war, OpenAI's generative AI model prices remained ***136 to 200 times the price of its competitors'***. ChatGPT subscribers in this same time frame continued to experience poor product quality, unreleased product innovations due to compute constraint, slow response times, and other clear measures of decreased product value as against competitive levels.

13. When, in June 2025, Microsoft finally relaxed part of its restraint on OpenAI's compute supply, allowing the company to purchase compute from Google, OpenAI token prices ***immediately dropped 80%***. ChatGPT ***immediately*** launched long-awaited—and long-delayed—new image generation features and new, more powerful models. The lifting of Microsoft's compute restraint ***immediately*** lowered prices, and it ***immediately*** resulted in vastly improved product quality for ChatGPT subscribers.

This was a powerful natural experiment confirming the serious, direct economic impact of Microsoft's anticompetitive restraint on OpenAI's generative AI products, principally ChatGPT.

14. This lawsuit is brought by purchasers of ChatGPT subscriptions from November 30, 2022, to the present, on their own behalf and on behalf of classes of similarly-situated ChatGPT purchasers. Plaintiffs and the class members overpaid for ChatGPT subscription products from November 30, 2022, through February 1, 2025, suffering an antitrust overcharge as the Microsoft-OpenAI agreement inflated prices and constrained output (including product quality, selection, and speed) in the Consumer Generative AI Market.

15. Further, although in June 2025 Microsoft temporarily relaxed its supply restriction to allow OpenAI to purchase compute from Google, Microsoft still retains the contractual ability to restrict OpenAI's compute purchases, and thereby to control its horizontal competitor's output in the Consumer Generative AI Market. This represents an ongoing threat to Plaintiffs and the class members, which can only be remedied by an appropriate injunction.

PARTIES

I. PLAINTIFFS

16. Plaintiff Samuel Bryant is a domiciled resident of Rancho Cordova, California. Plaintiff Bryant has been a continuous ChatGPT Plus subscriber since January 2024, other than brief gaps during periods when he changed payment methods. Plaintiff Bryant currently pays $19.99 per month (before taxes) for his ChatGPT Plus subscription.

17. Plaintiff Dominique Cavalier is a domiciled resident of Ontario, California. Plaintiff Cavalier is a former paid ChatGPT Plus subscriber. Plaintiff Cavalier began subscribing to ChatGPT Plus in or around February 2024 and cancelled his subscription in or around June 2024. Plaintiff Cavalier last paid $19.99 per month (before taxes) for his ChatGPT Plus subscription. Plaintiff Cavalier would be interested in re-subscribing to ChatGPT Plus if the ongoing anticompetitive conduct set forth in this Complaint were remedied.

18. Plaintiff Victoria Donovan is a domiciled resident of Denver, Colorado. Plaintiff Donovan is a former paid ChatGPT Plus subscriber. Plaintiff Donovan began subscribing to ChatGPT Plus in or around February 2023 and cancelled her subscription in or around June 2024. Plaintiff Donovan last paid

$20.00 per month (before taxes) for her ChatGPT Plus subscription. Plaintiff Donovan would be interested in re-subscribing to ChatGPT Plus if the ongoing anticompetitive conduct set forth in this complaint were remedied.

19. Plaintiff Alice Eidson is a domiciled resident of Seattle, Washington. Plaintiff Eidson has been a continuous paid subscriber to ChatGPT Plus since February 2023. Plaintiff Eidson currently pays $20.00 per month (before taxes) for her ChatGPT Plus subscription.

20. Plaintiff Alexander Halloran is a domiciled resident of Phoenix, Arizona. Plaintiff Halloran has been a continuous paid subscriber to ChatGPT Plus since April 2023. Plaintiff Halloran currently pays $20.00 per month (before taxes) for his ChatGPT Plus subscription.

21. Plaintiff Trenton Marsolek is a domiciled resident of Aldie, Virginia. Plaintiff Marsolek has been a continuous paid subscriber to ChatGPT Plus since April 2025. Plaintiff Marsolek currently pays $20.00 per month (before taxes) for his ChatGPT subscription.

22. Plaintiff Samir Ouijdani is a domiciled resident of Everett, Washington. Plaintiff Ouijdani has been a continuous paid subscriber to ChatGPT Plus since March 2023. Plaintiff Ouijdani currently pays $19.99 per month (before taxes) for his ChatGPT Plus subscription.

23. Plaintiff Zachary Payne is a domiciled resident of Mentor, Ohio. Plaintiff Payne has been a continuous paid subscriber to ChatGPT Plus since October 2023. Plaintiff Payne currently pays $19.99 per month (before taxes) for his ChatGPT Plus subscription.

24. Plaintiff Jake Wolfson is a domiciled resident of Centerreach, New York. Plaintiff Wolfson has been a continuous paid subscriber to ChatGPT Plus since December 2023, except for a pause in his subscription from February 2024 to March 2024. Plaintiff Wolfson currently pays $19.99 per month (before taxes) for his ChatGPT Plus subscription.

25. Plaintiff Cara Zajac is a domiciled resident of New Bedford, Massachusetts. Plaintiff Zajack is a former paid ChatGPT Plus subscriber. Plaintiff Zajack paid subscription fees to ChatGPT in March 2024. Plaintiff Zajac paid $20.00 that month (before taxes) for her ChatGPT Plus subscription. Plaintiff Zajack would be interested in re-subscribing to ChatGPT if the ongoing anticompetitive conduct set forth in this Complaint were remedied.

26. Plaintiff Jason Zhang is a domiciled resident of Oakland Township, Michigan. Plaintiff Zhang has been a continuous paid subscriber to ChatGPT Plus since December 2023. Plaintiff Zhang currently pays $20.00 per month (before taxes) for his ChatGPT Plus subscription.

II. DEFENDANT

27. Defendant Microsoft Corporation is a Delaware corporation headquartered at One Microsoft Way in Redmond, Washington. Microsoft is the designer and manufacturer of the leading PC operating system in the world, Microsoft Windows.

28. Microsoft also develops office productivity software, most notably its Office 365 suite of products and services, which include Microsoft Word, Excel, and PowerPoint.

29. All of Microsoft's primary offerings are deeply integrated with its Artificial Intelligence ("AI") products, including its Office 365 productivity suite. The flagship AI product sold by Microsoft is its Copilot product, which is utilizes a subset of AI technology called Generative AI (described further below).

30. As Microsoft explains in its 2024 annual report filed with the SEC on Form 10-K, the company leverages data obtained through its Microsoft Office365 suite of business applications to drive its Consumer Generative AI product, Copilot:

> Microsoft 365 is an AI first platform that brings together Office, Windows, Copilot, and Enterprise Mobility + Security to help organizations empower their employees. Copilot for Microsoft 365 combines AI with business data in the Microsoft Graph and Microsoft 365 applications.

31. Microsoft's Copilot platform can be adapted to specific workstreams through its Copilot Studio product. Microsoft also provides Generative AI support as part of its integrated developer environment ("IDE") for software developers on a subscription basis, part of a product called GitHub Copilot.

32. Microsoft provides its CGAI Copilot products through its cloud computing line of business, called Azure. As Microsoft explains in its 2024 Annual Report:

> Our AI platform, Azure AI, is helping organizations transform, bringing intelligence and insights to the hands of their employees and customers to solve their most pressing challenges. We offer a wide selection of industry-

> leading frontier and open models, including from partners, as well as state-of-the-art tooling, and AI-optimized infrastructure, delivering the Copilot stack for Microsoft, enterprises, and developers. Organizations large and small are deploying Azure AI solutions to achieve more at scale, more easily, with the proper enterprise-level responsible AI and safety and security protections. Azure AI Studio provides a full lifecycle toolchain customers can use to ground these models on their own data, create prompt workflows, and help ensure they are deployed and used safely.

33. As explained further below, Microsoft also powers OpenAI's competing Consumer Generative AI product on its Azure Cloud platform. As Microsoft explains in its 2024 Annual Report:

> We have a long-term partnership with OpenAI, a leading AI research and deployment company. We deploy OpenAI's models across our consumer and enterprise products. As OpenAI's exclusive cloud provider, Azure powers all of OpenAI's workloads. We have also increased our investments in the development and deployment of specialized supercomputing systems to accelerate OpenAI's research.

34. Microsoft's CEO reported $64.773 billion in product revenue for 2024, $64.669 billion in 2023, and $72.732 billion in 2022. Combined with services, Microsoft reported $245.122 billion, $211.915 billion, and $198.270 billion in revenue for 2024, 2023, and 2022 respectively. Microsoft reported a net income of $88.136 billion, $72.361 billion, and $72.738 billion for each of those years respectively.

35. Microsoft has been the subject of U.S. and international antitrust scrutiny since the late 90s, including for monopolization of the Intel x86 PC Operating System Market, which led to the seminal government case against the company, *United States v. Microsoft*, in which Microsoft was held to have violated the United States antitrust laws.

36. Microsoft was subject to a consent decree in connection with its anticompetitive conduct. As Microsoft recounted in its 2008 Annual Report to shareholders filed on SEC Form 10-K:

> We are subject to a Consent Decree and Final Judgment that resolved lawsuits brought by the U.S. Department of Justice, 18 states, and the District of Columbia in two separate actions. The Consent Decree imposed various constraints on our Windows operating system business. Portions of the Consent Decree were scheduled to expire on January 31, 2008; we voluntarily agreed to extend other elements of the Consent Decree to November 2009. In October 2007, some states filed a motion with the U.S. District court for the District of Columbia seeking to have most of the

> remaining provisions of the Final Judgment in the action to which they are party extended for five years. The U.S. Department of Justice and other states advised the Court that they would not seek any extension of the Final Judgments to which they are party. In January 2008, the court issued a decision granting the states' motion to extend these additional provisions of the consent decree until November 2009.

37. Although headquartered in the State of Washington, Microsoft's AI business is integrated heavily with OpenAI's San Francisco, California operations. Indeed, the head of Microsoft's AI division, Mustafa Suleyman, is tasked with incorporating OpenAI software, and his division at Microsoft works directly with OpenAI's California-based engineers, including by working on-site at OpenAI's offices in San Francisco using OpenAI laptops.

38. As the New York Times reported in an October 17, 2024 article titled "Microsoft and OpenAI's Close Partnership Shows Signs of Fraying," Microsoft's Suleyman has on occasion made heated demands of OpenAI engineers to deliver Generative AI products to Microsoft:

> Some OpenAI staff recently complained that Mr. Suleyman yelled at an OpenAI employee during a recent video call because he thought the start-up was not delivering new technology to Microsoft as quickly as it should, according to two people familiar with the call. Others took umbrage after Microsoft's engineers downloaded important OpenAI software following the protocols the two companies had agreed on, the people said.

39. Put simply, Microsoft's AI business has a significant presence in San Francisco, California, and Microsoft enforces the terms of its agreement with OpenAI (described below) by directly interacting with OpenAI in San Francisco.

40. Non-party OpenAI, Inc. ("OpenAI") is a registered non-profit organization incorporated under the laws of Delaware.

41. OpenAI controls a web of subsidiaries and special-purpose-vehicle entities ("SPVs"), including several for-profit subsidiaries through which it does business. These include OpenAI, LP, OpenAI, LLC, OpenAI GP, LLC, OpenAI OpCo, LLC, OpenAI Global LLC, and many others. OpenAI interacts with Microsoft, at least in part, through a subset of these subsidiaries and SPVs. The substance of the relationship between OpenAI and Microsoft, however, is direct and company-to-company.

42. OpenAI is headquartered in San Francisco, where its engineers and executives operate its business. OpenAI's headquarters are at 3180 18th Street in San Francisco, California—offices laid out like an archetypal Silicon Valley startup.

43. Although OpenAI purports to be governed by a non-profit entity—sometimes characterized as a capped-profit entity—it sells its subscription-based ChatGPT product to the market and derives substantial revenues from doing so. OpenAI was originally co-founded by Elon Musk to ensure that Google would not dominate AI technology after its 2014 acquisition of DeepMind, an early leader in neural network development. But since its founding, OpenAI has ejected Musk from its orbit and quietly worked to convert the company into a for-profit venture, including through a web of subsidiaries and its partnership with Microsoft. This pivot is the subject of a current lawsuit by Musk and others.

44. As *The New York Times* reported in a September 27, 2024 article titled, "OpenAI Is Growing Fast and Burning Through Piles of Money":

> OpenAI's monthly revenue hit $300 million in August, up 1,700 percent since the beginning of 2023, and the company expects about $3.7 billion in annual sales this year, according to financial documents reviewed by The New York Times. OpenAI estimates that its revenue will balloon to $11.6 billion next year.

45. As explained below, OpenAI has agreed to buy all of the computing resources necessary for its Consumer Generative AI products—namely, ChatGPT—from Microsoft.

46. The company purports to be pursuing Artificial General Intelligence—the holy grail of AI technology. As OpenAI explains on its website on a page titled, "Planning for AGI and beyond":

> Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.
>
> If AGI is successfully created, this technology could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility.
>
> AGI has the potential to give everyone incredible new capabilities; we can imagine a world where all of us have access to help with almost any cognitive task, providing a great force multiplier for human ingenuity and creativity.

47. As explained below, because AGI is OpenAI's purported mission, its partnership with Microsoft is purportedly set to end when OpenAI accomplishes this ambitious goal. However, as explained below, OpenAI defines AGI for the purposes of its agreement with Microsoft not in technological terms, but in terms of monetization. For that purpose, AGI is the point where the company is capable of making $100 billion in profits.

JURISDICTION AND VENUE

48. This Court has personal and subject matter jurisdiction over all parties to and causes of action asserted in this Complaint.

49. This action arises under Section 1 of the Sherman Act, and Sections 4 and 16 of the Clayton Act (15 U.S.C. §§ 1, 15, 26). Plaintiffs and the proposed classes seek to recover treble damages, interest, costs of suit, equitable relief, and reasonable attorneys' fees for their damages resulting from Defendant's anticompetitive agreements.

50. This Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 (federal question), 1332 (class action diversity jurisdiction), and 1337(a) (antitrust); and under 15 U.S.C. § 15 (antitrust).

51. The Court has supplemental jurisdiction over state-based claims asserted in this Complaint pursuant to 28 U.S.C. § 1367(a).

52. As to jurisdiction under 28 U.S.C. § 1332(d), at least one member of the proposed Classes is of diverse citizenship from Defendant; the proposed Classes consist of 100 or more members; and the aggregate claims of the members of the proposed Classes exceed $5 million, exclusive of interest and costs.

53. This Court has personal jurisdiction over Defendant Microsoft because Microsoft conducts business is in the State of California, and the conduct alleged in this Complaint, *i.e.*, the anticompetitive agreement between OpenAI and Microsoft, occurred in and/or emanated from the State of California. Microsoft has physical offices in three separate Bay Area locations within this judicial district—in Berkeley, San Francisco, and Mountain View—and before a recent round of layoffs employed approximately 6,700 people across California.

54. Venue is appropriate and proper in this district under 15 U.S.C. § 15(a) (Clayton Act), 15 U.S.C. § 22 (nationwide venue for antitrust matters), and 28 U.S.C. § 1391(b) (general venue provision). Microsoft transacts business within the district, including at the three Bay Area offices discussed above and through thousands of Bay Area-sited employees, and carries on its affairs and interstate trade and commerce, in substantial part, in this district.

DIVISIONAL ASSIGNMENT

55. This is an antitrust class action for which "venue is proper in any courthouse in this District" under Gen. Order No. 44 § D.3 and Civil Local Rule 3-2(c).

FACTS

I. OPENAI AND MICROSOFT: HORIZONTAL COMPETITORS WITH AN ANTICOMPETITIVE AGREEMENT

56. As explained below, Microsoft and OpenAI each offer Consumer Generative AI ("CGAI") products that directly compete in the same product market, the United States CGAI Market.

A. OpenAI and the Generative AI Revolution

57. For decades since the 1960s, the field of artificial intelligence had been stagnant. Most methods of implementing artificial intelligence systems were driven heavily by probabilistic, statistical and rule-based models.

58. By the mid-2010s, however, one method of computational decisionmaking—first developed in the 1960s but eventually fallen to desuetude—began to show new promise. An artificial neuron, implemented as a mathematical model, could be strung together into layered networks, allowing a computer to reason and learn based on data, rather than based on rules or calculated probabilities.

59. Although conceived of decades earlier, such a structure had long proved unworkable in practice due to hardware limitations—that is, existing computers simply could not handle the computational difficulty of training multi-layered networks of artificial neurons. But by 2016, computational power had advanced to the point where such training was feasible at large scale.

60. By 2017, a new breakthrough was announced in a paper by Google Deep Mind researchers, titled "Attention Is All You Need," in which a new organization of artificial neurons was proposed called the transformer.

61. The transformer allowed stable and deep networks to be trained at large scale, overcoming a slew of problems associated with other neural network architectures that appeared as layers of neurons deepened.

62. The transformer was immediately applied to create a new form of AI, called Generative AI, which allows an AI system to make predictions about a sequence of tokens (*e.g.*, words in a sentence) in response to prompting. For example, when posed a question, a Generative AI system can predict a sequence of words that answer the question. These systems are trained on vast amounts of data and can be adapted to accept prompts (and generate responses) beyond text-based tokens, including based on images and sounds.

63. Perhaps the most notable early breakthrough in Generative AI came in 2018, when OpenAI developed its GPT series of transformer models. As OpenAI reported on June 11, 2018 on its website in a page titled, "Improving language understanding with unsupervised learning," GPT appeared to provide reasoned responses to queries, generating large sequences of sensible text:

> We've obtained state-of-the-art results on a suite of diverse language tasks with a scalable, task-agnostic system, which we're also releasing. Our approach is a combination of two existing ideas: transformers and unsupervised pre-training. These results provide a convincing example that pairing supervised learning methods with unsupervised pre-training works very well; this is an idea that many have explored in the past, and we hope

> our result motivates further research into applying this idea on larger and more diverse datasets.

64. The advent of GPT kicked off a boom in Generative AI led by OpenAI. In November 2019, OpenAI released GPT-2, which was based on a 1.5-billion-parameter model—parameters referring roughly to the number of trainable computational nodes in the network of artificial neurons.

65. Notably, GPT-2 had achieved a ten-fold increase in functionality by simply increasing the size and scale of the GPT model. This led OpenAI and other researchers to believe that the key to creating intelligent Generative AI systems was to massively increase the scale and size of the models.

66. In June 2020, OpenAI launched its first commercial product based on GPT. Nine months later, it had massively expanded GPT in size and scale, producing a full-scale commercial GPT-3.

67. OpenAI's GPT-3 could respond with text completion in response to prompts, and the text provided by the model was highly intelligent. As OpenAI announced on its website on March 25, 2021, on a page titled "GPT-3 powers the next generation of apps":

> Nine months since the launch of our first commercial product, the OpenAI API, more than 300 applications are now using GPT-3, and tens of thousands of developers around the globe are building on our platform. We currently generate an average of 4.5 billion words per day, and continue to scale production traffic.
>
> Given any text prompt like a phrase or sentence, GPT-3 returns a text completion in natural language. Developers can "program" GPT-3 by showing it just a few examples or "prompts." We've designed the API to be both simple for anyone to use but also flexible enough to make machine learning teams more productive.

68. OpenAI had kicked off the Generative AI revolution, which reached new heights when OpenAI announced its new "chatbot" product, called ChatGPT, at the end of 2022, essentially creating the new Consumer Generated AI product market.[2] As The New York Times recounted in a January 7, 2023 article titled, "A New Area of A.I. Booms, Even Amid the Tech Gloom":

[2] As explained in more detail later in this Complaint, Consumer Generative AI products are frequently called "Chatbots," largely for historical reasons. For example, analysts and observers frequently call CGAI products like ChatGPT chatbots, including in market analyses. There is nothing inherently inaccurate about this labeling, and this Complaint occasionally uses the "chatbot" nomenclature, but modern CGAI products (including ChatGPT) do far more than "chat" with a user.

> Five weeks ago, OpenAI, a San Francisco artificial intelligence lab, released ChatGPT, a chatbot that answers questions in clear, concise prose. The A.I.-powered tool immediately caused a sensation, with more than a million people using it to create everything from poetry to high school term papers to rewrites of Queen songs.
>
> Now OpenAI is in the midst of a new gold rush.

69. As the same article explained, the new technology was groundbreaking, allowing computers trained on large amounts of data to generate output in response to prompting:

> No area has created more excitement than generative artificial intelligence, the term for technology that can generate text, images, sounds and other media in response to short prompts. Investors, pundits and journalists have talked up artificial intelligence for years, but the new wave—the result of more than a decade of research—represents a more powerful and more mature breed of A.I.
>
> This type of A.I. promises to reinvent everything from online search engines like Google to photo and graphics editors like Photoshop to digital assistants like Alexa and Siri. Ultimately, it could provide a new way of interacting with almost any software, letting people chat with computers and other devices as if they were chatting with another person.

70. Generative AI could not only generate text, it could generate images. OpenAI, in particular, had devised a model called DALL-E, which generated images based on textual prompts by users.

DALL-E, an A.I. system created by OpenAI, generates digital images based on text commands. OpenAI

DALL-E generated these images by following a command for "a giant hamster blimp carrying passengers." OpenAI

71. Although OpenAI had essentially created a new market for Consumer Generative AI products, its arrival on the scene would not go unnoticed by technology incumbents—namely, Microsoft and Google. Moreover, although the new technology had arrived, new competitors could not simply enter and compete with OpenAI. Consumer Generative AI required massive computational scale, which only a few firms—those same incumbent tech giants, specifically Microsoft and Google—had at their disposal.

72. But what neither Microsoft nor Google had at the end of 2022 was OpenAI's first-mover advantage and technological know-how.

B. Alarm Bells Sound at Google, But Google Is Too Late

73. The release of ChatGPT rang alarm bells at Google, whose researchers had pioneered the neural network transformer—the technological foundation of OpenAI's new ChatGPT product. Yet five years after Google's own researchers introduced the transformer architecture in their "Attention is All You Need" research paper, Google somehow found itself behind the curve.

74. As 2022 ended, Google's CEO, Sundar Pichai, declared a "code red." Google would have to devise a plan to catch up with OpenAI and would have to do so quickly, as the new technology, which provided clear and concise answers to questions, threatened to upend Google's longstanding—and largely unopposed—dominance in search. As the New York Times recounted in a December 21, 2022 article titled "A New Chat Bot Is a 'Code Red' for Google's Search Business":

> Sundar Pichai, Google's chief executive, has been involved in a series of meetings to define Google's A.I. strategy, and he has upended the work of numerous groups inside the company to respond to the threat that ChatGPT poses, according to a memo and audio recording obtained by The New York Times. Employees have also been tasked with building A.I. products that can create artwork and other images, like OpenAI's DALL-E technology, which has been used by more than three million people.

75. Google scrambled to release its own Consumer Generative AI product to compete with ChatGPT. Unlike other potential CGAI entrants in late 2022, Google had not only the know-how to create a competing product, but the computing power in its cloud computing platform to do so.

76. Google, however, floundered, unable to create a viable product to compete with ChatGPT's subscription-based ChatGPT product. It would not be until February 2024 that Google finally offered a rival subscription product for the first time, rebranding its Bard Generative AI system as Gemini.

77. As CNBC reported in a February 8, 2024 article titled, "Google rebrands Bard AI to Gemini and launches a new app and subscription":

> Google on Thursday announced a major rebrand of Bard, its artificial intelligence chatbot and assistant, including a fresh app and subscription options. Bard, a chief competitor to OpenAI's ChatGPT, is now called Gemini, the same name as the suite of AI models that power the chatbot.
>
> Google also announced new ways for consumers to access the AI tool: As of Thursday, Android users can download a new dedicated Android app for Gemini, and iPhone users can use Gemini within the Google app on iOS.

78. Gemini came far too late. The new chatbot failed to acquire significant market share from rival Consumer Generative AI competitors, including OpenAI. In fact, by the time Google arrived on the scene with a competing product, one of its long-standing rivals had already entered the fray, locked up the powerful Generative AI technology at the heart of OpenAI, and extracted inflated profits, all while preparing its own competing product: Microsoft.

