HARVARD | BUSINESS | SCHOOL

9-723-434
REV: JUNE 27, 2023

ANDY WU
MATT HIGGINS
MIAOMIAO ZHANG
HANG JIANG

AI Wars

Although ChatGPT still has plenty of room for improvement, its release led Google's management to declare a "code red." For Google, this was akin to pulling the fire alarm. Some fear the company may be approaching a moment that the biggest Silicon Valley outfits dread — the arrival of an enormous technological change that could upend the business.

— Nico Grant and Cade Metz in *The New York Times*, December 21, 2022

Now Google, the company that helped pioneer the modern era of artificial intelligence, finds its cautious approach to that very technology being tested by one of its oldest rivals. Last month Microsoft Corp. announced plans to infuse its Bing search engine with the technology behind the viral chatbot ChatGPT, which has wowed the world with its ability to converse in humanlike fashion.

— Miles Kruppa and Sam Schechner in *The Wall Street Journal*, March 7, 2023

In November 2022, amidst a growing consensus that generative artificial intelligence (AI) would shape the next generation of computing, San Francisco-based research lab OpenAI released ChatGPT, an AI-powered text-generating chatbot that was made available to the public at chat.openai.com. The release captured widespread attention, and in the months that followed OpenAI appeared to leap into the commercialization lead by offering a range of new generative AI products as both user-facing applications and developer facing application programing interfaces (APIs).

In January 2023, Microsoft and OpenAI signed a $10 billion deal extending their exclusive partnership on the underlying technology, Generative Pretrained Transformer (GPT), a type of large language model (LLM) built with transformer-based neural network architecture and capable of generating new text that was largely indistinguishable from human writing. Microsoft would continue to supply the startup with effectively unlimited computing power from its Azure cloud, while OpenAI's technology and brand recognition would help keep Microsoft at the center of the new generative AI boom. Microsoft signaled that it would soon begin deploying OpenAI's technologies throughout its suite of products, from its Microsoft 365 productivity apps to its search engine Bing.[1]

By July of 2023, Silicon Valley was looking to Google to see how the search giant and putative technical leader in AI would respond. Google researchers had invented the transformer architecture that made the generative breakthroughs demonstrated by GPT possible. Google had a decade of experience developing and deploying AI and machine learning (ML) technologies in its products, but much of their AI work happened in-house and behind the scenes. Breakthroughs in AI had been quietly supercharging Google products like Search and Ads for years, but most of the product work was internal and little of it had penetrated the public consciousness. In fact, until 2022, Google's leadership had been deliberately cautious about revealing the extent of their AI progress and opening Google's experimental AI tools to the public. Was generative AI really ready for user-facing applications? Was the public, not to mention the Google PR department, ready for the changes and controversies that more visible and active AI might unleash? What did Google have to gain, or lose, in this opening salvo of the AI wars? Most pressingly, how would Google respond to OpenAI and Microsoft's moves to commercialize artificial intelligence in the form of new generative products and services in what was becoming the biggest big tech narrative of 2023?

Google

Google's homegrown AI project, Google Brain, was started in 2011 as an exploratory collaboration between Google fellow Jeff Dean, Google researcher Greg Corrado, and visiting Stanford professor of computer science Andrew Ng.[2] Google Brain initially focused on the development of neural networks as a general-purpose AI technology. Google Brain developed DistBelief, a proprietary internal machine learning system for efficiently training deep learning neural networks.[3] Google internally refined DistBelief over time until finally releasing it to the public as an open-source developer platform, TensorFlow, in 2015.[4] TensorFlow was instrumental in the development of deep learning neural networks, both inside and outside Google. For years, TensorFlow was the most popular tool for artificial intelligence (AI) and machine learning (ML) applications in the world.

Since its inception, Google Brain's research and approach to AI had never been far from Google's products. A list of which Google products made use of Google Brain's AI and ML breakthroughs, or any details about how they were implemented, was not public knowledge, but public research papers and blog posts documented the use of machine learning in products such as Google Translate[5] and Google Maps,[6] among others.

Google Brain was not Google's only AI interest. Since 2011, Google acquired a number of AI companies, some rolling into existing teams at Google and others operating as subsidiaries. In March 2013, Google acquired DNNresearch, a deep neural networks startup founded by University of Toronto professor Geoffrey Hinton[a], one of the pioneering academics of the deep learning approach.[7] In April 2013, Google acquired Wavii, an iPhone app that used natural language processing and machine learning to convert content from the web into structured semantic knowledge by topic, after a reported bidding war with Apple.[8] Google continued to acquire startups with AI and ML expertise over the years, more than 30 since 2009, with AI-related acquisitions totaling over $3.8 billion in 2020.[9]

In January 2014, Google acquired UK-based AI lab DeepMind for over $500 million.[10] DeepMind, which was known for using games to test and train its AIs, made headlines when its AlphaGo program beat a human world-champion at Go—a complex strategy board game sometimes likened to chess—in 2016. DeepMind operated as an independent Alphabet company, organizationally distinct from the division that housed Brain, until April 2023 when the two were combined under the Google Research umbrella.[11]

---

[a] In May 2023, Hinton resigned from his role at Google to speak out about the dangers of the technology. In an interview, Hinton said, "I don't think they should scale this up more until they have understood whether they can control it."

In addition to creating the TensorFlow framework, Google made a number of important advances in the area of natural language processing (NLP), large language models (LLM), and pre-training tools that laid the groundwork for Generative Pre-trained Transformers (GPTs). In 2017, Google introduced a new network architecture, the Transformer, that relied on a new attention mechanism to train neural networks, "dispensing with [computationally expensive] recurrence and convolutions entirely."[12] In 2018, Google open-sourced BERT (Bidirectional Encoder Representations from Transformers), a technique for NLP pre-training that has been widely adopted by subsequent LLMs.[13] Transformers proved to be a crucial step in the emergence of high-quality LLM-powered chatbots such as GPT-3.[14]

Following OpenAI's 2020 announcement that GPT-3 would be licensed exclusively to Microsoft, Google ramped up its public work on LLMs.[15,b] In April 2022, Google researchers published the 540-billion parameter Pathways Language Model (PaLM) trained using a Google Brain system.[16]

In late 2022, Google invested $300 million to acquire 10% of the AI startup Anthropic and secure Anthropic's commitment to use Google Cloud as a preferred cloud provider.[17] Anthropic was founded as a public benefit corporation in 2021 by a group of former OpenAI employees led by Dario Amodei. Amodei, previously OpenAI's head of AI safety, and the rest of Anthropic's founding team reportedly left OpenAI because of "differences over the group's direction after it took a landmark $1bn investment from Microsoft in 2019."[18] Anthropic called its chatbot Claude.

In January 2023, reports indicated that Google would introduce a suite of generative AI products over the coming months.[19] Google added a mission statement to its AI website that summarized its view that the technology should be used conservatively and in open collaboration with others (see **Exhibit 1**). In March 2023, Google launched its new chatbot, Bard (see **Exhibit 2**).[20]

Competitors

OpenAI

Founded as a non-profit research institute in 2015, OpenAI set out to advance artificial general intelligence (AGI) "in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." OpenAI's founding research director was machine learning expert Ilya Sutskever, formerly of Google, and the group's founding co-chairs were Sam Altman and Elon Musk. At its founding, OpenAI received $1 billion in commitments from a group of prominent Silicon Valley investors and companies, including Altman, Musk, Reid Hoffman, Jessica Livingston, Peter Thiel, Amazon Web Services, Infosys, and YC Research.[21]

In February 2019, OpenAI released Generative Pre-trained Transformer 2 (GPT-2), a language model that could learn new tasks (such as composing "original" text in a particular style) through the use of self-attention, building on the experimental capabilities of the never-publicly-released GPT-1. Despite being a major breakthrough in the field of language modeling, few took notice outside of AI research circles.

In March 2019, OpenAI announced that it would restructure into two organizations: OpenAI Nonprofit, would remain a 501(c)(3) operating under its original charter, and OpenAI LP, a "capped-

---

[b] In October 2021, Google researchers published GLaM (Generalist Language Model) with 1.2 trillion parameters, approximately seven times larger than GPT-3. In January 2022, Google researchers published LaMDA (Language Models for Dialog Applications), a family of Transformer-based neural language models specialized for dialog with up to 137 billion parameters pre-trained on 1.56 trillion words of public dialog data and web text.

profit" partnership would be overseen by the nonprofit. The new structure intended to attract new investors without compromising its mission: profits for investors in the LP would be capped at 10 percent of investment.[22] In July 2019, Microsoft announced it would invest $1 billion in OpenAI and become the exclusive cloud provider for OpenAI, collaborating on a hardware and software platform to incorporate AGI within Microsoft Azure.[23]

In June 2020, the GPT-3 API was opened (in private beta) to select researchers.[24] Within weeks of its release, GPT-3 established itself as the most powerful and useful among large language models (LLM) and the first one to offer a glimpse of mainstream usability through a public-facing API.[25] The release of GPT-3 in the summer of 2020 marked a turning point in the development of language-based artificial intelligence. Shortly thereafter, the term "generative AI" began to appear regularly in the media.

