Natural Language Engineering (2024), **30**, pp. 201–213
doi:10.1017/S1351324923000554

CAMBRIDGE
UNIVERSITY PRESS

INDUSTRY WATCH

A year's a long time in generative AI

Robert Dale

Language Technology Group
Email: rdale@language-technology.com

(Received 28 November 2023; accepted 29 November 2023)

Abstract

A lot has happened since OpenAI released ChatGPT to the public in November 2022. We review how things unfolded over the course of the year, tracking significant events and announcements from the tech giants leading the generative AI race and from other players of note; along the way we note the wider impacts of the technology's progress.

Keywords: Generative AI; NLP industry

Introduction

As I write this, we're fast approaching the first anniversary of the release of OpenAI's ChatGPT, an event that made generative AI accessible to anyone with an internet connection. The interactive ChatGPT web interface was opened up to the public on 30th November 2022. It quickly went viral on the socials and reached over one million users within 5 days—an achievement frequently compared with, for example, Instagram taking 2.5 months to achieve one million downloads, and Netflix taking 3.5 years—and 100 million users within 2 months.

As most readers of this piece will be aware, the underlying technology had been around for quite a while: OpenAI's GPT-3, which underpinned the initial version of ChatGPT, was released back in June 2020. But the simple genius of wrapping up the technology in a chat interface made it easily available to every internet user, regardless of their technical expertise beyond conventional browser use.

As a consequence of this public awareness, AI and its potential impacts are now the fare of editorials in mainstream news publications across the globe; and nation states are vying for position in what has become a global race toward AGI and, at the same time, a race for determining how best to control the technology. A good excuse, then, to review the major events and developments over the course of the year leading up to ChatGPT's first birthday.[a]

The landscape

To provide some structure over the many relevant events that unfolded over the course of the year, we organize these as a month-by-month review under a number of headings. **The Majors** covers the activity of the indisputable leaders in the world of generative AI: OpenAI, Microsoft, and Google. **The Nippers**, so called because they nip at the heels of The Majors, refers to a number of smaller companies who punch above their weight in the generative AI space: Anthropic,

[a] The submission deadline for this article was 23rd November, so it doesn't report on events or developments subsequent to that date. I know, that sounds just like a ChatGPT caveat.

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

AI21 Labs, Character.AI, Cohere, and Perplexity, with occasional mentions of a few others. **The Laggards** refers—perhaps somewhat unfairly, I acknowledge—to other members of the big tech leaderboard who were caught a little flatfooted and have spent the year trying to catch up with The Majors: in particular, Amazon, Apple, IBM, and Meta. The bulk of visible activity has taken place in the US; under **The Rest of the World** we track significant events elsewhere, primarily in China but also in other places. And under **Ramifications** we record wider impacts and noteworthy consequences of, and reactions to, the technical developments.

So let's look at how the year has played out.

December 2022

The Majors: Following ChatGPT's release at the end of November, December sees wall-to-wall coverage in both the technical and the mainstream press: you can find representative coverage at VentureBeat, Fast Company, Wired, Vox, Harvard Business Review, and The New York Times. Everyone is impressed with ChatGPT's conversational abilities, but at the same time, most note that it's not a reliable source of truth, and tends to regurgitate hateful statements and biases; the Twitterverse is awash with examples of people tricking it into saying obviously silly things.

Meanwhile, OpenAI projects US$1B in revenue by 2024; and the question of whether something like ChatGPT might supplant Google search as a source of information is prominent. Google claims that its LLMs are just as capable as OpenAI's, explaining the lack of a similar offering by the need to move conservatively because of the reputational risk posed by the technology.

The Nippers: Smaller players are riding the wave of interest generated by OpenAI: Cohere releases a multilingual text-understanding LLM that works with more than 100 different languages; AI21 Labs launches its Wordtune iOS app, bringing users their own personal AI writing assistant on mobile devices; and You.com has already integrated a ChatGPT-like facility into its search engine.

Ramifications: Stack Overflow temporarily bans users from sharing responses generated by the chatbot because they often look right but are in fact wrong; and even Sam Altman, OpenAI's CEO, feels the need to tweet an assessment of ChatGPT as "fun creative inspiration; great! reliance for factual queries; not such a good idea".

