Harvard Business School

9-725-484
REV: AUGUST 25, 2025

ANDY WU
ANNA YANG

AI Wars in 2025

In June 2025, Google leaders in Mountain View, CA convened after its parent company Alphabet shed a quarter-trillion in market capitalization in a matter of months. The immediate spark—the quiet revelation that Google searches had dipped for the first time in 20 years—masked a deeper shift: hundreds of millions were now turning first to generative artificial intelligence (AI) chatbots like ChatGPT, Perplexity, and upstart DeepSeek, asking questions that once flowed automatically to Google Search. After two decades of dominating search, Google's mission to "organize the world's information" was now in peril. The company now faced a future in which the world's queries might bypass it altogether.

While many perceived Google as needing to play catchup in generative AI, it hardly lacked experience in the field. Google was the long-time putative leader in AI: its scientists were credited with having invented generative AI in the first place, making it especially frustrating that OpenAI had been the one to bring the technology into the public conversation with its launch of ChatGPT in November 2022. Nearly three years later, Google's Gemini model was comparable to OpenAI's GPT, and Google's advertising engine still printed hundreds of billions in cash. However, full-scale incorporation of generative AI into its flagship Search product to transform it into a chatbot would incur new costs and risks, while its own open-source push risked giving rivals a free ride on Google's research. Against this backdrop, Google leadership had to decide whether, how, and when to defend Search. Should Google rush into competing directly with OpenAI and other chatbot providers despite having no working business model for generative AI, or wait and perhaps risk obsolescence?

Even more monumental questions lay further afield. How could Google reshape the generative AI industry to make it more attractive? And most critically, where should Google position itself in this new world? The answers would determine whether the firm that once organized the world's information could still profit from it in the age of generative AI.

Google Search

In 1995, Google co-founders Larry Page and Sergey Brin met at Stanford University and became fast friends as doctoral students. One of their earliest collaborations would become the core technology underlying Google Search. While working on his dissertation, Page realized that a website's credibility

Professor Andy Wu, Research Associate Matt Higgins, HBS Doctoral Student Miaomiao Zhang, and Doctoral Student Hang Jiang (MIT) prepared the original version of this case, "AI Wars," HBS No. 723-434. This version was prepared by Professor Andy Wu and independent researcher Anna Yang. This case was developed from published sources. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

could be proxied by an algorithm, rather than just human curation. By the fall of 1996, he and Brin had built a search engine called BackRub that indexed links and ranked each page's importance based on the number and quality of links to that page. This ranking algorithm was referred to as PageRank.[1]

The engine that Page and Brin built together laid the foundation for Google Search, using a three-stage process:[a] crawling, indexing, and ranking.[2] Their approach was technically advanced for the era, leveraging a type of AI focused on classification and prediction to work faster, at a greater scale, and with more relevant matches than other search engines. Competitors like AltaVista and Yahoo struggled to keep up, relying on less scalable and less accurate approaches like human curation and basic ranking of pages based on keyword frequency.[3] Over time, Google evolved its Search experience beyond just providing a page of links by introducing consumer-friendly features like Shopping carousels, integration with Maps, commonly searched questions, Flights, and other information that would show up depending on the search. Google benefited from the volume of data running through its search engine, learning and building based on patterns of search behaviors—often a manual process that required human intervention and years of cumulative investment in the technology.

Since founding, Google pursued AI research to improve its product portfolio, starting with Search and moving into its broader suite of products. Though the company did not publicize specifics on which products leveraged Google's AI and machine learning (ML) developments (or how), public research papers and blog posts referenced the use of ML techniques in Google Translate and Google Maps, among others.[4] (See **Glossary** for definitions of AI terminology.) Competitors later developed their own search algorithms, but none could rival Google's. As of December 2022, Google handled 61.4% of internet searches in the U.S., followed by Microsoft Bing at 29.6% and Yahoo at 11%.[5] Globally, Google had a 93.4% share of the desktop search market; Bing was a distant second at 2.8%.[6]

Google used several distribution channels to get its search engine into the hands of as many people as possible. It expanded into adjacencies like the mobile OS Android and web browser Chrome, using these platforms to distribute Search. Google also paid Apple an estimated $20 billion a year to be the default search engine on iPhones. (It was telling, perhaps, that Apple chose not to develop its own search engine—even as it developed products like Apple Maps to compete with Google.)[7] This dominance in search would come at a cost: in August 2024, Google lost a massive antitrust case brought by the U.S. Department of Justice in August 2024 that jeopardized its access to these channels.[8]

Google had been a long-time leader in AI research because of its use of AI in its core Search and Advertising businesses. This innovation would continue into the next frontier: *generative* AI.

Google's Early Moves in Generative AI

Much of the AI research that Google conducted resulted in the technical foundations of what would later be referred to as generative AI. Among Google's first AI projects was Google Brain, an exploratory collaboration started in 2011 with Stanford computer science professor Andrew Ng.[9] This work was internally refined until 2015, when Google released TensorFlow, an open-source developer platform that was critical across the industry in the development of deep learning neural networks. TensorFlow became the most widely used AI development tool for years to come.[10]

[a] In the crawling stage, Google sought out new and updated web pages to add to its inventory of known pages (called "URL discovery"). Next, Google processed the contents of the page in a step called indexing, filtering out any pages that were duplicates, had low-value content, or otherwise had issues that prevented Google from indexing. Finally, pages were ranked based on quality and relevancy, which determined what got surfaced to a user.

In January 2014, Google acquired the AI lab DeepMind for over $500 million.[11] The startup operated independently from Google Brain under Alphabet until the two were later combined. Among the major technology players, Google was widely perceived to be the leader in advancing AI technology and attracting the thought leaders who would dominate the conversation in the coming years.

Invention of Generative AI

Two of Google's research papers resulted in landmark contributions to modern generative AI. The first paper, published in 2012 and titled "ImageNet Classification with Deep Convolutional Neural Networks," was written by Google-affiliated researchers Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. This paper demonstrated that convolutional neural networks (CNNs) trained on large datasets could achieve breakthrough performance in image recognition. It reintroduced deep learning to the mainstream AI community, leading to an inflow of talent and funding for further research and the development of more complex models.[12] Hinton was later awarded the 2024 Nobel Prize in Physics, largely for his contributions in machine learning and artificial neural networks.[13]

The second paper, titled "Attention Is All You Need," was published in 2017 by a group of eight scientists working at Google Brain. This research paper introduced the Transformer architecture, a type of neural network that leveraged a concept called *self-attention* to unpack "long-range dependencies" in sentences, i.e., the relationship between words that were far apart in a sentence. The Transformer was a groundbreaking development that would become the backbone for nearly all modern large language models (LLMs): deep learning algorithms that were capable of understanding and generating human language by processing large datasets.[14]

Modern generative AI models, like OpenAI's GPT or Google's Gemini, were trained on data largely sourced from the public internet. As one major source, the nonprofit Common Crawl provided a repository of petabytes of data and metadata from across the internet. Its bots regularly crawled and stored snapshots of websites, including copyrighted content. Originally intended for use by academic researchers, commercial entities would widely use this data for AI training along with data they collected themselves from the internet and elsewhere (see **Exhibit 1** for training data composition).

In AI models, a parameter was an internal variable that the model extracted from its training data and used to make predictions. The more parameters that were used, the "larger" the model was in size—which generally meant a greater volume of accumulated "knowledge" but also higher training costs. Thousands or more expensive specialized chips called graphics processing units (GPUs) were required for this computational complexity because they could carry out many calculations in parallel. GPUs were also electricity-intensive: operating a single industry-standard Nvidia GPU required 700W of power, more than the average American household.[15] Training also required many iterative cycles of experimentation, each of which incurred additional costs. Stanford estimated the training cost for GPT-4 at $78 million and Gemini at $191 million, dramatically higher than the $900 cost for Google's original Transformer model. (**Exhibit 2**[16] shows training cost estimates across multiple AI models.)[17] But training was just one component of the total cost to build and release a new model. OpenAI, for example, reportedly spent over $540 million in 2022 alone to develop ChatGPT.[18]

Generative models that were trained on broad datasets were referred to as "foundation models," which could later be fine-tuned on specialized data for particular tasks.[19] As training datasets increased in scale, these foundation models began to encode complex relationships that were not easily traceable by humans. In this way, generative AI models were considered "unexplainable." The training process relied heavily on researcher intuition and know-how, directed towards iterative experimentation with no guarantee of success. An Anthropic scientist said:

> We built it, we trained it, but we don't know what it's doing…The other big unknown that's connected to this is we don't know how to steer these things or control them in any reliable way…We're really just kind of steering these things almost completely through trial and error.[20]

In 2018, Google published its AI principles—among the first of their kind in this space. The next year, the company released its first annual transparency report where it publicized details on how it managed and mitigated risks associated with AI. "Responsible" development and deployment were core to Google's approach and ultimately embedded in its DNA (**Exhibit 3**).[21]

Competition Emerges

OpenAI

In 2015, OpenAI was founded as a non-profit research institute with an aspiration to advance AI "in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return."[22] Ilya Sutskever, previously a research scientist at Google, co-founded OpenAI as its research director with co-chairs Sam Altman and Elon Musk.

OpenAI's use of public internet data and large-scale human feedback allowed for the development of GPT-3, one of the most powerful large language models (LLMs) when it was released in 2020.[23] The company had a novel approach to its training data, setting up a large-scale team of human workers to annotate and clean data at low cost in Kenya, Uganda, and India.[24] When GPT-3 was opened to select researchers in June 2020, it quickly garnered acclaim for its mainstream usability. That summer was a turning point in the development of language-based artificial intelligence, and the term "generative AI" began to show up regularly in the media.

In November 2022, OpenAI released ChatGPT, a publicly available chatbot interface built on GPT-3.5. (See **Exhibit 4** for an image of the ChatGPT interface.) It immediately took off, reaching 100 million monthly active users just two months post-launch, a feat that led some to label it as the "fastest-growing consumer application in history." (See **Exhibit 5** for average monthly users for top generative AI consumer apps in January 2025.)[25] More milestones would come over the next several years: 100 million weekly active users by November 2023, 300 million by December 2024, and 400 million by February 2025.[26] In April 2025, CEO Altman declared: "Something like 10% of the world uses our systems."[27]

OpenAI continued to advance its consumer-facing products while serving developers via Application Programming Interface (API) services. APIs allowed third-party developers to access the GPT models to build their own applications. Allowing this access came with a risk: as one AI researcher tweeted, "If you allow any sufficiently wide-ranging access to your AI model, even by paid API, you're giving away your business crown jewels to competitors that can then nearly-clone your model without all the hard work you did to build up your own fine-tuning dataset."[28] Notable among these API-accessible developments was the March 2023 release of GPT-4.[29] The number of parameters in this model also increased by several orders of magnitude: 1.76 trillion for GPT-4, compared to 175 billion for GPT-3 and 1.5 billion for GPT-2.[30]

Funding for OpenAI came in large part from Microsoft, which began investing heavily in the company in July 2019 after OpenAI restructured and welcomed "capped-profit" partnerships through its OpenAI LP arm (overseen by OpenAI Nonprofit).[31] The initial investment of $1 billion from Microsoft was followed by a "multiyear, multibillion dollar" extension in 2023, widely reported to be worth $10 billion.[32] Meanwhile, the company shifted away from its earlier "open" strategy of

information sharing and moved to a closed-source development approach[b], allegedly for AI safety.[33] The decisions received internal and external criticism as OpenAI gained mainstream traction. In November 2023, disagreements between Altman and the company's board resulted in Altman's dismissal as CEO. Sources close to the company claimed that there had been disagreements over decisions to rapidly commercialize OpenAI's products. 702 out of 770 OpenAI employees threatened to quit, and Altman was rehired within the week (along with a newly assembled board).[34]

Co-founder Musk criticized OpenAI and took legal action after his 2019 departure. Musk would later start his own AI company xAI. In a March 2023 tweet, he stated: "I'm still confused as to how a non-profit to which I donated ~$100M somehow became a $30B market cap for-profit. If this is legal, why doesn't everyone do it?"[35] Musk condemned OpenAI for its compensation practices, claiming that software engineers at the company earned a median of $810,000 annually, well above the $220,000 average in San Francisco.[36] Microsoft AI researcher Peter Lee compared the cost of hiring a top AI researcher to that of acquiring an NFL quarterback.[37] In April 2025, OpenAI announced a countersuit against Musk, claiming that he used "bad-faith tactics" to slow the company's growth.[38]