In September 2020, Microsoft announced that it would exclusively license OpenAI's GPT-3 model. The terms of the exclusivity were such that OpenAI could continue to offer third-party developers input and output through its public-facing API, but only Microsoft would have access to the back end and be able to use GPT-3's data model and underlying code in its products.[26] To the community of AI researchers and observers who had hoped OpenAI would choose to open source the model, the Microsoft deal was both a disappointment and a confirmation of their worst fears about the change in OpenAI's non-profit status. One analyst headlined his post on the deal as, "How OpenAI Sold Its Soul for $1 Billion."[27] A reporter wrote: "It's not clear exactly how or if OpenAI's 'capped profit' structure will change things on a day-to-day level for researchers at the entity. But generally, we've never been able to rely on venture capitalists to better humanity."[28]

OpenAI launched ChatGPT, a chatbot interface built on GPT-3.5[c] that anyone could use, in November 2022. OpenAI reportedly spent more than $540 million developing ChatGPT in 2022 alone, a figure that reflected the high costs of training new large language models and helped explain OpenAI's continued reliance on Microsoft for computing power and cash.[29]

In January 2023, OpenAI and Microsoft announced a "multiyear, multibillion dollar" extension to their partnership with a new investment from Microsoft.[30] Terms of the deal were not disclosed, but Microsoft's investment was widely reported to be worth $10 billion, and rumors circulated that Microsoft would receive 75 percent of OpenAI's profits until it secured its investment return and a 49 percent stake in the company.[31] Microsoft would also become the exclusive cloud partner for OpenAI going forward and would begin deploying OpenAI's models in its enterprise products immediately. "In this next phase of our partnership, developers and organizations across industries will have access to the best AI infrastructure, models, and toolchain with Azure to build and run their applications," said Microsoft Chairman and CEO Satya Nadella.[32] Within months, Microsoft had deployed some of OpenAI's technology in its Bing search engine and announced plans to roll out more AI features across its portfolio of products. "The expectation from Satya is that we're pushing the envelope in A.I., and we're going to do that across our products," said Eric Boyd, the executive responsible for Microsoft's AI platform team, in an early 2023 interview.[33]

As OpenAI deepened its ties with Microsoft and pushed forward on the commercialization of generative AI, its transformation from non-profit to for-profit status rankled critics, including at least one of its original donors. In March 2023, Elon Musk tweeted: "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?"[34]

---

[c] OpenAI's major releases were numbered GPT-n: GPT-2 (February 2019), GPT-3 (June 2020), and GPT-4 (March 2023). GPT-3.5 was an informal designation, not an official release, reflecting the improvements made to the model between 2020 and late 2022.

OpenAI offered a number of its technologies to third-party developers through its API service. As of early 2023 these included Access GPT-3, DALL-E 2 (prompt-based image generation), and Codex (a set of tools for converting natural language to code). Several third-party applications had already begun building on OpenAI's platforms: GitHub Copilot (owned by Microsoft) drew on the Codex platform to create a powerful predictive autocomplete tool for programmers, and Duolingo used GPT-3.5 to interpret user input and provide French grammar corrections in its language instruction app.[35]

In March 2023, OpenAI released GPT-4 (see **Exhibit 3**). This next-generation model demonstrated improved conversational abilities, responsiveness to user steering, potential for image-based inputs, and safety precautions to prevent harmful advice or inappropriate content. With the release of GPT-4, OpenAI published benchmarks on the relative performance of GPT-3.5 and GPT-4 on a range of standardized tests (the Uniform Bar Exam, LSAT, GRE, and topic-specific AP tests from Chemistry to English Literature). In most tests, GPT-4 outperformed all other models, often significantly, achieving what OpenAI called "human-level performance on various professional and academic benchmarks."[36] Behind the scenes, GPT-4 was reported to be more computationally efficient and cost-effective than its predecessor, gains that OpenAI had presumably achieved through advances in training techniques and model architecture. Researchers reading through the technical documentation of GPT-4 in search of information on how those gains were achieved—a common practice with the release of a new model in the pre-commercial days of AI research labs—found only the following passage:

> Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.[37]

OpenAI Chief Scientist Ilya Sutskever underscored OpenAI's shift toward closed models, proprietary training methods, and the non-disclosure of training data in an interview:

> We were wrong. Flat out, we were wrong. If you believe, as we do, that at some point, AI – AGI – is going to be extremely, unbelievably potent, then it just does not make sense to open-source. It is a bad idea... I fully expect that in a few years it's going to be completely obvious to everyone that open-sourcing AI is just not wise.[38]

Microsoft

Microsoft had been working on the natural language component of artificial intelligence since the founding of Microsoft Research in 1990. The internal research division made an immediate splash by hiring away three of the top computational linguists of the era from rival IBM to start its NLP research group. Within a few years, the Microsoft had become a world leader in the development of grammar detection, spell check, and automatic translation tools.[39]

Advances in machine learning picked up when the cloud era got underway in the early 2010s. Satya Nadella was promoted to president of the Server and Tools Division in 2011, the division where Microsoft's then-nascent cloud initiative, Azure, was housed. In February 2014, Nadella took over the CEO role from his predecessor Steve Ballmer. That summer, Microsoft announced Azure ML, one of the first cloud services to offer a machine learning platform.[40] In the post-transformer deep-learning era (since 2017), Microsoft conducted advanced AI, ML, and LLM research primarily through its Turing program, a collaboration with academic researchers from around the world.[41] The Turing Natural Language Generation model (Turing-NLG), published in 2020, contained 17 billion parameters and outperformed other state-of-the-art models at the time.[42] In October 2021, Nvidia and Microsoft Research's Turing program combined their LLM efforts to publish Megatron-Turing NLG, the world's largest generative language model with 530 billion parameters.[43]

When Microsoft invested its first $1 billion in OpenAI in 2019, the headline was that Azure would become OpenAI's exclusive cloud provider.[44] One analyst noted, "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]

According to reporting from *The Information*, executives inside Microsoft were skeptical in 2019 that OpenAI would live up to the hype. Peter Lee, head of Microsoft Research, found it hard to believe that OpenAI could have accomplished in a few years what Microsoft researchers had been unable to do in a decade. Even Microsoft co-founder Bill Gates warned Nadella against the OpenAI investment. Nadella proceeded with caution, using the in-house Microsoft Research group to check OpenAI's work. Recounting the evolution of the relationship, the *Information* article continued:

> Over time, those doubts began to fade. When Microsoft researchers compared OpenAI's language models side by side with Microsoft's internal models, collectively dubbed Turing, it became undeniable that the startup had built something far more sophisticated.[46]

As the partnership progressed, Microsoft began to value OpenAI as more than just a big customer for Azure or a long-term R&D bet. The more confidence Nadella gained in OpenAI's generative AI capabilities, the more aggressively he pushed teams across the organization to integrate OpenAI's models into its products. Microsoft's subsequent investments, including a January 2023 investment reported to be worth $10 billion, reflected a growing confidence in its startup partner.[47]

Microsoft continued its own in-house AI efforts alongside integrating OpenAI technology. Microsoft researchers worked on large-scale models across the organization, including the user-facing implementations of GPT in Office (see **Exhibit 4**) and Bing, and the less user-facing machine-learning infrastructure being built in Azure. Prometheus, a Microsoft-developed model that helped merge the search and chat functions in Bing, was released in February 2023.[48] An article in *Wired* described the ongoing internal work at Microsoft:

> Microsoft has held back from going all-in on OpenAI's technology. Bing's conversational answers do not always draw on GPT-4, Ribas [Microsoft's CVP of Search and AI] says. For prompts that Microsoft's Prometheus system judges as simpler, Bing chat generates responses using Microsoft's homegrown Turing language models, which consume less computing power and are more affordable to operate than the bigger and more well-rounded GPT-4 model.[49]

Meta

Since its founding in 2013, Facebook AI Research (FAIR), later Meta AI, had been led by Yann LeCun, a French-American AI researcher and professor of computer science at NYU. LeCun and other Facebook researchers played an important role in advancing the theoretical model for Generative Adversarial Networks (GANs), pitting two neural networks (a generator and a discriminator) against each other to create an artificial intelligence co-evolutionary arms race capable of doing things like generating novel text and realistic-looking images.[50] Throughout the 2010s, FAIR also made advances in self-supervised learning (SSL) on large unstructured data sets and rapid text classification, inventing a framework called fastText, a simplified approach to text classification that could run on basic inexpensive hardware.[51] During this time, Facebook steadily implemented FAIR's research in a range of internal applications from newsfeed ranking, to content moderation, to language translation, to computer vision and image recognition.