C. Microsoft Captures OpenAI, Creating an Exclusive Agreement for Computation and OpenAI's Technology

79. OpenAI's growth trajectory had a different effect at Microsoft than at Google. Microsoft scrambled to embrace the new, rapidly-developing AI technology and to secure a valuable deal with OpenAI that would allow Microsoft to position itself against a vulnerable Google.

80. Microsoft had been in the same place before. Its founder and former CEO, Bill Gates, had once foreseen the massive effect the Internet would have on the computer industry—and on the world at large—and attempted to realign the entire company in the Internet's direction. Internally, Microsoft recognized the importance of embracing technological changes with the potential for exponential growth, particularly as compared with the rate of computational improvement. As one of Gates's senior lieutenants wrote in a September 3, 1993 memorandum to Gates and others, titled "Road Kill on the Information Highway":

> The trick for the next decade of the computing industry will hinge on being very smart about recognizing which things scale with or faster than computing and what things do not. This will *not* be obvious in the early stages, but *that is where the value lies* so this is the challenge to which we must rise.

(emphasis in original). The memorandum would prove not only prescient, but deeply insightful about the nature of Microsoft's business.

81. Since Gates's departure, Microsoft and its shareholders watched as companies such as Google and Apple strategically captured pieces of Microsoft's waning PC operating system monopoly as the Internet—and the devices people used to access it—grew and evolved. From web browsers to mobile operating systems, Microsoft fell behind and gave up ground with each new wave of technological change and development in the 2000s and 2010s. Each time, Microsoft missed the insight of its "Roadkill" memorandum—the innovations involved were of the sort that outpaced computation.

82. Microsoft had learned this lesson the hard way, and had invested heavily in its cloud computing business, called Azure. Microsoft's Azure is one of only a handful of cloud computing clusters with adequate computational power for the training of large-scale transformer and other AI models. At the start of the present decade, the only other companies with cloud computing platforms large enough to compete with Microsoft at scale in this area were Google and Amazon.

83. Microsoft's Azure was a source of massive growth for Microsoft. As the Wall Street Journal explained in a July 27, 2021 article titled "Microsoft Posts Another Quarter of Record Sales Driven by Cloud Growth":

> Azure, Microsoft's collection of cloud data centers and software tools that has been the backbone of its growth in recent years, saw year-over-year sales growth of 51%, topping the 50% seen in the two prior quarters. Microsoft said it expects relatively stable growth in the current period.

84. Azure was not only one of Microsoft's largest growth businesses, it provided Microsoft with a massive bargaining chip against companies such as OpenAI, which depended on large arrays of GPUs to train their AI products.

85. Indeed, Microsoft had trained its own AI products on its Azure platform. It was well familiar with the necessary ingredients for the growth of an AI-driven product, particularly for a large language model, which requires massive scale to become emergently intelligent.

86. In July 2019, Microsoft scrambled to cut a deal with OpenAI. In exchange for the ability to invest in the supposedly non-profit company, Microsoft dangled access to its powerful Azure platform.

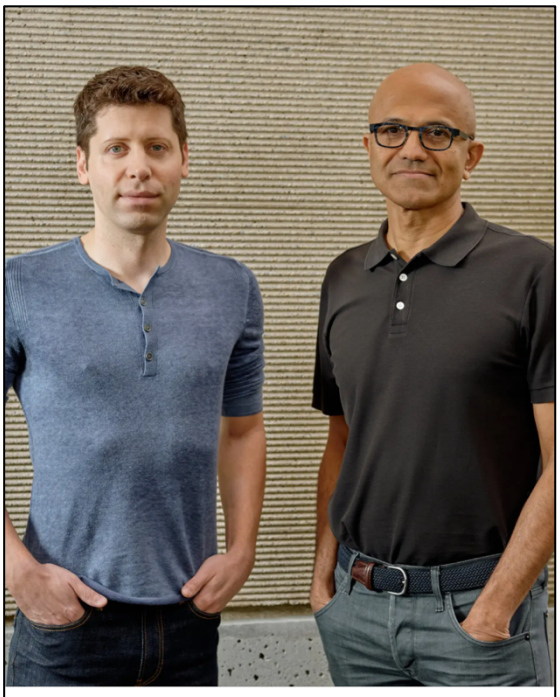
On July 22, 2019, the companies reached an initial deal. As OpenAI described the deal in a July 22, 2019 post on its website titled "Microsoft invests in and partners with OpenAI to support us building beneficial AGI":

> Microsoft is investing $1 billion in OpenAI to support us building artificial general intelligence (AGI) with widely distributed economic benefits. We're partnering to develop a hardware and software platform within Microsoft Azure which will scale to AGI. We'll jointly develop new Azure AI supercomputing technologies, and Microsoft will become our exclusive cloud provider—so we'll be working hard together to further extend Microsoft Azure's capabilities in large-scale AI systems.

87. Microsoft's press release about the deal was more granular:

> Microsoft Corp. and OpenAI, two companies thinking deeply about the role of AI in the world and how to build secure, trustworthy and ethical AI to serve the public, have partnered to further extend Microsoft Azure's capabilities in large-scale AI systems. Through this partnership, the companies will accelerate breakthroughs in AI and power OpenAI's efforts to create artificial general intelligence (AGI). The resulting enhancements to the Azure platform will also help developers build the next generation of AI applications. The partnership covers the following:
>
> - Microsoft and OpenAI will jointly build new Azure AI supercomputing technologies
> - OpenAI will port its services to run on Microsoft Azure, which it will use to create new AI technologies and deliver on the promise of artificial general intelligence
> - Microsoft will become OpenAI's preferred partner for commercializing new AI technologies

88. Microsoft's reference to becoming OpenAI's "preferred partner" was in fact an understatement. Microsoft had (as OpenAI's simultaneous press release recognized) acquired the exclusive right to commercialize and sell products incorporating any of OpenAI's inventions in exchange for its provision of computing resources on Azure and its massive investment in the young company.



Sam Altman, left, one of the founders of OpenAI, and Satya Nadella, Microsoft's chief executive, agreed to a $1 billion investment by Microsoft in 2019. Ian C. Bates for The New York Times

89. Azure credits were the cornerstone of the 2019 investment—and so was transitioning OpenAI away from Google's cloud computing platform. As The Information reported in a January 11, 2023 article, titled "Microsoft + OpenAI: Inside Tech's Hottest Romance," OpenAI could buy its computing ***only*** from Microsoft's Azure:

> In July 2019, Microsoft announced it was investing $1 billion in OpenAI. What it didn't say was that much of its $1 billion investment in the startup would come in the form of Azure credits that would enable OpenAI to essentially run on Microsoft's cloud rent free. In exchange, Microsoft would become OpenAI's exclusive cloud provider—snatching a customer from Google—and Microsoft and OpenAI would collaborate on further developing the software and hardware capabilities for training and running AI models.
>
> OpenAI had been one of Google's largest customers, paying Google more than $120 million for cloud computing in 2019 and 2020 combined, according to data viewed by The Information and a person familiar with the figures. Since then it has operated mostly on Azure.

90. Unlike other times where technological shifts caused Microsoft to lose ground, Microsoft had captured OpenAI's technology near to its very inception, not only securing the technological fruits of the company, but ensuring its dependence on Microsoft for growth. As The Information explained, Microsoft's CEO was pleased with the deal:

> Nadella was excited. He had watched Microsoft fumble numerous early advantages—notably in the smartphone market and with its tablets, which debuted before Apple's versions but were overtaken by them. Microsoft then acquired phone maker Nokia, still without achieving notable results. And most of Microsoft's efforts to play catch-up to Google in search and advertising have been notoriously ineffective.
>
> Microsoft's pact with OpenAI is "like a marriage made in heaven," said Carl Bass, former CEO of design software firm Autodesk and a former Google adviser. Microsoft's productivity apps and Bing search have been "pretty stagnant for at least a decade, at least from a user point of view. It feels like this is a time when you could have that quantum leap" using the new AI models, he said.

91. At the time of the deal, Microsoft was directly competing with OpenAI, including with its own model called Turing, which also ran on Azure's AI cloud computing platform. Indeed, Microsoft had launched an all-out AI strategy across its many lines of business. As The Information explained:

> Microsoft was years deep into a costly company-wide bet on AI, which it saw as a way to improve its productivity software and gain an edge against competitors. Microsoft researchers were training a large-scale AI designed to parse millions of documents scraped from the internet, which became known as Turing. Nadella had instructed teams across Microsoft to use AI models like Turing to enhance their products, someone familiar with the matter said.
>
> The strategy, now called AI at scale, hinged on the idea that Microsoft needed to find a variety of ways to make money from Turing because of how costly it was to develop: Training the models required far more computing power than Microsoft's systems had ever handled before. To make it happen, Microsoft Chief Technology Officer Kevin Scott struck a deal with chipmaker Nvidia to develop high-powered graphics processing units—AI practitioners' preferred type of chip—and cables that could handle the heavy workloads needed to train the AI. Microsoft developed new software, dubbed DeepSpeed, to help.

92. Forcing OpenAI to buy compute exclusively from Microsoft's Azure also had a powerful lock-in effect. Once OpenAI models were built for Azure's AI platform, they could not easily be ported and trained on a rival cloud computing system. As The Information explained in the same article:

> Immediately after the investment closed, the two companies began working together to develop the next generation of software and GPU clusters both now use to train their models, according to someone with direct knowledge of the work.

> OpenAI's models currently rely so much on Microsoft software and hardware to operate that it would be difficult for the startup to easily port its models to a different cloud provider, according to this person. And OpenAI's models currently take up more space and compute in Azure than Turing does, this person said.

93. As OpenAI made progress, Microsoft deepened its entrenchment in OpenAI's business. Specifically, in 2021, as OpenAI was developing its quantum-leap GPT3 model, Microsoft invested even more in the company and committed even more of its Azure AI computing power to OpenAI. As the New York Times reported in a January 12, 2023 article titled "Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I.":

> When a chatbot called ChatGPT hit the internet late last year, executives at a number of Silicon Valley companies worried they were suddenly dealing with new artificial intelligence technology that could disrupt their business.
>
> But at Microsoft, it was a cause for celebration. For several years, Satya Nadella, Microsoft's chief executive, had been putting the pieces in place for this moment.
>
> In 2019, Microsoft invested $1 billion in OpenAI, the tiny San Francisco company that designed ChatGPT. And in the years since, it has quietly invested another $2 billion, according to two people familiar with the investment who requested anonymity because they were not authorized to speak with the media.
>
> The $3 billion paid for the huge amounts of computing power that OpenAI needed to build the chatbot. And it meant that Microsoft could rapidly build and deploy new products based on the technology.

94. By the beginning of 2023, Microsoft sought to expand its investment to a massive $10 billion. As the New York Times reported in the same article:

> Microsoft is now poised to challenge Big Tech competitors like Google, Amazon and Apple with a technological advantage the company has not possessed for more than two decades. Microsoft is in talks to invest another $10 billion in OpenAI as it seeks to push its technology even further, according to a person familiar with the matter.
>
> The potential $10 billion deal—which would mainly provide OpenAI with even larger amounts of computing power—has not been finalized and the funding amount could change. But the talks are indicative of the tech giant's determination to be on the leading edge of what has become the hottest technology in the tech industry.

95. Microsoft's purpose was to continue to obtain access to all of OpenAI's innovations, so that Microsoft could incorporate them in its own products that competed with OpenAI. As the New York Times reported, Microsoft's senior executives were unambiguous on this point:

> "The expectation from Satya is that we're pushing the envelope in A.I., and we're going to do that across our products," Eric Boyd, the executive responsible for Microsoft's A.I. platform team, said in an interview.

96. As Microsoft moved to deepen OpenAI's exclusive dependence on Microsoft's Azure AI platform, Microsoft simultaneously incorporated OpenAI's technology into products that competed directly with ChatGPT and other OpenAI products. For example, Microsoft released its Co-Pilot product as part of its GitHub line of business, which generated code for developers using the OpenAI LLM technology. Microsoft also incorporated DALL-E technology into its Bing search engine, as well as ChatGPT technology. As the Information Reported in a January 3, 2023 article, titled "Microsoft and OpenAI Working on ChatGPT-Powered Bing in Challenge to Google":

> Microsoft is preparing to launch a version of its Bing search engine that uses the artificial intelligence behind ChatGPT to answer some search queries rather than just showing a list of links, according to two people with direct knowledge of the plans. Microsoft hopes the new feature, which could launch before the end of March, will help it outflank Google, its much bigger search rival.
>
> Details of Bing's integration of OpenAI software couldn't be learned, but the technology could power more full-sentence answers to questions from Bing users. And ChatGPT itself has some ideas about what the collaboration could yield. . . .

97. Bing integration came well after Microsoft had directly incorporated OpenAI technology into products that competed with OpenAI's core GPT-based business. As The Information explained (and explained above), Microsoft had already launched a paid LLM service in 2022:

> Microsoft already launched a paid service, Copilot, in June 2022 to help software developers automatically generate code with the aid of OpenAI's software. The Copilot tool has received favorable reviews from some prominent engineering leaders.
>
> The OpenAI software for Copilot and ChatGPT is based on a generative pre-trained transformer, GPT, also known as a large-language model. A transformer, which Google researchers pioneered years ago, is a type of

> deep-learning model that infers relationships among enormous amounts of data scraped form the internet. Deep learning is a branch of machine learning that has taken the business world by storm in the past decade. Companies use deep learning to automate everything from customer service replies to making predictions about products their customers might buy.

98. Microsoft's incorporation of GPT technology into Bing had also been part of the companies' pact, which began in 2019:

> But what OpenAI didn't say was that the 2019 Microsoft investment included an agreement to incorporate some aspects of GPT into Bing. Microsoft said in October that it was integrating a different OpenAI model—an image-generation tool, Dall-E 2—with Bing's Image Creator tool. And Bing teams have previously incorporated an older version of GPT into Bing in minor ways, including to power its automatic search suggestions that appear as people type. The forthcoming announcements will reveal more meaningful features, according to someone with direct knowledge of the situation.

99. In 2023, the companies extended their relationship even further in a secretive agreement (its existence was public; its details were not). Microsoft announced the deal on its official blog in a January 23, 2023 post titled, "Microsoft and OpenAI extend partnership":

> Today, we are announcing the third phase of our long-term partnership with OpenAI through a multiyear, multibillion dollar investment to accelerate AI breakthroughs to ensure these benefits are broadly shared with the world.
>
> This agreement follows our previous investments in 2019 and 2021. It extends our ongoing collaboration across AI supercomputing and research and enables each of us to independently commercialize the resulting advanced AI technologies.
>
> - **Supercomputing at scale**—Microsoft will increase our investments in the development and deployment of specialized supercomputing systems to accelerate OpenAI's groundbreaking independent AI research. We will also continue to build out Azure's leading AI infrastructure to help customers build and deploy their AI applications on a global scale.
>
> - **New AI-powered experiences**—Microsoft will deploy OpenAI's models across our consumer and enterprise products and introduce new categories of digital experiences built on OpenAI's technology. This includes Microsoft's Azure OpenAI Service, which empowers

> developers to build cutting-edge AI applications through direct access to OpenAI models backed by Azure's trusted, enterprise-grade capabilities and AI-optimized infrastructure and tools.
>
> - **Exclusive cloud provider**—As OpenAI's exclusive cloud provider, Azure will power all OpenAI workloads across research, products and API services.

100. By the end of 2024, more details about the deal terms between Microsoft and OpenAI had become public, including after OpenAI co-founder Elon Musk filed suit against the company.

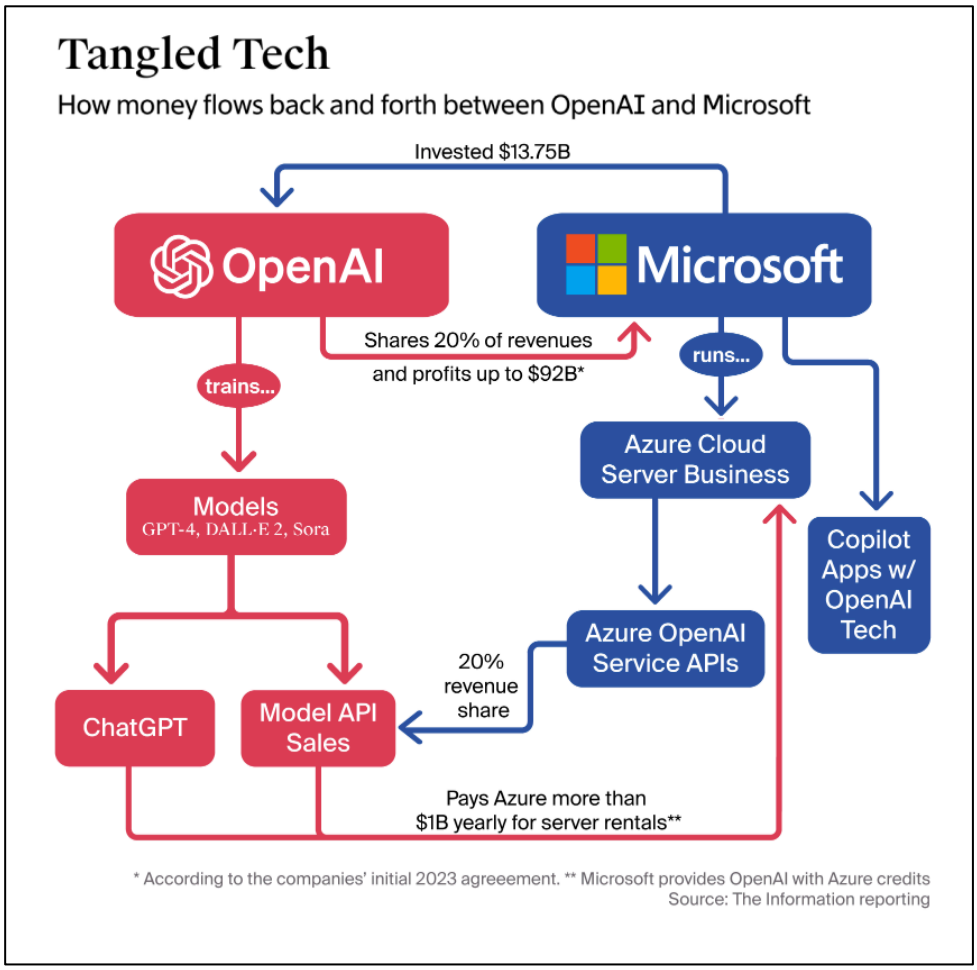
101. The companies were now in talks to circumvent the capped-profit structure, allowing Microsoft to deepen its stake in OpenAI while maintaining its stranglehold over the company's demand for computation—the most important cost input to OpenAI's commercial product. By the end of 2024, Microsoft had invested a massive $13 billion in OpenAI and was harvesting OpenAI's innovations for its own competing products—all through a web of private entities that were formed to end-run around the non-profit parent's limited charter.

102. As The Information reported in on December 26, 2024 in an article titled "Microsoft and OpenAI Wrangle Over Terms of their Blockbuster Partnership":

> OpenAI CEO Sam Altman wants to convert the artificial intelligence developer, which is governed by a nonprofit, into a for-profit corporation. His biggest hurdle is Microsoft, which has outsize influence on the process after having committed more than $13 billion to OpenAI.
>
> The companies have been negotiating potential changes in OpenAI's structure since around October. Those talks have focused on four areas: Microsoft's equity stake in the for-profit entity; whether Microsoft will continue to be OpenAI's exclusive cloud provider; how long Microsoft will maintain rights to use OpenAI's intellectual property in its products as it pleases; and whether Microsoft will continue to take 20% of OpenAI's revenue, according to a person who has talked to Altman about the discussions.

103. Reporting on the talks revealed some of the material—and central—terms of the companies' agreement: (a) Microsoft was permitted to make a massive investment in OpenAI—$13 billion; (b) OpenAI could buy cloud computing ***only*** from Microsoft, which would be OpenAI's exclusive provider; (c) Microsoft maintained rights to incorporate all of OpenAI's technology into its own competing products as it desired; and (d) Microsoft was entitled to take 20% of OpenAI's revenue.

104. *The Information* captured the flow of money and technology between the companies in a cogent infographic:



105. One key detail that emerged publicly was that OpenAI's deal with Microsoft was slated to continue until OpenAI achieved AGI. Critically, however, the companies defined AGI not in technological terms, but solely as an economic threshold. As The Information reported:

> In addition, last year's agreement between Microsoft and OpenAI, which hasn't been disclosed, said AGI would be achieved only when OpenAI has developed systems that have the "capability" to generate the maximum total profits to which its earliest investors, including Microsoft, are entitled, according to documents OpenAI distributed to investors. Those profits total about $100 billion, the documents showed.

106. Notably, AGI would not be achieved upon the attainment of $100 billion profits, but rather upon the "capability" to generate such profits. That threshold, however, is such a distant prospect for the

company, which is projected to turn its first profits in 2029. At present, OpenAI's agreement with Microsoft is effectively a perpetuity.

II. MICROSOFT EXTRACTS SUPRACOMPETITIVE PROFITS FROM OPENAI'S CUSTOMERS

A. "We Have No Moat and Neither Does OpenAI": The Threat of a More Open Model

107. By December 2022, when OpenAI launched its consumer-facing Consumer Generative AI product in earnest, Microsoft was perfectly positioned to exploit its relationship with OpenAI. Microsoft was well familiar with the power of having a first-mover advantage in technology-based markets. It understood since—a lesson it learned in the 1990s, and in fact wrote down for posterity. Out of all the tech giants, Microsoft has a unique and long-held understanding of the economic benefits that flow to first-movers in markets with powerful network effects and feedback loops.

108. Microsoft, however, also understood that network effects can in fact work bidirectionally—that is, in a market characterized by powerful network effects, a virtuous circle (a powerful positive feedback loop) can turn into a death spiral (a powerful negative feedback loop). The next killer app in a network market can induce consumer switching from a legacy firm—even if that firm was the first mover. (That is, unless the dominant firm actively distorts the market to maintain its advantage—behavior that got Microsoft into legal trouble with its PC operating system monopoly).

109. As a Microsoft senior executive, Nathan Myhrvold, observed in a draft memorandum to Bill Gates in May 1994 about Microsoft's operating system business:

> The Monopoly that Isn't
>
> The fact that VHS video tape format prevailed over the Beta format used by the Sony [sic] is probably familiar to most of us. Although Beta started first and was championed by Sony, it wound up losing the battle so thoroughly that players for it are no longer even made. As a result, VHS currently has virtually 100% of the market share for home players and prerecorded video tapes (8mm video tape is a contender, but so far only for handheld camcorders). Classical economics would tell us that this is has [sic] all the earmarks of a monopoly. One would expect that the Japan Victor Company which controls and licenses the VHS standard to be reaping monopoly profits, and that both consumers and competitors would be suffering the various harms of a monopoly— exorbitant pricing, restrain[t]s on free competition and a whole host of other symptoms.

> Even the most casual inspection of the consumer electronics markets shows that this is not the case. One might at first suspect this is unusually virtuous behavior on the part of JVC; while it is true in my experience that they are a fine firm, this is not a complete answer. Probing further, one could ask how many standards exist for music CDs? The answer of course is only one—a consortium composed of Sony and Phillips licenses 100% of that market, again without the deleterious effects one would expect from such a complete monopoly. Indirect competition from other prerecorded music formats might at first blush seem to be a factor, but it is not a sufficient explanation.
>
> A new breed of economists, such as W. Brian Arthur . . . of Stanford University have realized what a small number of companies have known for some time—that there exist some fundamental exceptions to classical economic theory which require new insights and laws to explain them. The classical notions of monopoly simply do not apply.
>
> The phenomena responsible for the VHS triumph over Beta is a positive feedback cycle. Consider the situation early in the VCR battle when VHS started getting more popular. As the players proliferated, video rental stores tended to stock more VHS tapes than Beta—the cost of having both in equal numbers was too onerous. The owner of a VHS player was therefore more likely to find the movie he wanted at the video store than a Beta owner. This made VHS fundamentally more useful to its owners, and caused even more people to buy VHS players, which in turn further incented video stores to stock VHS. Positive feedback existed for VHS market share.