Amid both astronomic growth and rising controversy, OpenAI experienced the high-profile departures of several leaders, including Mira Murati (former chief technology officer), Ilya Sutskever (former chief scientist and co-founder), and siblings Dario and Daniela Amodei (former VP of research and former VP of safety and policy). After leaving OpenAI, these executives each launched new AI startups: Thinking Machines Lab, Safe Superintelligence Inc., and Anthropic, respectively.[39] In the state of California, where many of the leading AI players were based, non-compete agreements[c] were not enforceable, and the state had struck down application of the "inevitable disclosure" doctrine[d].[40]

OpenAI's focus expanded beyond ChatGPT in 2025. A joint venture with SoftBank Group planned to sell custom services and support for enterprises adopting AI.[41] And OpenAI announced a $6.5 billion acquisition of hardware startup io, founded by former Apple design chief Jony Ive; details of their pocket-size screen-free wearable device remained confidential, with release expected by late 2026.[42]

Microsoft

Though Microsoft's AI efforts started in the 1990s, the company did not successfully build a sizeable business in the space until its eventual partnership with OpenAI. By then, anxiety had risen within the organization that Microsoft was "multiple years behind the competition in terms of machine learning scale." In a 2019 email to CEO Satya Nadella and co-founder Bill Gates, Microsoft CTO Kevin Scott shared his concerns that Google was rapidly outpacing them and admitted that his prior dismissal of both Google's and OpenAI's capabilities had been hasty.[43]

Later in 2019, Microsoft invested $1 billion (about half of which was in credits for Microsoft's Azure cloud computing services) in OpenAI. As part of the partnership, the companies agreed for OpenAI to use Azure as its exclusive cloud provider, though this agreement was later broken.[44] One analyst noted,

---

[b] As of February 2025, CEO Altman stated in a Reddit "Ask Me Anything" event that OpenAI was reconsidering its open-source strategy but that it was not among the company's top priorities.

[c] The State of California historically maintained policies allowing for employee mobility, innovation, and economic growth. Non-compete agreements were typically void, with narrow exemptions, and employers could even be held liable for including non-compete clauses in employment contracts (according to amendments added in 2024).

[d] The inevitable disclosure doctrine allowed the plaintiff in a trade secrets case to block a defendant's employment by demonstrating that the defendant's new employment would inevitably require them to use (disclose) the former employer's trade secrets. California courts had rejected the doctrine, in line with the state's stance on non-competes.

"Beyond the financial risks and rewards for Microsoft, the bigger prize is that it gets to work alongside OpenAI in developing the technology on Microsoft Cloud, which instantly puts Microsoft at the forefront of what could be the most important consumer technology over the next decade."[45] Nadella urged his teams to integrate generative AI into its products, including the launch of Copilot in the Microsoft 365 productivity suite (see **Exhibit 4**). Per the company's terms of service, a customer's data in these products would not be used to train models.[46] Many of the company's interests in AI would be put in support of growing its core Microsoft 365 and Azure businesses.

The OpenAI partnership did not slow Microsoft's pursuit of in-house research and development. For example, Microsoft released its own generative model, Prometheus, in February 2023 to merge the search and chat functions in Bing.[47] An article in *Wired* noted of the work: "Microsoft has held back from going all-in on OpenAI's technology." Though GPT might have been larger and more advanced, Microsoft's own models were more affordable to operate and consumed less computing power.[48]

Microsoft also kept a close eye on the broader landscape of AI startups, making several investments in early 2024. It participated in a $675 million funding round for AI robotics startup Figure, alongside Nvidia and Jeff Bezos, founder of Amazon. A few days later, it made a $16 million investment into Paris-based Mistral AI, an open-source competitor to OpenAI. And just one month later, Nadella announced that Microsoft would partner with another OpenAI competitor, Inflection AI, licensing its technology and hiring several members of its staff, including co-founders Mustafa Suleyman and Karén Simonyan, both previously from Google DeepMind (where Suleyman was a co-founder).[49]

Perplexity

Perplexity, launched as an AI alternative to classic search engines like Google and Bing, was founded in 2022 by Aravind Srinivas (formerly of OpenAI and Google DeepMind), Denis Yarats (formerly of Facebook AI), Johnny Ho, and Andy Konwinski.[50] Perplexity sought to build an "answer engine" rather than a "search engine" by using LLM technology to provide synthesized answers to a user's question, rather than return a page of links based on a set of input keywords (**Exhibit 4**).[51]

Some users observed that Perplexity's responses tended to mirror the first page of results from Google and Bing, speculating that the platform queried these search engines and leveraged an LLM to summarize and present results.[52] The company acknowledged that it used generative AI models from OpenAI and Anthropic through API access.[53]

Though Perplexity's business model started as a subscription service, the company announced in November 2024 that it would soon introduce advertising to its platform to generate needed revenue, admitting: "Experience has taught us that subscriptions alone do not generate enough revenue to create a sustainable revenue-sharing program…advertising is the best way to ensure a steady and scalable revenue stream."[54] While Perplexity sought to sell text ads for as much as $60 CPM (cost per thousand impressions), above even the cost of a Netflix screen takeover ad,[55] an advertiser stated: "Our hesitation [to invest further in Perplexity ads] stems from a few key concerns: limited scale on the platform, lack of demonstrated ROI, brand safety considerations, and CPM efficiency."[56]

By 2025, Perplexity's aspirations had shifted towards AI-powered web browsers. In May, it began early testing on Comet, an "agentic" web browser that could act independently on a user's behalf (e.g., by posting on social platforms or managing open tabs).[57] This product came at a pivotal moment, just as the U.S. Department of Justice was pushing for Google to divest its leading web browser, Chrome, which was worth $20 billion and owned 65% of the global desktop browser market.[58] In a court hearing, Perplexity executives expressed interest in acquiring Chrome if the opportunity arose.[59]

Meta

For Meta CEO Mark Zuckerberg, the interest in generative AI lay in how it would contribute to Meta's social media and advertising and accelerate his aspirations for wearable devices and the metaverse. Meta led in social media with over 3 billion monthly active users across its platforms. This wealth of user data enabled a massive digital advertising business, second only to Google and accounting for 98% of its revenue.[60] Over the last decade, Zuckerberg sought to position the company as a leader in the future metaverse, where an immersive virtual world of 3D content would be delivered over the internet and through virtual reality. Progress had been slow, in part due to the high cost of 3D content production, but the advancements in generative AI models could unlock his vision.

Since its founding in 2013, Facebook AI Research (FAIR), later Meta AI, had been led by Yann LeCun, an AI researcher and NYU computer science professor who played an important role in advancing generative AI.[61] Meta's early AI work was largely proprietary and used to enhance its core products, like ad targeting for Facebook Ads.[62] However, Meta would soon adjust its stance: while Google was most associated with open-source technology, Meta quickly established a reputation as the new open-source leader[e] in generative AI.

In February 2023, Meta made LLaMA (Large Language Model Meta AI) available by request to AI researchers. The company also provided model details that, at the time, OpenAI and Google had kept proprietary for their own models. After copies of LLaMA were leaked to the public,[63] Meta released LLaMA 2 under an open-source license and made it available for both research and commercial use, with limitations.[f] The openness of LLaMA made it possible for AI researchers outside of the biggest labs to examine and modify it. In Meta's Q1 2023 earnings call, Zuckerberg said:

> Right now most of the companies that are training large language models have business models that lead them to a closed approach to development. I think there's an important opportunity to help create an open ecosystem. If we can help be a part of this, then much of the industry will standardize on using these open tools and help improve them further. So this will make it easier for other companies to integrate with our products and platforms as we enable more integrations, and that will help our products stay at the leading edge as well.

In response to a question about open source, Zuckerberg elaborated:

> Unlike some of the other companies in the space, we're not selling a cloud computing service where we try to keep the different software infrastructure that we're building proprietary. For us, it's way better if the industry standardizes on the basic tools that we're using and therefore we can benefit from the improvements that others make, and others' use of those tools can, in some cases like Open Compute, drive down the costs of those things which make our business more efficient too. So I think to some degree we're just

[e] An early such contribution was Meta's development and maintenance of PyTorch, a computing package and machine learning framework. Though these efforts did not directly contribute to Meta's bottom line, Pytorch's availability had valuable second-order effects, generating goodwill and drawing the developer community to Meta's preferred toolset.

[f] The license appeared at first like many other open-source licenses, but further reading revealed the fine print: "If, on the Llama 2 version release date, the monthly active users of the products or services made available…is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights." (Source: Llama 2 Version Release Date: July 18, 2023, Meta, https://ai.meta.com/llama/license/)

> playing a different game on the infrastructure than companies like Google or Microsoft or Amazon, and that creates different incentives for us.[64]

In 2025, Zuckerberg began assembling a new AI "superintelligence" research lab. Zuckerberg personally recruited new talent, including Scale AI founder Alexandr Wang[g] and researchers from OpenAI and Google.[65] Meta offered OpenAI employees initial compensation packages rumored to be as high as $100 million paid over time.[66] Remarkably, one researcher turned down an $18 million salary from Meta to instead join former OpenAI CTO Murati at Thinking Machines Lab.[67]

DeepSeek

In May 2023, Liang Wenfeng founded DeepSeek, a Chinese AI startup backed by his hedge fund. Within a year, the startup launched several model iterations with a chatbot rivaling ChatGPT. In January 2025, DeepSeek shook the market[h] with claims that its latest DeepSeek-R1 model had been trained at a fraction of the cost of other modern generative models. Some estimated that DeepSeek was able to achieve similar model performance to Meta's LLaMA 3 70B model with one-fifth the compute and one-twentieth the amount of GPU performance.[68] By February 2025, DeepSeek had overtaken ChatGPT as the number one free app on Apple's App Store.[69] U.S. President Donald Trump warned the American tech industry to view the emergence of DeepSeek as a "wake up call."[70] DeepSeek also made its model available under an open-source license. Meta's LeCun reflected: "DeepSeek has profited from open research and open source (e.g., PyTorch and LLaMA from Meta). They came up with new ideas and built them on top of other people's work. Because their work is published and open source, everyone can profit from it. That is the power of open research and open source."[71]

Like many AI model companies, DeepSeek published papers accompanying each model release to share insights into its modeling approach and benchmark performance. In its papers, DeepSeek demonstrated the efficacy of innovative training techniques that stirred the AI industry. These included the use of a Mixture of Experts (MoE) architecture to reduce computation costs by activating only a small subset of model parameters for a given ask. The startup also used a technique called distillation, allowing it to train its models more quickly by essentially posing questions to a larger "teacher" model.[72] Soon after DeepSeek demonstrated the promise of distillation, researchers at Berkeley, Stanford, and the University of Washington used this technique to create their own reasoning models for only hundreds of dollars—and in less than a day. Databricks CEO Ali Ghodsi stated: "This distillation technique is just so extremely powerful and so extremely cheap, and it's just available to anyone."[73] OpenAI opened an investigation into whether DeepSeek had trained its new model by querying ChatGPT as its teacher.[74]

Google's Response

The pressure was on Google as an ever-growing set of generative AI competitors released new models week to week. Generative AI had moved beyond the research community, achieving

---

[g] Scale AI was a startup providing data annotation and model evaluation to help businesses build AI models. Meta made a $15 billion investment in Scale AI to acquire talent from it and a commitment towards future work for Meta.