In August 2022, Meta made its chatbot prototype Blenderbot 3 available to the public. Although Blenderbot 3 was built on Meta's open-source OPT-175B, released in the midst of the early GPT-3 hype, and preceded ChatGPT by several months, the chatbot from Meta didn't garner the same widespread attention or enthusiasm. Researchers and reviewers found Blenderbot underwhelming compared to the early version of GPT-3, as *Vox* noted in its headline from August 2022, "Why is Meta's new AI chatbot so bad?"[52]

Like Google's extensive in-house AI work, much of what Meta did in AI was never made public, so an observer could only guess at the full extent of their work based on their published papers and public contributions to open-source projects. In November 2022, Meta AI announced CICERO, an AI agent that had achieved human-level performance at the strategy game Diplomacy. CICERO was a language model integrated with strategic reasoning algorithms that enable effective negotiation and cooperation with human players.[53]

On the open-source front, Meta developed and maintained PyTorch, a computing package and machine learning framework based on the open-source Torch package for Python. Like TensorFlow, PyTorch provided a pre-built set of developer tools that could be used to quickly set up and train deep-learning neural networks. Like Google's investment in TensorFlow and other company-supported open-source projects, Meta's investment in PyTorch ecosystem did not directly contribute to Meta's bottom line, but its availability had valuable second-order effects, generating goodwill and drawing the developer community to Meta's preferred toolset. Meta's PyTorch-based internal tools (which remained proprietary) nonetheless benefited from the rapid innovation open source made possible, strengthening Meta's appeal as a hub of innovation and an employer of top AI/ML engineers. In 2022, Meta transferred control of PyTorch to the Linux Foundation.[54]

In May 2022, Meta released OPT-175B (Open Pretrained Transformer), a large language model with 175 billion parameters under a non-commercial GPL 3.0 license. In February 2023, Facebook released LLaMA (Large Language Model Meta AI), also under a non-commercial GPL 3.0 license, and shared it with the AI research community. LLaMA was released in four sizes: 7 billion, 13 billion, 33 billion, and 65 billion parameters, making it more flexible for researchers with different computational capacities. LLaMA's 13 billion-parameter model outperformed GPT-3 on most benchmarks, and LLaMA-65B was competitive with the best LLMs in the world. [55] Meta also released details about how the model had been built and trained, including model weights that were proprietary for comparable models at OpenAI and Google. Though LLaMA was not released as open source, copies of it leaked shortly after it was shared with researchers. Within a few days, LLaMA had effectively become open source.[56] The relative openness of LLaMA made it possible for AI researchers outside of the biggest labs to examine how a genuinely advanced LLM was constructed.

As of May 2023, Meta AI did not offer any commercial API services to developers and had not announced any plans to do so. In fact, Meta looked to be doubling down on its approach of using proprietary internal tools built on open-source technologies, as they had done with PyTorch and Open Compute.[d] In Meta's Q1 2023 earnings call, CEO Mark 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.

---

[d] Open Compute was a collaborative community project seeded by Facebook in 2009 to design a more energy-efficient data center. In 2011, the Open Compute Project Foundation was founded to "apply the benefits of open source and open collaboration to [networking and data center] hardware." (opencompute.org/about).

> 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 during the Q&A, 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 playing a different game on the infrastructure than companies like Google or Microsoft or Amazon, and that creates different incentives for us.[57]

Stability AI

Stability AI was founded in 2019 by British hedge fund manager-turned-AI-entrepreneur Emad Mostaque. Stability AI's flagship product, the open-source image generator Stable Diffusion, was announced to the public in August 2022.[58] In October 2022, Stability AI raised $101 million in a funding round led by Coatue and Lightspeed Venture Partners with participation from O'Shaughnessy Ventures at a $1 billion post-money valuation.[59] In an interview with *TechCrunch*, Mostaque emphasized Stability AI's independence: "Nobody has any voting rights except our employees—no billionaires, big funds, governments or anyone else with control of the company or the communities we support. We're completely independent."[60]

Stability AI was not entirely independent of big tech at the technical level—cloud computing was an expensive necessity for AI development, and Stability ran a cluster of over 4,000 Nvidia A100 GPUs on Amazon Web Services (AWS). In November 2022, Stability AI announced that AWS would be its preferred cloud provider and that Stability AI's open-source models would be made available through Amazon SageMaker, AWS's end-to-end machine learning service.[61] Though Stability AI was small and the commercialization of generative AI was early, the emergence of Stable Diffusion as a credible open-source alternative to OpenAI's proprietary DALL-E 2 made it valuable to several companies experimenting with image generation.

In December 2022, Apple's machine learning group announced that it had created its own custom Stable Diffusion model and was optimizing it to run on-device with the Apple Neural Engine, a processor customization available only on Apple Silicon.[62] Apple's quick move to optimize its own processors for on-device machine learning capabilities suggested to some analysts the possibility of a world in which custom processors and local hardware optimizations could become a factor in the emerging market for AI and ML products.[63]

Hugging Face

Hugging Face was founded by Clément Delangue, Julien Chaumond, and Thomas Wolf in 2016. Delangue, based in Paris, had previously sold a machine-learning-based image-recognition technology company to Google. Hugging Face's first product was a chatbot aimed at engaging teenagers, but the company soon pivoted to focus on building a platform for natural language processing technologies, becoming the go-to Github repository and a key tool for NLP developers. Hugging Face's Transformers library was widely adopted by AI and ML researchers as early as 2018, carving out a niche in the back end of almost every chat-based product that emerged after it. In May 2022, Hugging Face raised a $100 million Series C round led by Lux Capital with Sequoia and Coatue participating.[64]

Hugging Face's largely open-source approach appeared to be modeled after GitHub, originally an open-source developer-tools and developer-community startup that flourished by making a free tool so useful that it became a de facto industry standard. After GitHub built a userbase of over 27 million developers, Microsoft acquired it in 2018 for $7.5 billion.

Its flagship product HF Hub was a free and open-source community where AI developers could host their models, datasets, and private repositories. Hugging Face also offered paid products including Spaces Hardware (a range of compute-on-demand services), Inference Endpoints (a scalable production environment easy to deploy from HF Hub), Autotrain (a tool to automatically train, evaluate, and deploy ML models from existing data), Pro (a monthly subscription with exclusive features for the HF community), and Enterprise (bespoke services for enterprise customers).[65]

Hugging Face led the development of 176B Bloom, an open-source autoregressive LLM created by the BigScience project, a volunteer community comprised of more than 1,000 researchers from 60 countries and more than 250 institutions.[66] Bloom was trained on 1.6TB of pre-processed text, about 350 billion unique tokens.[e] As impressive as Bloom was, one developer who assessed Bloom vs. GPT-3 on a range of tasks concluded: "GPT-3 performed far better than Bloom in every single test, making it the obvious choice if you're looking for an AI model to help with your project."[67]

Business Models for AI

Google had several options for responding to Microsoft and OpenAI's recent moves to commercialize generative AI. The first, and most obvious, was to release a competing standalone chatbot. Google already had the core generative AI technology in place and could move quickly to bring a new customer-facing product to market, provided it could meet Google's high standards for trust and safety. Another way of bringing generative AI into Google products would be to implement it into the existing results page of Google Search, a delicate proposition considering that the results page was already heavily monetized, and heavily negotiated, territory. Google could also insert generative AI tools into the sidebars and menus of its suite of productivity software, Google Workspace. Outside of consumer-facing applications, Google could appeal to organizations interested in AI by selling services, such as managed hosting and APIs for developers or by licensing models to developers. Although new technologies are often adopted before the business model is clear, any business model for generative AI would have to account for the enormous costs of running cutting-edge generative models on the cloud, costs that even a company as large as Google had to take seriously.

Applications for End Users

Chatbot A straightforward path for Google would be to launch a standalone chatbot application. ChatGPT, initially based on the GPT-3.5 model, wowed the world upon its release in November 2022. ChatGPT was free to use on the web as part of what OpenAI called a "research preview," under which free access was limited and subject to demand restrictions (a paid tier with unlimited access, ChatGPT Plus, was soon introduced for $20 a month). In January 2023, just two month after its release, ChatGPT had reached 100 million active users, the fastest app ever to reach that

---

[e] A token was a semantic unit for processing text that could vary in size or length depending on the training objective. A very general rule of thumb was that one thousand tokens were equivalent to roughly 750 words.

milestone.[68] On the back of the surprisingly strong demand in December 2022, OpenAI projected that ChatGPT would make $200 million in 2023 and $1 billion in 2024.[69] ChatGPT Plus, a premium version of ChatGPT with enhanced features and priority access, was introduced in February 2023 as a $20 a month subscription.[70]

March 2023 was a busy month in chatbot news: Google began rolling out limited access to its own chatbot, Bard,[71] Chinese search giant Baidu released Ernie (Enhanced Representation through Knowledge Integration),[72] and a Google-backed chatbot startup, Anthropic, launched a chatbot called Claude.[73] ChatGPT, Ernie, Bard, and Claude were intended as general use tools, but a number of other AI chatbot startups had chosen to focus on specific use cases, ranging from psychotherapy (Woebot) to healthcare diagnosis (Ada) to virtual companionship (Replika).