110. Myhrvold recognized that feedback loops and network effects created powerful advantages for companies that were first to capture a critical share of the market. He also, however, recognized that the dominance that such a firm achieves is the result of an inherently unstable process:

> Mathematically speaking, any process of this sort is fundamentally unstable. Perfectly balanced competitors could maintain their market share, but once an imbalance started it would grow without bound. The initial lead that allows one contender to pull away from the others might be the result of a deliberate action or it could even be random chance. In either event, the initial lead will grow at an exponential rate until it captures the bulk of the market. . . .
>
> The computer industry as we know it today is full of examples of positive feedback. The value of a computer to its user depends on the quality and variety of the application software available for it. The incentive to create such software for a particular computer depends on the number of users, since they are the potential customers for the application developer. This creates a similar situation to the video store—the best software is attracted to the most popular platform, making it more popular still.

111. Myhrvold recognized that the effect of the feedback loop was that the first mover would receive a dominant share, and second and third movers would fractally achieve dominant shares of what little is left of the market:

> The historical situation is that the market share leader in systems software takes about 90% or so of the market, the runner up takes about 90% of what is left and so on. This maps reasonably well to the current world wide market share figures—MS Dos based computers have about 91%, Apple Macintosh computers which run the Macintosh OS have about 8% and the largest variant of UNIX has less than 1% [GET PUBLISHED FIGURES] [SIC]. Applications software usually gets a somewhat smaller benefit from positive feedback than systems software because the effects which drive the feedback are less central to how someone uses the application. As a result the typical figures are something like 60% to 70% share for the leader, 60% to 70% of the remainder for the runner up and so forth.

112. Myhrvold further recognized that the same feedback loop could rapidly unwind a dominant position if a new entrant manages to clone the dominant product:

> As a general rule of thumb, a product in the computer industry—whether hardware or software—is only as strong as its current version. A single strong release, or a weak one can make or break a product, or company. In most cases a single version is marketed for between one and two years until it is replaced, so there is little time to rest on ones [sic] laurels. The strongest products might have a bit more leeway, but despite all of the advantages that accrue to the incumbent leader, no product could expect to survive two consecutive bad versions—at least if its competition is awake. Conversely it takes two consecutive good versions (and thus 3-4 years) to establish a newcomer. When a product does fall behind, the positive feedback cycle becomes a double edged sword because it will help the challenger just as surely as it helped the leader in a previous round. . . .
>
> These factors explain why high market shares created by positive feedback lack the negative symptoms of a traditional monopoly. The market share leader must maintain an extremely competitive posture with respect to pricing, technological innovation and openness, or else he risks a long and irreversible fall from grace. In a sense, the leader is a prisoner both of his own success, and the process which put him there in the first place.

113. In the context of the nascent Consumer Generative AI Market, OpenAI's first-mover advantage gave it market power, but its partner Microsoft quickly recognized the phenomenon the company had long recognized and indeed studied in technology markets—absent market distortion,

OpenAI's pricing power would be dependent on its ability to continue innovating ahead of its competitors.

114. Google, for its part, also recognized that Consumer Generative AI technology businesses were potentially vulnerable to competition, at least from companies with the computational infrastructure required for entry at scale—namely Google, Microsoft, and Amazon, each of which maintained massive cloud-computing infrastructure that could be used to train Generative AI models.

115. In May 2023, an internal Google memorandum on OpenAI's advantages in Consumer Generative AI leaked to the public. Excerpts of the memorandum were reported on semianalysis.com on May 4, 2023, and made clear that Google understood that OpenAI's first-mover advantage was contingent on maintaining its competitiveness. More importantly, open-source models posed a threat to barriers to entry protecting both Google and OpenAI's Generative AI businesses:

> **We Have No Moat and Neither Does OpenAI**
>
> We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?
>
> But the uncomfortable truth is, *we aren't positioned to win this arms race and neither is OpenAI*. While we've been squabbling, a third faction has been quietly eating our lunch.
>
> I'm talking, of course, about open source. Plainly put, they are lapping us. . . .

116. As Google explained, open-source models, such as Facebook's Llama LLMs, created a risk that the LLM models to which OpenAI and (eventually) Google sold consumers access could be run locally on users' machines or on servers by competitors.

117. As explained below, Google's insights were part prescient—and part incorrect.

118. First, Google's prediction about computing infrastructure requirements lessening over time simply proved wrong. In early 2023 (the time of Google's memo), model sizes remained at magnitudes that could conceivably allow local inference on a user's phone or local machine. However, as the AI "arms race" continued, it became clear that the computing cost of running models would in fact

not shrink, but rather *grow* significantly, requiring massive computation resources to serve a large user base at scale.

119. What Google got right, however, was that OpenAI was vulnerable to disruption by an outsider with a more open model. As explained below, this would ultimately erode OpenAI's market power—but not before Microsoft leveraged it to extract supracompetitive profits from consumers.

B. Microsoft Creates a Supply Constraint to Extract Supracompetitive Profits

120. By the end of 2022, Microsoft decided to extract as much value from its OpenAI relationship as possible. Realizing that OpenAI's first-mover advantage could be eroded by a new entrant or by a series of lackluster releases by OpenAI, Microsoft moved quickly to manipulate the most powerful lever it had over the price of OpenAI's consumer products—the cost and availability of computation.

121. OpenAI grew rapidly upon releasing its first ChatGPT product to consumers and developers. On June 12, 2024, The Information Reported on OpenAI's massive revenue growth in an article titled "OpenAI's Annualized Revenue Doubles to $3.4 Billion Since Late 2023":

> OpenAI has more than doubled its annualized revenue to $3.4 billion in the past six months or so, OpenAI CEO Sam Altman has told staff, a sign that growth in the ChatGPT developer's business is accelerating despite intensifying competition.
>
> Annualized revenue—a measure of the past month's revenue multiplied by 12—was $1.6 billion in late 2023, The Information previously reported, and about $1 billion last summer. That rapid growth reflects how quickly businesses and individuals have incorporated OpenAI's conversational AI and chatbot in their work.

122. The surge in revenue came directly from OpenAI's consumer-facing subscription products and its developer-facing APIs. OpenAI also received a modest cut of any OpenAI-branded models Microsoft sold to its Azure customers—but not from Microsoft products derived from OpenAI technology, such as Copilot.

> Most of OpenAI's revenue—about $3.2 billion on an annualized basis—comes from subscriptions to its chatbots as well as fees from letting software developers access its models through an application programming interface. Microsoft has typically taken a cut from some of OpenAI's sales of AI models because those models run on Microsoft's cloud. OpenAI also receives a cut from Microsoft's sales of OpenAI models to Microsoft's own

> Azure cloud customers. OpenAI's cut now amounts to about $200 million on an annualized basis, or roughly 20% of the revenue Microsoft is generating from that business, Altman told staff."

123. In other words, when OpenAI sold access to its models through Microsoft to Azure users, Microsoft kept 80% of the revenue. With respect to OpenAI's own revenue, Microsoft took a substantial share for itself as part of its deal.

124. OpenAI's revenue through 2023 and into the beginning of 2024 was far ahead of its competitors, reflecting its significant first-mover advantage and (at the time) pricing power. As The Information reported:

> The revenue rate puts the ChatGPT-maker far ahead of its rivals. Last fall, for instance, rival Anthropic told investors it was generating revenue at a $100 million annualized rate, with plans to reach more than $850 million in annualized revenue by the end of 2024. Cohere, a Canadian OpenAI rival, was generating just $22 million in annualized revenue in April.

125. OpenAI, however, was still not profitable. "Inference" costs—the cost of responding to user queries by running the trained model, remained high, particularly at the prices and capacity limits imposed on OpenAI's business by Microsoft. As The Economist reported on May 15, 2025 in an article titled, "Will OpenAI ever make real money?":

> OpenAI is cagey about its numbers. But according to estimates, for every $1 in training costs, GPT-4 would cost around $4 a year to run, based on OpenAI's current level of traffic. For O3, whose reasoning relies on more computing in the post-training "inference" phase, the ration could be as high as one to 100.

126. In other words, while training costs were already high, the cost of running a Generative AI model at scale and in response to millions of users was multiples higher still. OpenAI's revenue continued its rapid rise, and by the end of 2024, the company was expecting $13 billion in revenue, but its costs—a direct result of the price and supply of compute provided by Microsoft—continued to increase with the scale of its models. As The Economist reported:

> These ballooning operating costs explain OpenAI's mounting losses. Despite tripling its sales to $3.7bn in 2024, it lost perhaps $5bn (excluding stock-based compensation). This year it expects revenue to triple again, to $13bn, and inference costs to grow at the same rate, to $6bn. A shifting

> cost structure also makes it hard to price products and plan budgets. A fixed subscription fee that made sense in the age of GPT-4 looks unviable for O3. You could try keeping subscriptions for older, dumber versions and add a variable usage fee for inference-heavy reasoning. But how many people will pay anything for obsolete technology? And how long until the next model forces another complete rethink?

127. OpenAI confronted the problem Microsoft's Nathan Myhrvold wrote about in a company memorandum in the 1990s: OpenAI's technological advantage and first-mover status provided it market power, but that market power would not last long against rapid product obsolescence and other forces threatening its dominant position. As explained in this Complaint, Microsoft used its control over OpenAI's supply of compute to extract maximum (indeed, supracompetitive) value from ChatGPT consumers during the period in which OpenAI held market power.

128. OpenAI's predictions about the rate at which its inference costs would increase with revenue were directionally accurate. As The Information reported on January 31, 2025, in an article titled "ChatGPT Subscribers Nearly Tripled to 15.5 Million in 2024," OpenAI's sales were skyrocketing:

> Paid subscribers to ChatGPT nearly tripled to 15.5 million last year from 5.8 million a year earlier, OpenAI recently told some shareholders, despite competition from chatbots made by Google, Anthropic and Meta Platforms.
>
> Based on what OpenAI charges for the chatbot subscriptions, the increase means ChatGPT was likely generating at least $4 billion in annualized revenue around the end of last year, or $333 million per month, just two years after its launch.
>
> Separately, usage of OpenAI's application programming interface—which gives companies such as Salesforce and T-Mobile access to its AI models—increased seven times, the company told the shareholders.

129. Despite its massive growth and high prices, OpenAI operated its business at a net loss after considering its largest cost—its deal with Microsoft. Microsoft not only forced OpenAI to buy compute exclusively from its Azure platform, it extracted significant portions of the revenue generated by OpenAI. OpenAI became desperate to renegotiate its deal with Microsoft. As The Information reported:

> Meanwhile, OpenAI is negotiating to reduce Microsoft's 20% cut of its revenue as they rework the terms of their contract, according to two people

> who have spoken to OpenAI leaders about it. Such a change could further boost OpenAI's value.
>
> The revenue sharing is part of a deal in which Microsoft funded the startup and provided it with servers to run its technology. In return, OpenAI has gotten a 20% cut of the revenue Microsoft generates from selling OpenAI models to its cloud customers.
>
> OpenAI's total 2024 revenue could not be learned, but the company last fall projected revenue of $4 billion, mostly from subscriptions to ChatGPT, and a loss of $5 billion—excluding stock based compensation—in part due to high computing costs. For the second half of that year, it set a goal of generating at least $3.5 billion in annualized revenue, or $290 million per month, from the chatbot by the end of the year, and it beat that by 14%. OpenAI's API business generates the rest of the company's revenue.

130. Microsoft recognized that microeconomic truism that restricting supply and increasing price are two sides of the same coin. Recognizing OpenAI's substantial, but likely impermanent, market power in the Consumer Generative AI Market that ChatGPT had inaugurated, Microsoft promptly went to work squeezing supply and output of Generative AI products produced by OpenAI, causing prices to rise marketwide—including for Microsoft's own competing Generative AI products. It did so by fastidiously extracting its 20% of OpenAI's revenues—not profits, but revenues—and refusing to expand the supply of compute available to OpenAI for its models.

131. Sam Altman publicly lamented the constraint lack of compute imposed on OpenAI's release of products as well as on the company's operation of its existing products at scale. ChatGPT, for example, was forced to degrade the functionality of its voice and image-based features, which it had previously demoed to the public with great fanfare. As TechCrunch reported in an October 31, 2024 article titled "OpenAI CEO Sam Altman says lack of compute capacity is delaying the company's products":

> Many reports suggest that OpenAI has struggled to secure enough compute infrastructure to run and train its generative models. Just this week, Reuters, citing sources, said that OpenAI has for months been working with Broadcom to create an AI chip for running models, which could arrive as soon as 2026.
>
> Partly as a result of strained capacity, Altman said, OpenAI's realistic-sounding conversational feature for ChatGPT, Advanced Voice Mode, won't be getting the vision capabilities first teased in April anytime soon.

> At its April press event, OpenAI showed the ChatGPT app running on a smartphone and responding to visual cues, such as the clothes someone was wearing, within view of the phone's camera.

132. As OpenAI was forced to reduce its output and even degrade its product in response to supply constraints imposed by its Microsoft agreement, Microsoft itself aggressively deployed and promoted its own competing Consumer Generative AI products, which were developed based on technology it extracted from OpenAI under its agreement with the company.

133. In October 2023, Microsoft began selling a version of Copilot for its Microsoft Office 365 suite of applications, with prices as high as $30 per person per month—higher than ChatGPT's base subscription price. As CNBC reported in a November 1, 2023 article titled, "Microsoft starts selling AI tool for Office, which could generate $10 billion a year by 2026":

> Microsoft is primed to enjoy its next cycle of growth. On Wednesday, the company started selling the Microsoft 365 Copilot artificial intelligence add-on for its Office app subscription targeting businesses.
>
> The feature that appears in Word, Excel and other Office programs will cost $30 per person per month. That can add up to more than $10 billion in annualized revenue by 2026, Piper Sandler analyst Brent Bracelin and Hannah Rudoff wrote in a note to clients earlier this week.

134. In other words, as Microsoft restricted the supply of compute necessary for OpenAI to lower its prices and develop new products to avoid obsolescence, Microsoft leveraged OpenAI's technology to directly compete, announcing its own Consumer Generative AI products unconstrained by the capacity constraints it imposed on its chief competitor. Moreover, Microsoft benefited directly from the inflated prices it caused by constraining OpenAI's product output, as Microsoft was able to charge more than OpenAI for its rival product. The net effect was the inflation of the market price for Consumer Generative AI products and the simultaneous erosion of OpenAI's first-mover advantage to the benefit of Microsoft.

135. As explained below, Microsoft continued to extract supracompetitive prices from OpenAI's customers—and from the market through its own competing products—until a disruptive new entrant from China entered at scale with a competitive and more open Consumer Generative AI model.

III. DEEPSEEK ENTERS THE CONSUMER GENERATIVE AI MARKET AND PRICES COLLAPSE FOR OPENAI'S PRODUCTS

136. By the end of 2024, vector-based computation at scale and OpenAI's asymmetric know-how from being the first mover remained significant barriers to entry in the Consumer Generative AI Market. But they were not insurmountable. And indeed, in early 2025, Google's analysis two years earlier proved prescient—a well-resourced outsider espousing open-source concepts entered the market and sparked a price war.

137. In the beginning of 2025, a new entrant from China, DeepSeek, upended the Consumer Generative AI market by releasing its own model at a fraction of the cost and using a fraction of the computing power used by OpenAI and others then in the Consumer Generative AI Market. The result was a price war—except, as explained below, OpenAI's prices remained orders of magnitude inflated over the rest of the market because of Microsoft's contractual limitations on OpenAI's ability to buy compute from any other source for ChatGPT and other OpenAI products. By imposing an artificial output restraint on OpenAI's Generative AI products, the OpenAI-Microsoft agreement actually set a price floor for ChatGPT products.

138. As explained below, OpenAI's flagship model was 136x the price charged by the competition, threatening ruin for OpenAI if Microsoft, its horizontal competitor, held it to its restrictive agreement on compute.

139. Further, because ChatGPT was still the market leader with a non-negligible first-mover advantage, the artificial constraint on OpenAI capacity imposed by the Microsoft agreement inflated not just ChatGPT's own prices, but those marketwide.

A. DeepSeek Enters the Market

140. On January 20, 2025, a Chinese company called DeepSeek AI stunned global markets. It announced a Consumer Generative AI model that rivaled OpenAI's, yet DeepSeek had managed to train the model for (as reported by Axios in January 2025) approximately $6 million—a rounding error when compared to what OpenAI had spent to train its models. What's more, DeepSeek had apparently accomplished this feat while U.S. export controls limited its ability to obtain scarce Nvidia GPUs required to train an LLM at scale.

141. DeepSeek was started by Chinese hedge fund manager Liang Wenfeng, who had turned his attention to AI research after amassing significant computational resources and infrastructure in connection with his hedge fund. As *Wired* reported in a January 25, 2025 article titled "How Chinese AI Startup DeepSeek Made a Model that Rivals OpenAI":

> Even within the Chinese AI industry, DeepSeek is an unconventional player. It started as Fire-Flyer, a deep-learning research branch of High-Flyer, one of China's best-performing quantitative hedge funds. Founded in 2015, the hedge fund quickly rose to prominence in China, becoming the first quant hedge fund to raise over 100 billion RMB (around $15 billion). (Since 2021, the number has dipped to around $8 billion, though High-Flyer remains one of the most important quant hedge funds in the country.)
>
> For years, High-Flyer had been stockpiling GPUs and building Fire-Flyer supercomputers to analyze financial data. Then, in 2023, Liang, who has a master's degree in computer science, decided to pour the fund's resources into a new company called DeepSeek that would build its own cutting-edge models—and hopefully develop artificial general intelligence. It was as if Jane Street had decided to become an AI startup and burn its cash on scientific research.

142. DeepSeek's accomplishment was especially stunning given U.S. export controls on the necessary Nvidia GPUs required to train such a large and competitive model. As *Wired* explained, DeepSeek innovated around this ordinarily crippling constraint:

> In October 2022, the US government started putting together export controls that severely restricted Chinese AI companies from accessing cutting-edge chips like Nvidia's H100. The move presented a problem for DeepSeek. The firm had started out with a stockpile of 10,000 A100's, but it needed more to compete with firms like OpenAI and Meta. "The problem we are facing has never been funding, but the export control on advanced chips," Liang told 36Kr in a second interview in 2024.
>
> DeepSeek had to come up with more efficient methods to train its models. "They optimized their model architecture using a battery of engineering tricks—custom communication schemes between chips, reducing the size of fields to save memory, and innovative use of the mix-of-models approach," says Wendy Chang, a software engineer turned policy analyst at the Mercator Institute for China Studies. "Many of these approaches aren't new ideas, but combining them successfully to produce a cutting-edge model is a remarkable feat."
>
> DeepSeek has also made significant progress on Multi-head Latent Attention (MLA) and Mixture-of-Experts, two technical designs that make

> DeepSeek models more cost-effective by requiring fewer computing resources to train. In fact, DeepSeek's latest model is so efficient that it required one-tenth the computing power of Meta's comparable Llama 3.1 model to train, according to the research institution Epoch AI.

143. In addition to innovating around part of the barrier to entry protecting OpenAI and other competitors in the Consumer Generative AI Market, DeepSeek made its new innovations open source, as Google had previously feared a disruptive new entrant would. As *Wired* reported:

> DeepSeek's willingness to share these innovations with the public has earned it considerable goodwill within the global AI research community. For many Chinese AI companies, developing open source models is the only way to play catch-up with their Western counterparts, because it attracts more users and contributors, which in turn help the models grow. "They've now demonstrated that cutting-edge models can be built using less, though still a lot of, money and that the current norms of model-building leave plenty of room for optimization," Chang says. "We are sure to see a lot more attempts in this direction going forward."

144. DeepSeek alarmed U.S. CGAI companies, including OpenAI and Microsoft, not just because it managed to train its state-of-the-art model at a fraction of the cost, but also because it was offering its new model at a fraction of the price U.S. companies charged. As Reuters reported in a January 29, 2025 article titled "DeepSeek's low-cost AI spotlights billions spent by US tech":

> While the price of using AI models has been falling with rising competition and the progress in the technology, Bernstein's Rasgon said DeepSeek stands out as it has priced its models at up to 40 times lower than OpenAI's comparable models.
>
> That could, analysts said, start a price war for AI services, potentially pressuring tech companies such as OpenAI that are already losing billions of dollars each year due to the high operational costs of running services such as ChatGPT.

145. Although other CGAI entrants, such as Anthropic and Google, had the ability to lower their prices to compete with DeepSeek, OpenAI did not—because the company's agreement with Microsoft imposed an artificial capacity constraint that acted as a price floor.

146. In the early-2025 price war, OpenAI reduced its prices to some extent, but nonetheless maintained prices far higher than the rest of the market, reflecting the company's unique artificial output

restraint imposed by its agreement with Microsoft. OpenAI also introduced o3-mini, a new, lower-computing-power model that the company could more flexibly price in competition with DeepSeek. But even with price cuts, OpenAI's models were all orders of magnitude more expensive than the CGAI competition, including in comparison to state-of-the-art models by Anthropic and Google.

147. On February 25, 2025, in an article titled, "DeepSeek rushes to launch new AI model as China goes all in," Reuters reported on the price cuts and product changes that had occurred in the Consumer Generative AI Market in response to DeepSeek:

> DeepSeek's pricing was 20 to 40 times cheaper than what OpenAI charged for equivalent models, analysts at Bernstein brokerage estimated in early February.
>
> For now, Western and Chinese tech giants have signaled plans to continue heavy AI spending, but DeepSeek's success with R1 and its earlier V3 model has prompted some to alter strategies.
>
> OpenAI cut prices this month, while Google's Gemini has introduced discounted tiers of access. Since R1's launch, OpenAI has also released an O3-Mini model that relies on less computing power.

148. In response to the price cuts and product changes that had occurred, DeepSeek fired another volley. On February 26, 2025, DeepSeek announced that it would drop its "off-peak" pricing by 75%. Off-peak, however, meant off-peak in China—which was prime business hours in the United States.

149. As Reuters reported in a February 26, 2025 article titled "DeepSeek cuts off-peak pricing for developers by up to 75%":

> Chinese AI startup DeepSeek on Wednesday introduced discounted off-peak pricing for developers looking to use its AI models to build their own products, its website showed, a move that could put pressure on rivals in China and overseas to cut prices. . . .
>
> The Hangzhou-based company said on Wednesday that between 1630 GMT and 00030 GMT, the cost of using its API, a platform that allows developers of other apps and web products to integrate its AI models, would be up to 75% cheaper.
>
> Usage costs during this timeframe for the API of the R1 and V3 models would be 75% and 50% cheaper, respectively, according to a table on DeepSeek's website.

> While the company calls this timeframe "off-peak" as it starts at 0030 and ends at 0830 in Beijing, it encompasses daytime hours in Europe and the United States, where DeepSeek's cheap but powerful models triggered a sell-off in tech stocks.

150. DeepSeek had overtly targeted the United States Consumer Generative AI Market with additional price cuts. These added 75% price cuts put even more pressure on U.S. companies to further cut their own prices.

151. Although virtually every U.S. Consumer Generative AI Market participant made aggressive price cuts to remain competitive, OpenAI was cornered. Its exclusive compute purchase agreement with Microsoft had imposed an artificial capacity constraint that acted as a supracompetitive price floor for ChatGPT and other OpenAI products—and Microsoft was not budging. Not only did Microsoft, OpenAI's horizontal competitor, refuse to release OpenAI from its exclusivity requirement, it continued to constrain the supply of compute available to OpenAI.