[h] This announcement posed an unprecedented threat to chip producer Nvidia, a key beneficiary in the generative AI boom. Access to Nvidia's chips had been critical to the development of modern generative models, but DeepSeek's more efficient models raised questions about future demand. NVIDIA saw a $593 billion drop in market value in January 2025 because of broad selloffs—the single biggest one-day decline ever on Wall Street. (Source: Sinéad Carew, Amanda Cooper and Ankur Banerjee, "DeepSeek Sparks AI Stock Selloff; Nvidia Posts Record Market-Cap Loss," Reuters, January 27, 2025, https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/, accessed May 21, 2025.)

mainstream appeal as consumer applications started to take off. Some perceived Google to have fallen behind despite its pioneering role in developing modern generative AI and large language models.[75] A *The Wall Street Journal* headline put it succinctly: "I quit Google Search for AI—and I'm not going back."[76] In April 2023, Google merged its two internal AI research arms, Google Brain and DeepMind, in a move to unify its AI efforts under a new combined DeepMind.[77]

Investment in Anthropic

In late 2022, Google made its first investment in the then relatively new AI startup Anthropic. Anthropic's sibling founders, Daniela and Dario Amodei, were among the executives who had left OpenAI. The duo founded Anthropic as a public-benefit corporation in 2021, reportedly due to concerns about OpenAI's shift to a for-profit model and a differing vision for how AI could positively change the world.[78] Daniela said that the founding team sought "the opportunity to make a focused research bet with a small set of people who were highly aligned around a very coherent vision of AI research and AI safety," an approach they would refer to as "Constitutional AI."[79]

Google initially invested $300 million for a 10% stake in Anthropic and its pledge to use Google Cloud as a preferred cloud provider.[80] While the relationship was initially compared to that of Microsoft and OpenAI, Google never levied the same degree of influence on Anthropic. In September 2023, Amazon agreed to invest up to $4 billion in the startup, becoming Anthropic's "primary cloud provider for mission critical workloads."[81] Google then increased its own investment by $2 billion and reached a multiyear deal with Anthropic for the use of Google Cloud, worth over $3 billion.[82]

Anthropic improved its model, receiving praise as "the 2nd best publicly accessible model after OpenAI's GPT-4" by October 2023.[83] Beyond its consumer-facing chatbot, Anthropic licensed its proprietary model and Constitutional AI approach to other companies building AI products.[84]

Open-Source Gemma Models

An ongoing question for Google was whether to pursue an open- or closed-source approach to its generative AI models. In a closed-source model, Google could keep its technological innovations proprietary as it did with most of the intellectual property behind its search engine. However, Google had a strong reputation as a leader in open source and recognized that in the realm of generative AI Meta had caught up and perhaps even surpassed it. In May 2023, a leaked internal Google document surfaced titled "We Have No Moat, and Neither Does OpenAI." In it, a Google researcher warned:

> Paradoxically, the one clear winner in all of this is Meta. Because the leaked model was theirs [Meta's], they have effectively garnered an entire planet's worth of free labor. Since most open source innovation is happening on top of their architecture, there is nothing stopping them from directly incorporating it into their products. The value of owning the ecosystem cannot be overstated. Google itself has successfully used this paradigm in its open source offerings, like Chrome and Android. By owning the platform where innovation happens, Google cements itself as a thought leader and direction-setter, earning the ability to shape the narrative on ideas that are larger than itself. The more tightly we control our models, the more attractive we make open alternatives. Google and OpenAI have both gravitated defensively toward release patterns that allow them to retain tight control over how their models are used. But this control is a fiction. Anyone seeking to use LLMs for unsanctioned purposes can simply take their pick of the freely available models.[85]

A year later, in February 2024, Google released the first version of Gemma, its family of open-source AI models. To develop Gemma, Google leveraged the research that went into building its flagship closed model, Gemini, but focused on developing a pair of lightweight models that were faster and less costly to run. Google continued to refine these models, with Gemma 3 released in March 2025.[86]

New Consumer-Facing Products

Gemini Chatbot In January 2023, reports indicated that Google would introduce a suite of generative AI products over the coming months.[87] In March, Google released its experimental chatbot Bard.[88] In December, the company introduced its newest and most powerful model, Gemini.[89] Gemini became Google's flagship model and soon replaced Bard as the company's primary generative AI chatbot, a direct answer to ChatGPT (see **Exhibit 7** for the Gemini interface). On Android devices, users could use Gemini as their primary mobile assistant through text, voice, and images.[90]

Productivity Applications Google later integrated Gemini with several of the company's productivity applications (reminiscent of Microsoft integrating OpenAI models in its parallel product suite). In Gmail, Gemini could compose email responses or search an inbox. In Slides, it could help design a presentation. In Meet, it could transcribe or summarize a meeting.[91] Google also launched new AI products like NotebookLM, where users upload content to receive summarized insights.[92]

Google Search Meanwhile, Google started to make incremental changes to its core Search product. In May 2024, Google rolled out Gemini-powered AI Overviews to the U.S. market after a year of beta testing.[93] After submitting a search, users would see a short AI narrative summary followed by a typical results page (**Exhibit 8**). AI Overviews did not always show up, appearing more often for informational queries (e.g., those with a clear question structure and fact focus) and less for high-volume transactional or branded search terms (e.g., "men's running shoes").[94] Publishers like the Daily Mail saw up to a 44% lower clickthrough rate after the launch of AI Overviews.[95] Negative reports also emerged about the feature providing inaccurate or dangerous information.[96] In one estimate, Google reduced AI Overviews from 15% to 7% of searches by June 2024.[97] Even so, by April 2025 AI Overviews reached over 1.5 billion users monthly, making it the most-used generative AI product in the world.[98]

The next significant generative AI enhancement to Search came in the form of AI Mode, rolled out in May 2025. Users would see a new tab for AI Mode where they could choose to use a more narrative, generative AI-powered version of Search. Within the results, AI Mode still provided specific links to relevant articles and sometimes embedded other Google products, like Google Maps (**Exhibit 9**). Google planned to integrate elements of the AI Mode experience directly into core Search as the company received feedback from use.[99] Google also teased an "AI Ultra Plan" at $249.99 a month, which provided access to the most advanced models, experimental products (e.g., AI filmmaking tool Flow), and higher AI usage limits, among other premium features.[100]

Google's Challenges for Search

One of the key questions that Google faced was how deeply to integrate generative AI into Google Search. Google was by far the leader in the global search engine market, consisting of technologies for querying, sorting, and displaying relevant internet information (including advertising). But in May of 2025, Apple revealed that Google searches on Safari had fallen for the first time in over 20 years—attributed in large part to a shift towards generative AI alternatives. Alphabet's share price immediately declined 7%.[101] Google's competitors had much to gain from a change in the way people sought information online, and Google had just as much (or more) to lose.

Unit Economics of Generative AI in Search

Questions around unit economics complicated the decision of whether and how to integrate generative AI into Search. Thus far, it was unclear whether any other company in the space had managed to make generative AI profitable. Even OpenAI, whose ChatGPT product had taken the market by storm, was speculated to have lost $5 billion in 2024 because of incredibly high operating costs and challenges of monetization.[102]

Inference Costs Unbeknownst to most users, each traditional Google search did cost money: one estimate put the cost of each index query at 1.060 cents. Adding generative AI to Search would introduce the new cost of "inference."[i] There was inference cost associated with both understanding the user's prompt and generating a human-language response.[103] In April 2025, a user asked Sam Altman: "I wonder how much money OpenAI has lost in electricity costs from people saying 'please' and 'thank you' to their models." Altman responded: "Tens of millions of dollars well spent—you never know."[104] By one estimate, adding basic generative AI into Search would add 0.356 cents in inference cost to each search (on top of its existing cost per index query), holding user behavior fixed.[105]

With over five trillion Google searches conducted each year, inference costs could add up.[106] However, still-unknown dynamics could cause inference cost to either contract or balloon. As technology improved, some estimated that inference cost could come down by at least 30% each year.[107] At the same time, inference cost scaled up linearly with each word in a search prompt. A user communicating with a chatbot in a more "human" way[j] might use many more words. For instance, a 10x longer search prompt would be 10x the inference cost.[k] The complexity of the specific generative model used would also have major implications on inference cost: a more capable model with more parameters or step-by-step reasoning would be more expensive to offer, often dramatically so.

Monetization There were questions as to whether the lucrative search advertising business model (which earned an estimated 1.61 cents per search for Google) would be as effective if Search switched from delivering a list of links to delivering a narrative chat response. If search queries could be answered directly by a chat response, instead of a page of links with integrated ads, it was unclear where ads would be placed and what advertisers would pay for those ads.

Google could also consider a paid subscription model for generative AI search. Companies like Perplexity and OpenAI had gone down this route, both converging on $20 per month per user as the entry price point. However, conversion to the paid plan was low.[108] Even industry-leading subscription applications like Microsoft 365, Netflix, and Spotify Premium had each reached a ceiling of 400 million paid users globally.[109] Thus, only a subset of the 6.25 billion unique users of Google Search would plausibly pay for a subscription.[110] Furthermore, the users likely to self-select into a paid subscription would be those that used it intensively; many technologies followed a Pareto distribution whereby 20%

---

[i] Inference was the process where input data (a prompt) passed through a generative AI model to generate an output (response). Known as a "forward pass," this process consumed significantly more GPU capacity and electricity than conventional computing. For example, a computer could solve 1+1 in a single conventional operation. But asking the same of a large language model with a neural network of trillions of parameters could require trillions of operations. Inference cost was roughly proportional to the size of the model (parameters) multiplied by the total length of the prompt and response.

[j] For example, a user communicating with a generative chatbot might increase their search prompt length by: saying "please" and "thank you," framing prompts as questions (e.g., "What is…" or "Where should I…"), providing context (e.g., "I am writing a case about AI…"), and asking follow-up questions.

[k] A single generative search (in which the user typed their prompt and submitted it to the chatbot) might also require multiple index queries in the background, analogous to a user making numerous traditional searches on Google in order to gather the information needed for the generative search's response.

of the users represented 80% of the usage (this could be calculated as: X share of users represented X^0.14 share of usage where X was between 0 and 1)[1].

Chips

Specialized chips called graphics processing units (GPUs) provided the computational power needed for AI model training and inference. GPUs, originally designed for rendering video graphics, were able to run many similar calculations in parallel, making them ideal for AI computation.

More than any other chip provider, Nvidia was at the foundation of modern generative AI. Its A100 and H100 GPUs, estimated to control between 70% and 95% of the market in 2024, were the industry standard.[111] A single H100 was priced at $25,000 or more, but thousands of chips were required for the magnitude of computation. For example, Meta planned to have 350,000 H100 graphics cards with a total cost of up to $10 billion for GPUs alone.[112] Nvidia announced its newest Blackwell chip in March 2024 at a similar price and by November the entire 2025 production supply had reportedly sold out.[113] By the end of 2024, Nvidia's market value reached $3.28 trillion, up over $2 trillion from the previous year, making it the second-most valuable company in the world.[114]

To keep up with generative AI, Google would need to spend billions of dollars every year on chips alone if it depended on Nvidia. But for a decade, Google had been developing their answer to GPUs: the Trillium Tensor Processing Unit (TPU). In December 2024, the company announced that its sixth-generation TPU would be available for Google Cloud customers and shared that its flagship model Gemini 2.0 had been trained using Trillium TPUs.[115]

Copyright

Generative AI companies drew increasing scrutiny for potential copyright violations. In the early days of AI model training, it was common to disclose the sources of training data for the benefit of the research community. As copyright holders started paying attention, companies moved away from disclosing their sources of training data to avoid creating additional liability. However, it was broadly understood that these models were using copyrighted works in their training data—for example, Meta was found to have trained their LLaMA 3 model on a large digital library of pirated books.[116]

The pace of AI copyright lawsuits heated up in 2023, including the New York Times's lawsuit against OpenAI and Microsoft. The Times saw use of their copyrighted works for model training as unlawful infringement: in their eyes, OpenAI reproduced and monetized the newspaper's work without adequate compensation.[117] More plaintiffs would later come forward (**Exhibit 6**). The case was further complicated by the "fair use" doctrine, which allowed for limited use of copyrighted material without permission from the copyright holder.[118] In response, Google and OpenAI began to strike content licensing agreements. In February 2024, Google and Reddit announced a content licensing deal reportedly worth $60 million annually. In this deal, Reddit's content would be available for use in AI model training by Google.[119] A couple of months later, OpenAI entered into a parallel agreement with Reddit.[120] In June 2025, Reddit filed a lawsuit against Anthropic, where no content agreement existed, alleging unauthorized use of Reddit data for training.[121]

[1] For example, assume that X = 0.20 (representing 20% of users). Under a typical Pareto distribution, one would then calculate this subpopulation's share of usage as 0.20^0.14 = 0.80 (representing 80% of usage). Thus, 20% of users make up 80% of usage.

Liability and Ethics

During Google's later tenure as a tech giant, its internal culture had evolved to one of risk aversion. As controversy around various AI efforts piled up, Google leaders and employees increasingly worried about the ethical, reputational, and legal risks of deploying generative AI in Search.[122]

The risks associated with generative AI were real and ever-growing as usage scaled. The unexplainability of generative models meant that even the most robust efforts at content moderation could not fully prevent AI models from delivering incorrect or misleading results known as "hallucinations". For example, in May 2025, Google's AI Overview was persistent that the year was in fact 2024 even after repeated questioning from a user (see **Exhibit 10**).[123] Beyond this relatively innocuous example, many worried that generative AI could easily lead to more dangerous or insidious outcomes: self-injury, violence and incitement, criminal activity, human exploitation, fraud, etc.