In February 2023, Replika—which had advertised its NSFW chat features before deciding to move the service in a different direction—removed the ability of its AI companions to send romantically-charged and sexually-suggestive messages. This sudden change in boundaries disappointed some users, breaching the continuity of the relationship and generating a round of light news stories about lost AI lovers and the real-life pain some users experienced as a result.[74]

As competitors rolled out their own general purpose chatbots, OpenAI was busy adding the infrastructure to make ChatGPT into a platform. OpenAI announced in March 2023 that ChatGPT would support plug-ins, allowing the app to tap into external sites for additional information or capabilities (see **Exhibits 5** and **6**).[75] Third-party sites like Instacart, Expedia, OpenTable, and Wolfram signed on to create plug-ins that would grant ChatGPT access to their systems. ChatGPT alone could generate a recipe; ChatGPT with Instacart could plan a meal and have all the ingredients delivered to your doorstep. The prospect of plug-ins created an interesting new potential source of revenue for OpenAI: the sale of ChatGPT's defaults for various functional categories.

Amidst the chatbot hype, important questions remained for Google. Would generative chatbots be embraced as a general-purpose application? Would Google risk its reputation when its chatbot provided inaccurate or even offensive responses? And perhaps more importantly, it was unclear whether a chatbot would be a viable business. Would regular customers be willing to pay a monthly subscription for such an application? Or would there be a way to integrate advertising and monetizable analytics into chat responses?

Search Google also needed to consider whether to respond to Microsoft's challenge in search. Microsoft's Bing was the first to implement generative AI features directly in its search application, raising questions about the future of the search engine space. The global search engine market—a range of technologies for querying, sorting, and displaying relevant information to internet searchers (including advertising)—was valued at $167 billion in 2021.[76] As of December 2022, Google handled 61.4% of internet searchers in the U.S., with Microsoft handling 29.6% and Yahoo 11%.[77] Globally, Google had a 93.4% share of the desktop search market; Bing was a distant second at 2.8%.[78] Microsoft had a lot to gain from a change in the way people sought information online, and Google had just as much (or more) to lose.

However, Google faced two key issues that made it tough to decide whether and how to integrate generative AI into search. First, generative AI would make each search query dramatically more expensive. One analyst estimated that ham-fisting the technology into every Google search would incur $100 billion in capital expenditures alone for server and networking hardware. If Google had the same operating costs as OpenAI, each Google search would cost 0.36 cents more per query (on top of its existing cost of 1.41 cents per query), leading to an increase $694,444 in daily operation costs if applied to every search.[79] Google would suffer a reduction of $36 billion in operating income by implementing this technology across the board for all Google search queries (**Exhibit 7** details the analyst estimate of the cost impact to Google Search).[80]

Second, it was unclear whether the lucrative search advertising business model would be effective if Google switched from delivering links (with ads embedded among those links) to delivering narrative chat responses. Google had long relied on AI and machine learning to optimize its ad-based business model, selling advertising alongside the organic search results that appeared on the results page, but the AI was secondary to the provision of links to the user. If search queries could be answered directly by a chat feature instead of a page of links and ads, it was unclear where ads would be placed and what advertisers would pay for those ads.

In an April 2023 interview, Google CEO Sundar Pichai indicated that chat functionality would eventually be integrated into Search:

> We've been using LLMs to improve search quality and the search experience, but I think we will bring natively the modern LLM capabilities in search. We are working to make sure it works well for users—they have a high bar, and we want to meet that bar. But yes, will people be able to ask questions to Google and engage with LLMs in the context of search? Absolutely.[81]

Productivity Software Google could integrate generative AI into its productivity applications in Google Workspace. The global office productivity software market—a bundle of applications typically including word processing, spreadsheets, presentations, email, calendar, notetaking, and basic databases—was valued at $23.5 billion in 2021.[82] In 2020 Microsoft 365 had a 54.7% share of the global enterprise productivity suite market, whereas Google Workspace had a 26.3% share.[83] Microsoft 365 was the number one bestselling SaaS application of all time. The "Business Standard" tier of Google Workspace was priced at $144 per user per year.[84] The equivalent tier of Microsoft 365 was priced at $150 per user per year.[85] For Microsoft, its long-standing suite of productivity applications was more than a huge source of revenue; Microsoft used 365 to cross-sell many of its other products like Azure cloud computing and a bevy of other enterprise applications.

Microsoft's vision for the integration of generative AI into its productivity software was laid out in Satya Nadella's keynote at "The Future of Work with AI" event in March 2023:

> You could say we've been using AI on autopilot, and now this next generation of AI, we're moving from autopilot to copilot... As we build this next generation of AI, we made a conscious design choice to put human agency both at a premium and at the center of the product. For the first time we have the access to AI that is as empowering as it is powerful.[86]

Throughout the event, Microsoft's examples and demos showed AI-powered copilots assisting humans with tasks in helpful and intuitive ways: triaging emails in Outlook, expanding bullet points into a written memo in Word, or analyzing and highlighting trends in an Excel spreadsheet in response to text-based queries such as "What contributed to the decline in sales growth this period?" Microsoft Business Chat, an enterprise chatbot integrated into Teams, can draw on data from across the organization's suite of Microsoft apps to surface data and insights in response to text queries, harnessing the power of stored organizational knowledge.[87]

API for Developers

As an alternative to or in addition to offering AI-infused applications to end users, Google could sell access to an API for third-party developers who would develop their own applications for users by accessing Google's AI. An Application Programming Interface (API) was a software feature for input and output that could be used by third-party developers to integrate functionality from the provider into their own applications or services. In this business model, the provider of the API charged a fee for access to the API, either on a subscription or usage-based model. The payments company Stripe and communication-tools company Twilio had risen to success using an API business model. In June 2020, OpenAI launched an API for third-party developers to use its models, noting in its announcement:

> In addition to being a revenue source to help us cover costs in pursuit of our mission, the API has pushed us to sharpen our focus on general-purpose AI technology—advancing the technology, making it usable, and considering its impacts in the real world. We hope that the API will greatly lower the barrier to producing beneficial AI-powered products, resulting in tools and services that are hard to imagine today.[88]

As influential as Google had been in the research and development of AI infrastructure, it did not offer any public-facing commercial API service for AI in 2021 or early 2022. In August 2022, Google began inviting select researchers to access some Google AI technologies through a limited API called AI Test Kitchen.[89] In early 2023, Google began opening some of its closed projects and services to outside users in limited beta. The sudden opening of Google's private AI garden was widely viewed as a competitive response to OpenAI's recent moves to publicize the technology in a consumer-facing way, opening the door to the possibility of a new market for generative AI applications and an API for developers.[90]

Developments throughout 2023 continued to create uncertainty about how AI would be monetized. In March, Stanford researchers released Alpaca 7B, an instruction-following model fine-tuned from Meta's open-source LLaMA 7B model. The Stanford team published complete details of their data generation and training processes, allowing anyone with an interest in how cutting-edge LLMs were created to follow and reproduce their work.[91] Upon the news, an AI researcher tweeted:

> I don't think people realize what a big deal it is that Stanford retrained a LLaMA model, into an instruction-following form, by **cheaply** fine-tuning it on inputs and outputs [from one of the most powerful versions of OpenAI GPT-3]. It means: 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…
>
> In other words: The AI companies that make profits will be ones that either have a competitive moat not based on the capabilities of their model, OR those which don't expose the underlying inputs and outputs of their model to customers, OR can successfully sue any competitor that engages in [model cloning].[92]

Licensing the Model to Developers

Finally, Google could license its underlying AI models to third-party developers under either a closed or open license.

Closed License When an entity retains exclusive control over its intellectual property, but grants access to customers with certain terms on how it can be used, this is a closed license. Closed-license business models focus on protecting the company's intellectual property and generating revenue from direct sales or licensing fees. The underlying code or technology is typically owned by the company selling the license (the licensor), and customers (the licensees) pay that company a fee to use it under terms and conditions spelled out in a licensing agreement.

Google could license its AI model to other companies and in essence serve as the OpenAI party in the OpenAI-Microsoft relationship but for all other companies not named Microsoft. Microsoft's deal with OpenAI was a combination of a closed license fee paid by Microsoft to OpenAI and a sizable strategic investment, some of which came in the form of cash and some of which came in the form of grants of Microsoft resources (such as Azure computing power). Though the full terms of the deal were not made public, Microsoft's September 2020 announcement specified that Microsoft was *exclusively* licensing OpenAI's GPT-3, restricting OpenAI's ability to license its model to other rival tech companies, but not restricting OpenAI's ability to sell access to the model via API.[93] If Google went down this route, it could decide to license broadly or exclusively to a few companies.