January 2023

The Majors: The big news in January drops toward the end of the month but had been foreshadowed earlier: Microsoft's investment of US$10 billion in OpenAI, building on its earlier funding of US$1B. One analyst tells Yahoo Finance that ChatGPT could be a US$600 billion opportunity for Microsoft. Rumor has it that the company intends to use OpenAI's technology in its Office suite.

Meanwhile, OpenAI releases ChatGPT Professional, a paid version of the chatbot, at US$42 per month, and shares its plans for improving ChatGPT behavior, allowing more user customization. It also releases a tool to help you determine whether text was more likely written by a human or AI. Along the way, ChatGPT passes a US medical licencing exam and an MBA exam.

Google is concerned: it's revealed that the month before, Larry Page and Sergey Brin, Google's founders, held several meetings with company executives to discuss how to respond; and we hear that Google may soon demo an AI search chatbot.

The Nippers: Character.AI, a chatbot startup founded by two former Google researchers, indicates that it wants to raise as much as US$250m in new funding. And initiating a trend that hasn't let up over the past year, a slew of companies announce products or services powered by generative AI.

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

Ramifications: Mistrust of ChatGPT abounds. In its call for submissions, the International Conference on Machine Learning indicates that "papers that include text generated from LLMs like ChatGPT are prohibited"; and Springer Nature, the world's largest academic publisher, announces that software like ChatGPT can't be credited as an author. Meanwhile, ChatGPT is indeed being listed as author on published research papers; a survey finds that around 17% of Stanford students are using ChatGPT to assist with their fall quarter assignments and exams; a survey of 1000 US college students finds that 30% use ChatGPT for written homework; and CNET pauses its use of AI-written articles after being notified of serious errors.

February 2023

The Majors: OpenAI meets with conservative criticism of "wokeness" and responds by publishing its rules for answering culture war queries; a number of major news outlets complain about OpenAI using their stories as data without payment; and OpenAI reportedly pays millions for the ai.com domain name. Microsoft launches a new version of its Bing search engine running on a variant of the technology underlying ChatGPT and then adds voice-driven Bing Chat on Android and iOS. But the main theme of the month is Google's reaction: pressured by ChatGPT's success and positive press, and in a move seen by some Google staff as hasty, it announces the forthcoming release of Google Bard. Google shares subsequently drop US$100B after Bard makes a mistake in a demo. Semianalysis estimates that, if Google is forced to shift all of its search traffic to a conversational model, its annual operating expenses will be raised by US$36B.
The Nippers: Anthropic receives a US$300m investment from Google; Elon Musk announces he's looking into creating an alternative to ChatGPT; and You.com launches YouChat 2.0, a multimodal conversational AI that can serve up charts, images, videos, tables, graphs, text, or code embedded in its responses to user queries.
The Laggards: Meta enters the fray by releasing LLaMA, a 65-billion-parameter LLM that runs on a single GPU, but makes it accessible only to approved researchers, government organizations, and members of civil society; and Mark Zuckerberg says "AI personas" are coming to WhatsApp, Messenger, and Instagram.
The Rest of the World: China's Tencent sets up a team to develop a ChatGPT-like product; e-commerce giant JD.com says it will launch a ChatGPT-style product called ChatJD; and a number of other Chinese companies are rushing to prove they have tech similar to ChatGPT. Companies are also cutting deals with Baidu ahead of the expected launch in March of its generative AI chatbot.
Ramifications: A student uses ChatGPT to cheat in an AI Ethics class. Meanwhile, AI detection tools are struggling to keep up with ChatGPT.

March 2023

The Majors: OpenAI makes available the ChatGPT API and reduces the cost by 10× compared to GPT-3.5. GPT-4 is launched: the accompanying 98-page technical paper is critiqued by many for being short on important details. Toward the end of the month, the company launches third-party plugins that allow ChatGPT to retrieve real-time information and to connect to APIs to perform actions on behalf of users. Meanwhile, Microsoft reveals that the AI-enabled Bing search engine was built on top of GPT-4. The company unveils Copilot, a new AI-powered tool that sits alongside Microsoft 365 and helps users with generating text in documents, creating PowerPoint presentations and formatting Excel data.