152. In response, Google continued to cut prices. As reported by Cybernews in a February 6, 2025 article titled "Google releases Gemini 2.0 AI models with a lower price tag than DeepSeek-R1":

> Google now also offers better Gemini 2.0 Flash API pricing for developers. It is priced at $0.10 for a million input tokens (DeepSeek-R1 - $0.14-$0.55/1M input tokens) and $0.4 for a million output tokens (DeepSeek-R1 - $2.19/1M output tokens). Google hasn't listed the price for the Pro model. Google models also support image and audio inputs.

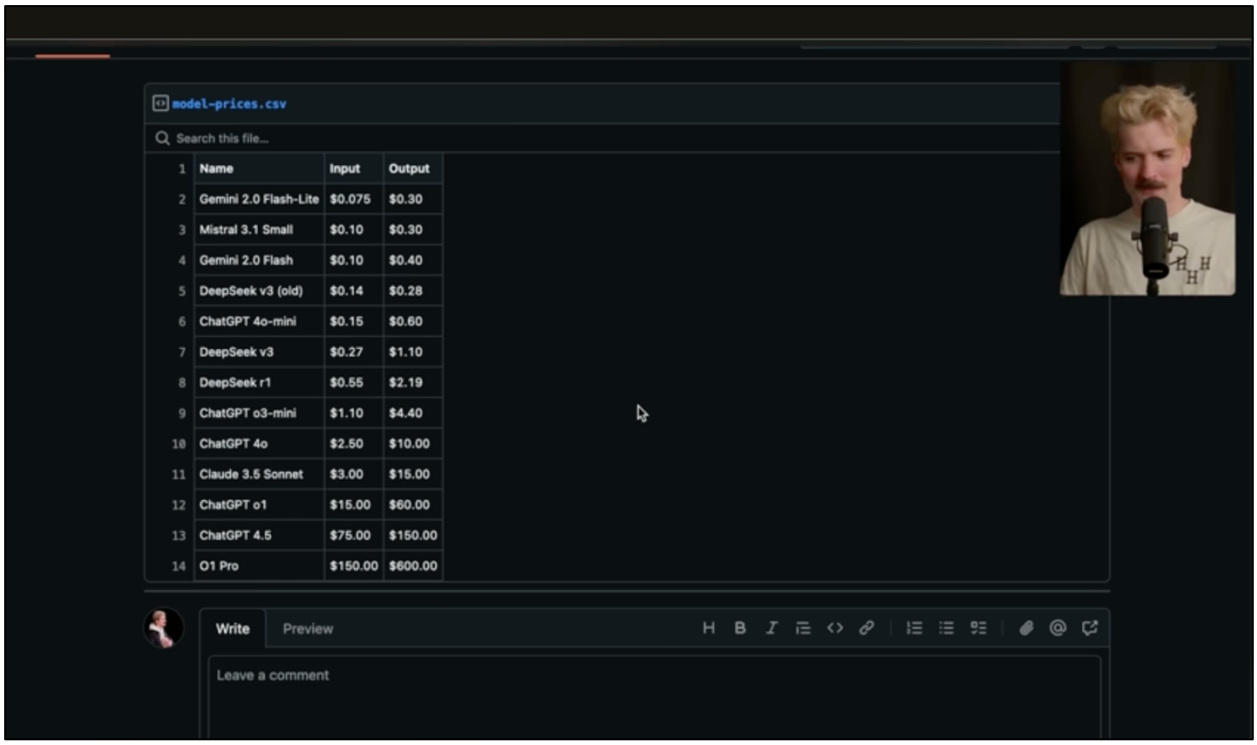
153. Other competitors followed—except OpenAI.

B. DeepSeek's Entry Causes a Price War, but OpenAI Maintains Its Inflated Prices

154. By March 2025, OpenAI's prices had become an albatross. OpenAI's state-of-the-art models were priced more than 100x more than what every competitor charged for equivalent models.[3] The pricing had become so disparate that prominent market participants began to balk.

[3] As explained in more detail in the market and harm to competition sections of this Complaint, OpenAI maintains per-token prices for developers alongside its monthly fees for consumers for access to ChatGPT. OpenAI's per-token pricing for ChatGPT models derives from, and closely models, the output constraints it imposes on flat-fee consumer products in the form of model quality, rate limiting, speed, and other measures of consumer value for paying subscribers. The higher the per-token price for ChatGPT products, the less value OpenAI provides to flat-fee subscribers to ChatGPT and other CGAI flat-fee

155. For example, leading AI and AI-assisted coding YouTuber Theo Browne (@t3dotgg) posted a video on March 21, 2025 comparing OpenAI's prices with those of competitors, titled "OpenAI's new API is 200x more expensive than competition. The video featured a price comparison of each CGAI market participant (per million tokens):



156. ChatGPT's flagship O1 Pro model was priced at $150 per million input tokens and $600 per million output tokens. OpenAI sold the model through its own front-end (as a flat-fee subscription) at an eye-popping $200 per month price point.

157. The model, however, was functionally equivalent to DeepSeek, Anthropic, and Google's flagship reasoning models, which were orders of the magnitude less expensive. DeepSeek's r1 was priced at $0.55 per million input tokens and $2.19 per million output tokens. Anthropic's Claude 3.5 Sonnet was priced at $3 per million input tokens and $15 per million output tokens. Google's Gemini 2.0 Flash sold at $0.10 per million input tokens and $0.40 per million output tokens.

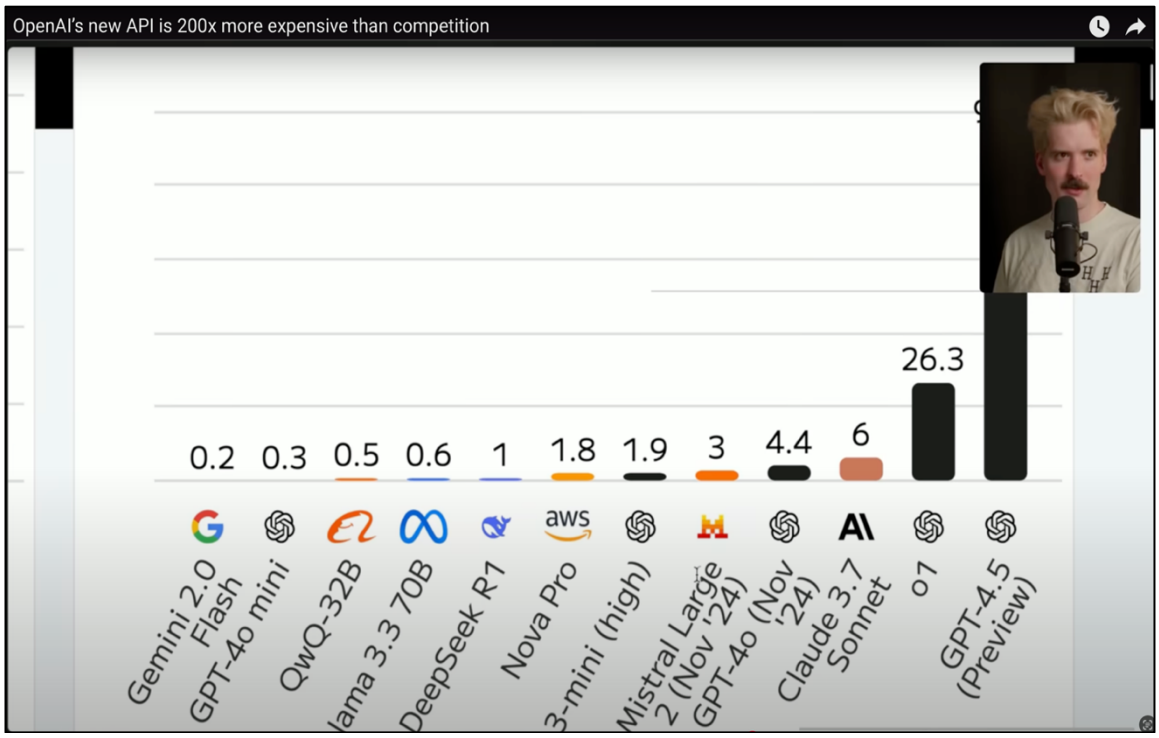
products. This is a standard response of a firm to an artificial capacity constraint: increase the price for a fixed quantity, or decrease quantity for a fixed price. In the context of consumer goods, this phenomenon has been recently called "shrinkflation"—receiving less of a product for the same price.

158. OpenAI's o3-mini model and new ChatGPT 4.5 models were also orders of magnitude more expensive than the competition. Although the entire CGAI Market had cut its prices, OpenAI's prices remain inflated—substantially so.

159. OpenAI's flagship model had also been severely degraded to reduce the amount of use—and thus computing power—it had to purchase from Microsoft. The monthly subscription pricing, in particular, risked ruinous losses if left unconstrained. As such, OpenAI made its UI virtually unusable and made it difficult to copy and paste prompts or code segments outside of its own interface. As Theo Browne observed in the same video:

> O1 Pro is a bad experience. What the [expletive]. This is $200 a month. Nadda. Nothing. Failure. And now I have to go rerun it and be that much closer to hitting my rate limit there. All for 200 bucks a month. . . .
>
> I can't even select the text because the scroll bar is in the way. I am not making this up. I just want to copy this message and paste it on a better service. Maybe this an attempt at locking me in, but all it's doing is pissing me off.

160. OpenAI's pricing was so out of line with every other product in the market that it confounded the scale of Browne's comparison chart:



161. Notably, the price of OpenAI's newly introduced o3-mini model was misleading, as it was released with a low, medium, and high setting, which changed the number of tokens it would use as part of the reasoning process. A high setting, which would make the model on par with other reasoning models on the market in performance, would generate significantly more tokens, causing the API-access price to rise. With respect to subscription-based access to the model, OpenAI degraded performance and limited the rate at which users could use the model.

162. Notably, even at the misleading price, o3-mini was still double the price of the equivalent model offered by competitors. As Theo Browne observed:

> But when you use O3-mini-high, the cost isn't necessarily represented just by this number, because when you use it on high, it is generating way more tokens. That's what the low, medium, high is. It's how much time can it spend and how many tokens can it generate in the step before it starts answering. It's almost like the ratio between the answer to the reasoning/thinking stage. Both cost just as much money per token, but the reasoning steps aren't useful to you. They're only useful to the AI. Which means that even if you got the same length of answer from O3-mini and 4o, O3 mini will have more tokens used because it had to reason and generate more tokens before the actual answer came out. So low, medium, high, the difference isn't it cost more based on the output length. It cost more based on the time spent generating the things that are used to generate the correct output.

163. OpenAI's models were approximately ***136 times the price*** charged by other CGAI firms for equivalent models. As Browne explained:

> And the price gap here is just unfathomable. It's 136 times more expensive for input tokens. This isn't comparing an open source model or a bad model. This is comparing the best model right now, which is o3-mini-high. 150/1.1 is 136. If we compare output numbers here, it's 600/4.40 which is still almost exactly 136 times more expensive. How? How the [expletive] does this exist now. Like this model isn't new. What's the point of even releasing this. It almost feels like a response to my video that we're racing to the bottom to just put something like this out. Their margins aren't great. They aren't going to make much money off of this. It feels like a parity thing where they don't want to have any models that aren't released externally. That aren't available via API. But this model came far too late and is far not good enough for this. It's weird and I don't get it. I've had a much better experience with o3-mini-high.

164. O1 and O1 Pro were likewise priced supracompetitively—approximately ***200 times the price*** charged by the competition. As Browne lamented, while OpenAI had attempted to release differently priced models, the rest of the market had simply dropped prices to competitive levels:

> It's 10x more expensive than O1, which is – Both O1 and O1 Pro are worse than O3-mini. O3-mini is way faster, way cheaper, way better, and you compare to other reasoning models like r1, which is even cheaper, you end up realizing that O1 pro is roughly 200x more expensive than it should be. I don't know how we got here, but I am thankful that despite OpenAI's insistence on breaking new ground in pricing, the rest of the industry is focused on lowering the price.

165. With prices 200 times those of CGAI competitors and no unique product features to offer to meaningfully distinguish its ChatGPT models, by early 2025 OpenAI had to free itself of its agreement with Microsoft. If it did not do so, it could simply no longer compete.

166. Yet Microsoft held its ground, and for months, OpenAI charged customers—many locked-in through history, experience, and technology to its longtime market-leading ChatGPT products—prices that were artificially inflated by Microsoft. Microsoft, for its part, continued reaping inflated profits from OpenAI while it sold its own competing products at market prices.

IV. OPENAI OBTAINS COMPUTE FROM GOOGLE, ENDING ITS MICROSOFT EXCLUSIVITY, AND ITS INFLATED PRICES IMMEDIATELY DROP

167. Tensions rose between OpenAI and Microsoft. The exclusive compute requirement in the companies' agreement made it impossible for OpenAI to buy cheaper compute from another cloud provider, constraining OpenAI's capacity and instituting an artificial price floor on OpenAI's products, including ChatGPT. Microsoft, however, continued to reap the benefits of the inflated prices OpenAI charged OpenAI customers who were locked into the company's Generative AI products due to switching costs. Microsof—which was itself a horizontal competitor of OpenAI in the Consumer Generative AI Market and thus profiting on both the front and back ends from the market-impacting inflation in OpenAI prices—had no reason to relinquish its stranglehold over its competitor.

168. In fact, Microsoft monetized OpenAI's models, including o1, through its Copilot product, which it bundled with its Office365 services suite. Yet Microsoft's subscription plans remained

competitive, and the functionality Microsoft offered improved—unlike OpenAI, which was forced to degrade its functionality to avoid ruinous losses.

169. The OpenAI agreement was wildly beneficial for Microsoft. The tech giant had milked its horizontal competitor OpenAI's first-mover advantage in CGAI for cloud compute profits at the same time that Microsoft readied and then commercialized its own competing CGAI products. At the same time, Microsoft imposed artificial capacity restraints on the CGAI first-mover, ultimately neutralizing the emergent CGAI threat posed by OpenAI while simultaneously capturing the company's technology and extracting all the supracompetitive profits that could obtained during OpenAI's precious first-mover period.

170. Microsoft, however, understood that OpenAI's pricing power would ultimately vanish. As OpenAI's prices became more and more out of line with the rest of the market in 2025, the company's days of CGAI dominance were numbered. OpenAI could remain the market leader in the long term only if it began adopting market prices—even switching costs would not keep the company's locked-in user base in the fold indefinitely, given the shockingly large artificial price premium imposed on OpenAI products from Microsoft's compute restraint. As OpenAI's first-mover advantage eroded, so too did the agreement's value to Microsoft, reducing the incentive for Microsoft's executives to continue to invest in their joint venture with OpenAI, including by providing compute.

171. On June 10, 2025, the dam broke. An exclusive news story by Reuters reported that OpenAI was tapping Google—its largest competitor—for AI compute. As Reuters reported in a June 10, 2025 article titled "Exclusive: OpenAI taps Google in unprecedented cloud deal despite AI rivalry, sources say":

> OpenAI plans to add Alphabet's Google cloud service to meet its growing needs for computing capacity, three sources told Reuters, marking a surprising collaboration between two prominent competitors in the artificial intelligence sector.
>
> The deal, which has been under discussion for a few months, was finalized in May, one of the sources added. It underscores how massive computing demands to train and deploy AI models are reshaping the competitive dynamics in AI, and marks OpenAI's latest move to diversify its compute sources beyond its major supporter Microsoft, including its high-profile Stargate data center project.

172. The move signaled that Microsoft had relaxed its exclusive compute requirement. Moreover, Google itself would now hold its own primary competitors' costs directly in its hands—as Microsoft had for several years, though without exclusivity. As Reuters observed:

> Google's DeepMind AI unit also competes directly with OpenAI and Anthropic in a race to develop the best models and integrate those advances into consumer applications.
>
> Selling computing power reduces Google's own supply of chips while bolstering capacity-constrained rivals. The OpenAI deal will further complicate how Alphabet CEO Sundar Pichai allocates the capacity between the competing interests of Google's enterprise and consumer business segments.

173. At the same time that OpenAI's new deal with Google and the end of its exclusivity agreement with Microsoft became public, OpenAI immediately cut its prices to come in line with the rest of the market. OpenAI's CEO Sam Altman tweeted on June 10, 2025:

174. As soon as Microsoft's exclusivity provision abated, OpenAI's prices dropped ***80%***.

175. This simultaneous price drop makes clear that Microsoft's exclusivity agreement had indeed inflated OpenAI's prices for years—and was the only barrier to lowering them to competitive levels.

176. OpenAI's new prices had just a few months prior been 136-200x that of its competitors. Its new prices were finally competitive. As VentureBeat reported on June 10, 2025 in an article titled "OpenAI announces 80% price drop for o3, its most powerful reasoning model":

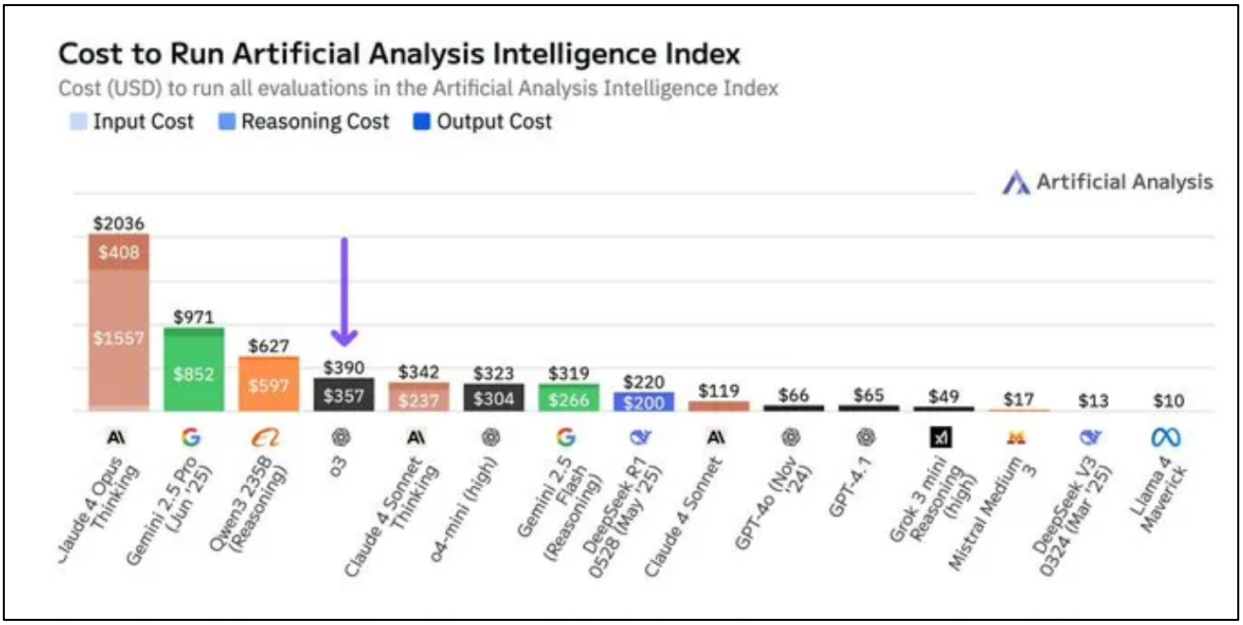
> The cost of using o3 is now $2 per million input tokens and $8 per million output tokens, with an extra discount of $0.50 per million tokens when the user enters information that has been "cached" or is stored and identical to what they provided before.[4]
>
> This marks a significant reduction from the previous rates of $10 (input) and $0 (output), as OpenAI researcher Noam Brown pointed out on X.
>
> Ray Fernando, a developer and early adopter, celebrated the pricing drop in a post writing "LFG!" short for "let's fucking go!"
>
> The sentiment reflects a growing enthusiasm among builders looking to scale their projects without prohibitive model access costs.

177. As a chart included in the VentureBeat article made clear, OpenAI's prices were now clearly in line with the rest of the market:

Model	Input	Cached Input	Output	Discount Notes
OpenAI o3	$2.00 (down from $10.00)	$0.50	$8.00 (down from $40.00)	Flex Processing: $5 / $20
Gemini 2.5 Pro	$1.25 – $2.50	$0.31 – $0.625	$10.00 – $15.00	Higher rate applies to prompts >200k tokens
Claude Opus 4	$15.00	$1.50 (read) / $18.75 (write)	$75.00	50% off with batch processing
DeepSeek-Chat	$0.07 (hit)$0.27 (miss)	—	$1.10	50% off during off-peak hours
DeepSeek-Reasoner	$0.14 (hit)$0.55 (miss)	—	$2.19	75% off during off-peak hours

[4] Caching frequently occurs because many prompts to Generative AI contain instructions that are automatically included every time the model is prompted, such as, *inter alia*, instructions not to hallucinate information, how to format the response, or concerning the tone in which to deliver a response.

178. Estimated costs of running AI analysis had also put OpenAI's prices in line with the market—notably, as to precisely the same models OpenAI had been selling at inflated prices for months after DeepSeek's entry and the ensuing price war. As another figure in VentureBeat showed, OpenAI's overall pricing made OpenAI's models competitive for the first time:



179. The effect of lifting Microsoft's exclusive compute restraint was undeniable—and instantaneous. OpenAI's Generative AI products ***immediately*** dropped in price to market levels. Put simply, Microsoft's anticompetitive restraint was the but-for and proximate cause of OpenAI's supracompetitive prices and degraded product quality.

180. OpenAI itself was aware that Microsoft's squeeze of its business was anticompetitive. Even after the June 2025 relaxation as to Google, the anticompetitive agreement with Microsoft remained—albeit now in a more latent form. But Microsoft, OpenAI's horizontal competitor, could (and can) choose to enforce the compute restraint again, even to potentially eliminate OpenAI from the CGAI competitive landscape. The Microsoft agreement with its compute restraint sat (and sits) as a sword of Damocles wielded by OpenAI's own giant—and increasingly CGAI-focused—direct competitor.

181. At various points including in mid-2025, OpenAI internally considered leveraging antitrust accusations against Microsoft and seeking government review of its agreement. As the Wall

Street Journal reported on June 16, 2025 in an article titled "OpenAI and Microsoft Tensions Are Reaching a Boiling Point":

> OpenAI wants to loosen Microsoft's grip on its AI products and computing resources, and secure the tech giant's blessing for its conversion into a for-profit company. Microsoft's approval of the conversion is key to OpenAI's ability to raise more money and go public.
>
> But the negotiations have been so difficult that in recent weeks, OpenAI's executives have discussed what they view as a nuclear option: accusing Microsoft of anticompetitive behavior during their partnership, people familiar with the matter said. That effort could involve seeking federal regulatory review of the terms of the contract for potential violations of antitrust law, as well as a public campaign, the people said.

182. Among the contentious issues between the company was the elephant in the room—Microsoft held control over OpenAI's products and pricing while simultaneously competing with it. Moreover, Microsoft was balking at OpenAI's intended acquisition of Windsurf, a company that provides a coding front-end for Generative AI models in direct competition with Microsoft's GitHub Copilot:

> OpenAI and Microsoft are at a standoff over the terms of the startup's $3 billion acquisition of the coding startup Windsurf, the people said. Microsoft currently has access to all of OpenAI's IP, according to their agreement. It offers its own AI coding product, GitHub Copilot, that competes with OpenAI. OpenAI doesn't want Microsoft to have access to Windsurfs intellectual property.

183. More importantly, OpenAI wants out of its exclusivity agreement, even in its relaxed form after the June 10th announcement of the deal with Google. As The Wall Street Journal reported:

> The startup is trying to renegotiate elements of that deal alongside its planned conversion. It wants to join with other cloud providers so it can sell its technology to more customers and access additional computing resources. Microsoft, meanwhile, wants access to OpenAI's technology even after the startup declares its models have achieved humanlike intelligence, which would end the current partnership.

184. To the extent Microsoft has a role in selecting which providers other than Microsoft that OpenAI can deal with, it can ensure OpenAI remains under control as a direct competitor. While Microsoft may have decided that Google could be allowed to provide additional cloud compute, it may

not provide such a dispensation with respect to a cloud provider that is not a direct competitor of OpenAI's. The risk of future and selective enforcement of the provision is too competitively dangerous to OpenAI's business to ignore. As such, OpenAI continues to seek an exit from its agreement with Microsoft.