Open License When a licensor allows others to freely use, modify, and distribute its intellectual property (sometimes subject to certain conditions such as attribution or non-commercial use), this is an open license. Open licenses were popularized by the open-source software community, where interesting projects brought together diverse collectives of contributors who had a joint interest in maintaining and advancing the project for collective and societal gain. There were degrees of openness in open licenses: GPL (GNU General Public License) and AGPL (GNU Affero General Public License) inhabit the more restrictive end of the spectrum while Apache, BSD, and MIT occupy the more permissive end. Whichever license was used, open licenses made direct monetization more difficult. Competitors could simply copy the code; where a more restrictive license prohibited that, competitors could often reengineer the idea in a copyright-safe way from the available code. In spite of these challenges, open-source software grew in popularity as leaders of open-license projects found ways to build profitable businesses around them. Rather than monetizing the intellectual property directly through sales or licensing, open-license businesses monetized their work indirectly by offering premium products built on top of the open-licensed tool, or adjacent paid services such as consulting or support. Open-license products could scale quickly because of their low cost to users, and businesses could eventually monetize those larger user bases by offering additional products that the next generation of users would be willing to pay for. Google had experience doing this with past open-source projects like Android and TensorFlow. While anyone could use the code for TensorFlow under the Apache 2.0 license, using the software effectively requires a range of infrastructure that can be easily purchased through and deployed on Google Cloud Platform. The widespread availability of TensorFlow under a permissive open license contributed to Google's cloud services business.

Monetizing open source by driving customers to managed services, however, was not a foolproof strategy in the ultra-competitive world of cloud computing. By design, open-source tools can be used on any infrastructure, and Amazon's AWS offered a range of cloud computing tools that were competitive with, and often cheaper than, their Google Cloud or Microsoft Azure equivalents. Amazon SageMaker, a popular choice for deploying and training ML models in the cloud launched in 2017, was already established as the market leader in cloud-based ML model training and fine-tuning. The modular nature of AWS made it easy to add or subtract the infrastructure needed to store the data and train the model. Unless a company was already running on Google Cloud, using SageMaker on AWS for fine-tuning models usually made more sense.[94]

In May 2023 a leaked internal document from Google surfaced on the technology blog *SemiAnalysis* that showed the company grappling with the strong presence of open source in generative AI. The document, titled "We Have No Moat, and Neither Does OpenAI," was apparently an internal memo intended to influence Google's strategy for generative AI. The author laid out a number of propositions about the changing market for generative AI, chief among them that the biggest winner in the turn to open source that defined the spring of 2023 was Meta, for reasons that Google should have understood:

> 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 memo continues:

> 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.[95]

Conclusion

Large language models like ChatGPT required a lot of computing power to train and run. Though ChatGPT and products like it were sometimes presented to the public as "free," the cost of running them—and especially opening them to the public at scale—was substantial, even for a large tech company like Microsoft or Google. OpenAI spent millions of dollars each year to keep ChatGPT running. While OpenAI had opened a world of possibility by introducing generative LLMs to the public, its business model journey was far from complete. Over time, generative AI would need to mature into a revenue-generating business, just as cloud computing had over the previous decade, if the early excitement and investment were to continue. While there was no question about the vast potential value of generative AI, Google and its competitors faced serious questions about what the business and monetization of generative AI would look like.

Exhibit 1 Google's AI Philosophy (Excerpt)

Why we focus on AI (and to what end)

We're excited about the transformational power of AI and its helpful new applications…

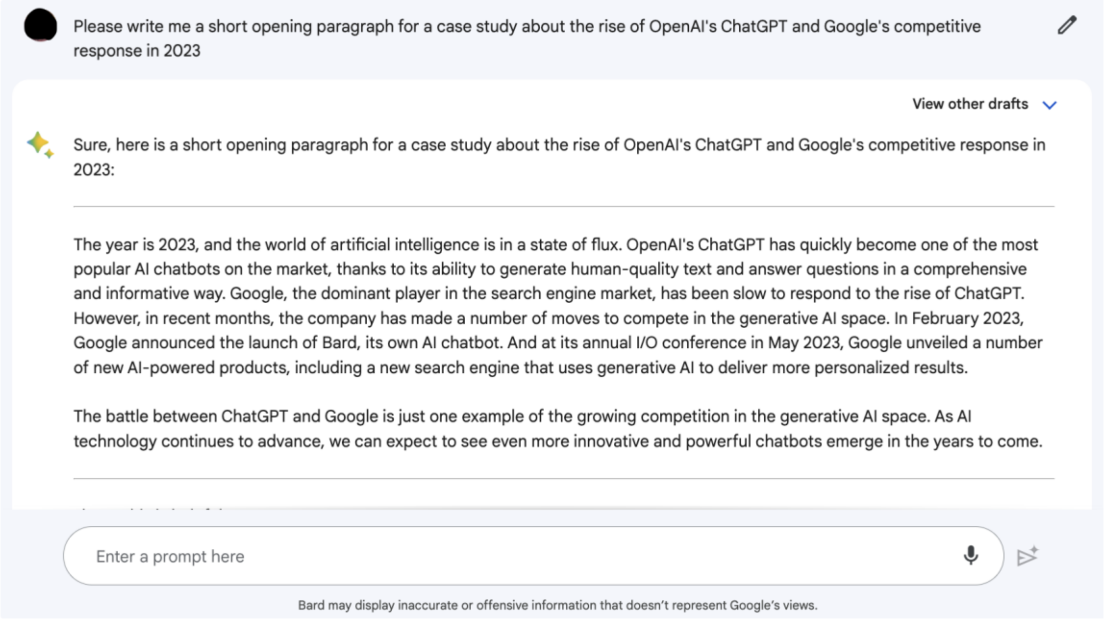
At the same time, we understand that AI, as a still-emerging technology, poses various and evolving complexities and risks. Our development and use of AI must address these risks. That's why we as a company consider it an imperative to pursue AI responsibly. We are committed to leading and setting the standard in developing and shipping useful and beneficial applications, applying ethical principles grounded in human values, and evolving our approaches as we learn from research, experience, users, and the wider community.

– James Manyika, Jeff Dean, Demis Hassabis, Marian Croak and Sundar Pichai

January 16, 2023

Source: Excerpted from Google AI, "Why we focus on AI (and to what end)," January 16, 2023, https://ai.google/our-focus/, accessed July 2023.

Exhibit 2 Google Bard Interface and Sample Output



Please write me a short opening paragraph for a case study about the rise of OpenAI's ChatGPT and Google's competitive response in 2023

Sure, here is a short opening paragraph for a case study about the rise of OpenAI's ChatGPT and Google's competitive response in 2023:

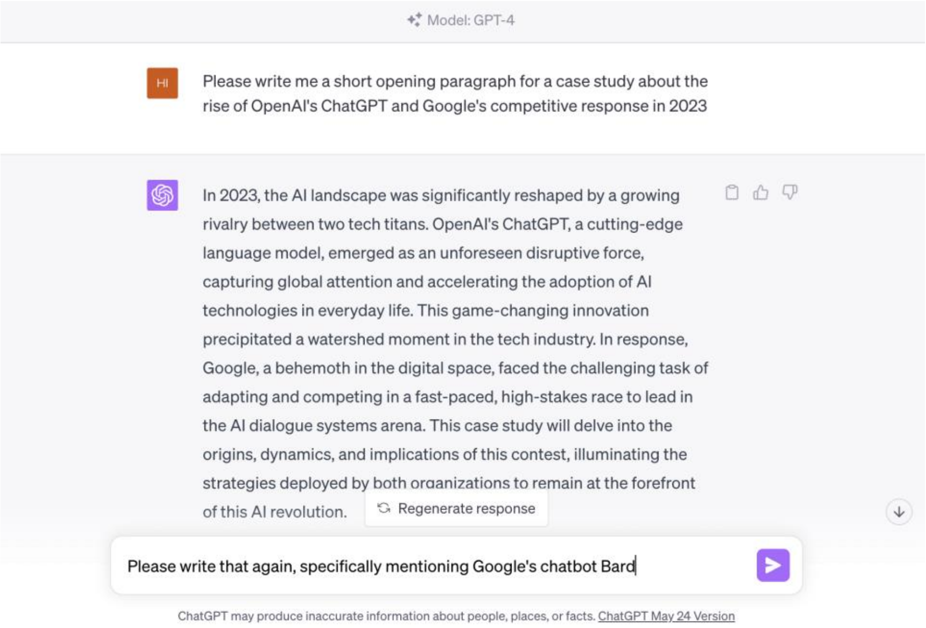
The year is 2023, and the world of artificial intelligence is in a state of flux. OpenAI's ChatGPT has quickly become one of the most popular AI chatbots on the market, thanks to its ability to generate human-quality text and answer questions in a comprehensive and informative way. Google, the dominant player in the search engine market, has been slow to respond to the rise of ChatGPT. However, in recent months, the company has made a number of moves to compete in the generative AI space. In February 2023, Google announced the launch of Bard, its own AI chatbot. And at its annual I/O conference in May 2023, Google unveiled a number of new AI-powered products, including a new search engine that uses generative AI to deliver more personalized results.