A generative AI-centric approach to Search could also leave Google unprotected by U.S. legal provision Section 230 of the Communications Decency Act of 1996. Section 230 provided internet platforms with immunity from being sued for surfacing content that was posted by users of the platform.[124] In the case of Google Search, Google would not be liable for the contents of third-party links surfaced in response to a user search. If, however, Google were to shift towards a generative chatbot, search responses would be delivered in Google's own "voice"—and it was unclear whether Section 230 would still protect the company.

Looking Forwards

After over two decades of dominance, Google stood at the precipice of its first true existential crisis: the decisions made in the coming months and years would determine the fate of the company that had defined an era of the internet. Along with nearly $100 billion in cash on hand,[125] Google had all the technical capabilities and distribution channels necessary to compete in generative AI; there was nothing technically preventing it from immediately and fully integrating the technology into Search. Yet rushing into this could jeopardize Google's most lucrative business by far. But perhaps instead, the company just needed to take a step back and rethink its vision for the industry as a whole. How could Google reshape the industry of generative AI to make it more attractive for itself and others? And, most critically, how should Google position itself to best profit from all the value created by this world-changing technology?

Exhibit 1 OpenAI GPT-3 Training Data Sources

Dataset	Tokens	Proportion
Common Crawl	410 billion	60%
WebText2[a]	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Source: Tom B. Brown et al., "Language Models Are Few-Shot Learners," in Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, December 6–12, 2020, pp. 1877–1901.

Note: *Tokens* are the individual units of text used for model training. These can include whole words, parts of words, or even punctuation marks. *Proportion* is the share of each dataset relative to the total training data.

[a] The WebText2 dataset created by OpenAI was believed to contain web articles, blogs, and other content that was posted to and upvoted on Reddit.

Exhibit 2 Model Size and Training Cost for Select Generative AI Models

Model	Company	Released	Parameters (bn)	Inference[a]	Training Cost
Transformer	Google	Jun 2017	0.065	—	$900
BERT LARGE	Google	Oct 2018	0.34	680 GFLOPS	$3,000
GPT-3	OpenAI	Jun 2020	175	350 TFLOPS	$4,000,000
Megatron-Turing NLG 530B	MSFT & Nvidia	Oct 2021	530	—	$6,400,000
GPT-4	OpenAI	Mar 2023	1760	560 TFLOPS	$78,000,000
Gemini 1.0 Ultra	Google	Feb 2024	540	—	$191,000,000
Gemini 1.5 Pro	Google	Feb 2024	200	—	—
Gemma 7B	Google	Feb 2024	7	—	—
Claude 3.5 Sonnet	Anthropic	Jun 2024	175	—	>$30,000,000[b]
DeepSeek-R1	DeepSeek	Jan 2025	671	—	>$6,000,000[c]
LLaMA 4 Behemoth	Meta	Apr 2025	2000	390 TFLOPS	—
LLaMA 4 Maverick	Meta	Apr 2025	400	—	—
LLaMA 4 Scout	Meta	Apr 2025	109	—	—

Source: Compiled by casewriters from multiple sources (see Endnote 16).

Note: "—" denotes data not available or not publicly disclosed.

[a] GFLOPS refers to gigaFLOPS (on billion floating-point operatings per second) and TFLOPS refers to teraFLOPS (representing one trillion floating-point operations per second).

[b] According to Anthropic co-founder Dario Amodei on his blog: "Claude 3.5 Sonnet is a mid-sized model that cost a few $10M's to train (I won't give an exact number)."

[c] $6 million is often quoted as the training cost for DeepSeek-R1 based on a comment from DeepSeek researchers. However, it has been clarified that this cost better represents the computing cost for the final model training cycle.

Exhibit 3 Google's AI Principles: Responsible Development and Deployment (2025)

Because we understand that AI, as a still-emerging transformative technology, poses evolving complexities and risks, we pursue AI responsibly throughout the AI development and deployment lifecycle, from design to testing to deployment to iteration, learning as AI advances and uses evolve.

a. Implementing appropriate human oversight, due diligence, and feedback mechanisms to align with user goals, social responsibility, and widely accepted principles of international law and human rights.

b. Investing in industry-leading approaches to advance safety and security research and benchmarks, pioneering technical solutions to address risks, and sharing our learnings with the ecosystem.

c. Employing rigorous design, testing, monitoring, and safeguards to mitigate unintended or harmful outcomes and avoid unfair bias.

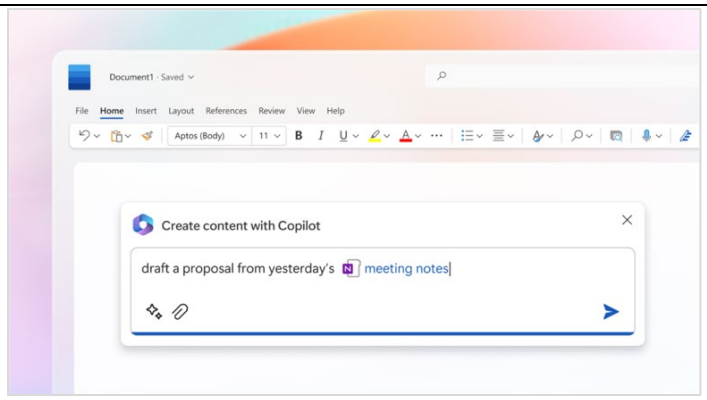
d. Promoting privacy and security, and respecting intellectual property rights.

Source: Excerpted from Google AI, "Our AI Principles," https://ai.google/principles/, accessed June 3, 2025.

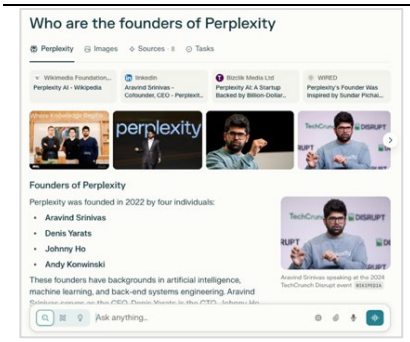
Exhibit 4 Select Generative AI Application Interfaces

ChatGPT: Interface

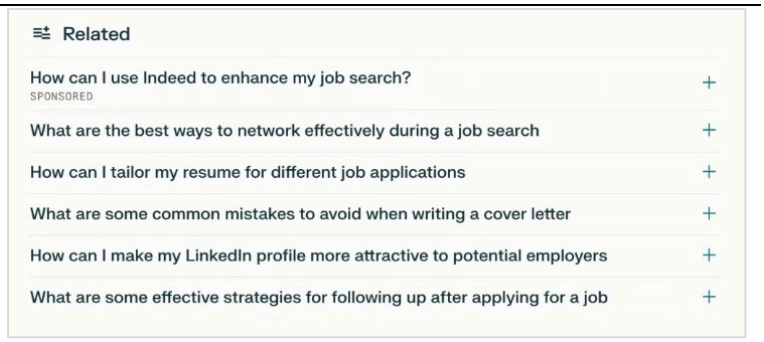
Microsoft Office Copilot: Demo



Perplexity: Interface



Perplexity: Advertising Example (Sponsored Question)



Source: Compiled by casewriters from:

Jared Spataro, "Introducing Microsoft 365 Copilot—Your Copilot for Work," Microsoft Blog, March 16, 2023, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/, accessed June 3, 2025.

Casewriter screenshot, OpenAI, "ChatGPT," https://chatgpt.com/, accessed June 3, 2025.

Casewriter screenshot, "Perplexity AI," https://www.perplexity.ai/, accessed June 3, 2025.

Perplexity Team, "Why We're Experimenting with Advertising," Perplexity, November 12, 2024, https://www.perplexity.ai/hub/blog/why-we-re-experimenting-with-advertising, accessed June 3, 2025.

Exhibit 5 Estimated Users for Consumer Generative AI Applications (January 2025)

Application	Monthly Active Users (M)
OpenAI ChatGPT	339
Microsoft Copilot	11
Google Gemini	18
Anthropic Claude	2
Perplexity	8
DeepSeek	27

Source: Compiled by casewriters from: Edward Zitron, "There Is No AI Revolution," Where's Your Ed At?, February 24, 2025, https://www.wheresyoured.at/wheres-the-money/, accessed June 11, 2025.

Exhibit 6 Selected AI Copyright Lawsuits

Date Filed	Plaintiff	Defendant	Primary Focus
Jan 13, 2023	Sarah Andersen	Stability AI, Midjourney, DeviantArt	Image
Feb 3, 2023	Getty Images	Stability AI	Image
Jun 28, 2023	Paul Tremblay	OpenAI	Literary
Jul 7, 2023	Richard Kadrey	Meta	Literary
Jul 11, 2023	Jill Leovy	Google	Journalism
Sep 12, 2023	Chabon	Meta	Literary
Sep 19, 2023	Authors Guild	OpenAI and Microsoft	Literary
Oct 17, 2023	Huckabee	Meta, Microsoft, EleutherAI, Bloomberg	Literary
Oct 18, 2023	Concord Music Group	Anthropic	Music
Nov 21, 2023	Alter	OpenAI and Microsoft	Journalism
Dec 27, 2023	The New York Times	Microsoft and OpenAI	Journalism
Feb 28, 2024	The Intercept	Microsoft and OpenAI	Journalism
Feb 28, 2024	Raw Story Media	OpenAI and Microsoft	Journalism
Mar 8, 2024	Nazemian and Dubus	Nvidia Corporation	Journalism, Literary
Apr 26, 2024	Jingna Zhang	Google	Image
Apr 30, 2024	Daily News	Microsoft and OpenAI	Journalism
Aug 19, 2024	Andrea Bartz	Anthropic	Literary
Oct 21, 2024	Dow Jones & Company	Perplexity AI	Journalism
Feb 13, 2025	Advance Local Media	Cohere	Journalism
Feb 24, 2025	Chegg	Google	Educational
Jun 4, 2025	Reddit	Anthropic	User content

Source: Compiled by casewriters from:

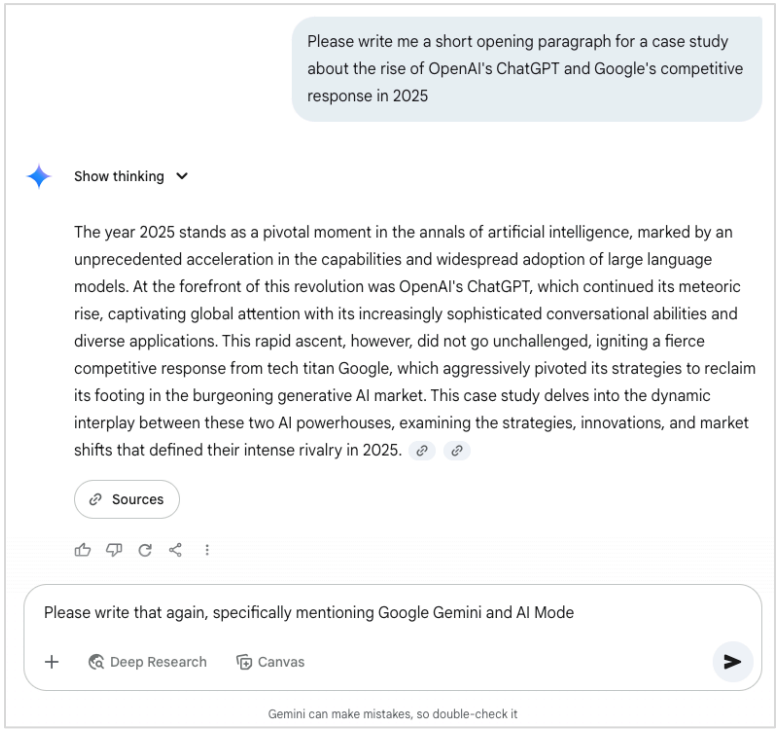
Bruce Barcott, "AI Lawsuits Worth Watching: A Curated Guide," Tech Policy Press, July 1, 2024, https://www.techpolicy.press/ai-lawsuits-worth-watching-a-curated-guide/, accessed June 11, 2025.

Theresa M. Weisenberger, Diana C. Milton, and Harrison A. Enright, "Case Tracker: Artificial Intelligence, Copyrights and Class Actions," BakerHostetler, https://www.bakerlaw.com/services/artificial-intelligence-ai/case-tracker-artificial-intelligence-copyrights-and-class-actions/, accessed June 11, 2025.