Back at Google HQ, an internal directive requires generative AI to be incorporated into all of its biggest products within months. Just ahead of Microsoft's Copilot announcement, Google unveils a collection of generative AI tools for Gmail and Google Docs that will automatically create drafts

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

and adds new AI capabilities to its business products, including Google Cloud and a new API for developers. Pitched as an early experiment that lets you collaborate with generative AI, Bard is made available for public beta testing in the US and UK.
The Nippers: Anthropic raises another US$300m; Character.AI raises US$150m in a Series A round; and Perplexity AI raises US$25.6m to develop its conversational search engine. AI21 Labs announces Jurassic-2, its latest generative AI model, as well as a line of task-specific APIs that offer specialized reading and writing functions; Anthropic launches Claude, a competitor to ChatGPT; and Character.AI announces the early preview release of their new AI model, C1.2, which offers customizable and personalized AI experiences for users.
The Laggards: Meta's LLaMA language model is leaked to 4chan.
The Rest of the World: Baidu launches the beta version of its Ernie Bot but disappoints investors with no live preview, causing a drop of 6.4% in its stock value; but by the following day, positive reviews for its Chinese language capabilities cause the company's stock to jump 15%. Testing by Reuters demonstrates that Ernie won't answer questions about Chinese President Xi Jinping.

The EU issues a 20m euro RFP for LLMs that respect European values, and TechMonitor argues that the UK urgently needs to develop its own LLM to help its companies compete.
Ramifications: Voices of concern about AI's potential for harm get louder and more organized: the Future of Life Institute issues an open letter demanding a 6-month pause in generative AI development, signed by a number of worthies in the space. Reactions are mixed.

April 2023

The Majors: OpenAI plans to launch ChatGPT Business, a subscription tier that caters to enterprise customers' data control needs, and unveils changes to ChatGPT that address privacy concerns by giving users more control of their data and chat history. The company also raises just over US$300m in a tender offer from VC firms at a valuation of US$27B to US$29B, separate from Microsoft's US$10B investment earlier in the year. Microsoft is developing its own AI chip, named Athena, in a bid to improve chatbot performance and save money. Google Cloud announces limited access to its Med-PaLM 2 medical LLM, which aims to more accurately and safely answer medical questions. Google employees criticize Bard as "a pathological liar" and potentially dangerous. Meanwhile, Google's search quality raters have reportedly been instructed to prioritize rating chatbot prompt responses over the quality of search results, potentially impacting search quality.
The Nippers: Anthropic plans to raise up to US$5B over the next 2 years to challenge OpenAI; Cohere partners with conversational AI firm LivePerson to adapt its LLM technology for enterprise deployments; Perplexity announces new chatbot features including personalized search options and the ability to save dialog threads, and launches an Apple iOS app; and Elon Musk is reportedly pursuing a generative AI project within Twitter, said to be in the early stages of development and using 10,000 GPUs.
The Laggards: Amazon is building a new and more generalized LLM to power Alexa and improve its personal assistant capabilities. AWS announces Amazon Bedrock, a new service that allows customers to access a variety of generative AI models including Amazon's Titan. Zuckerberg says Meta will introduce generative AI across all of its products.
The Rest of the World: Alibaba and Huawei are set to launch generative AI chatbots.

May 2023

The Majors: OpenAI CEO Sam Altman embarks on a world tour, aiming to reassure AI doomers and to warn of the danger of overregulation. OpenAI releases a free ChatGPT iOS app; it surpasses half a million downloads in its first 6 days, outperforming other AI chatbot apps and Microsoft's

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

Bing and Edge apps. Microsoft opens its Bing chatbot to all users, including new features such as multimodal support, conversation histories, and closer integration with third parties such as Wolfram Alpha. Microsoft also announces the Microsoft 365 Copilot Early Access Program, an invitation-only, paid preview that will roll out to an initial wave of 600 customers worldwide. And Microsoft's new Windows Copilot will offer "centralized AI assistance" in a desktop sidebar.