The battle between ChatGPT and Google is just one example of the growing competition in the generative AI space. As AI technology continues to advance, we can expect to see even more innovative and powerful chatbots emerge in the years to come.

Bard may display inaccurate or offensive information that doesn't represent Google's views.

Source: Casewriter screenshot, Google, "Bard," https://bard.google.com/, accessed June 2023.

Exhibit 3 ChatGPT Interface and Sample Output with GPT-4



Please write me a short opening paragraph for a case study about the rise of OpenAI's ChatGPT and Google's competitive response in 2023

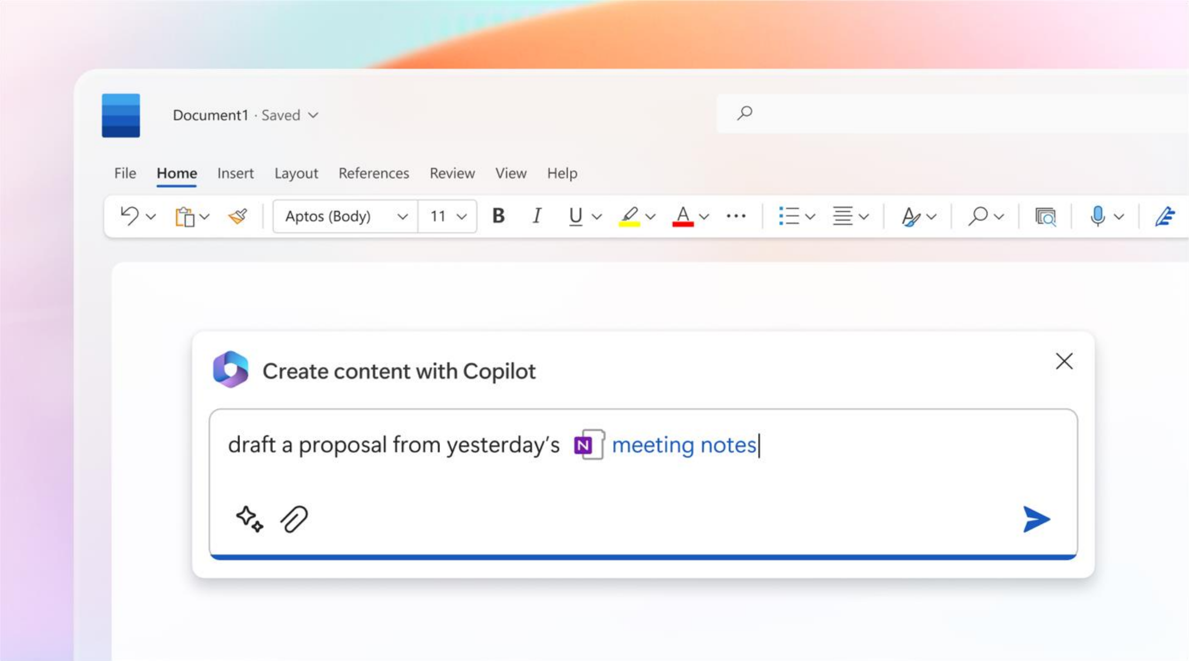
In 2023, the AI landscape was significantly reshaped by a growing rivalry between two tech titans. OpenAI's ChatGPT, a cutting-edge language model, emerged as an unforeseen disruptive force, capturing global attention and accelerating the adoption of AI technologies in everyday life. This game-changing innovation precipitated a watershed moment in the tech industry. In response, Google, a behemoth in the digital space, faced the challenging task of adapting and competing in a fast-paced, high-stakes race to lead in the AI dialogue systems arena. This case study will delve into the origins, dynamics, and implications of this contest, illuminating the strategies deployed by both organizations to remain at the forefront of this AI revolution.

Please write that again, specifically mentioning Google's chatbot Bard

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Source: Casewriter screenshot, OpenAI, "ChatGPT," https://chat.openai.com/chat, accessed June 2023.

Exhibit 4 Microsoft Office Copilot Demo

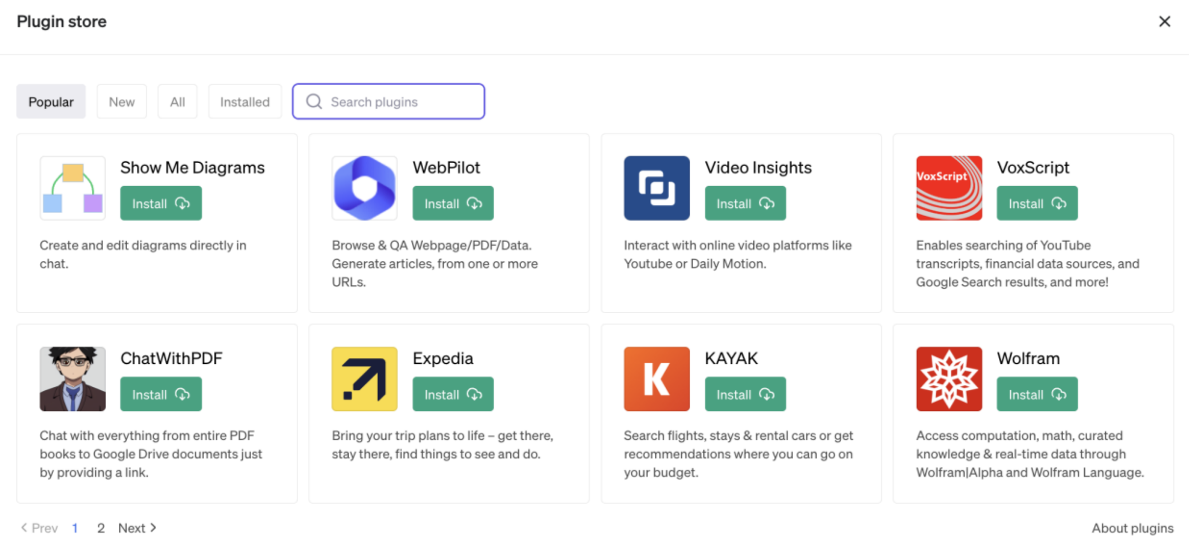


Create content with Copilot

draft a proposal from yesterday's meeting notes

Source: Microsoft, "Introducing Microsoft 365 Copilot — Your Copilot for Work," March 16, 2023, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/.

Exhibit 5 ChatGPT Plugin Store



Show Me Diagrams
Create and edit diagrams directly in chat.

WebPilot
Browse & QA Webpage/PDF/Data. Generate articles, from one or more URLs.

Video Insights
Interact with online video platforms like Youtube or Daily Motion.

VoxScript
Enables searching of YouTube transcripts, financial data sources, and Google Search results, and more!

ChatWithPDF
Chat with everything from entire PDF books to Google Drive documents just by providing a link.

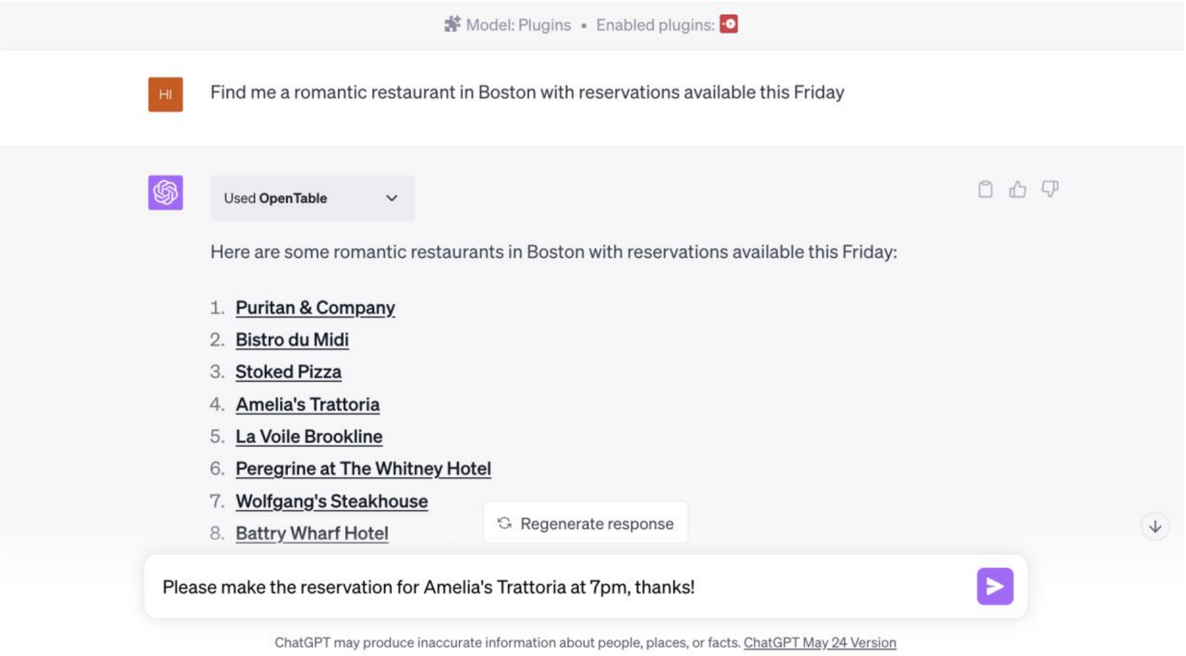
Expedia
Bring your trip plans to life – get there, stay there, find things to see and do.