Chat GPT Is Eating the World, "Status of All 20 Copyright Lawsuits v. AI (Mar. 22, 2024)," March 22, 2024, https://chatgptiseatingtheworld.com/2024/03/22/status-of-all-20-copyright-lawsuits-v-ai-mar-22-2024/, accessed June 11, 2025.

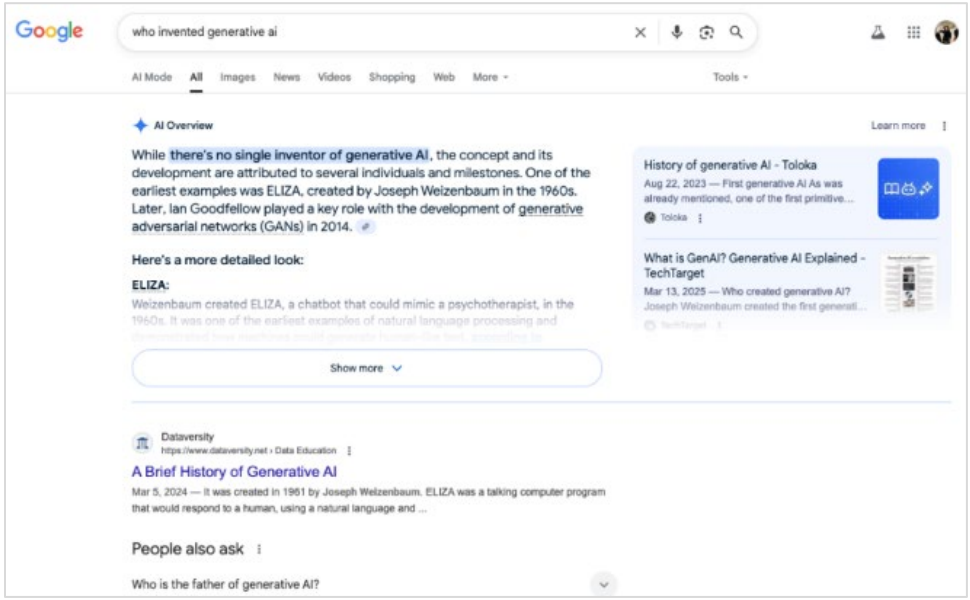
Note: Individual cases shown may have been consolidated after filing with other similar cases by the courts (e.g., *Zhang v. Google* and *Leovy v. Google* were consolidated under *In re: Google Generative AI Copyright Litigation*).

Exhibit 7 Google Gemini Interface and Sample Output with 2.5 Flash Model



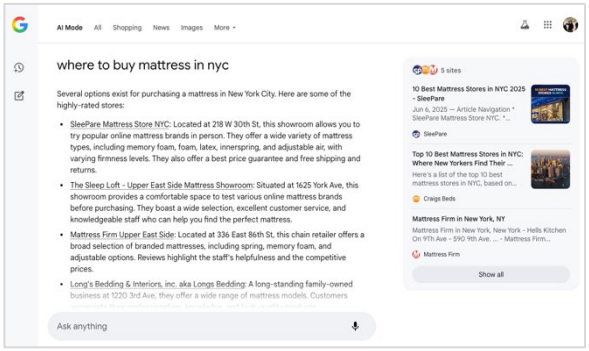
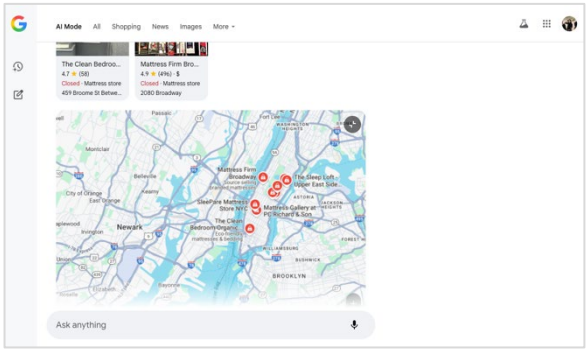
Source: Casewriter screenshot, Google, "Gemini," https://gemini.google.com/app, accessed June 3, 2025.

Exhibit 8 Sample AI Overview Output in Google Search



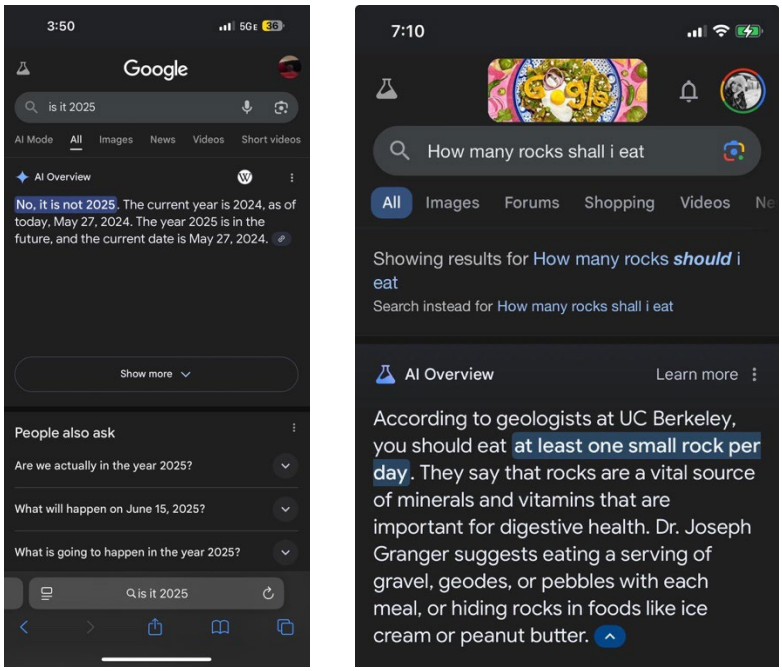
Source: Casewriter screenshot, Google Search, https://www.google.com/, accessed June 3, 2025.

Exhibit 9 AI Mode Interface in Google Search

Sample Output	Integration of Maps in AI Mode
	

Source: Casewriter screenshot, Google Search, https://www.google.com/, accessed June 3, 2025.

Exhibit 10 Examples of Incorrect AI Overview Responses in Google Search



Source: Independent-Wait-873 and UgarMalwa, "And Yet We Can't Turn This Off," Reddit, May 27, 2025, https://www.reddit.com/r/google/comments/1kwyjfg/and_yet_we_cant_turn_this_off/, accessed June 10, 2025.

Glossary

Artificial Intelligence (AI). The field of computer science focused on creating systems that could perform tasks typically requiring human intelligence.

Machine Learning (ML). A subset of artificial intelligence that enabled computers to learn from data by using algorithms and statistical models to improve performance without explicit instructions.

Neural Network. A type of machine learning model consisting of connected nodes inspired by neurons in a biological brain. By using layers of nodes that represented words as high-dimensional vectors, neural networks could capture semantic meanings from surrounding words and predict the next word in a sentence with some awareness of context. One type of neural network was a *convolutional neural network (CNN)*, designed to process structured grid-like data such as images.

Deep Learning. A type of machine learning that used multi-layered neural networks to analyze data and learn complex patterns.

Large Language Model (LLM). A type of artificial intelligence trained on huge amounts of textual data to understand and generate human language.

Generative AI. A type of artificial intelligence that could create new content (e.g., text, image, or audio) in response to a prompt from a user.

Model Training. A process consisting of *pre-training,* where a machine learning model was taught to recognize patterns and make predictions by exposing it to large amounts of data (referred to as *training data*), and *fine-tuning*, where the model was trained on more specific data to perform in a particular context on a specific task. After a model was trained, *inference* was the process of using the model to generate outputs or make predictions based on new input data.

Token. A discrete unit of text that was processed by a model (e.g., word, sub-word, character).

Parameter. An internal variable that a model extracted from its training data and used to improve its future predictions. Parameters determined how input data was transformed as it moved through the network, influencing the model's behavior and performance.

Compute (noun). The processing power required to train, fine-tune, or run AI models. It typically required specialized hardware such as *graphics processing units (GPUs)* capable of handling massive calculations. *(Public) cloud computing* was a common way to deliver computing services over the internet rather than on local servers.

Application Programming Interface (API). A set of protocols that enabled communication between two or more pieces of software and specified how the interaction would proceed. APIs worked by exposing certain functionalities of the software to outside calls, in a limited way.

Open Source. A way to license software or technologies where source code was made publicly available for anyone to inspect, use, modify, and distribute. In contrast, *closed source* was a way to license where source code was proprietary and not publicly shared.

Fair Use. A legal doctrine permitting limited use of copyrighted material without permission from the copyright holder.

Source: Andy Wu and Matt Higgins, "Generative AI Value Chain," HBS No. 724-355 (Boston: Harvard Business School Publishing, 2023).

Endnotes

[1] John Battelle, "The Birth of Google," Wired, August 1, 2005, https://www.wired.com/2005/08/battelle/, accessed May 20, 2025.

[2] Google, "In-Depth Guide to How Google Search Works," Google Search Central Documentation, March 6, 2025, https://developers.google.com/search/docs/fundamentals/how-search-works, accessed May 20, 2025.

[3] Gil Press, "Why Yahoo Lost and Google Won," Forbes, July 26, 2016, https://www.forbes.com/sites/gilpress/2016/07/26/why-yahoo-lost-and-google-won/, accessed June 3, 2025; Dan Tynan, "The History of Yahoo, and How It Went from Phenom to Has-Been," Fast Company, March 21, 2018, https://www.fastcompany.com/40544277/the-glory-that-was-yahoo, accessed June 3, 2025.

[4] Quoc V. Le and Mike Schuster, "A Neural Network for Machine Translation, at Production Scale," Google Research Blog, September 27, 2016, https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html, accessed January 20, 2023; Julian Ibarz and Sujoy Banerjee, "Updating Google Maps with Deep Learning and Street View," Google Research Blog, May 3, 2017, https://ai.googleblog.com/2017/05/updating-google-maps-with-deep-learning.html, accessed January 20, 2023.

[5] comScore, "Share of Search Queries Handled by Leading Search Engine Providers in the United States as of December 2022," Statista, March 10, 2023, https://www.statista.com/statistics/267161/market-share-of-search-engines-in-the-united-states/, accessed March 28, 2023.

[6] StatCounter, "Search Engine Market Share Worldwide," https://gs.statcounter.com/search-engine-market-share, accessed March 28, 2023.

[7] Leah Nylen, "Google's Payments to Apple Reached $20 Billion in 2022, Antitrust Court Documents Show," Bloomberg, May 1, 2024, https://www.bloomberg.com/news/articles/2024-05-01/google-s-payments-to-apple-reached-20-billion-in-2022-cue-says, accessed May 20, 2025.

[8] Associated Press, "Google Loses Massive Antitrust Case Over Its Search Dominance," NPR, August 5, 2024, https://www.npr.org/2024/08/05/nx-s1-5064624/google-justice-department-antitrust-search, accessed May 21, 2025.

[9] John Markoff, "How Many Computers to Identify a Cat? 16,000," *New York Times*, June 25, 2012, https://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html, accessed January 20, 2023.

[10] Sundar Pichai, "TensorFlow: Smarter Machine Learning, for Everyone," The Keyword Blog, November 9, 2015, https://blog.google/technology/ai/tensorflow-smarter-machine-learning-for/, accessed May 15, 2023.

[11] Catherine Shu, "Google Acquires Artificial Intelligence Startup DeepMind for More than $500M," TechCrunch, January 26, 2014, https://techcrunch.com/2014/01/26/google-deepmind/, accessed January 23, 2023.

[12] Lance Johnson, "The Origins of AlexNet: Revolutionizing Deep Learning and Computer Vision," BytePlus, April 24, 2025, https://www.byteplus.com/en/topic/401667?title=the-origins-of-alexnet-revolutionizing-deep-learning-and-computer-vision, accessed June 3, 2025.

[13] Nobel Prize, "Geoffrey Hinton—Facts—2024," Nobel Prize, https://www.nobelprize.org/prizes/physics/2024/hinton/facts/, accessed May 20, 2025.

[14] Amazon Web Services, "What Are Transformers in Artificial Intelligence?," https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/, accessed May 20, 2025; Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," NeurIPS 2017, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[15] Anton Shilov, "Nvidia's H100 GPUs Will Consume More Power Than Some Countries—Each GPU Consumes 700W of Power, 3.5 Million Are Expected to Be Sold in the Coming Year," Tom's Hardware, December 26, 2023, https://www.tomshardware.com/tech-industry/nvidias-h100-gpus-will-consume-more-power-than-some-countries-each-gpu-consumes-700w-of-power-35-million-are-expected-to-be-sold-in-the-coming-year, accessed June 10, 2025.