But the big news is Google's new Search Generative Experience (SGE), which integrates into search results a new feature called the "AI snapshot," synthesizing a range of information from web searches into a summary and links to help answer more complex query types. Then there's Duet AI, the label for a collection of AI-powered productivity tools across its Workspace apps, including an image generation feature in Google Slides and AI-generated responses to emails in Gmail's "Help me write" mode. But a leaked internal document allegedly from a Google researcher claims that open-source AI is outcompeting both Google AI and OpenAI.

The Nippers: Inflection AI announces the release of Pi, its personal AI that serves as a companion by offering conversations, friendly advice, and concise information. Anthropic expands its context window from 9K to 100K tokens, allowing the Claude LLM to digest and analyze hundreds of pages of content, and raises US$450m in a Series C funding round, with participation from Google, Salesforce, and Zoom.

The Laggards: IBM appears belatedly on the scene with the launch of Watsonx, a platform that delivers tools for building AI models and access to pretrained models to generate computer code and text. We hear that Amazon plans to add ChatGPT-style search to its online store.

The Rest of the World: Baidu is preparing to integrate its generative AI LM, Ernie 3.5, into its search engine and chatbot application, Ernie Bot. And it looks like China's AI industry is facing a chip shortage, with only around 40,000 of Nvidia's data-center-grade A100 GPUs available in the country.

Ramifications: A group of prominent AI researchers, engineers, and CEOs sign a 22-word statement warning of the existential threat they believe AI poses to humanity, calling for mitigating the risk of extinction from AI to be a global priority. And a lawyer representing a man who sued an airline relies on ChatGPT to help prepare a court filing; it creates fictitious legal decisions and citations, causing the lawyer to face potential sanctions from the judge.

June 2023

The Majors: It's reported that OpenAI warned Microsoft about integrating GPT-4 into Bing Chat too early due to inaccurate and unpredictable responses. OpenAI plans to launch a marketplace for AI models built on top of its AI technology, allowing developers to sell their models to other businesses. Microsoft is ending support for Cortana as a standalone app on Windows and instead prioritizing Windows Copilot to take on all that Cortana used to do. Microsoft also brings Bing Chat's voice mode to Edge on desktop, allowing users to ask questions in multiple languages using their voice. Google Bard offers location-specific information by accessing a user's precise location and using it to provide more relevant data. Meanwhile, Google delays Bard's EU launch over privacy concerns; Google's SGE is found to be frustrating for users due to the wait time and cluttered results; and it's revealed that Google DeepMind is developing a system called Gemini that combines the reinforcement learning used in AlphaGo with LLMs similar to GPT-4 to give the system new capabilities such as planning and problem-solving.

The Nippers: Cohere announces US$270m in new capital as part of its Series C financing. Inflection secures US$1.3B in funding with Microsoft and Bill Gates among its investors; and it unveils its LLM, Inflection-1, which is said to be roughly of GPT-3.5 size and, according to the company, competitive or superior to other models on this tier.

The Laggards: Meta showcases new generative AI technologies for its consumer products, including the earlier-promised AI chatbots for Messenger and WhatsApp, as well as internal-only

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

products, like an AI productivity assistant and an experimental interface for interacting with AI agents. And Meta plans to offer its AI models for free commercial use, shifting the AI landscape toward open-source alternatives. The company also unveils Voicebox, a generative model that can perform multiple tasks, including speech editing and noise removal, and is trained to synthesize speech across six languages. Amazon is testing generative AI to summarize product reviews in order to save shoppers from wading through unreliable reviews.

The Rest of the World: Alibaba has integrated its ChatGPT-like AI into its meeting assistant Tingwu and plans to add the technology to its DingTalk collaboration platform. Tencent launches its LLM service for corporate clients, aimed at various sectors including finance and media. And Baidu's Ernie now has plugins. European AI startup Mistral AI aims to develop cutting-edge, open-source AI models that prioritize data security, privacy, and auditability, aiming to become the leading AI tool for Europe.