KAYAK
Search flights, stays & rental cars or get recommendations where you can go on your budget.

Wolfram
Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.

Source: Casewriter screenshot, OpenAI, "ChatGPT Plugin Store," https://chat.openai.com/?model=gpt-4-plugins, accessed June 2023.

Exhibit 6 ChatGPT's OpenTable Plugin in Action



Find me a romantic restaurant in Boston with reservations available this Friday

Here are some romantic restaurants in Boston with reservations available this Friday:

1. Puritan & Company
2. Bistro du Midi
3. Stoked Pizza
4. Amelia's Trattoria
5. La Voile Brookline
6. Peregrine at The Whitney Hotel
7. Wolfgang's Steakhouse
8. Battry Wharf Hotel

Please make the reservation for Amelia's Trattoria at 7pm, thanks!

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

Source: Casewriter screenshot, OpenAI, "ChatGPT," https://chat.openai.com/chat, accessed June 2023.

Exhibit 7 Analyst Estimate of Google Search Cost Structure

Metric	Google Search in 2022	Google Search with ChatGPT
Revenue per query	$0.01610	$0.01610
Cost per query	$0.01060	$0.01416
Incremental cost per query	-	*$0.00356*
Income per query	**$0.00550**	**$0.00194**
Queries per year	10.1 trillion queries	10.1 trillion queries
Annual Revenue	$162.5 billion	$162.5 billion
Annual Cost	$107.0 billion	$142.9 billion
Incremental Cost	-	*$35.9 billion*
Operating Income	**$55.5 billion**	**$19.5 billion**

Source: Excerpted from: Dylan Patel and Afzal Ahmad, "Peeling the Onion's Layers—Large Language Models Search Architecture and Cost," SemiAnalysis, February 13, 2023, https://www.semianalysis.com/p/peeling-the-onions-layers-large-language, accessed April 2023.

Note: Assumes 320,000 queries per second and 31,536,000 seconds per year.

Endnotes

[1] Microsoft, "Microsoft and OpenAI Extend Partnership," Microsoft Blog, January 23, 2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/, accessed January 27, 2023.

[2] 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.

[3] Jeffrey Dean et al., "Large Scale Distributed Deep Networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, edited by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf

[4] 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.

[5] 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.

[6] 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.

[7] University of Toronto, "U of T Neural Networks Start-up Acquired by Google," University of Toronto Media Room, March 12, 2013, https://media.utoronto.ca/media-releases/u-of-t-neural-networks-start-up-acquired-by-google/.

Cade Metz, "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead," New York Times, May 4, 2023, https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html, accessed June 22, 2023.

[8] Alexia Tsotsis, "Google Buys Wavii for North of $30 Million," TechCrunch, April 23, 2013, https://techcrunch.com/2013/04/23/google-buys-wavii-for-north-of-30-million/, accessed January 23, 2023.

[9] Aaron Hurst, "Google Revealed to Have Acquired the Most AI Startups since 2009," Information Age, February 18, 2020, https://www.information-age.com/google-revealed-acquired-most-ai-startups-since-2009-15415/, accessed January 26, 2023.

[10] 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.

[11] Sundar Pichai, "Google DeepMind: Bringing Together Two World-Class AI Teams," Google Keyword Blog, April 20, 2023, https://blog.google/technology/ai/april-ai-update/, accessed May 3, 2023.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," arXiv.org, https://arxiv.org/pdf/1706.03762.pdf, accessed June 14, 2023.

[13] Jacob Devlin and Ming-Wei Chang, "Open Sourcing BERT: State-of-the-Art Pre-Training for Natural Language Processing," Google Research Blog, November 2, 2018, https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html, accessed January 23, 2023.

[14] Dale Markowitz, "Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5," Dale on AI, May 6, 2021, https://daleonai.com/transformers-explained, accessed June 14, 2023.

[15] Andrew M. Dai and Nan Du, "More Efficient In-Context Learning with GLaM," Google Research Blog, December 9, 2021, https://ai.googleblog.com/2021/12/more-efficient-in-context-learning-with.html, accessed January 23, 2023.

Heng-Tze Cheng and Romal Thoppilan, "LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything," Google Research Blog, January 21, 2022, https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html, accessed January 23, 2023.

[16] Sharan Narang and Aakanksha Chowdhery, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," Google Research Blog, April 4, 2022, https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html, accessed January 23, 2023.

[17] 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.

[18] Richard Waters and Miles Kruppa, "Rebel AI Group Raises Record Cash after Machine Learning Schism," *Financial Times*, May 28, 2021, https://www.ft.com/content/8de92f3a-228e-4bb8-961f-96f2dce70ebb, accessed May 5, 2023.

[19] 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.

[20] 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.

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

[22] OpenAI, "OpenAI LP," OpenAI Blog. March 11, 2019, https://openai.com/blog/openai-lp/, accessed January 20, 2023.

[23] OpenAI, "Microsoft Invests in and Partners with OpenAI to Support Us Building Beneficial AGI," July 22, 2019, https://openai.com/blog/microsoft/, accessed January 20, 2023.

[24] OpenAI, "OpenAI API," OpenAI Blog, June 11, 2020, https://openai.com/blog/openai-api/, accessed January 20, 2023.

[25] 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.

[26] Karen Hao, "OpenAI is Giving Microsoft Exclusive Access to Its GPT-3 Language Model," MIT Technology Review, September 30, 2020, https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/, accessed January 20, 2023.

[27] Alberto Romero, "How OpenAI Sold Its Soul for $1 Billion," Medium, August 30, 2021, https://onezero.medium.com/openai-sold-its-soul-for-1-billion-cf35ff9e8cd4, accessed January 20, 2023.

[28] Caroline Haskins, "OpenAI's Mission to Benefit Humanity Now Includes Seeking Profit," Motherboard: Tech by Vice, March 12, 2019, https://www.vice.com/en/article/kzdyme/openais-mission-to-benefit-humanity-now-includes-seeking-profit, accessed March 15, 2023.

[29] 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.

[30] Microsoft, "Microsoft and OpenAI Extend Partnership," Microsoft Blog, January 23, 2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/, accessed January 27, 2023.

[31] 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, 223.

[32] Microsoft, "Microsoft and OpenAI Extend Partnership," Microsoft Blog, January 23, 2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/, accessed January 27, 2023.

[33] Cade Metz and Karen Weise, "Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I.," *New York Times*, January 12, 2023, https://www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html, accessed January 27, 2023.

[34] 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.

[35] OpenAI, "Build Next-Gen Apps with OpenAI's Powerful Models," https://openai.com/api/, accessed January 20, 2023.

[36] OpenAI, "GPT-4," March 14, 2023, https://openai.com/research/gpt-4, accessed May 8, 2023.

[37] OpenAI, "GPT-4 Technical Report," https://cdn.openai.com/papers/gpt-4.pdf, accessed May 8, 2023.

[38] James Vincent, "OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong'," The Verge, March 15, 2023, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview, accessed May 8, 2023.

[39] Rob Knies, “The Meteoric Rise of Microsoft Research: An Oral History,” Microsoft Research Blog, September 26, 2006, https://www.microsoft.com/en-us/research/blog/meteoric-rise-microsoft-research-oral-history/, accessed May 18, 2023..

[40] Ron Miller, “Microsoft Announces Azure ML, Cloud-based Machine Learning Platform That Can Predict Future Events,” TechCrunch, June 16, 2014, https://techcrunch.com/2014/06/16/microsoft-announces-azure-ml-cloud-based-machine-learning-platform-that-can-predict-future-events/, accessed May 18, 2023.

[41] Microsoft Turing, “About,” https://turing.microsoft.com/about, accessed May 18, 2023.

[42] Corby Rosset, “Turing-NLG: A 17-Billion-Parameter Language Model by Microsoft,” Microsoft Research Blog, February 13, 2020, https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, accessed May 18, 2023.

[43] Paresh Kharya and Ali Alvi, “Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World’s Largest and Most Powerful Generative Language Model,” Nvidia Developer Blog, October 11, 2021, https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/, accessed May 18, 2023.