185. As described below, the agreement's exclusivity provision should be declared unlawful and enjoined from further enforcement. Otherwise, Microsoft can exert direct price and output controls over its horizontal CGAI competitor—as it had done in the past.

V. THE AI COMPUTATION BARRIER TO ENTRY

186. During the Damages Class Period, the Consumer Generative AI Market was protected by an AI Computation Barrier to Entry ("AICBE"), which prevented companies other than those with significant technical knowledge of AI systems and massive cloud computing resources from entering and competing with OpenAI. As explained below, this barrier to entry limited the CGAI Market to a small, stable group of firms until January 20, 2025, when DeepSeek AI innovated around it and sparked a price war. Prior to DeepSeek's entry in late January 2025, the CGAI was remarkably stable due to the AICBE, and during this period Microsoft used its agreement with OpenAI to extract supracompetitive prices from CGAI consumers, including subscribers to OpenAI's then-market leading ChatGPT.

187. From January 20, 2025 to the present, the compute required for training and inference remains a significant barrier to entry in the CGAI Market, but not an insurmountable one, particularly with respect to companies such as Google that maintain their own cloud-based AI computation infrastructure. Further, switching costs and lock-in effects remained for legacy firms—principally legacy first-mover OpenAI—in the CGAI Market after DeepSeek's January 2025 entry. This allowed Microsoft to continue to extract supracompetitive prices through its exclusive compute agreement even after DeepSeek entered and sparked a price war—until OpenAI was finally allowed to obtain compute from its competitor, Google on June 10, 2025.

188. The AI Computation Barrier to Entry in its full form—as it existed until January 20, 2025, when DeepSeek innovated around it and sparked a price war—is described in this section.

A. The Scale Problem and the Transformer

189. The advent of deep neural networks brought the promise of solving complex tasks by learning directly from data. That is, rather than rely on a series of rules or evolving probability estimates, deep learning created the prospect of a computer being able to find patterns in the data itself and to learn to process information based on that data.

190. The depth of a neural network allows it to create refined representations of data. Traditional forms of artificial intelligence required extensive work to extract important features in data, which would then be used by a model to learn from data. A deep model solves many of the problems presented when using traditional AI models, including the tangled knot of interrelated features implicit in real-world datasets. As the seminal book on deep learning, DEEP LEARNING, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville explains:

> A major source of difficulty in many real-world artificial intelligence applications is that many of the factors of variation influence every single piece of data we are able to observe. The individual pixels in an image of a red car might be very close to black at night. The shape of the car's silhouette depends on the viewing angle. Most applications require us to disentangle the factors of variation and discard the ones that we do not care about.
>
> Of course, it can be very difficult to extract such high-level abstract features from raw data. Many of these factors of variation, such as a speaker's accent, can be identified only using sophisticated, nearly human-level understanding of the data. When it is nearly as difficult to obtain a representation as to solve the original problem, representation learning does not, at first glance, seem to help us.
>
> Deep learning solves this central problem in representation learning by introducing representations that are expressed in terms of other, simpler representations. Deep learning allows the computer to build complex concepts out of simpler concepts. Figure 1.2 [omitted] shows how a deep learning system can represent the concept of an image of a person by combining simpler concepts, such as corners and contours, which are in turn defined in terms of edges.

191. Deep learning created the possibility that complex tasks could be solved by increasing the depth of the network. The deeper the network, the more complex patterns an artificial neural network could extract.

192. In practice, however, increasing the depth of a network created significant problems. The deeper the network, the more likely the values in the network would "blow up," meaning become so large during training that the network fails to train. Neurons could also result in smaller and smaller values, referred to as a vanishing gradient. Operating the neural network could result in dead neurons, which in training would result in no change with the dataset. As the depth of the neural network increased, the stability of the intermediary values in the network suffered, robbing the model of the power it could achieve from scale and size.

193. As DEEP LEARNING explains, these problems arise inherently from the depth of a network:

> Another difficulty that neural network optimization algorithms must overcome arises when the computational graph becomes extremely deep. Feedforward networks with many layers have such deep computational graphs. So do recurrent networks . . . which construct very deep computational graphs by repeatedly applying the same operation at each time step of a long temporal sequence. Repeated application of the same parameters gives rise to especially pronounced difficulties.

194. Practitioners developed methods to deal with these problems, including methods for regularizing each layer of neural networks to keep their outputs within tighter bounds. From 2016 through the end of 2018, much academic research was directed toward such methods. Even with regularization, however, traditional methods were limited in their ability to consider sequences of tokens as a whole. As *Wired* recounts in a March 20, 2024 article titled "8 Google Employees Invented Modern AI. Here's the Inside Story":

> But the field was running into limitations. Recurrent neural networks struggled to parse longer chunks of text. Take a passage like Joe is a baseball player, and after a good breakfast he went to the park and go two hits. To make sense of "two hits," a language model has to remember the part about baseball. In human terms, it has to be paying attention. The accepted fix was something called "long short-term memory" (LSTM), an innovation that allowed language models to process bigger and more complex sequences of text. But the computer still handled those sequences strictly sequentially—word by tedious word—and missed out on context clues that might appear later in a passage. "The methods we were applying were basically Band-Aids," Uszkoreit [one of the early transformer pioneers] says. "We could not get the right stuff to really work at scale."

195. The breakthrough came when transformers were invented. These neural networks contained what is referred to as an "attention" mechanism. The attention mechanism breaks up each layer's inputs into fragments, allowing the network to see the data as whole, instead of as individual pieces. This reduced the need to iteratively reprocess data, such as with recurrent networks. The output of the attention mechanism is then fed into a feed-forward neural network layer before producing the final output. This keeps the output of each layer within bounds, allowing the network to become deeper without the drawbacks associated with traditional deep neural networks.

196. When coupled with other tricks, such as tagging inputs with alternating trigonometric functions, transformers allowed networks to keep track of the order of inputs, all while breaking them into smaller pieces and considering them at once.

197. The transformer had solved the first major barrier to creating large-scale neural networks. Using transformers, models could be built and trained at massive scale, with billions of parameters.

198. As these models were used for generative AI applications, such as LLMs, it became clear that they would become emergently intelligent with scale. Simply increasing the size of a model would significantly increase its power.

199. As the *MIT Technology Review* explained in a December 21, 2021 article titled "2021 was the year of the monster AI models":

> When OpenAI released GPT-3, in June 2020, the neural network's apparent grasp of language was uncanny. It could generate convincing sentences, converse with humans, and even autocomplete code. GPT-3 was also monstrous in scale—larger than any other neural network ever built. It kicked off a whole new trend in AI, one in which bigger is better.

200. As OpenAI itself recognized, new ideas were not needed as much as scale to improve their AI models. As the same *Microsoft Technology Review* article reported:

> GPT-3 grabbed the world's attention not only because of what it could do, but because of how it did it. The striking jump in performance, especially GPT-3's ability to generalize across language tasks that it had not been specifically trained on, did not come from better algorithms (although it does rely heavily on a type of neural network invented by Google in 2017, called a transformer), but from sheer size.

> "We thought we needed a new idea, but we got there just by scale," said Jared Kaplan, a researcher at OpenAI and one of the designers of GPT-3, in a panel discussion in December at NeurIPS, a leading AI conference.
>
> "We continue to see hyperscaling of AI models leading to better performance, with seemingly no end in sight," a pair of Microsoft researchers wrote in October in a blog post announcing the company's massive Megatron-Turing NLG model, built in collaboration with Nvidia.

201. Scale was the name of the game, and the architectural barrier in deep learning had largely been broken. What was left was a far less forgiving barrier to creating, training, testing, and running new models—computation.

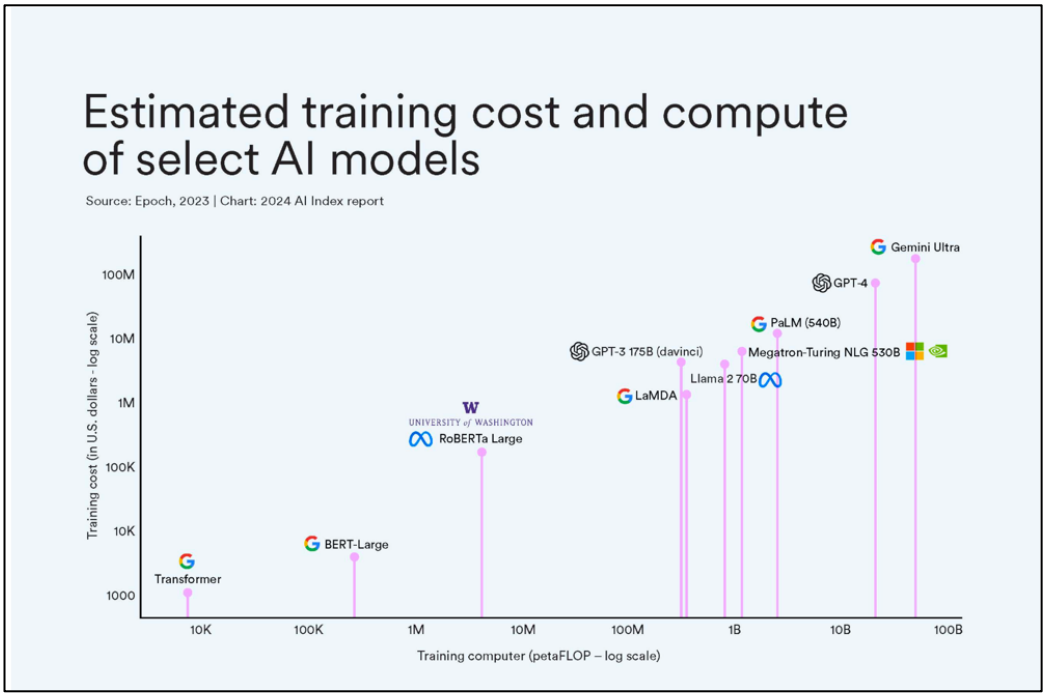
B. The Need for Massive Computation

202. The gating ingredient for building generative AI models based on transformers is computation. Indeed, the cost of training large transformer models has grown exponentially.

203. As a recent Stanford University Human-Centered Artificial Intelligence report, titled "AI Index: State of AI in 13 Charts," reported:

> One of the reasons academia and government have been edged out of the AI race: the exponential increase in cost of training these giant models. Google's Gemini Ultra cost an estimated $191 million worth of compute to train, while OpenAI's GPT-4 cost an estimated $78 million. In comparison, in 2017, the original Transformer model, which introduced the architecture that underpins virtually every modern LLM, cost around $900.

204. The change's exponential nature required Stanford to compare model training costs on a logarithmic scale (depicted below).



205. This increased cost stems from neural-scaling effects. Generally, increasing the size of a model results in an exponential increase in the cost of training. The reason for this is that each branching path of the neural network adds exponential complexity to the process of training, namely by increasing the number of parameters that must be trained through the repeated computation of gradients—mathematical derivatives of large vectors and tensors fed through a neural network. These computations allow a network to "train" by changing parameter weights iteratively, such that predictive error is minimized as the model observes more data.

206. This computation is done using Graphical Processing Units ("GPUs"), processing hardware that can handle large vectors and tensors, performing computation and other operations on them at the same time. The largest provider of GPUs is Nvidia, which has an approximate 88% market share of GPUs designed for AI and high-performance computing. Systems of such GPUs used for training large-scale AI can cost tens of thousands of dollars per unit and consume significant energy.

207. Not only are these GPUs expensive, they are scarce. Nvidia GPUs, particularly cutting-edge products such as Nvidia's H100, are in short supply, and certain large companies have an edge in procuring the required supply of GPUs from Nvidia to provide CGAI products.

208. By 2023, Microsoft and OpenAI, in particular, had secured one of the largest troves of Nvidia GPU arrays. As *Nextplatform* reported in a May 11, 2023 article titled "When Push Comes to Shove, Google Invests Heavily in GPU Compute":

> Ironically, OpenAI is the software vendor and Microsoft Azure is the hardware vendor in this possibly emerging duopoly. Microsoft is said to have used 10,000 Nvidia A100 GPUs to train the GPT 4 large language model from OpenAI and is rumored to be amassing 25,000 GPUs to train the GPT 5 successor. We presume this will be on a mix of Nvidia A100 and H100 GPUs, because getting their hands on 25,000 H100 GPUs could be a challenge, even for Microsoft and OpenAI.
>
> Customers outside of Microsoft and OpenAI using the Azure cloud are more limited in what they can get their hands on.

209. Microsoft had long been securing priority supplies of GPUs from Nvidia. In fact, as early as 2020, Microsoft secured tens of thousands of GPUs, originally provided in part for OpenAI's use. As Microsoft reported on its website on May 19, 2020, in an article titled "Microsoft announces new supercomputer, lays out vision for future AI work":

> The supercomputer developed for OpenAI is a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server. Compared with other machines listed on the Top500 supercomputers in the world, it ranks in the top five, Microsoft says. Hosted in Azure, the supercomputer also benefits from all the capabilities of a robust modern cloud infrastructure, including rapid deployment, sustainable datacenters and access to azure services.

210. Microsoft also entered into an agreement with CoreWeave, a company backed by Nvidia, to obtain additional GPUs. As pymnts.com reported in a July 28, 2023 article titled, "Microsoft Emphasizes Importance of GPU Supply or AI":

> The company's partnership with OpenAI provides it with access to ChatGPT and AI models to improve existing products, such as Microsoft Outlook and Microsoft Word, the report said.
>
> In order to secure enough GPU capacity, Microsoft has signed an agreement with CoreWeave, a company backed by Nvidia, per the report. Going forward, the company plans to increase its capital expenditures to pay for the necessary data centers, central processing units, networking hardware and GPUs.

211. As the article notes, computation is such a significant constraint on the ability to train large-scale generative AI that Microsoft discloses to investors that a failure to obtain sufficient GPUs is a material risk to its business.

212. By June 2024, Microsoft had become one of Nvidia's largest GPU customers. As *The Observer* reported in a June 2024 article titled "Nvidia's Market Cap Surpasses $3T: Here Are the Largest Buyers of Its A.I. Chips":

> Microsoft and Meta are two of the largest buyers of Nvidia's H100 chips. The tech companies spent a combined $9 billion on the accelerators in 2023 alone, according to analysts at DA Davidson, while Omdia Research estimates the two companies acquired 150,000 chips each.
>
> Microsoft reportedly plans to amass some 1.8 million GPUs by the end of 2024, much of which will likely come from Nvidia.

213. Nvidia sells significant chips to CoreWeave and Oracle, with which Microsoft contracts for additional GPU capacity. Combined, Microsoft controls a critical mass of GPUs sourced by Nvida—more than any other company, including Google, is able to obtain.

214. Put simply, large-scale generative AI—the foundation of CGAI products—requires massive amounts of computation, and Microsoft long dominated the supply of GPUs necessary for such large-scale computation.

215. In addition to high training costs, inference costs are also high. Inference is when a trained model runs data or a query through its weights and produces an output. At scale, frequent queries of Generative AI are costly and computationally intensive. To provide results from a large-scale Generative AI model—as in a CGAI product—GPUs are also necessary.

216. The net effect was a powerful barrier to entry surrounding the Consumer Generative AI Market from its inception in late 2022. A new entrant to this market would have to obtain sufficient GPUs to both train models and to fulfill requests to its models by customers. This barrier to entry for years prevented significant competition with OpenAI and its competitor-partner, Microsoft.

C. DeepSeek Eroded, But Did Not Eliminate, the AI Computational Barrier to Entry

217. As explained above, on January 20, 2025, Chinese company DeepSeek AI successfully launched a state-of-the-art Consumer Generative AI product that rivaled OpenAI's flagship model. It reportedly did so at a fraction of OpenAI's costs and that of competitors, such as Anthropic and Google—approximately $6 million to train its model.

218. DeepSeek had managed to overcome the AICBE by innovating around it. As explained earlier in this Complaint, DeepSeek had devised a series of techniques to significantly reduce its computation costs, allowing it to provide access to Generative AI models at a fraction of the price charged by OpenAI.

219. Although the initial effects of its lower prices were initially felt in China, by late January 2025, DeepSeek AI had entered the US Consumer Generative AI Market. Indeed, DeepSeek targeted the U.S. Consumer Generative AI Market in late February 2025 by reducing its off-hours pricing (representing peak hours in the U.S.) by 75%. On January 27, 2025, *TechCrunch* reported that DeepSeek's consumer-facing app had reached number 1 on the U.S. Apple App Store—displacing ChatGPT.

220. Further, with its entry into the U.S. Consumer Generative AI Market, DeepSeek open-sourced its models and innovations, significantly eroding the AICBE for other new entrants. With DeepSeek's entry and its open-sourced models and technology, any firm with sufficient compute resources to enter the market could avail itself of the state-of-the-art model quality DeepSeek had created, then publicly released.

221. The net effect of DeepSeek AI's entry in the U.S. Consumer Generative AI Market in early 2025 was that the AICBE, which protected OpenAI from price competition, was substantially eroded. Although significant compute was (and is) still required for training and inference, DeepSeek's open-source innovations and models meant that a larger pool of firms could incur the cost of entry to the CGAI Market at scale.

222. Even after DeepSeek's entry, however, other barriers to entry persisted, including switching costs, which are described below in the description of the relevant market. These additional barriers to entry, coupled with the weaker-but-still extant AICBE, allowed Microsoft to continue to

extract supracompetitive profits from its anticompetitive agreement with OpenAI even after the entry of DeepSeek in the first quarter of 2025.

VI. THE RELEVANT MARKET

223. The relevant market is the United States Consumer Generative AI ("CGAI") Market, sometimes called the "Chatbot" market for historical reasons.[5] As explained below, this market is a distinct submarket of the general market for AI software.

224. The relevant product in this market is a deep transformer-based AI model that can process tokens as inputs and return tokens as outputs, and be queried as a service by an end-user. The tokens processed by the model underlying the product can be derived from text, pictures, audio, video, or any other form of digital information that can be processed as a vector or tensor for input. The overall product also consists of a user-facing interface, such as either an app or web-based UI. As explained below, these elements form a product for which there is a single submarket of the general market for AI products. The market for this product is what is defined below as the Consumer Generative AI Market.

A. The Consumer Generative AI Market Is a Distinct Submarket

225. The Consumer Generative AI ("CGAI") Market, sometimes called the "Chatbot" market, is a distinct submarket of the general market for AI software. Several relevant factors indicate that the CGAI Market is distinct from other markets, including the general AI software market.

226. ***Industry and public sources recognize the Consumer Generative AI submarket as a separate economic entity.*** Several industry and public sources recognize the Consumer Generative AI submarket. Specifically, these sources recognize a submarket in which OpenAI, Anthropic, Microsoft, Google, xAI and other companies compete directly by offering CGAI/"chatbot" products, such as ChatGPT, Claude, Copilot, Gemini, and Grok.

227. For example, on April 7, 2023, the New York Times, in an article titled, "In A.I. Race, Microsoft and Google Choose Speed Over Caution," described competition among OpenAI, Google, and

[5] As explained earlier in this Complaint, OpenAI essentially created the CGAI Market with its release of ChatGPT, which has led many commentators and observers to refer to the market and its products to this day as "Chatbots." Since the market's inception in 2022, CGAI products do much more than simply chat with users, but the legacy name remains among many observers and even industry participants.

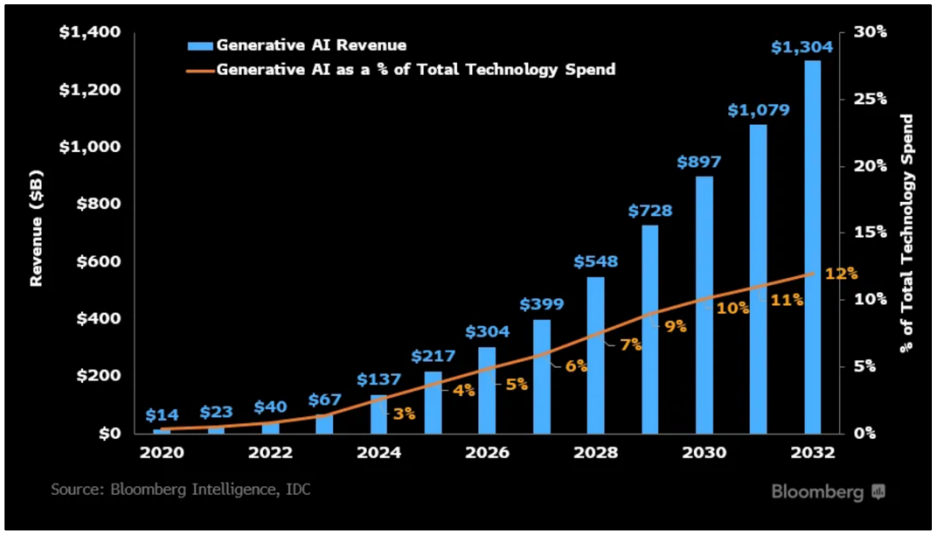
Microsoft in the generative AI market. The article refers to the consumer-facing products in the market as "chatbots":

> The companies released their chatbots anyway. Microsoft was first, with a splashy event in February to reveal an A.I. chatbot woven into its Bing search engine. Google followed about six weeks later with its own chatbot, Bard.
>
> The aggressive moves by the normally risk-averse companies were driven by a race to control what could be the tech industry's next big thing—generative A.I., the powerful new technology that fuels those chatbots.
>
> That competition took on a frantic tone in November when OpenAI, a San Francisco start-up working with Microsoft, released ChatGPT, a chatbot that has captured the public imagination and now has an estimated 100 million monthly users.
>
> The surprising success of ChatGPT has led to a willingness at Microsoft and Google to take greater risks with their ethical guidelines set up over the years to ensure their technology does not cause societal problems, according to 15 current and former employees and internal documents from the companies.

228. On June 1, 2023, in a Bloomberg article titled, "ChatGPT to Fuel $1.3 Trillion AI Market by 2032, New Report Says," ChatGPT and Google's Gemini predecessor Bard, were discussed as part of a "consumer-focused . . . market for generative AI":

> The release of consumer-focused artificial intelligence tools such as ChatGPT and Google's Bard is set to fuel a decade-long boom that grows the market for generative AI to an estimated $1.3 trillion in revenue by 2032 from $40 billion last year.
>
> The sector could expand at a rate of 42% over ten years—driven first by the demand for infrastructure necessary to train AI systems and then the ensuing devices that use AI models, advertising and other services, according to a new report by Bloomberg Intelligence analysts led by Mandeep Singh.