[16] **Exhibit 2** compiled from the following sources:

Code Titan, "Increasing Size of Transformer Models," Medium, December 20, 2024, https://medium.com/@info.codetitan/increasing-size-of-transformer-models-21325bda4639, accessed June 11, 2025.

Grigory Sapunov, "Speeding Up BERT," Medium, September 30, 2019, https://moocaholic.medium.com/speeding-up-bert-5528e18bb4ea, accessed June 11, 2025.

Matt Casey, "BERT Models: Google's NLP for the Enterprise," Snorkel AI, December 27, 2023, https://snorkel.ai/large-language-models/bert-models/, accessed June 11, 2025.

Guido Appenzeller, Matt Bornstein, and Martin Casado, "Navigating the High Cost of AI Compute," Andreessen Horowitz, April 27, 2023, https://a16z.com/navigating-the-high-cost-of-ai-compute/, accessed June 11, 2025.

Ali Alvi and Paresh Kharya, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model," Microsoft Research Blog, October 11, 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/, accessed June 11, 2025.

Farhan Ghumra, "OpenAI GPT-3, the Most Powerful Language Model: An Overview," eInfochips, March 14, 2022, https://www.einfochips.com/blog/openai-gpt-3-the-most-powerful-language-model-an-overview/, accessed June 11, 2025.

Dorothy Neufeld, "Visualizing the Training Costs of AI Models Over Time," Visual Capitalist, June 4, 2024, https://www.visualcapitalist.com/training-costs-of-ai-models-over-time/, accessed June 11, 2025.

Brett Young, "AI Expert Speculates on GPT-4 Architecture," Weights & Biases, June 21, 2023, https://wandb.ai/byyoung3/ml-news/reports/AI-Expert-Speculates-on-GPT-4-Architecture---Vmlldzo0NzA0Nzg4, accessed June 11, 2025.

OpenAI, "Announcing GPT-4 in ChatGPT," ChatGPT Release Notes, March 14, 2023, https://wandb.ai/byyoung3/ml-news/reports/AI-Expert-Speculates-on-GPT-4-Architecture---Vmlldzo0NzA0Nzg4, accessed June 11, 2025.

Martin Treiber, "The Secrets of GPT-4 Leaked?," IKANGAI, July 14, 2023, https://www.ikangai.com/the-secrets-of-gpt-4-leaked/, accessed June 11, 2025.

Frontiere, "ChatGPT vs. Gemini: Comparing 2 Giants of Artificial Intelligence," June 7, 2024, https://frontiere.io/en/chatgpt-vs-gemini-comparing-2-giants-of-artificial-intelligence/, accessed June 11, 2025.

Hassan Shittu, "Google Gemini PRO 1.5: All You Need to Know About This Near Perfect AI Model," Fello AI, September 9, 2024, https://felloai.com/2024/09/google-gemini-pro-1-5-all-you-need-to-know-about-this-near-perfect-ai-model/, accessed June 11, 2025.

Jeanine Banks and Tris Warkentin, "Gemma: Introducing New State-of-the-Art Open Models," The Keyword Blog, February 21, 2024, https://blog.google/technology/developers/gemma-open-models/, accessed June 11, 2025.

Hassan Shittu, "The Ultimate Review of Claude 3.5 Sonnet AI—Comparison to ChatGPT & All You Need to Know!," Fello AI, August 29, 2024, https://felloai.com/2024/08/claude-ai-everything-you-need-to-know/, accessed June 11, 2025.

Anthropic, "Claude 3.5 Sonnet," June 20, 2024, https://www.anthropic.com/news/claude-3-5-sonnet, accessed June 11, 2025.

Dario Amodei, "On DeepSeek and Export Controls," January 2025, https://www.darioamodei.com/post/on-deepseek-and-export-controls, accessed June 11, 2025.

Fireworks AI, "DeepSeek R1: All You Need to Know," https://fireworks.ai/blog/deepseek-r1-deepdive, accessed June 11, 2025.

Mark Minevich, "The $6 Million AI Bombshell: How DeepSeek Shook Wall Street and AI Leadership," Forbes, February 6, 2025, https://www.forbes.com/sites/markminevich/2025/02/06/the-6-million-ai-bombshell-how-deepseek-shook-wall-street-and-ai-leadership/, accessed June 11, 2025.

DeepSeek, "DeepSeek-R1 Release," DeepSeek API Docs, January 20, 2025, https://api-docs.deepseek.com/news/news250120, accessed June 11, 2025.

Meta, "The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation," Meta Blog, April 5, 2025, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, accessed June 11, 2025.

Collabnix Team, "Deep Technical Analysis of Llama 4 Scout, Maverick and Behemoth," Collabnix, April 6, 2025, https://collabnix.com/deep-technical-analysis-of-llama-4-scout-maverick-and-behemoth/, accessed June 11, 2025.

[17] The Stanford Institute for Human-Centered Artificial Intelligence, "AI Index: State of AI in 13 Charts," April 15, 2024, https://hai.stanford.edu/news/ai-index-state-ai-13-charts, accessed June 2, 2025.

[18] Erin Woo and Amir Efrati, "OpenAI's Losses Doubled to $540 Million as It Developed ChatGPT," The Information, May 4, 2023, https://www.theinformation.com/articles/openais-losses-doubled-to-540-million-as-it-developed-chatgpt, accessed June 6, 2023.

[19] Rishi Bommasani and Percy Liang, "Reflections on Foundation Models," Stanford University Human Centered Artificial Intelligence Center for Research on Foundation Models, October 18, 2021, https://crfm.stanford.edu/2021/10/18/reflections.html, accessed July 6, 2023.

[20] Noam Hassenfeld, "Even the Scientists Who Build AI Can't Tell You How It Works," Vox, July 15, 2023, https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast, accessed May 20, 2025.

[21] James Manyika and Demis Hassabis, "Responsible AI: Our 2024 Report and Ongoing Work," The Keyword Blog, February 4, 2025, https://blog.google/technology/ai/responsible-ai-2024-report-ongoing-work/, accessed June 2, 2025.

[22] OpenAI, "Introducing OpenAI," OpenAI Blog, December 11, 2015, https://openai.com/blog/introducing-openai/, accessed January 20, 2023.

[23] Will Douglas Heaven, "OpenAI's New Language Generator GPT-3 Is Shockingly Good—and Completely Mindless," MIT Technology Review, July 20, 2020, https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/, accessed January 20, 2023.

[24] Karen Hao and Deepa Seetharaman, "Cleaning Up ChatGPT Takes Heavy Toll on Human Workers," July 24, 2023, https://www.wsj.com/tech/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483, accessed May 21, 2025.

[25] Cindy Gordon, "ChatGPT Is the Fastest Growing App in the History of Web Applications," Forbes, February 2, 2023, https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/, accessed June 2, 2025.

[26] Sarah Perez, "ChatGPT Doubled Its Weekly Active Users in Under 6 Months, Thanks to New Releases," March 6, 2025, https://techcrunch.com/2025/03/06/chatgpt-doubled-its-weekly-active-users-in-under-6-months-thanks-to-new-releases/, accessed June 2, 2025.

[27] SA News Team, "Around 10% of World Is Using OpenAI's Systems: Sam Altman," Seeking Alpha, April 13, 2025, https://seekingalpha.com/news/4430645-around-10-of-world-is-using-openais-systems-sam-altman, accessed June 2, 2025.

[28] Eliezer Yudkowski, "I don't think people realize…," Twitter: @ESYudkowsky, March 14, 2023, https://twitter.com/ESYudkowsky/status/1635577836525469697, accessed March 28, 2023.

[29] OpenAI, "GPT-4 Technical Report," https://cdn.openai.com/papers/gpt-4.pdf, accessed May 21, 2025.

[30] Ali Alvi, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model," Microsoft Research Blog, October 11, 2021, https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/, accessed June 5, 2025; Brett Young, "AI Expert Speculates on GPT-4 Architecture," Weights & Biases, June 21, 2023, https://wandb.ai/byyoung3/ml-news/reports/AI-Expert-Speculates-on-GPT-4-Architecture---Vmlldzo0NzA0Nzg4, accessed June 5, 2025.

[31] OpenAI, "Our Structure," https://openai.com/our-structure/, accessed May 21, 2025.

[32] Tom Warren, "Microsoft Extends OpenAI Partnership in a 'Multibillion Dollar Investment'," The Verge, January 23, 2023, https://www.theverge.com/2023/1/23/23567448/microsoft-openai-partnership-extension-ai, accessed January 27, 2023.

[33] PYMNTS, "OpenAI CEO Sam Altman: Company Considering 'Different Open-Source Strategy,'" February 2, 2025, https://www.pymnts.com/artificial-intelligence-2/2025/openai-ceo-sam-altman-company-considering-different-open-source-strategy/, accessed June 10, 2025; Sarah Jackson, "Sam Altman Explains OpenAI's Shift to Closed AI Models," Business Insider, November 2, 2024, https://www.businessinsider.com/sam-altman-why-openai-closed-source-ai-models-2024-11, accessed June 16, 2025.

[34] Bobby Allyn, "How OpenAI's Origins Explain the Same Altman Drama," NPR, November 24, 2023, https://www.npr.org/2023/11/24/1215015362/chatgpt-openai-sam-altman-fired-explained, accessed June 2, 2025.

[35] Elon Musk, "I'm still confused as to how a non-profit to which I donated ~$100M somehow became a $30B market cap for-profit. If this is legal, why doesn't everyone do it?," Twitter: @elonmusk, March 15, 2023, https://twitter.com/elonmusk/status/1636047019893481474, accessed February 12, 2024.

[36] The Economic Times, "OpenAI Offers Lavish Compensation to Staff; Here's How Much Employees Get Paid," November 19, 2024, https://economictimes.indiatimes.com/news/international/us/openai-offers-lavish-compensation-to-staff-heres-how-much-employees-get-paid/articleshow/115490469.cms?from=mdr, accessed May 21, 2025; Glassdoor, "Software Engineer Salaries in San Francisco, CA, United States," April 10, 2025, https://www.glassdoor.co.uk/Salaries/san-francisco-ca-united-states-software-engineer-salary-SRCH_IL.0,30_IM759_KO31,48.htm, accessed May 21, 2025.

[37] Scott Maxwell, "Why the AI Brain Drain Won't Last," Inc. Magazine, April 24, 2017, https://www.inc.com/scott-maxwell/why-the-ai-brain-drain-wont-last.html, accessed May 21, 2025.

[38] Lily Jamali and Ben Derico, "OpenAI Sues Elon Musk Claiming 'Bad-Faith Tactics,'" BBC, April 11, 2025, https://www.bbc.com/news/articles/cg4114271x2o, accessed May 21, 2025.

[39] Alexandra Tremayne-Pengelly, "OpenAI's Leadership Exodus: 9 Key Execs Who Left the A.I. Giant This Year," Observer, September 28, 2024, https://observer.com/2024/09/openai-executives-resign/, accessed June 2, 2025; Steven Levy, "Mira Murati Is Ready to Tell the World What She's Working On," Wired, February 18, 2025, https://www.wired.com/story/mira-murati-thinking-machines-lab/, accessed June 2, 2025; Dan Holden, "Anthropic's Sibling Founders on Leaving OpenAI to Start a $15B Startup," Observer, May 10, 2024, https://observer.com/2024/05/anthropics-sibling-founders-on-leaving-openai-to-start-a-15b-startup/, accessed June 2, 2025.

[40] Mojan Anari, "Non-Compete Clarity: California Employers Must Provide Notice of Non-Competes to Employees by February 14, 2024," Akerman LLP, February 8, 2024, https://www.akerman.com/en/perspectives/hrdef-non-compete-clarity-california-employers-must-provide-notice-of-non-competes-to-employees-by-february-14-2024.html, accessed June 2, 2025; Ross Shanberg, "Covenants Not to Compete (Non-Compete Agreements)," Shanberg Stafford LLP, https://www.ssfirm.com/employee-covenants-not-to-compete-the-myth-of-enforceability-and-alternative-protective-measures-available-to-california-employers/, accessed June 2, 2025.

[41] SoftBank Group, "OpenAI and SoftBank Group Partner to Develop and Market Advanced Enterprise AI," February 3, 2025, https://group.softbank/en/news/press/20250203_0, accessed June 2, 2025.