[44] Mohanbir Swahney, “The Real Reasons Microsoft Invested in OpenAI,” Forbes, July 24, 2019, https://www.forbes.com/sites/mohanbirsawhney/2019/07/24/the-real-reasons-microsoft-invested-in-openai/, accessed May 19, 2023.

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

[46] Aaron Holmes, “How Microsoft Swallowed Its Pride to Make a Massive Bet on OpenAI,” The Information, May 25, 2023, https://www.theinformation.com/articles/how-microsoft-swallowed-its-pride-to-make-a-massive-bet-on-openai, accessed June 6, 2023.

[47] Ibid.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Kelsey Piper, “Why Is Meta’s New AI Chatbot So Bad?” Vox, August 21, 2022, https://www.vox.com/future-perfect/23307252/meta-facebook-bad-ai-chatbot-blenderbot, accessed January 27, 2023.

[53] 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

[54] Meta, “Announcing the PyTorch Foundation to Accelerate Progress in AI Research,” Meta Newsroom, September 12, 2022, https://about.fb.com/news/2022/09/pytorch-foundation-to-accelerate-progress-in-ai-research/, accessed June 14, 2023.

[55] Meta, “LLaMA: Open and Efficient Foundation Language Models,” February 24, 2023, Meta Newsroom, https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/, accessed March 23, 2023.

[56] Richard MacManus, “Why Developers Are Flocking to LLaMA, Meta’s Open Source LLM,” The New Stack, May 5, 2023, https://thenewstack.io/why-open-source-developers-are-using-llama-metas-ai-model/, accessed May 12, 2023.

[57] Meta Platforms Inc., “First Quarter 2023 Results Follow Up Call,” 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.

[58] "Emad Mostaque, "Stable Diffusion Launch Announcement," Stability AI Blog, August 10, 2022, https://web.archive.org/web/20220905105009/https://stability.ai/blog/stable-diffusion-announcement, accessed June 22, 2023.

[59] Mureji Fatunde and Crystal Tse, "Stability AI Raises Seed Round at $1 Billion Value," Bloomberg, October 17, 2022, https://www.bloomberg.com/news/articles/2022-10-17/digital-media-firm-stability-ai-raises-funds-at-1-billion-value, accessed January 30, 2023.

[60] Kyle Wiggers, "Stability AI, the Startup behind Stable Diffusion, Raises $101M," TechCrunch, October 17, 2022, https://techcrunch.com/2022/10/17/stability-ai-the-startup-behind-stable-diffusion-raises-101m/, accessed January 30, 2023.

[61] Amazon, "Stability AI Selects AWS as Its Preferred Cloud Provider to Build Artificial Intelligence for the Future," Amazon Press Center, November 30, 2022, https://press.aboutamazon.com/2022/11/stability-ai-selects-aws-as-its-preferred-cloud-provider-to-build-artificial-intelligence-for-the-future, accessed January 30, 2023.

[62] Atila Orhon, Michael Siracusa, and Aseem Wadhwa, "Stable Diffusion with Core ML on Apple Silicon," December 2022, https://machinelearning.apple.com/research/stable-diffusion-coreml-apple-silicon, accessed January 30, 2023.

[63] Ben Thompson, "AI and the Big Five," Stratechery, January 9, 2023, https://stratechery.com/2023/ai-and-the-big-five/, accessed January 30, 2023.

[64] Hugging Face, "We Raised $100 Million for Open & Collaborative Machine Learning," Hugging Face Blog, May 9, 2022, https://huggingface.co/blog/series-c, accessed January 30, 2023.

[65] "Pricing," Hugging Face, https://huggingface.co/pricing, accessed January 30, 2023.

[66] "BigScience Large Open-science Open-access Multilingual Language Model," Hugging Face, Version 1.3, July 2022, https://huggingface.co/bigscience/bloom, accessed March 23, 2023.

[67] Dylan Kaplan, "GPT-3 vs Bloom [Real Code & Real Results]," Enjoy Machine Learning, January 25, 2023, https://enjoymachinelearning.com/blog/gpt-3-vs-bloom, accessed March 23, 2023.

[68] David Curry, "ChatGPT Revenue and Usage Statistics (2023)," Business of Apps, February 20, 2023, https://www.businessofapps.com/data/chatgpt-statistics/, accessed March 28, 2023.

[69] Jeffrey Dastin, Krystal Hu and Paresh Dave, "Exclusive: ChatGPT Owner OpenAI projects $1 Billion in Revenue by 2024," Reuters, December 15, 2022, https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/, accessed March 28, 2023.

[70] OpenAI, "Introducing ChatGPT Plus," February 1, 2023, https://openai.com/blog/chatgpt-plus, accessed March 28, 2023.

[71] 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 March 22, 2023.

[72] Chang Che and John Liu, "China's Answer to ChatGPT Gets an Artificial Debut and Disappoints," *New York Times*, March 16, 2023, https://www.nytimes.com/2023/03/16/world/asia/china-baidu-chatgpt-ernie.html, accessed March 22, 2023.

[73] Emma Roth, "Google-Backed Anthropic Launches Claude, an AI Chatbot That's Easier to Talk to," The Verge, March 14, 2023, https://www.theverge.com/2023/3/14/23640056/anthropic-ai-chatbot-claude-google-launch, accessed March 28, 2023.

[74] Ben Thompson, "Apple and AI, Lobotomized Lovers, XR Companions," Stratechery, March 21, 2023, https://stratechery.com/2023/apple-and-ai-lobotomized-lovers-xr-companions, accessed May 10, 2023.

[75] OpenAI, "ChatGPT Plugins," March 23, 2023, https://openai.com/blog/chatgpt-plugins, accessed May 10, 2023.

[76] Business Research Insights, "Search Engine Market Size, Share, Growth and Industry Analysis," https://www.businessresearchinsights.com/market-reports/search-engine-market-101546, accessed March 28, 2023.

[77] 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.

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

[79] 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.

[80] Dylan Patel and Afzal Ahmad, "Peeling the Onion's Layers—Large Language Models Search Architecture and Cost," February 13, 2023, SemiAnalysis.com, https://www.semianalysis.com/p/peeling-the-onions-layers-large-language, accessed March 24, 2023.

[81] Miles Kruppa, "Google CEO on AI, the Future of Search, Efficiency and Battling Microsoft," *Wall Street Journal*, April 8, 2023, https://www.wsj.com/articles/google-ceo-sundar-pichai-interview-ai-search-8ed7d2b4, accessed May 8, 2023.

[82] Grandview Research, "Office Software Market Size & Share Report, 2022–2028," https://www.grandviewresearch.com/industry-analysis/office-software-market-report, accessed March 28, 2023.

[83] Gartner, "Market Share: Enterprise Infrastructure Software, Worldwide, 2020," September 2021.

[84] Google, "Google Workspace: Pricing," https://workspace.google.com/pricing.html, accessed March 28, 2023.

[85] Microsoft, "Microsoft Office 365 for Business: Pricing: Compare All," https://www.microsoft.com/en-us/microsoft-365/business/compare-all-microsoft-365-business-products, accessed March 28, 2023.

[86] Satya Nadella, "The Future of Work with AI," Microsoft, March 16, 2023, https://news.microsoft.com/wp-content/uploads/prod/2023/04/Satya-Nadella-The-Future-of-Work-with-AI.pdf, accessed May 10, 2023.

[87] 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 May 10, 2023.

[88] Greg Brockman, Mira Murati, Peter Welinder, and OpenAI, "OpenAI API," June 11, 2020, https://openai.com/blog/openai-api, accessed May 9, 2023.

[89] Tris Warkentin and Josh Woodward, "Join Us in the AI Test Kitchen," Google Keyword Blog, August 25, 2022, https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/, accessed January 27, 2023.

[90] Jennifer Elias, "Google Announced Bard A.I. in Response to ChatGPT," CNBC, February 6, 2023, https://www.cnbc.com/2023/02/06/google-announces-bard-ai-in-response-to-chatgpt.html, accessed May 8, 2023.

[91] Rohan Taori et al, "Alpaca: A Strong, Replicable Instruction-Following Model," Stanford University Center for Research on Foundation Models, https://crfm.stanford.edu/2023/03/13/alpaca.html, accessed March 28, 2023.

[92] Eliezer Yudkowski, March 14, 2023, https://twitter.com/ESYudkowsky/status/1635577836525469697, accessed March 28, 2023.

[93] Kevin Scott, "Microsoft teams up with OpenAI to exclusively license GPT-3 language model," Official Microsoft Blog, September 22, 2020, https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/, accessed May 10, 2023.

[94] Disha Misal, "Google Datalab vs Amazon SageMaker: Which Cloud Platform Is Best for Your ML Project," Analytics India Magazine, July 11, 2019, https://analyticsindiamag.com/google-datalab-vs-amazon-sagemaker-which-cloud-platform-is-best-for-your-ml-project/, accessed May 10, 2023.

[95] 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.