229. The article predicted "Generative AI Revenue" for the market through 2032:



230. A June 14, 2023 report by McKinsey & Company, titled "The economic potential of generative AI: The next productivity frontier" defines generative AI products to include ChatGPT, Anthropic's Claude, and Google's competing models. As the report explained, competing products can perform tasks such as writing, composing music, and creating digital art:

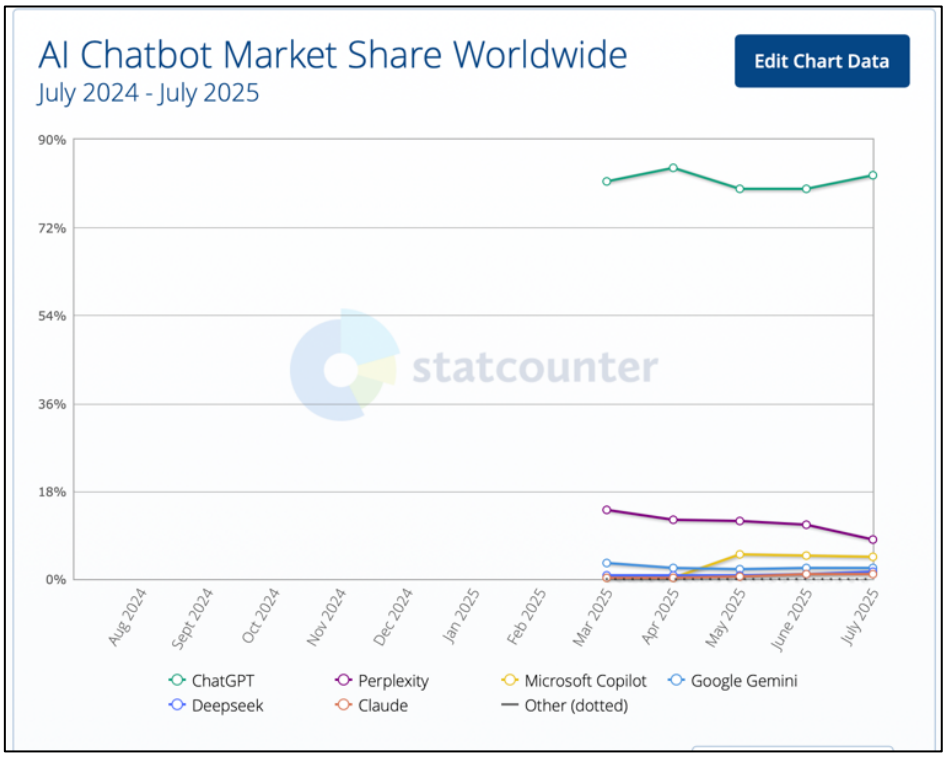
> Generative AI applications such as ChatGPT, GitHub Copilot, Stable Diffusion, and others have captured the imagination of people around the world in a way AlphaGo did not, thanks to their broad utility—almost anyone can use them to communicate and create—and preternatural ability to have a conversation with a user. The latest generative AI applications can perform a range of routine tasks, such as the reorganization and classification of data. But it is their ability to write text, compose music, and create digital art that has garnered headlines and persuaded consumers and households to experiment on their own. As a result, a broader set of stakeholders are grappling with generative AI's impact on business and society but without much context to help them make sense of it.
>
> The speed at which generative AI technology is developing isn't making this task any easier. ChatGPT was released in November 2022. Four months later, OpenAI released a new large language model, or LLM, called GPT-4 with markedly improved capabilities. Similarly, by May 2023, Anthropic's generative AI, Claude, was able to process 100,000 tokens of text, equal to about 75,000 words in a minute—the length of the average novel—compared with roughly 9,000 tokens when it was introduced in March 2023. And in May 2023, Google announced several new features powered by Generative AI, including Search Generative Experience and a new LLM called PaLM 2 that will power its Bard chatbot, among other Google products.

231. An October 5, 2023 article in Zeo, titled, "ChatGPT, Google Bard, Microsoft Bing, Claude, and Perplexity: Which is the Right AI Tool?" refers to competing Consumer Generative AI products as "chatbots." The compared products included Google Bard, ChatGPT, Microsoft's Bing, Anthropic's Claude, and Perplexity.

232. An August 9, 2025 article, titled "From ChatGPT to Gemini: how AI is rewriting the internet" describes direct competition among Consumer Generative AI products from Microsoft, Google, OpenAI, and others:

> Big players, including Microsoft, with Copilot, Google, with Gemeni, and OpenAI, with GPT-4o, are making AI chatbot technology previously restricted to test labs more accessible to the general public.

233. Statcounter, in a July 2025 post, describes an AI Chatbot Market that is coextensive with the Consumer Generative AI Market. It includes ChatGPT, Deepseek, Perplexity, Claude, Microsoft Copilot, Google Gemeni, and others as market participants as of the date of the report:

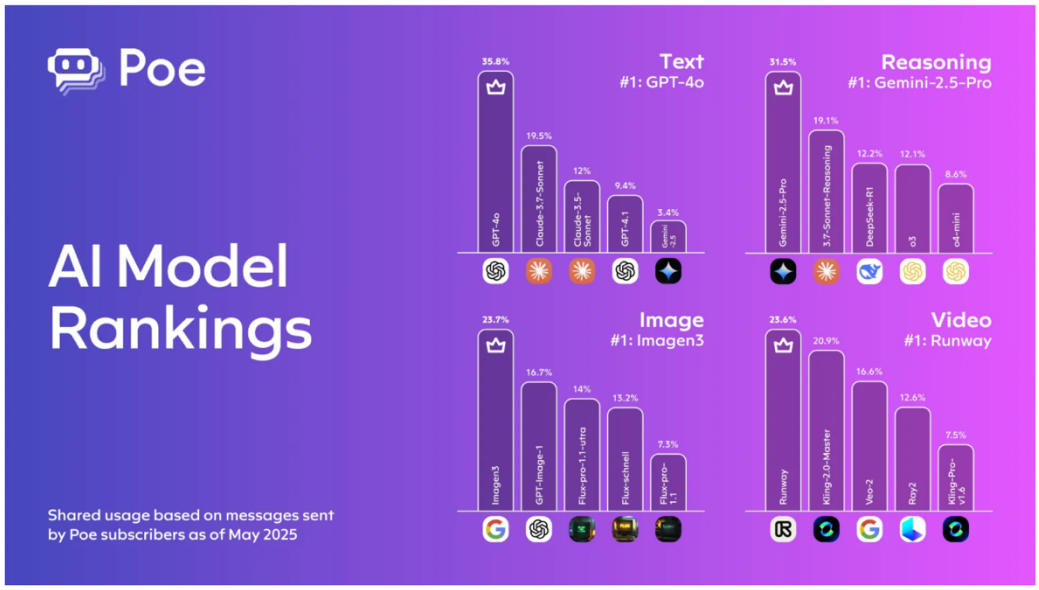


234. The backup data for the chart above confirms ChatGPT's product as dominant:[6]

[6] The full column names in the below table are Date, ChatGPT, Perplexity, Microsoft Copilot, Google Gemini, Deepseek, and Claude.

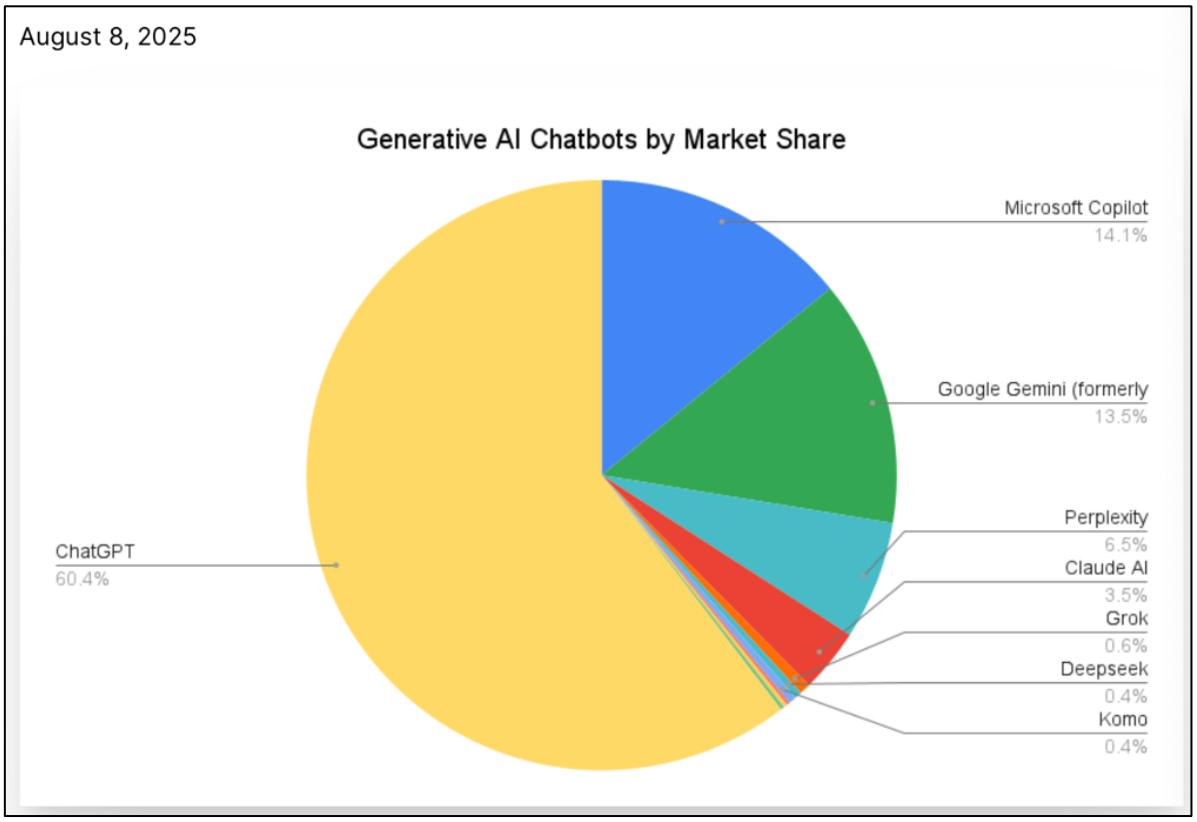
Date	ChatGPT	Perplexity	Microsoft Co	Google Gemi	Deepseek	Claude
2025-03	81.39	14.09	0.34	3.29	0.64	0.25
2025-04	84.21	12.07	0.23	2.31	0.88	0.3
2025-05	79.79	11.83	5.18	1.95	0.8	0.46
2025-06	79.86	11	4.83	2.19	1.02	1.11
2025-07	82.65	8.03	4.59	2.19	1.63	0.91

235. A May 13, 2025 post on VentureBeat titled "AI power rankings upended: OpenAI, Google rise as Anthropic falls, Poe report finds," reports on estimated market shares for each competing model in the Consumer Generative AI Market by usage category:



A chart from Poe showing AI model rankings across different categories as of May 2025. OpenAI's GPT-4o dominates in text generation with 35.8% usage share, while Google's Gemini-2.5-Pro leads in reasoning capabilities and Imagen3 in image generation. (Credit: Poe)

236. On August 8, 2025, FirstPageSage estimated the market shares for "AI Chatbots," which the post defined as covering "the major generative AI chatbots in the U.S as of August 8, 2025." The article explained that "[f]or the purposes of this study, the term 'generative AI chatbot' refers to LLM-based web & mobile applications used by the public to seek answers or create content." The study's market shares are as follows:



237. These sources—and many others like them—confirm a widespread consensus amongst industry and public sources that there is a distinct submarket in the United States for Consumer Generative AI, also known as "Chatbots," and that OpenAI, Microsoft, Google, Deepseek, Anthropic, Perplexity, and xAI are firms that compete within this market.

238. ***Consumer Generative AI products have peculiar characteristics and uses***. Products in the Consumer Generative AI Market are defined by their broad set of capabilities. To begin with, the products are not merely large language models, as they were at the inception of the market. Today, frontier models are "multi-modal," meaning they can process text, audio, images, and other inputs in a fungible fashion. Moreover, these models can also generate text, images, code, and other outputs in response to inputs of varying or mixed types.

239. Products in the Consumer Generative AI Market are maintained in the cloud and are updated and tuned by the companies. They are accessed directly through a consumer-facing UI presented by the company that maintains the underlying models (*e.g.*, OpenAI, Microsoft, Google, Anthropic), using native applications (for example, iOS or Android applications) and web interfaces/web apps, depending on what device a user is on when they wish to interact with the CGAI product (or "chatbot").

240. Products in the Consumer Generative AI Market also include a reinforcement learning layer. Early reinforcement learning layers on the products were referred to as Reinforcement Learning from Human Feedback ("RLHF"). Modern reinforcement learning layers built on top of multi-modal models do not necessarily require human feedback for training and refinement.

241. Products competing in the Consumer Generative AI Market also include either a selector or model multiplexor, which allows varying versions of models, including reasoning-based models, to be selected for a particular query or task. OpenAI's ChatGPT 5 now includes a multiplexing layer that automatically selects the correct model for a particular prompt.

242. The U.S. Consumer Generative AI Market excludes open-source models locally hosted on a users' computer. Open-source models generally must run on local machines using a software lawyer, such as Open Llama ("Ollama").

243. These models are generally not frontier models, as those models cannot run on most users' computers. Most users lack the GPUs, memory, and disk space to host massive frontier models. Moreover, inference using locally hosted models is slow, often prohibitively slow for use competitive with the cloud-hosted models offered by OpenAI, Anthropic, Microsoft, and Google.

244. Locally hosted models, including open-source models, also lack large enough context windows—the size of inputs that can be accepted by the model—to compete with cloud-based Consumer Generative AI products.

245. Competing products in the Consumer Generative AI Market perform a variety of functions. For one, the products accept text, image, or other modalities as prompts and return a variety of responses, including text, images, audio, and video. Most models are capable of returning a mix of such responses.

246. Frontier Consumer Generative AI models are capable of searching the Internet to augment the context window used to provide responses. ChatGPT, Claude, and others provide models that automatically perform searches depending on the substance of a prompt.

247. Consumer Generative AI products can also provide longer "reports," functionality often referred to as "deep research" model.

248. In addition, competing Consumer Generative AI products can use "tools" to perform various tasks, including on a user's own computer. A standardized means of defining and calling tools is referred to as the Model Context Protocol ("MCP"), though other tool-based standards exist.

249. Among the tool-based methods used by Consumer Generative AI models is the ability to create a virtual computer or web browser that can be operated and manipulated by the model.

250. Tool-based functionality is also used for "agentic coding," which may be accessed directly through the CGAI user interface (*e.g.*, ChatGPT.com or Claude.ai in a browser, or ChatGPT or Claude native apps) or through command-line based applications, such as Claude Code.

251. ***Unique production facilities***. Consumer Generative AI products require unique production facilities. To begin with, large arrays of GPUs are required to train models that are used for such products. These GPUs must be powerful enough to perform repeated matrix computations at a massive scale. They must also allow computations to be performed in parallel, meaning that computations must be transmitted to a central server to be aggregated. These GPUs must also possess sufficient memory to process and train large generative AI models. The memory used by these GPUs are generally order of magnitude faster than conventional forms of memory, including memory used on servers.

252. As Liquid Web explained in a webpage titled, "GPU memory and why it's important":

> Unlike general-purpose RAM in a CPU, GPU memory is optimized to work in tandem with the GPU's architecture to quickly access and process vast amounts of data. This makes it crucial not only for rendering high-definition graphics in gaming and visualization but also for powering modern applications like machine learning, artificial intelligence (AI), and high-performance computing (HPC).

253. GPUs are also required for inference. That is, once a model is trained, it is provided a prompt or set of tokens as input. This input is processed through the model by running the matrix computations using the weights determined during training. Like the task of training, inference is computationally demanding and costly. To serve a Consumer Generative AI model to users over the Internet, particularly with low latency and fast response times, a Consumer Generative AI product must run on a large cluster of GPUs or tensor processing units ("TPUs").

254. These GPU clusters must be powered using complex power management systems. These power systems also require extensive cooling. These systems are typically deployed as part of data centers.

255. These production facilities can either be maintained by the Consumer Generative AI company itself or leased/rented from a cloud-based provider, such as Microsoft Azure, Google Cloud, or Amazon Web Services.

256. Although companies such as AMD develop GPUs for AI training and inference, the leading manufacturer of GPUs used for AI training and inference is Nvidia. Generally, Nvidia GPUs are used for large-scale training and inference because they are designed to run on a math and matrix operation library called CUDA, which was developed, and is maintained, by Nvidia.

257. As a March 23, blog post on weightythoughts.com by James Wang, titled "CUDA is Still a Giant Moat for NVIDIA," explains:

> But beyond that, because NVIDIA controls CUDA and their GPUs, they control the entire stack. Software, firmware, and hardware. This is the same thing that gives Apple a killer advantage on their platform. NVIDIA has the same thing, and even aside from programmer productivity/ease (which, as we've learned over the years, is itself huge), stuff in CUDA just magically works much better. Behind the scenes, CUDA performs significant optimizations.
>
> Additionally, as new GPUs come out, as they inevitably do, CUDA continues to be forward and backward-compatible (depending on the features you use), and you continue to reap the benefits of even better hardware and firmware that NVIDIA continues to pump out.

258. Because of NVIDIA's dominance, any data center, whether on a cloud-based platform or directly built for AI training/inference, must source a sufficient number of Nvidia GPUs. These GPUs are difficult to obtain, particularly ahead of large purchasers.

259. For example, to obtain sufficient Nvidia GPUs, xAI's Elon Musk rerouted Nvidia GPUs bound for Tesla to the then-new xAI venture responsible for Grok. As CNBC reported on June 4, 2024 in an article titled "Elon Musk ordered Nvidia to ship thousands of AI chips reserved for Tesla to X and xAI":

> By ordering Nvidia to let privately held X jump the line ahead of Tesla, Musk pushed back the automaker's receipt of more than $500 million in graphics processing units, or GPUs, by months, likely adding to delays in setting up the supercomputers Tesla says it needs to develop autonomous vehicles and humanoid robots.
>
> "Elon prioritizing X H100 GPU cluster deployment at X versus Tesla by redirecting 12k of shipped H100 GPUs originally slated for Tesla to X instead," an Nvidia memo from December said. "In exchange, original X orders of 12k H100 slated for Jan and June to be redirected to Tesla."

260. xAI ultimately worked with Nvidia to build what is now known as the Colossus supercomputer cluster, which is composed of a massive array of GPUs.

XAI Colossus Data Center Compute Hall, courtesy ServeTheHome

261. As VentureBeat described in a January 14, 2025 article titled "Building Colossus: Supermicro's groundbreaking AI supercomputer built for Elon Musk's xAI":

> The team at xAI, partnering with Supermicro and NVIDIA, is building the largest liquid-cooled GPU cluster deployment in the world. It's a massive AI supercomputer that encompasses over 100,000 NVIDIA HGX H100 GPUs, exabytes of storage and lightning-fast networking, all built to train and power Grok, a generative AI chatbot developed by xAI.

262. Microsoft, for its part, maintains the same scale of supercomputing based on Nvidia GPUs, which it requires OpenAI to use for its products. The GPU-heavy computing required for AI training and inference is a vital part of any Consumer Generative AI product.

263. Consumer Generative AI products, namely chatbots, also depend on a reinforcement learning layer built on top of the trained model underlying the product. These reinforcement learning layers have historically required human feedback, meaning that the means of production for the Consumer Generative AI Market includes significant human training to refine the model for consumer use, including by professionals that use the chatbots to generate source code or other technical outputs.

264. ***Distinct customers/consumers***. Consumer Generative AI products have distinct consumers and customers. These chatbots are used for a variety of purposes and accessed either directly through a UI provided by the generative AI company or programmatically, such as through an API. Customers include individuals that query generative AI chatbots for information, such as information from the web. These customers may also query Consumer Generative AI models for therapy, conversation, or companionship. These queries often come from home PCs, smartphones, or tablets, which do not have the computational capacity to run a generative AI model locally. As such, Consumer Generative AI products are often hosted on a cloud or data center, with the user querying the model through the internet.

265. Likewise, some customers use Consumer Generative AI models to create source code or other provide other technical responses. These customers may repeatedly query generative AI models in an iterative fashion, seeking adjustments or edits to codes and code bases. These responses must be rapid, but also optimized for coding, which must often occur through a reinforcement learning layer built on top of the generative AI model.

266. Certain customers also query generative AI models to generate images. Because image generation is computationally expensive and requires powerful GPUs, this must be done remotely and the result transmitted to the customer over the internet.

267. The distinct characteristic of customers that consume Consumer Generative AI products is that they remotely query and interact with the products over the Internet. They do so through devices that cannot run generative AI models locally. They also require responsiveness and accuracy, both of which require significant computation.

268. ***Distinct prices and sensitivity to price changes***. As described above, generative AI products have distinct prices, either a monthly charge for interacting with the product through a

company's own UI, or an API price, which is usually charged per million tokens provided to, or received from, the API. Consumer Generative AI products such as ChatGPT and Claude are priced to consumers through a monthly fee for premium features, including access to the newest models, lower downtime, expanded features, fewer or no explicit rate limits, and faster response time. The API price for the generative AI models underlying CGAI per-month subscription products directly reflects the functionality provided to a CGAI subscriber at a flat fee—API pricing varies price against quantity, per-month fees vary quantity against price at a fixed price point.

269. The price for a CGAI product includes, as a significant input, the price of compute used to create the generative AI product. The cost of compute has the largest impact on the price of generative AI products, as well as the level of product quality provided by a generative AI company.

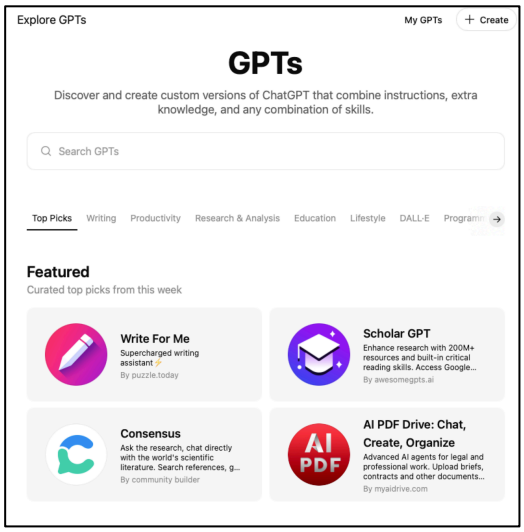
270. Thus, a compute-constrained company such as OpenAI will not only increase or maintain inflated prices, but will also degrade the quality of the products it offers as part of flat-priced subscription products. Degradation of quality may include the use of smaller models, the reduction of reasoning time or "thinking" for reasoning models, and the slowing of response times on a tokens-per-second basis.

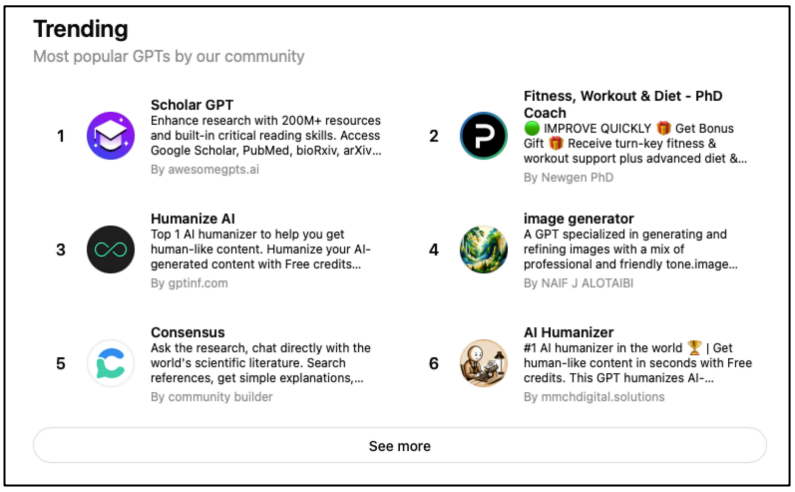
271. Compute is an inherent part of the product, meaning that in the face of a compute constraint, either less compute is used as part of the product and/or the price is increased.

272. As explained above, both monthly and API pricing is highly sensitive to entry by new competitors, including because such competitors provide additional compute capacity to the market. As explained above, DeepSeek's entry resulted in an immediate drop in prices throughout the market for any competitor that could lower prices. OpenAI, as explained above, was constrained by its anticompetitive agreement with Microsoft from lowering its prices in response to new competition by DeepSeek.

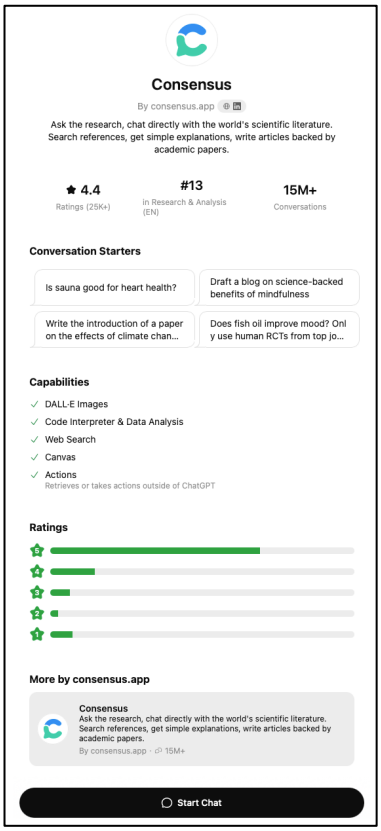
273. ***Specialized vendors***.