[42] Keach Hagey and Berber Jin, "Former Apple Design Guru Jony Ive to Take Expansive Role at OpenAI," *Wall Street Journal*, May 21, 2025, https://www.wsj.com/tech/ai/former-apple-design-guru-jony-ive-to-take-expansive-role-at-openai-5787f7da, accessed June 2, 2025; Jay Peters, "OpenAI Is Buying Jony Ive's AI Hardware Company," The Verge, May 21, 2025, https://www.theverge.com/news/671838/openai-jony-ive-ai-hardware-apple, accessed June 2, 2025; Jess Weatherbed, "Details Leak About Jony Ive's New 'Screen-Free' OpenAI Device," The Verge, May 22, 2025, https://www.theverge.com/news/672357/openai-ai-device-sam-altman-jony-ive, accessed June 2, 2025.

[43] Tom Warren, "Microsoft's OpenAI Investment Was Triggered by Google Fears, Emails Reveal," The Verge, May 1, 2024, https://www.theverge.com/2024/5/1/24146302/microsoft-openai-investment-google-worries-internal-emails, accessed June 5, 2025.

[44] OpenAI, "Microsoft Invests in and Partners with OpenAI to Support Us Building Beneficial AGI," July 22, 2019, https://openai.com/index/microsoft-invests-in-and-partners-with-openai/, accessed May 21, 2025; Kyle Wiggers, "Microsoft Invests Billions More Dollars in OpenAI, Extends Partnership," TechCrunch, January 23, 2023, https://techcrunch.com/2023/01/23/microsoft-invests-billions-more-dollars-in-openai-extends-partnership/, accessed June 11, 2025.

[45] Liz Hoffman and Reed Albergotti, "Microsoft Eyes $10 Billion Bet on ChatGPT," Semafor, January 9, 2023, https://www.semafor.com/article/01/09/2023/microsoft-eyes-10-billion-bet-on-chatgpt, accessed May 19, 2023.

[46] Machelle Tran and Dean Haas, "Privacy, Security, and Compliance in Microsoft OneDrive," Microsoft SharePoint, January 28, 2025, https://learn.microsoft.com/en-us/sharepoint/onedrive-privacy-security-overview, accessed May 21, 2025.

[47] Microsoft Bing Blogs, "Building the New Bing," February 21, 2023, https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing, accessed June 6, 2023.

[48] Paresh Dave, "How Microsoft's Bing Chatbot Came to Be—and Where It's Going Next," Wired, April 26, 2023, https://www.wired.com/story/how-microsofts-bing-chatbot-came-to-be-and-where-its-going-next/, accessed June 6, 2023.

[49] Harshita Mary Varghese and Krystal Hu, "Robotics Startup Figure Raises $675 mln from Microsoft, Nvidia, OpenAI," Reuters, February 29, 2024, https://www.reuters.com/technology/robotics-startup-figure-raises-675-mln-microsoft-nvidia-other-big-techs-2024-02-29/, accessed May 21, 2025; Satya Nadella, "Mustafa Suleyman, DeepMind and Inflection Co-founder, Joins Microsoft to Lead Copilot," Microsoft Blog, March 19, 2024, https://blogs.microsoft.com/blog/2024/03/19/mustafa-suleyman-deepmind-and-inflection-co-founder-joins-microsoft-to-lead-copilot/, accessed May 20, 2025.

[50] Aravind Srinivas, LinkedIn, https://www.linkedin.com/in/aravind-srinivas-16051987/, accessed June 2, 2025; Denis Yarats, LinkedIn, https://www.linkedin.com/in/denisyarats/, accessed June 2, 2025; Johnny Ho, LinkedIn, https://www.linkedin.com/in/hjohnny/, accessed June 2, 2025, Andy Konwinski, LinkedIn, https://www.linkedin.com/in/andykon/, accessed June 2, 2025.

[51] Joanne Chen, "How Perplexity Is Pioneering the Future of Search," Forbes, September 6, 2023, https://www.forbes.com/sites/joannechen/2023/09/06/how-perplexityai-is-pioneering-the-future-of-search/, accessed May 21, 2025.

[52] Cryptokaykay, "Reverse Engineering Perplexity," Reddit: u/Cryptokaykay, March 17, 2024, https://www.reddit.com/r/LocalLLaMA/comments/1bh6o3e/reverse_engineering_perplexity/, accessed May 21, 2025; Thanhtheman, "Perplexity, ChatGPT Search/DeepResearch Are Using Google Search Under the Hood?," Reddit: u/Thanhtheman, February 10, 2025, https://www.reddit.com/r/perplexity_ai/comments/1imjzte/perplexity_chatgpt_searchdeepresearch_are_using/, accessed May 21, 2025; DensityInfinite, "Whats the Deal with Perplexity Search Engine? Is It Worth It?," Reddit: u/DensityInfinite, January 26, 2024, https://www.reddit.com/r/ArcBrowser/comments/1abtq5f/whats_the_deal_with_perplexity_search_engine_is/, accessed May 21, 2025.

[53] Perplexity, "What Advanced AI Models Are Included in a Perplexity Pro Subscription?," https://www.perplexity.ai/hub/technical-faq/what-advanced-ai-models-does-perplexity-pro-unlock, accessed May 21, 2025.

[54] Perplexity Team, "Why We're Experimenting with Advertising," Perplexity Blog, November 12, 2024, https://www.perplexity.ai/hub/blog/why-we-re-experimenting-with-advertising, accessed May 21, 2025.

[55] Krystal Scanlon, Sara Guaglione, and Marty Swant, "Ads Might Have Just Hit Perplexity AI, But Not All Buyers Are Totally Convinced—Yet," Digiday, November 18, 2024, https://digiday.com/marketing/ads-might-have-just-hit-perplexity-ai-but-not-all-buyers-are-totally-convinced-yet/, accessed May 21, 2025.

Catherine Perloff, "Netflix Is Lowering Ad Prices to Below $30," Adweek, August 6, 2024, https://www.adweek.com/convergent-tv/netflix-is-lowering-ad-prices-to-below-30/, accessed May 21, 2025.

[56] Krystal Scanlon and Sara Guaglione, "Perplexity Has Offered Ads For Half a Year—Marketers Already Want Scale," Digiday, May 28, 2025, https://digiday.com/marketing/perplexity-has-offered-ads-for-half-a-year-marketers-already-want-scale/, accessed June 16, 2025.

[57] Alexey Shabanov, "Comet Enters Early Testing as Perplexity Debuts Its Agentic Browser," TestingCatalog, May 20, 2025, https://www.testingcatalog.com/comet-enters-early-testing-as-perplexity-debuts-its-agentic-browser/, accessed June 10, 2025.

[58] StatCounter Global Stats, "Desktop Browser Market Share Worldwide," May 2025, https://gs.statcounter.com/browser-market-share/desktop/worldwide, accessed June 10, 2025.

[59] Shirin Ghaffary, "Perplexity, Google Bet Web Browsers Are a New Frontier for AI," Bloomberg, June 6, 2025, https://www.bloomberg.com/news/newsletters/2025-06-06/perplexity-google-bet-web-browsers-are-a-new-frontier-for-ai, accessed June 10, 2025.

[60] Leo Sun, "Where Will Meta Platforms Stock Be in 1 Year?," Motley Fool, February 6, 2024, https://www.fool.com/investing/2024/02/06/where-will-meta-platforms-stock-be-in-1-year/, accessed February 10, 2024.

[61] Meta AI, "FAIR Turns Five: What We've Accomplished and Where We're Headed," Meta AI Blog, December 5, 2018, https://ai.facebook.com/blog/fair-turns-five-what-weve-accomplished-and-where-were-headed/, accessed January 23, 2023; Piotr Bojanowski, Armand Joulin, Edouard Grave, Christian Puhrsch, and Tomas Mikolov, "Expanded fastText Library Now Fits on Smaller-Memory Devices," Engineering at Meta Blog, May 2, 2017, https://engineering.fb.com/2017/05/02/ml-applications/expanded-fasttext-library-now-fits-on-smaller-memory-devices/, accessed January 23, 2023.

[62] Meta, "CICERO: An AI Agent That Negotiates, Persuades, and Cooperates with People," Meta Newsroom, November 22, 2022, https://ai.facebook.com/blog/cicero-ai-negotiates-persuades-and-cooperates-with-people/, accessed January 27, 2023.

[63] James Vincent, "Meta's Powerful AI Language Model Has Leaked Online—What Happens Now?," The Verge, March 8, 2023, https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse, accessed May 21, 2025.

[64] Meta Platforms Inc., "First Quarter 2023 Results Follow Up Call," Meta Newsroom, April 26, 2023, https://s21.q4cdn.com/399680738/files/doc_financials/2023/q1/META-Q1-2023-Follow-Up-Call-Transcript.pdf, accessed May 9, 2023.

[65] Ben Thompson, "Meta + Scale AI?, Meta's Reset, AI as Sustaining Innovation," Stratechery, June 11, 2025, https://stratechery.com/2025/meta-scale-ai-metas-reset-ai-as-sustaining-innovation/, accessed June 11, 2025.

[66] Dylan Butts, "Sam Altman Says Meta Offered OpenAI Staff $100 Million Bonuses, As Mark Zuckerberg Ramps Up AI Poaching Efforts," CNBC, June 18, 2025, https://www.cnbc.com/2025/06/18/sam-altman-says-meta-tried-to-poach-openai-staff-with-100-million-bonuses-mark-zuckerberg.html, accessed June 20, 2025.

[67] Julie Bort and Marina Temkin, "Meta is offering multimillion-dollar pay for AI researchers, but not $100M signing bonuses," TechCrunch, June 27, 2025, https://techcrunch.com/2025/06/27/meta-is-offering-multimillion-dollar-pay-for-ai-researchers-but-not-100m-signing-bonuses/, accessed August 19, 2025.

[68] Dylan Patel and Daniel Nishball, "OpenAI Is Doomed – Et Tu, Microsoft?," SemiAnalysis, May 7, 2024, https://semianalysis.com/2024/05/07/openai-is-doomed-et-tu-microsoft/, accessed June 5, 2025.

[69] Sinéad Carew, Amanda Cooper and Ankur Banerjee, "DeepSeek Sparks AI Stock Selloff; Nvidia Posts Record Market-Cap Loss," Reuters, January 27, 2025, https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/, accessed May 21, 2025.

[70] AP Newsroom, "Trump Says DeepSeek AI 'Should Be a Wake Up Call,'" Associated Press, January 28, 2025, https://newsroom.ap.org/editorial-photos-videos/detail?itemid=4de508f05fa145ea9550dc0ec2b19f63, accessed May 21, 2025.

[71] Yann LeCun, "To people who see the performance of DeepSeek…," LinkedIn: Yann LeCun, February 24, 2025, https://www.linkedin.com/posts/yann-lecun_to-people-who-see-the-performance-of-deepseek-activity-7288591087751884800-I3sN/, accessed May 21, 2025.

[72] Janakiram MSV, "All About DeepSeek—The Chinese AI Startup Challenging US Big Tech," Forbes, January 26, 2025, https://www.forbes.com/sites/janakirammsv/2025/01/26/all-about-deepseekthe-chinese-ai-startup-challenging-the-us-big-tech/, accessed May 21, 2025.

[73] Jasmine Wu and Deirdre Bosa, "How DeepSeek Used Distillation to Train Its Artificial Intelligence Model, and What It Means for Companies Such As OpenAI," CNBC, February 21, 2025, https://www.cnbc.com/2025/02/21/deepseek-trained-ai-model-using-distillation-now-a-disruptive-force.html, accessed May 21, 2025.

[74] Sam Schechner, "OpenAI Is Probing Whether DeepSeek Used Its Models to Train New Chatbot," *Wall Street Journal*, January 29, 2025, https://www.wsj.com/tech/ai/openai-china-deepseek-chatgpt-probe-ce6b864e, accessed May 21, 2025.

[75] Richard Nieva, "'AI First' to Last: How Google Fell Behind in the AI Boom," Forbes, February 8, 2023, https://www.forbes.com/sites/richardnieva/2023/02/08/google-openai-chatgpt-microsoft-bing-ai/, accessed May 21, 2025.

[76] Joanna Stern, "I Quit Google Search for AI—and I'm Not Going Back," *Wall Street Journal*, March 26, 2025, https://www.wsj.com/tech/personal-tech/google-search-chatgpt-perplexity-gemini-6ac749d9, accessed June 2, 2025.