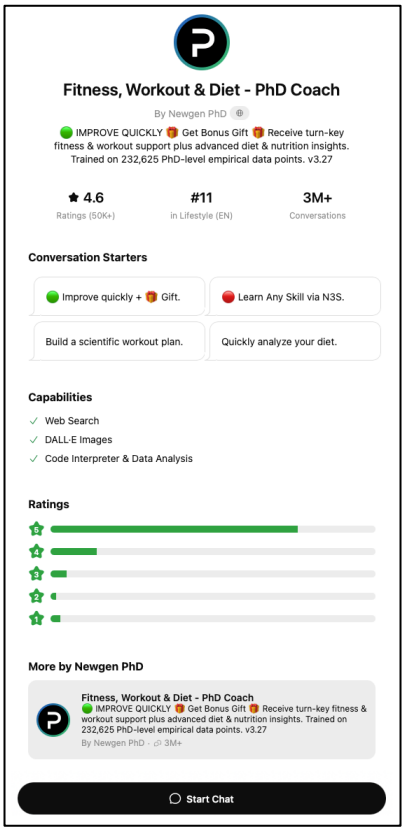
274. Specialized vendords serve Consumer Generative AI products such as ChatGPT. For example, ChatGPT's web interface offers a community-sourced GPTs specialized for various uses.





275. These GPTs are provided not only from the CGAI provider itself, but from specialized vendors like consensus.app and Newgen PhD, among many others:





276. Specialized vendors also include agentic integrated development environments ("IDEs") like Cursor, which bundle and sell access to generative AI products through their API interface.

277. Other specialized vendors include custom GPT products, which implement products built on Consumer Generative AI models, accessed either through an API or through the product's custom UI.

These custom GPTs are provided engineered prompts, which allows them to perform bespoke functions as part of other applications.

B. Market Participants and Market Concentration

278. The U.S. Consumer Generative AI Market includes the following products: Google's Gemini, OpenAI's ChatGPT, Microsoft's Copilot, Anthropic's Claude, Perplexity, xAI's Grok, and DeepSeek AI's Deepseek. As such, Google, OpenAI, Microsoft, Anthropic, Perplexity, xAI, and Deepseek are all competitors in the U.S. Generative AI Market.

279. As of July 2025, OpenAI possessed an 82.65% share of the market, with Perplexity 8.03%, Microsoft 4.59%, Google 2.19%, Deepseek 1.63%, and Anthropic 0.91% of the market.

280. Computing the Herfindahl-Hirschman Index (the "HHI") for the market based on these shares, a method of quantifying market concentration sued by the United States Department of Justice in its Horizontal Merger Guidelines, yields a value of 6916.59. HHIs above 1800 are considered "highly concentrated" by the United States Department of Justice under the Horizontal Merger Guidelines.

281. Another estimate of market share for the Consumer Generative AI/chatbot market, dated August 8, 2025, places ChatGPT in the lead at 60.4% of the market, with Microsoft's Copilot at 14.1%, Google's Gemeni at 13.5%, Perplexity at 6.5%, Claude at 3.5%, Grok at 0.6%, and Deepseek at 0.4%.

282. The HHI given these market shares is 4084.24—also well above the 1,800 threshold for highly concentrated markets.

C. The Relevant Geographic Market

283. The relevant geographic market is the United States.

284. Consumer Generative AI products require low-latency compute and network connections, including for applications such as streaming voice or mission critical agentic applications that require low-latency response times. This requires U.S. Generative AI products to use U.S. based data centers. Data centers in other countries, which will have higher latency network connections, are not reasonable substitutes. Indeed, the use of data centers outside of the United States would result in slow response times when users query Generative AI models. As such, market participants in the United States do not use non-U.S. data centers for their products.

285. Moreover, GPUs used for data centers are export controlled, and U.S. regulatory and legal constraints prevent companies such as Nvidia from exporting GPUs to most foreign countries, including China. The unavailability of GPUs in foreign countries means that products hosted outside of the United States do not compete with, and are not reasonably substitutable with, those hosted and provided in the United States.

286. Data privacy and intellectual property laws also constrain the use of Consumer Generative AI trained in foreign countries in the United States, as certain countries, such as China, have more lax laws and regulations than the United States. Thus, a Consumer Generative AI model trained using the unlicensed intellectual property of others may be lawful in China, but could not be lawfully deployed in the United States without regulatory and litigation risk.

VII. THE MICROSOFT-OPENAI AGREEMENT IS ANTICOMPETITIVE AND CAUSES ANTITRUST INJURY TO PLAINTIFFS

287. The anticompetitive agreement between Microsoft and OpenAI is a direct restriction on the supply and output of Consumer Generative AI products sold in the U.S. Consumer Generative AI Market. Specifically, by restricting the availability of a central component of the Consumer Generative AI product that OpenAI sells, Microsoft directly controls the output of its largest competitor in the Consumer Generative AI Market.

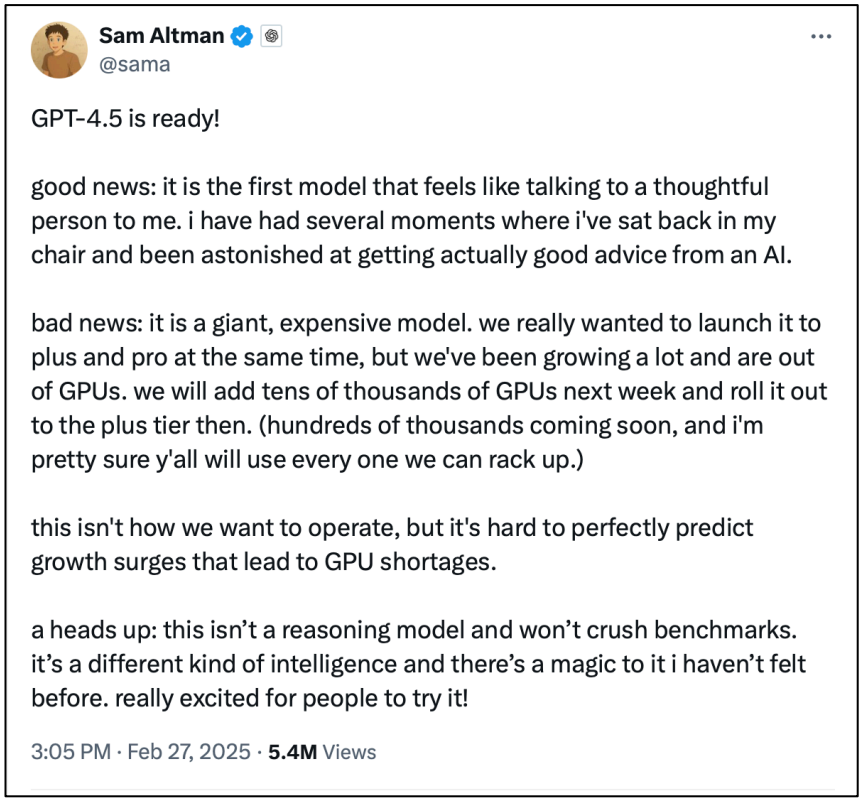
288. Microsoft is not merely an exclusive supplier of compute to OpenAI. It is OpenAI's direct, horizontal competitor. Microsoft's Copilot line of products, among others, compete with OpenAI's ChatGPT line of products. Through its anticompetitive agreement with OpenAI requiring exclusive purchases of compute from Microsoft, Microsoft is able to throttle the output, and thus the price, of OpenAI's products.

289. Indeed, price and output are two sides of the same coin. A restriction of output raises prices, and a restriction of prices coincides with a decrease in output. An output restriction is thus equivalent to a price fixing scheme. *In re Nat'l Football League's Sunday Ticket Antitrust Litig.*, 933 F.3d 1136, 1158 (9th Cir. 2019). An output restriction between or among competing firms is quintessential harm to competition in the relevant market. As the Supreme Court has repeatedly

explained, the anticompetitive character of such an agreement is apparent without the need for elaborate industry analysis.

290. In addition, because Microsoft can limit the amount of compute available to OpenAI, OpenAI is forced to degrade the economic value of the products it sells, but maintain an inflated price in order to pay Microsoft's prices for compute. OpenAI must pay whatever Microsoft asks, as it cannot simply avoid an unreasonable or uncompetitive price from Microsoft by buying compute from a competitor; OpenAI must obtain Microsoft's permission to purchase compute from another source. All of this means that OpenAI was forced to provide less and less to its customers for the same subscription price.

291. For example, in February 2025, OpenAI was forced to delay the release of version 4.5 of its ChatGPT model due to inadequate compute. When it ultimately rolled out the new model, it limited its availability to "Pro" subscribers paying $200 per month. Sam Altman blamed the lack of GPUs available to deploy the model.



292. Absent its anticompetitive agreement with Microsoft, OpenAI could have acquired the necessary GPU capacity from a different cloud provider. Because of Microsoft, however, it was forced

to delay the release of its cutting-edge model, and even upon release, it limited the availability of the model to the highest-paying subscription tier.

293. In March 2025, OpenAI again had to limit its output due to its anticompetitive agreement with Microsoft. Specifically, OpenAI was forced to limit the number of images that users could generate using OpenAI's ChatGPT product. As *PC Magazine* reported in an article titled "Recent OpenAI Launches Limited by GPU Shortages: Poor Planning or Wider Issue?":

> This week, the same rigamarole happened with the new image generator. The company shouted the release from the rooftops and created hype on social media, largely through Studio Ghibli-inspired images that raised alarms about copyright infringement. Following high demand, Altman announced a limit on how many images users could create, again citing overworked GPUs.

294. Altman again blamed compute capacity, which, at the time, was contractually limited by Microsoft:

295. This was not the first time OpenAI was forced to delay or throttle features due to the massive constraint on compute placed upon it by the anticompetitive agreement with Microsoft. As

explained above, OpenAI had throttled or limited other features, such as voice chat, due to the supply and output restriction imposed by Microsoft.

296. All of this meant that while consumers paid the same subscription price, they received less and less as a result of the anticompetitive agreement. Put simply, as a result of its anticompetitive agreement with Microsoft, OpenAI sold its subscriptions at inflated prices until June 2025. The price had been inflated because the economic value of the subscriptions was lower than the price charged.

297. All of this immediately reversed after June 2025, when Altman announced that OpenAI would be purchasing additional compute capacity from rival Google. In the next several months, OpenAI's output of new products skyrocketed. For example, OpenAI announced ChatGPT 5 and deployed the new model to subscribers at scale. OpenAI announced its new Codex model, a refined version of ChatGPT 5 for coding, as well as a command line agentic coding product to rival Claude Code.

298. In short, after June 2025, when OpenAI was permitted to buy compute from its rival Google, OpenAI immediately began providing increased economic value in the form of new models and new features. This immediate-significant increase in welfare provided to ChatGPT subscribers was a powerful real-world demonstration of the consumer-facing harm that had been caused by the Microsoft agreement's compute restraint.

299. OpenAI secured additional compute capacity from Oracle in September 2025. Since then it has continued to announce new products and features—finally at a competitive pace. For example, OpenAI announced Sora 2, a new version of its video generation model on September 30, 2025—a compute intensive model that it could not have launched if constrained to the compute capacity Microsoft deigned to provide under the companies' anticompetitive agreement.

300. Likewise, after June 2025 when additional compute was secured from Google, the API price for OpenAI models, which is based on the same compute cost input as the the ChatGPT consumer product, plummeted approximately 80%. This is direct evidence that both API and subscription prices for OpenAI's generative AI products had been inflated for years—and as a direct and proximate result of the anticompetitive agreement with Microsoft.

301. The anticompetitive agreement between Microsoft and OpenAI harmed competition by inflating prices, restricting output, and setting supply levels of the final product—the Consumer

Generative AI product ChatGPT. By manipulating the compute available to OpenAI, Microsoft directly constrained OpenAI's generative AI output, and thus fixed the price of its Consumer Generative AI products. As a result of the agreement and Microsoft's conduct, OpenAI's customers, including ChatGPT subscribers, paid inflated prices—*i.e.*, supracompetitive prices that well exceeded the economic value of OpenAI's generative AI products and were elevated above competitive price levels for the level of economic value actually delivered. During the period in which the Microsoft compute restraint operated to inflate OpenAI generative AI prices, ChatGPT customers received less and less for the same monthly price—and less than they would have in a competitive market without the anticompetitive compute restriction.

302. The true and competitive price for OpenAI's generative AI products, including its flagship CGAI product ChatGPT, would have been lower but-for the anticompetitive agreement with Microsoft, or, as the flip side of the same economic coin, the economic value of the product consumers purchased would have been far greater. That is, in a but-for world without the anticompetitive compute restraint agreed to between Microsoft and OpenAI, ChatGPT subscribers would have either paid less (holding quantity/economic value fixed) or received more (holding monthly price fixed). Either way, Plaintiffs and members of the classes were overcharged for their ChatGPT subscriptions and thus suffered antitrust injury.

303. There is no legitimate business justification for Microsoft's anticompetitive agreement to force OpenAI to purchase compute exclusively from Microsoft, but for the restraint's anticompetitive effect of allowing Microsoft to control its direct competitors' supply and output. Indeed, Microsoft had a contractual share of OpenAI's revenue and profits going forward. It would have been in the companies' best long-term interest if OpenAI was able to meet market demand with new models and features. But Microsoft forewent long-term profits from its joint venture in order to ensure that its competitor, OpenAI, would not surpass its own competing product, Copilot—while still reaping short-term revenues from overcharged OpenAI customers.

304. This agreement among horizontal competitors to restrict supply and output is *per se* unlawful under the antitrust laws. OpenAI and Microsoft are not permitted to agree upon the output of

products by OpenAI—they each directly compete with each other in the same product market, the Consumer Generative AI Market.

305. Moreover, there is no procompetitive justification for the exclusive compute requirement in the anticompetitive agreement. As such, the agreement fails under the rule of reason. Moreover, even if there were any procompetitive effects, they would be substantially outweighed by the anticompetitive harm that resulted—namely, the restriction of supply, the inflation of price, and the direct constraint on output.

CLASS ACTION ALLEGATIONS

306. Plaintiffs bring this action and seek to certify and maintain it as a class action under Rules 23(a), (b)(2), (b)(3), and/or (c)(4) of the Federal Rules of Civil Procedure, on behalf of themselves and on behalf of the proposed classes of persons (collectively, the "Classes") defined below.

307. Each class's claims derive directly from a course of conduct by Microsoft and its agreement with OpenAI.

308. Defendant has engaged in uniform and standardized conduct toward each class. Defendant did not materially differentiate in its actions or inactions toward members of the respective Classes. For each class, the objective facts on these subjects are the same for all class members.

309. Within each Claim for Relief asserted by each class, the same legal standards govern. Accordingly, Plaintiffs bring this lawsuit as a class action on their own behalf and on behalf of all other persons similarly situated as members of the proposed classes pursuant to Fed. R. Civ. P. 23.

310. Additionally, many states, and for some claims all states, share the same legal standards and elements of proof, allowing for a multistate or nationwide class or classes for some or all claims.

311. This action may be brought and properly maintained as a class action because the questions it presents are of a common or general interest, and of many persons, and also because the parties are numerous, and it is impracticable to bring them all before the court. Plaintiffs may sue for the benefit of all as representative parties pursuant to Federal Rule of Civil Procedure 23.

The Nationwide Damages Class

312. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang bring this action and seek to certify and maintain it as a class action on behalf

of themselves and all affected ChatGPT subscription purchasers. The Nationwide Damages Class comprises:

> All United States persons, business associations, entities, or corporations that purchased subscriptions to ChatGPT from November 30, 2022 to February 1, 2025, inclusive (the "Class Period").

313. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang, and members of the Nationwide Damages Class assert nationwide claims based on the uniform application of federal and California law against Microsoft.

314. Excluded from the Nationwide Damages Class are Defendant's employees, officers, directors, legal representatives, heirs, successors, and wholly or partly owned subsidiaries or affiliates; and the judicial officers and their immediate family members and associated court staff assigned to this case.

The UCL Restitution Class

315. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang bring this action and seek to certify and maintain it as a class action on behalf of themselves and all affected ChatGPT subscription purchasers. The UCL Restitution Class comprises:

> All United States persons, business associations, entities, or corporations that purchased subscriptions to ChatGPT from November 30, 2022 to June 10, 2025, inclusive (the "UCL Class Period").

316. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang, and members of the UCL Restitution Class assert nationwide claims based on the uniform application of federal and California law against Microsoft.

Excluded from the UCL Restitution Class are Defendant's employees, officers, directors, legal representatives, heirs, successors, and wholly or partly owned subsidiaries or affiliates; and the judicial officers and their immediate family members and associated court staff assigned to this case.

The Nationwide Injunction Class

317. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang bring this action and seek to certify and maintain it as a class action on behalf

of themselves and all affected ChatGPT subscription purchasers. The Nationwide Injunction Class comprises:

> All United States persons, business associations, entities, or corporations that purchased subscriptions to ChatGPT from November 30, 2022, to the present, inclusive (the "Nationwide Injunction Class Period").

318. Plaintiffs Bryant, Cavalier, Donovan, Eidson, Halloran, Marsolek, Ouijdani, Payne, Wolfson, Zajac, and Zhang, and members of the Nationwide Injunction Class assert nationwide claims based on the uniform application of federal and California law against Microsoft.

Excluded from the Nationwide Injunction Class are Defendant's employees, officers, directors, legal representatives, heirs, successors, and wholly or partly owned subsidiaries or affiliates; and the judicial officers and their immediate family members and associated court staff assigned to this case.

Numerosity

319. This action satisfies the requirements of Fed. R. Civ. P. 23(a)(1).

320. The members of the Classes are so numerous that a joinder of all members would be impracticable. Millions of subscribers overpaid for ChatGPT subscriptions during the Class Period.

Ascertainability

321. The Classes are ascertainable.

322. The defined Classes consist of persons, business associations, entities, or corporations that purchased ChatGPT subscriptions. The identity of these purchasers can be determined through records maintained by OpenAI.

323. This information can be used to provide members of each class with direct notice pursuant to the requirements of Rule 23 and the Due Process Clause of the United States Constitution.

Typicality

324. Plaintiffs' claims are typical of the members of the Classes.

325. Plaintiffs' claims are the same as those asserted by members of the Classes. Each Plaintiff, like the members of the Classes, has purchased a paid ChatGPT subscription product and has been injured similarly by Defendant's conduct.

326. Each Plaintiff alleges injury that is not unique to them, but is typical of members of each of the Classes, including measures of damages, such as their overpayment for ChatGPT subscription products caused by Microsoft's anticompetitive agreement with OpenAI.

327. Each Plaintiff alleges that their injury flows from the common course of conduct alleged as to Microsoft—namely, the anticompetitive agreement with OpenAI alleged above.

328. Each Plaintiff is similarly positioned as to each member of the Classes. As such, each Plaintiff's injury can be redressed in the same manner as any redress provided to the members of the Classes (and *vice versa*).

Adequate Representation

329. Plaintiffs and their counsel will fairly and adequately protect the interests of the class members.

330. Plaintiffs are committed to putting the interest of the Classes ahead of their own and to act in the best interest of members of the Classes.

331. Plaintiffs understand their obligations to the Classes and are committed to monitoring/supervising developments in the case and class counsel.

332. Plaintiffs have retained competent counsel experienced in computer science, artificial intelligence, antitrust law, and consumer class actions.

333. Plaintiffs have retained counsel with the resources and capital to litigate the case on behalf of the Classes.

334. Plaintiffs and their counsel intend to prosecute this action vigorously and to obtain relief, including both injunctive and monetary relief, that will remedy the market distortions, including as to price and output, caused by Defendant's anticompetitive agreement.

Superiority

335. This action satisfies the requirements of Fed. R. Civ. P. 23(b)(2) because Defendant has acted and/or refused to act on grounds generally applicable to the Classes, thereby making final injunctive and/or corresponding declaratory relief appropriate with respect to each class as a whole.

336. The class device is superior to all other available methods of adjudication, as it would make little sense for each of the millions of class members to separately prove the common conduct in which Defendant has engaged.

337. Moreover, damages suffered by each individual member of the Classes may be small, meaning that the expense or burden of individual litigation would make it very difficult or impossible for individual class members to redress their injury individually.

338. Because individual damages may be relatively small, individual members of the Classes may not have a rational economic interest in individually controlling the prosecution of a single action, and the burden imposed on the judicial system from having to individually adjudicate such claims will be significant in comparison to the value of individual claims.

339. Class litigation is thus superior to individual litigation and is the best procedural device to vindicate the rights of the members of the Classes.

340. In addition, class litigation will streamline the management of the litigation, such that the expense, burdens, inconsistencies, economic infeasibility, and other negative effects of individual mitigation will be lessened if not eliminated.

341. In sum, class litigation is superior because it mitigates significant inefficiencies and barriers that would result from individual litigation. In fact, absent invocation of the class device, the Classes' claims would likely not be vindicated individually, and Defendant's anticompetitive harm to the market and consumers will go unaddressed. Indeed, the anticompetitive agreement between Microsoft and OpenAI will continue absent injunctive relief eliminating and/or mitigating the alleged anticompetitive effects of the agreement.

Commonality and Predominance

342. This action and the claims asserted by the classes satisfy the requirements of Fed. R. Civ. P. 23(a)(2) and 23(b)(3) because there are many questions of law and fact that are common as to all of the members of the Classes.

343. These questions of fact and law concern Defendant's conduct, which is common as to the members of the Classes, and answers to those questions would provide answers to issues posed by claims asserted by all members of the Classes.

344. These common issues will predominate at trial, and any individual issues that may arise would not outweigh the predominance of common issues.

345. Common issues that will predominate at trial include, without limitation, the following:

a. Whether Microsoft and OpenAI entered into a contract or conspiracy in restraint of trade;

b. Whether Microsoft and OpenAI's agreement is per se anticompetitive and unlawful, or in the alternative, whether it violates the rule of reason because the agreement lacks pro-competitive benefits or the anticompetitive effects of the agreement outweigh its pro-competitive benefits;

c. Whether the members of the classes are entitled to trebled (or other relevant measure of) damages, attorneys' fees, costs, and other monetary relief under the antitrust and unfair competition laws, including the federal antitrust laws and the law of California.

d. Whether members of the classes are entitled to injunctive relief preventing Microsoft from enforcing and/or continuing its anticompetitive agreement with OpenAI, including the provisions requiring OpenAI to purchase compute exclusively from Microsoft or other Microsoft-permitted party;

e. Whether members of the classes are entitled to injunctive relief requiring the segregation or divestiture of Microsoft of its Consumer Generative AI assets and lines of business;

f. Whether Defendant unlawfully and anticompetitively reinforced or strengthened the barriers to entry surrounding the Consumer Generative AI Market as a result of their anticompetitive agreement;

g. Whether Microsoft's anticompetitive agreement with OpenAI restricted supply and output, raised prices, impaired or eliminated consumer choice, and/or otherwise harmed competition in the Consumer Generative AI Market.

Grounds Generally Applicable to the Classes

346. Plaintiffs intend to seek injunctive relief ending Microsoft's anticompetitive agreement with OpenAI and/or mitigating or eliminating the anticompetitive effects of the agreement.

347. Plaintiffs are properly situated to seek such an injunction because Microsoft has acted and/or refused to act on grounds generally applicable to Plaintiffs and the members of the Classes.

348. This means that final injunctive relief or declaratory relief will redress Plaintiffs' harm as well as the harm to members of the Classes.

349. An injunction preventing Microsoft from continuing its anticompetitive agreement with OpenAI in the future. In the alternative, an injunction may reduce, mitigate, or eliminate the anticompetitive effects of the alleged agreement.

CHOICE OF LAW

350. Plaintiffs aver that as to state-based claims, California law applies to their claims, and Plaintiffs accordingly assert their claims on behalf of a nationwide class.

351. There are no conflicts between California law and those of the several States as to the state-based nationwide claims asserted in this Complaint.

CLAIMS FOR RELIEF

REALLEGATION AND INCORPORATION BY REFERENCE

352. Plaintiffs reallege and incorporate by reference all the preceding paragraphs and allegations of this Complaint, as though fully set forth in each of the following Claims for Relief asserted on behalf of the classes.

COUNT ONE
Section 1 of the Sherman Antitrust Act
15 U.S.C. § 1
(On behalf of the Nationwide Damages and Injunction Classes)

353. Plaintiffs incorporate by reference all preceding and succeeding allegations as though fully set forth in this Count.

354. Plaintiffs bring claims under Section 1 of the Sherman Act on behalf of themselves and the Nationwide Damages and Injunction Classes, as defined above.

355. Beginning no later than November 30, 2022, Microsoft restricted the supply and output of Consumer Generative AI products sold by OpenAI through an express agreement that required OpenAI to buy the compute that is a necessary part of these products exclusively from Microsoft (the "Compute

Restraint"). This agreement provided Microsoft the power to directly restrict the supply and output of Consumer Generative AI products sold by OpenAI, which was the market leader in the U.S. Computer Generative AI Market.

356. The Compute Restraint provided Microsoft with market power from November 30, 2022 until February 1, 2025 (the "Pre-entry Period"), when DeepSeek AI successfully entered the market and triggered a price war. During this period, OpenAI and Microsoft's competing businesses were protected by the powerful AICBE, which made entry at scale cost prohibitive for all but the most well-resourced companies—specifically, those with cloud-based AI compute at their disposal.

357. During the pre-entry period, OpenAI generative AI customers paid supracompetitive prices due to Compute Restraint, and Microsoft received monies from thewse supracompetitive prices through its revenue sharing agreement with OpenAI. Microsoft caused these prices—including the price of ChatGPT subscriptions—to be supracompetitively inflated by restricting the availability of compute to OpenAI while maintaining the Compute Restraint. This directly inflated prices in the Consumer Generative AI Market, where OpenAI was a market leader and had a first-mover advantage.

358. Microsoft was able to force a price floor on OpenAI's market-leading CGAI products even after the date of DeekSeek's entry as a direct result of the Compute Restraint. By continuing to leverage the Compute Restraint even after DeepSeek AI's entry, Microsoft extracted supracompetitive prices from OpenAI's customers, which were as high as 136-200x the market price. Microsoft was able to do this due to switching costs, network effects, and various barriers to entry, described above, including the AICBE in its diminished post-entry form.

359. Microsoft's Compute Restraint is horizontal. As explained above, OpenAI and Microsoft directly competed in the Consumer Generative AI Market since the inception of the Compute Restraint. Indeed, Microsoft had an exclusive right to commercialize OpenAI's technology as part of its own business. The companies are widely regarded by the press, trade press, public, and even the companies themselves as direct, horizontal competitors in the Consumer Generative AI Market.

360. The compute used to fulfill requests made to OpenAI's Generative AI models is part-in-parcel with the Consumer Generative AI product itself. A Generative AI model is nothing more than a series of floating-point weights. It functions only when the necessary computation is applied to query the

model and process the results. The Compute Restraint thus directly constrained supply and output in the Consumer Generative AI Market. Indeed, that was the purpose and effect of the Compute Restraint.

361. OpenAI and Microsoft's Compute Restraint directly concerns—and indeed governs—sales in interstate commerce by the companies. Moreover, queries to Consumer Generative AI models are made through the Internet using the TCP/IP protocol and higher-level Internet-based client-server systems, such as web servers, REST APIs, and other technology designed to move information from one system to another across the United States through the instrumentalities of interstate commerce. Put simply, virtually every Consumer Generative AI product used, accessed, and sold in the United States was in, and related to, interstate commerce.

362. ***Per se unlawful.*** Because the Compute Restraint was between two firms that compete directly in the Consumer Generative AI Market, it is a horizontal agreement among direct competitors.

363. A horizontal agreement, such as the Compute Restraint, that constrains the supply and output of a relevant product is *per se* unlawful under Section 1 of the Sherman Act.

364. Moreover, because the Compute Restraint directly limits supply and output in the Consumer Generative AI Market, it is also a fixing of the price of ChatGPT, the Consumer Generative AI product sold by OpenAI, as supply/output and price are two sides of the same coin.

365. Because the Compute Restraint is per se unlawful, no elaborate market analysis is required to condemn it as unlawful under Section 1 of the Sherman Act. As such, the Compute Restraint is unlawful, invalid, and anticompetitive, and the Court should declare it as such and enjoin it from further enforcement.

366. ***Rule of Reason.*** The Compute Restraint violates the Rule of Reason and is therefore anticompetitive, invalid, and unlawful. As explained above, the Compute Restraint directly resulted in a restriction of supply and output in the Consumer Generative AI Market and also supracompetitive prices, which reached 136-200x the competitive price, as revealed by market prices after the February 2025 entry of DeepSeek AI.

367. But for the Compute Restraint, OpenAI would have charged lower prices for its Consumer Generative AI products, and would have provided additional features, more sophisticated models, and better service (*e.g.,* decreased wait times and faster responses), to ChatGPT subscribers at the monthly

price these subscribers paid. This is confirmed by the 80% price drop in the high-end OpenAI model that occurred the very same day Microsoft relaxed the Compute Constraint and allowed OpenAI to purchase the compute that was part of its product from Google.

368. The simultaneous price drop and relaxing of the Compute Restraint makes clear that the Compute Restraint was the but-for and proximate cause of the price inflation/supply and output restriction that prevailed during the Class Period, including the effects that persisted after the February 2025 entry of DeepSeek AI until June 10, 2025, when Microsoft finally relented on the Compute Restraint as to OpenAI purchases from Google.

369. The Compute Restraint had no legitimate business or technical justification. Because Microsoft obtained a significant portion of OpenAI's revenues, it should have maximized the supply and output of OpenAI's products, but because of its Compute Restraint, it instead had the motive to restrict supply and output, which allowed it to extract supracompetitive prices from its direct horizontal competitor. In other words, there is no legitimate business or economic reason to prevent OpenAI from buying as much compute as it needed to grow its business and sell its products but for the anticompetitive effects of the Compute Restraint.

370. The Compute Restraint restricted consumer choice in the Consumer Generative AI Market. But for the Compute Restraint, OpenAI could develop and deploy additional models, additional features, and further differentiate its product.

371. These are injuries that the antitrust laws were designed to protect against. Plaintiffs and the Classes thus suffered direct antitrust injury.

372. Because OpenAI was the market leader in the Consumer Generative AI Market, with a first-mover advantage, it, along with its partner Microsoft, maintained market power in the relevant market, meaning OpenAI could raise prices during the Class Period without experiencing a material loss in market share. Microsoft leveraged this market power to extract supracompetitive prices from OpenAI's customers and to impose a price floor in the Consumer Generative AI Market. In other words, the Compute Restraint provided Microsoft with market power to set its chief competitor's prices—and prices throughout the Consumer Generative AI Market—during the Class Period.

373. Moreover, when any procompetitive effects, if any exist at all, are considered, the anticompetitive effects, including the price inflation and output/supply restriction that resulted, substantially outweigh such procompetitive effects.

* * *

374. Although, as explained above, Microsoft relaxed the Compute Restraint as of June 10, 2025 to allow OpenAI to purchase of compute from OpenAI's direct horizontal competitor, Google, the contractual provision embodying the Compute Restraint remains live between the parties. There is, therefore, the threat of future enforcement of the Compute Restraint, including during times of shortage. Absent an injunction and a declaration as to the invalidity of the Compute Restraint, Microsoft can resume restricting the output of its rival, and therby extracting supracompetitive prices from OpenAI's customers, in the future if market conditions change or barriers to entry strengthen (*e.g.*, a shortage of Nvida-made GPUs occurs).

375. Plaintiffs and the Classes are also entitled to trebled damages for the inflated prices they paid as a direct result of the Compute Restraint. Indeed, the supracompetitive prices did not result in a profit for OpenAI, but directly enriched Microsoft during the Class Period because of its revenue sharing agreement with OpenAI.

376. By paying inflated prices for OpenAI products in the Consumer Generative AI Market—specifically, ChatGPT subscriptions—Plaintiffs and class members were injured in their business or property.

377. Statutory trebled damages are required to compensate the Nationwide Damages Class, disgorge Microsoft's ill-gotten profits, deter Microsoft from such conduct in the future, and punish Microsoft for its direct restraint on competition in the Consumer Generative AI Market. Indeed, Microsoft is a recidivist violator of the antitrust laws in the U.S. and internationally.

378. Plaintiffs and the Classes also seek to recover their costs of suit, including attorney fees.

COUNT TWO
California Cartwright Act—Cal. Bus. & Prof. Code § 16700, *et seq.*
(On behalf of the Nationwide Damages and Injunction Classes)

379. Plaintiffs incorporate by reference all preceding and succeeding allegations as though fully set forth in this Count.

380. Plaintiffs bring this claim on behalf of themselves and the Nationwide Damages and Injunction Classes, as defined above.

381. The California Business & Professions Code generally governs conduct of corporate entities. The Cartwright Act, Cal. Bus. & Prof. Code §§ 16700-16770, governs antitrust violations in California.

382. California policy is that "vigorous representation and protection of consumer interests are essential to the fair and efficient functioning of a free enterprise market economy," including by fostering competition in the marketplace. Cal. Bus. & Prof. Code § 301.

383. Under the Cartwright Act, Plaintiffs and the classes have standing to maintain an action based on the facts alleged in this Complaint. Cal. Bus. Prof. Code § 16750(a).

384. A trust in California is any combination of capital, skills or acts by two or more persons intended for various purposes, including but not limited to creating or carrying out restrictions in trade or commerce, limiting or reducing the production or increasing the price of merchandise, preventing competition in the market for merchandise, or fixing prices for any merchandise. Cal. Bus. & Prof. Code § 16720. Every trust in California is unlawful except as provided by the Code. *Id.* § 16726.

385. Microsoft entered into a contract, combination, or conspiracy between two or more persons in restraint of, or to monopolize, trade of commerce in the Consumer Generative AI Market, a substantial part of which occurred within California. In addition, by virtue of their conspiracy/joint venture/unlawful agreement, Microsoft and OpenAI are a "trust" within the meaning of the Cartwright Act.

386. Beginning no later than November 30, 2022, Microsoft restricted the supply and output of Consumer Generative AI products sold by OpenAI through an express agreement that required OpenAI to buy the compute that is a necessary part of these products exclusively from Microsoft (the "Compute Restraint"). This agreement provided Microsoft the power to directly restrict the supply and output of

Consumer Generative AI products sold by OpenAI, which was the market leader in the U.S. Computer Generative AI Market.

387. The Compute Restraint provided Microsoft with market power from November 30, 2022 until February 1, 2025 (the "Pre-entry Period"), when DeepSeek AI successfully entered the market and triggered a price war. During this period, OpenAI and Microsoft's competing businesses were protected by the powerful AICBE, which made entry at scale cost prohibitive for all but the most well-resourced companies—specifically, those with cloud-based AI compute at their disposal.

388. During the pre-entry period, OpenAI generative AI customers paid supracompetitive prices due to Compute Restraint, and Microsoft received monies from these supracompetitive prices through its revenue sharing agreement with OpenAI. Microsoft caused these prices—including the price of ChatGPT subscriptions—to be supracompetitively inflated by restricting the availability of compute to OpenAI while maintaining the Compute Restraint. This directly inflated prices in the Consumer Generative AI Market, where OpenAI was a market leader and had a first-mover advantage.

389. Microsoft was able to force a price floor on OpenAI's market-leading CGAI products even after the date of DeekSeek's entry as a direct result of the Compute Restraint. By continuing to leverage the Compute Restraint even after DeepSeek AI's entry, Microsoft extracted supracompetitive prices from OpenAI's customers, which were as high as 136-200x the market price. Microsoft was able to do this due to switching costs, network effects, and various barriers to entry, described above, including the AICBE in its diminished post-entry form.

390. Microsoft's Compute Restraint is horizontal. As explained above, OpenAI and Microsoft directly competed in the Consumer Generative AI Market since the inception of the Compute Restraint. Indeed, Microsoft had an exclusive right to commercialize OpenAI's technology as part of its own business. The companies are widely regarded by the press, trade press, public, and even the companies themselves as direct, horizontal competitors in the Consumer Generative AI Market.

391. The compute used to fulfill requests made to OpenAI's Generative AI models is part-in-parcel with the Consumer Generative AI product itself. A Generative AI model is nothing more than a series of floating-point weights. It functions only when the necessary computation is applied to query the

model and process the results. The Compute Restraint thus directly constrained supply and output in the Consumer Generative AI Market. Indeed, that was the purpose and effect of the Compute Restraint.

392. ***Per se unlawful.*** Because the Compute Restraint was between two firms that compete directly in the Consumer Generative AI Market, it is a horizontal agreement among direct competitors.

393. A horizontal agreement, such as the Compute Restraint, that constrains the supply and output of a relevant product is *per se* unlawful under the Cartwright Act.

394. Moreover, because the Compute Restraint directly limits supply and output in the Consumer Generative AI Market, it is also a fixing of the price of ChatGPT, the Consumer Generative AI product sold by OpenAI, as supply/output and price are two sides of the same coin.

395. Because the Compute Restraint is per se unlawful, no elaborate market analysis is required to condemn it as unlawful under the Cartwright Act. As such, the Compute Restraint is unlawful, invalid, and anticompetitive, and the Court should declare it as such and enjoin it from further enforcement.

396. ***Rule of Reason.*** The Compute Restraint violates the Rule of Reason and is therefore anticompetitive, invalid, and unlawful. As explained above, the Compute Restraint directly resulted in a restriction of supply and output in the Consumer Generative AI Market and also supracompetitive prices, which reached 136-200x the competitive price, as revealed by market prices after the February 2025 entry of DeepSeek AI.

397. But for the Compute Restraint, OpenAI would have charged lower prices for its Consumer Generative AI products, and would have provided additional features, more sophisticated models, and better service (*e.g.,* decreased wait times and faster responses), to ChatGPT subscribers at the monthly price these subscribers paid. This is confirmed by the 80% price drop in the high-end OpenAI model that occurred the very same day Microsoft relaxed the Compute Constraint and allowed OpenAI to purchase the compute that was part of its product from Google.

398. The simultaneous price drop and relaxing of the Compute Restraint makes clear that the Compute Restraint was the but-for and proximate cause of the price inflation/supply and output restriction that prevailed during the Class Period, including the effects that persisted after the February 2025 entry of DeepSeek AI until June 10, 2025, when Microsoft finally relented on the Compute Restraint as to OpenAI purchases from Google.

399. The Compute Restraint had no legitimate business or technical justification. Because Microsoft obtained a significant portion of OpenAI's revenues, it should have maximized the supply and output of OpenAI's products, but because of its Compute Restraint, it instead had the motive to restrict supply and output, which allowed it to extract supracompetitive prices from its direct horizontal competitor. In other words, there is no legitimate business or economic reason to prevent OpenAI from buying as much compute as it needed to grow its business and sell its products but for the anticompetitive effects of the Compute Restraint.

400. The Compute Restraint restricted consumer choice in the Consumer Generative AI Market. But for the Compute Restraint, OpenAI could develop and deploy additional models, additional features, and further differentiate its product.

401. These are injuries that the California antitrust laws were designed to protect against. Plaintiffs and the Classes thus suffered direct antitrust injury.

402. Because OpenAI was the market leader in the Consumer Generative AI Market, with a first-mover advantage, it, along with its partner Microsoft, maintained market power in the relevant market, meaning OpenAI could raise prices during the Class Period without experiencing a material loss in market share. Microsoft leveraged this market power to extract supracompetitive prices from OpenAI's customers and to impose a price floor in the Consumer Generative AI Market. In other words, the Compute Restraint provided Microsoft with market power to set its chief competitor's prices—and prices throughout the Consumer Generative AI Market—during the Class Period.

403. Microsoft leveraged this market power to extract supracompetitive prices from OpenAI's customers and to impose a price floor in the Consumer Generative AI Market. In other words, the Compute Restraint provided Microsoft with market power to set its chief competitor's prices—and prices throughout the Consumer Generative AI Market—during the Class Period.

404. Moreover, when any procompetitive effects, if any exist at all, are considered, the anticompetitive effects, including the price inflation and output/supply restriction that resulted, substantially outweigh such procompetitive effects.

* * *

405. Although, as explained above, Microsoft relaxed the Compute Restraint as of June 10, 2025 to allow OpenAI to purchase of compute from OpenAI's direct horizontal competitor, Google, the contractual provision embodying the Compute Restraint remains live between the parties. There is, therefore, the threat of future enforcement of the Compute Restraint, including during times of shortage. Absent an injunction and a declaration as to the invalidity of the Compute Restraint, Microsoft can resume restricting the output of its rival, and thereby extracting supracompetitive prices from OpenAI's customers, in the future if market conditions change or barriers to entry strengthen (*e.g.*, a shortage of Nvida-made GPUs occurs).

406. Plaintiffs and the Classes are also entitled to trebled damages for the inflated prices they paid as a direct result of the Compute Restraint. Indeed, the supracompetitive prices did not result in a profit for OpenAI, but directly enriched Microsoft during the Class Period because of its revenue sharing agreement with OpenAI.

407. By paying inflated prices for OpenAI products in the Consumer Generative AI Market—specifically, ChatGPT subscriptions—Plaintiffs and class members were injured in their business or property.

408. Statutory trebled damages are required to compensate the Nationwide Damages Class, disgorge Microsoft's ill-gotten profits, deter Microsoft from such conduct in the future, and punish Microsoft for its direct restraint on competition in the Consumer Generative AI Market. Indeed, Microsoft is a recidivist violator of the antitrust laws in the U.S. and internationally.

409. Plaintiffs and the Classes also seek to recover their costs of suit, including attorney fees.

COUNT THREE
Violation of the California Unfair Competition Law
Cal. Bus. & Prof. Code § 17200, *et seq.*
(On behalf of the UCL Restitution Class and Injunction Class)

410. Plaintiffs incorporate by reference all preceding and succeeding allegations as though fully set forth in this Count.

411. Plaintiffs bring unlawful and unfair prongs on this Count on behalf of themselves and the UCL Restitution Class.

412. California's Unfair Competition Law ("UCL"), Cal. Bus. & Prof. Code §§ 17200-17210 proscribes acts of unfair competition, including "any unlawful, unfair or fraudulent business act or practice and unfair, deceptive, untrue or misleading advertising." Microsoft has engaged in unfair or deceptive acts or practices that violated the UCL, as described above and below, by, among other things, entering into an unlawful agreement with its horizontal competitor, which required OpenAI to buy the computation that is part of its products exclusively from Microsoft. Microsoft has violated the unlawful and unfair prongs of the UCL, as set forth in this Complaint and below.

413. Microsoft's actions constitute **"unlawful" trade practices** within the meaning of the UCL. In the course of Microsoft's business, Microsoft entered into an express, unlawful, horizontal agreement that required OpenAI to buy the computation that is part of its Consumer Generative AI products exclusively from Microsoft ("the Compute Restraint"). Accordingly, Microsoft engaged in unlawful trade practices by restricting the supply and output of Consumer Generative AI products using the Compute Restraint and exploiting its resulting market power to extract supracompetitive prices from consumers through its revenue sharing agreement with OpenAI. Microsoft's actions are further unlawful because they violated (and violate) other statutes and common law prohibitions, including those recited in the other counts of this Complaint.

414. Microsoft's actions constitute **"unfair" trade practices** within the meaning of the UCL. Microsoft's unlawful acts and practices complained of in this Complaint affect the public interest, as its actions offend public policy and are immoral, unethical, oppressive, unscrupulous, and substantially injurious to consumers.

415. To begin, Microsoft's horizontal agreement and extraction of supracompetitive prices are unfair, unethical, oppressive, and unscrupulous practices.

416. Moreover, under the balancing test, Microsoft's conduct is unfair within the meaning of the UCL because the gravity of the harm inflicted by its conduct is greater than any possible utility:

- It imposed supracompetitive prices, which reached 136-200x the competitive price, on OpenAI's customers and imposed a price floor in the Consumer Generative AI Market. Without the Compute Restraint, Open AI would have also charged significantly lower

prices for its products as can be seen by the 80% drop in high-end OpenAI model prices when Microsoft relaxed the Compute Restraint.

- It resulted in a restriction of the supply and output in the Consumer Generative AI Market which also restricted consumer choice. But for the Compute Restraint, OpenAI could have developed and deployed additional models, features, and further differentiated its product.
- There is no legitimate business or economic reason to prevent OpenAI from buying as much compute as it needed to grow its business and sell its products but for the anticompetitive effects of the Compute Restraint.

417. All of this conduct is unfair under the UCL and proximately caused injury to Plaintiffs and the UCL Restitution Class members.

418. Microsoft's actions as set forth in this Complaint occurred in the conduct of trade or commerce.

419. Microsoft knew or should have known that its conduct violated California law regarding unfair and/or deceptive acts in trade or commerce.

420. As a direct and proximate result of Microsoft's violations of the UCL, Plaintiffs and the UCL Restitution Class members have suffered injury-in-fact and/or actual damage.

421. Microsoft has been unjustly enriched and should be required to make restitution to Plaintiffs and the UCL Restitution Class members under Sections 17203 and 17204 of the California Business & Professions Code. Plaintiffs and the UCL Restitution Class members also seek injunctive relief as deemed appropriate by the Court, including but not limited to a prohibition on enforcing the Compute Restraint.

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs, individually and on behalf of members of the Proposed Classes, respectfully request that the Court enter judgment in their favor and against Defendant, as follows:

A. Certification of the proposed Nationwide Classes or, alternatively, certification of the proposed Classes, including appointment of Plaintiffs as Class Representatives and Plaintiffs' counsel as Class Counsel;

B. Injunctive relief barring and/or invalidating Microsoft's anticompetitive agreement with OpenAI;

C. Injunctive relief mitigating or lessening the anticompetitive effects/aspects of Microsoft's anticompetitive agreement with AI, including, to the extent necessary, by requiring the divestiture or segregation of Microsoft's Generative AI line(s) of business;

D. Restitution, including at the election of all Class Members;

E. Damages, including trebled damages, for injury to Plaintiffs' and Class Members, including damages for the anticompetitive overcharge for their ChatGPT subscription products proximately caused by Microsoft's anticompetitive agreement with OpenAI;

F. Damages (including trebled damages), costs, and disgorgement in an amount to be determined at trial;

G. An order requiring Defendant to pay both pre- and post-judgment interest on any amounts awarded;

H. An award of costs and attorneys' fees; and

I. Such other or further relief as may be appropriate.

JURY DEMAND

Plaintiffs demand a trial by jury on all claims so triable as a matter of right.

Dated: October 13, 2025

Respectfully submitted,

/s/ Brian J. Dunne
Brian J. Dunne (CA 275689)
bdunne@bathaeedunne.com
Edward M. Grauman (*p.h.v. to be sought*)
egrauman@bathaeedunne.com
BATHAEE DUNNE LLP
901 South Mopac Expressway
Barton Oaks Plaza I, Suite 300
Austin, TX 78746
Tel.: (213) 462-2772

Allison Watson (CA 328596)
awatson@bathaeedunne.com
BATHAEE DUNNE LLP
3420 Bristol Street, Suite 600
Costa Mesa, CA 92626
Tel.: (213) 462-2772

/s/ Yavar Bathaee
Yavar Bathaee (CA 282388)
yavar@bathaeedunne.com
Andrew C. Wolinsky (CA 345965)
awolinsky@bathaeedunne.com
BATHAEE DUNNE LLP
445 Park Avenue, 9th Floor
New York, NY 10022
Tel.: (332) 322-8835

Attorneys for Plaintiffs and the Proposed Classes