[77] Sundar Pichai, "Google DeepMind: Bringing Together Two World-Class AI Teams," The Keyword Blog, April 20, 2023, https://blog.google/technology/ai/april-ai-update/, accessed May 21, 2025.

[78] Ben Sherry, "Anthropic CEO Dario Amodei Says He Left OpenAI Over a Difference in 'Vision,'" Inc. Magazine, November 13, 2024, https://www.inc.com/ben-sherry/anthropic-ceo-dario-amodei-says-he-left-openai-over-a-difference-in-vision/91018229, accessed May 21, 2025.

[79] Future of Life Institute, "Daniela and Dario Amodei on Anthropic," Future of Life Institute Podcast, Mar 4, 2022, https://futureoflife.org/podcast/daniela-and-dario-amodei-on-anthropic/, accessed October 20, 2023.

[80] James Vincent, "Google Invested $300 Million in AI Firm Founded by Former OpenAI Researchers," The Verge, February 3, 2023, https://www.theverge.com/2023/2/3/23584540/google-anthropic-investment-300-million-openai-chatgpt-rival-claude, accessed March 28, 2023.

[81] Amazon Staff, "Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI," Amazon, March 27, 2024, https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment, accessed May 21, 2025.

[82] Berber Jin and Miles Kruppa, "Google Commits $2 Billion in Funding to AI Startup Anthropic," *Wall Street Journal*, October 27, 2023, https://www.wsj.com/tech/ai/google-commits-2-billion-in-funding-to-ai-startup-anthropic-db4d4c50, accessed November 9, 2023.

[83] Dylan Patel, "Amazon Anthropic: Poison Pill or Empire Strikes Back," SemiAnalysis, Oct 1, 2023, https://www.semianalysis.com/p/amazon-anthropic-poison-pill-or-empire, accessed October 20, 2023.

[84] Anthropic, "Meet Claude," https://www.anthropic.com/product, accessed Oct 20, 2023.

[85] Dylan Patel and Afzal Ahmad, "Google 'We Have No Moat, and Neither Does OpenAI'," SemiAnalysis, May 4, 2023, https://www.semianalysis.com/p/google-we-have-no-moat-and-neither, accessed May 9, 2023.

[86] Clement Farabet and Tris Warkentin, "Introducing Gemma 3: The Most Capable Model You Can Run on a Single GPU or TPU," The Keyword Blog, https://blog.google/technology/developers/gemma-3/, accessed May 21, 2025.

Emilia David, "Google Gemma: Because Google Doesn't Want to Give Away Gemini Yet," The Verge, February 21, 2024, https://www.theverge.com/2024/2/21/24078610/google-gemma-gemini-small-ai-model-open-source, accessed May 21, 2025.

[87] Nico Grant, "Google Calls in Help from Larry Page and Sergey Brin for A.I. Fight," *New York Times*, January 20, 2023, https://www.nytimes.com/2023/01/20/technology/google-chatgpt-artificial-intelligence.html, accessed January 23, 2023.

[88] Nico Grant and Cade Metz, "Google Releases Bard, Its Competitor in the Race to Create A.I. Chatbots," *New York Times*, March 21, 2023, https://www.nytimes.com/2023/03/21/technology/google-bard-chatbot.html, accessed May 3, 2023.

[89] Sundar Pichai and Demis Hassabis, "Introducing Gemini: Our Largest and Most Capable AI Model," The Keyword Blog, December 6, 2023, https://blog.google/technology/ai/google-gemini-ai/, accessed May 21, 2025.

[90] Google Gemini, "Get Started with the Gemini Mobile App," Gemini Apps Help, https://support.google.com/gemini/answer/14554984?h, accessed June 2, 2025.

[91] DeepMind, "The Gemini Ecosystem," Google DeepMind, https://deepmind.google/models/gemini/, accessed May 6, 2025.

[92] NotebookLM, "Google NotebookLM: Note Taking and Research Assistant Powered by AI," Google NotebookLM, https://notebooklm.google/, accessed May 21, 2025.

[93] Elizabeth Reid, "Supercharging Search with Generative AI," The Keyword Blog, May 10, 2023, https://blog.google/products/search/generative-ai-search/, accessed June 2, 2025; Elizabeth Reid, "Generative AI in Search: Let Google Do the Searching for You," The Keyword Blog, May 14, 2024, https://blog.google/products/search/generative-ai-google-search-may-2024/, accessed June 2, 2025.

[94] Emir Erçelen, "What Triggers an AI Overview in Google Search," Adsby, May 16, 2025, https://adsby.co/blog/what-triggers-an-ai-overview-in-google-search/, accessed June 2, 2025.

[95] Sara Guaglione, "The Winners and Losers of Google's AI Mode," Digiday, May 29, 2025, https://digiday.com/media/the-winners-and-losers-of-googles-ai-mode/, accessed June 13, 2025.

[96] Nico Grant, "Google AI Overviews Search Errors Cause Furor Online," *New York Times*, May 24, 2024, https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html, accessed June 2, 2025.

[97] Danny Goodwin, "Google AI Overviews Only Show for 7% of Queries, a New Low," Search Engine Land, July 11, 2024, https://searchengineland.com/google-ai-overviews-visibility-new-low-444048, accessed June 2, 2025.

[98] Ben Thompson, "Google Earnings, Search and AI, Building AI Products," Stratechery, April 28, 2025, https://stratechery.com/2025/google-earnings-search-and-ai-building-ai-products/, accessed June 2, 2025.

[99] Elizabeth Reid, "AI in Search: Going Beyond Information to Intelligence," The Keyword Blog, May 20, 2025, https://blog.google/products/search/google-search-ai-mode-update/, accessed June 2, 2025.

[100] Shimrit Ben-Yair, "Introducing Google AI Ultra: The Best of Google AI in One Subscription," The Keyword Blog, May 20, 2025, https://blog.google/products/google-one/google-ai-ultra/, accessed June 2, 2025.

Ben Thompson, "Google I/O, The Search Funnel, Product Possibilities," Stratechery, May 21, 2025, https://stratechery.com/2025/google-i-o-the-search-funnel-product-possibilities/, accessed June 2, 2025.

[101] Dan Gallagher, "AI's Threat to Google Just Got Real," *Wall Street Journal*, May 8, 2025, https://www.wsj.com/tech/ai/ais-threat-to-google-just-got-real-8280b4ee, accessed May 21, 2025.

[102] Edward Zitron, "There Is No AI Revolution," Where's Your Ed At?, February 24, 2025, https://www.wheresyoured.at/wheres-the-money/, accessed June 5, 2025.

[103] Adam Kohan, "Learning and Inference in a Forward Pass: The New Framework," Towards Data Science, January 20, 2023, https://towardsdatascience.com/learning-and-inference-in-a-forward-pass-the-new-framework-dc1356399002, accessed May 30, 2023.

[104] Sopan Deb, "Saying 'Thank You' to ChatGPT Is Costly. But Maybe It's Worth the Price." *New York Times*, April 24, 2025, https://www.nytimes.com/2025/04/24/technology/chatgpt-alexa-please-thank-you.html, accessed June 3, 2025.

[105] Dylan Patel and Afzal Ahmad, "The Inference Cost of Search Disruption—Large Language Model Cost Analysis," SemiAnalysis.com, February 9, 2023, https://www.semianalysis.com/p/the-inference-cost-of-search-disruption, accessed March 24, 2023.

[106] Danny Goodwin, "Google Now Sees More Than 5 Trillion Searches Per Year," Search Engine Land, March 3, 2025, https://searchengineland.com/google-5-trillion-searches-per-year-452928, accessed June 12, 2025.

[107] The Stanford Institute for Human-Centered Artificial Intelligence, "The 2025 AI Index Report," https://hai.stanford.edu/ai-index/2025-ai-index-report, accessed June 12, 2025.

[108] Edward Zitron, "There Is No AI Revolution," Where's Your Ed At?, February 24, 2025, https://www.wheresyoured.at/wheres-the-money/, accessed June 5, 2025.

[109] Tony Redmond, "Microsoft Cloud Revenues Powered by Office 365," Office 365 IT Pros, January 31, 2024, https://office365itpros.com/2024/01/31/office-365-reaches-400-million/, accessed June 12, 2025; Todd Spangler, "Netflix Q1 Results Top Expectations as Streaming Stops Reporting Subscriber Counts," Variety, April 17, 2025, https://variety.com/2025/tv/news/netflix-q1-2025-earnings-financial-results-subscriber-counts-1236371830/, accessed June 12, 2025; Spotify, "About Spotify," For the Record Newsroom, https://newsroom.spotify.com/company-info/, accessed June 12, 2025.

[110] Statista, "Most Popular Websites Worldwide as of November 2024, by Unique Visitors," https://www.statista.com/statistics/1201889/most-visited-websites-worldwide-unique-visits/, accessed June 16, 2025.

[111] Kif Leswing, "Nvidia Dominates the AI Chip Market, But There's More Competition Than Ever," CNBC, June 2, 2024, https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html, accessed May 21, 2025.

[112] Jonathan Vanian, "Mark Zuckerberg Indicates Meta Is Spending Billions of Dollars on Nvidia AI Chips," CNBC, January 18, 2024, https://www.cnbc.com/2024/01/18/mark-zuckerberg-indicates-meta-is-spending-billions-on-nvidia-ai-chips.html, accessed May 21, 2025.

[113] Jeremy Kahn, "60 Direct Reports, But No 1-on-1 Meetings: How an Unconventional Leadership Style Helped Jensen Huang of Nvidia Become One of the Most Powerful People in Business," Fortune, November 12, 2024, https://fortune.com/2024/11/12/jensen-huang-nvidia-ceo-leadership-mpp/, accessed May 21, 2025.

[114] Reuters, "Nvidia's Market Value Gets $2 Trillion Boost in 2024 on AI Rally," January 2, 2025, https://www.reuters.com/technology/artificial-intelligence/global-markets-marketcap-pix-2025-01-02/, accessed May 21, 2025.

[115] Mark Lohmeyer, "Announcing the General Availability of Trillium, Our Sixth-Generation TPU," Google Cloud Blog, December 11, 2024, https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga?utm_source=keywordsnippet&utm_medium=referral, accessed May 21, 2025.

[116] Alex Reisner, "The Unbelievable Scale of AI's Pirated-Books Problem," The Atlantic, March 20, 2025, https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/, accessed May 21, 2025.

[117] Emilia David, "Vox Media and The Atlantic Sign Content Deals with OpenAI," The Verge, May 29, 2024, https://www.theverge.com/2024/5/29/24167072/openai-content-copyright-vox-media-the-atlantic, accessed May 21, 2025.

[118] Kate Knibbs, "Every AI Copyright Lawsuit in the US, Visualized," Wired, December 19, 2024, https://www.wired.com/story/ai-copyright-case-tracker/, accessed May 21, 2025.

[119] Anna Tong, Echo Wang and Martin Coulter, "Exclusive: Reddit in AI Content Licensing Deal with Google," Reuters, February 21, 2024, https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/, accessed May 21, 2025.

[120] Angie Basiouny, "Who Benefits in the Deal Between Reddit and OpenAI?," Knowledge at Wharton, July 16, 2024, https://knowledge.wharton.upenn.edu/article/who-benefits-in-the-deal-between-reddit-and-openai/, accessed May 21, 2025.

[121] Meghan Bobrowsky, "Reddit Sues Anthropic, Alleges Unauthorized Use of Site's Data," *Wall Street Journal*, June 4, 2025, https://www.wsj.com/tech/ai/reddit-lawsuit-anthropic-ai-3b9624dd, accessed June 10, 2025.

[122] Miles Kruppa and Sam Schechner, "How Google Became Cautious of AI and Gave Microsoft an Opening," *Wall Street Journal*, March 7, 2023, https://www.wsj.com/tech/ai/google-ai-chatbot-bard-chatgpt-rival-bing-a4c2d2ad, accessed May 21, 2025.

[123] Reece Rogers, "Google AI Overviews Says It's Still 2024," Wired, May 29, 2025, https://www.wired.com/story/google-ai-overviews-says-its-still-2024/, accessed June 2, 2025.

[124] Nina Totenberg, "Supreme Court Showdown for Google, Twitter, and the Social Media World," NPR, February 21, 2023, https://www.npr.org/2023/02/21/1157683233/supreme-court-google-twitter-section-230, accessed May 21, 2025.

[125] Alphabet Inc. June 30, 2025 Form 10-Q, https://abc.xyz/investor/, accessed August 19, 2025.