July 2023

The Majors: It's reported that ChatGPT experienced a decline in website traffic and unique visitors in June. And the company has taken down its AI text classifier, which was meant to detect whether text was AI-generated or written by a human, due to its low accuracy rate. On the upside, OpenAI's new custom instructions feature allows ChatGPT users to provide personal information and preferences to improve the AI's responses, making it a more efficient and personalized virtual assistant. And OpenAI is assembling a team to focus on the problem of superintelligence alignment, dedicating 20% of their compute power over the next 4 years to tackle this critical issue. GPT-4's details are leaked, revealing that it has over 1.8 trillion parameters, utilizes a mixture of experts model, is trained on 13 trillion tokens, and has a training cost of approximately US$63m.

Bing Chat now supports visual search, allowing users to upload photos and ask the AI search engine questions based on the images. Microsoft is also releasing Bing Chat Enterprise, an AI-powered chat tool that addresses the privacy, security, and data protection concerns associated with generative AI tool use in businesses. Microsoft 365 Copilot, its new corporate AI tools, will cost US$30 per user per month in addition to existing subscription fees.

Google is planning to give Google Assistant a generative face-lift by incorporating LLMs. Reports persist that Google's Bard is trained and improved by thousands of underpaid and overworked external contractors who review and assess its answers, raising concerns about the quality of the AI's responses.

The Nippers: Anthropic releases Claude 2, an improved LLM that can code, analyze text, and write compositions, with enhancements in maths and reasoning, as well as a reduced likelihood of generating harmful outputs. The company is also developing a variant of Claude that is purposely deceptive, with the goal of understanding and preventing deception in AI systems, as part of its mission to prioritize safety in AI development. Elon Musk announces the formation of xAI, a company aiming to understand the nature of the universe and compete with ChatGPT. AI21 Labs launches a plug-and-play generative AI engine called Contextual Answers, which can be embedded into digital assets to implement LLM technology on organizational data, allowing users to retrieve information through a conversational experience. SAP invests directly in Cohere, Anthropic and Aleph Alpha. Cohere partners with Amazon Web Services to bring its enterprise AI models to Amazon Bedrock.

The Laggards: IBM's generative AI foundation model, watsonx, is now generally available after a 2-month beta. Meta is making the second version of LLaMA available free of charge and partnering with Microsoft to provide increased access to its foundational AI technologies. Apple is reportedly developing its own AI-powered chatbot called Apple GPT, using its own LM framework called Ajax, and is planning to make a significant AI-related announcement next year; Apple's market value increases by US$71B on the news.

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

The Rest of the World: In China, Baidu's Ernie Bot joins iFlytek's Spark as local ChatGPT alternatives on Apple's mainland App Store.
Ramifications: Comedian Sarah Silverman, along with authors Christopher Golden and Richard Kadrey, are suing OpenAI and Meta in separate lawsuits over claims of copyright infringement involving their works being used to train AI models without consent. And in what are possibly the first signs of the Majors' constant feature-creep destroying the business models of smaller players, Jasper AI, a generative AI startup that raised US$125m the year before, is laying off employees and refocusing on servicing the marketing industry.

August 2023

The Majors: Wolfram Research partners with OpenAI to develop a plugin that combines the capabilities of ChatGPT with math-solving technology. OpenAI files a trademark application for GPT-5 with the USPTO, covering a range of categories related to LMs and ML. And the company launches ChatGPT Enterprise, a secure and customizable AI assistant designed for enterprise use, offering features like privacy, advanced data analysis, longer input processing, and unlimited access to GPT-4. OpenAI is reported to be earning US$80m per month, on track to generate over US$1B in annual revenue, potentially recovering its US$540m loss from the previous year. Microsoft confirms that Bing Chat will soon be available in third-party browsers, allowing it to compete with other browsers' built-in tools. Analytics India Magazine reports that Microsoft introduced Azure ChatGPT as an enterprise option but then denied its existence, causing confusion and strain in its partnership with OpenAI. Microsoft begins to roll out Bing Chat Enterprise in the Windows Copilot Preview for eligible commercial customers. Google introduces updates to its Search Generative Experience, all aimed at helping users better learn and understand information they find on the web; this includes testing the inclusion of links directly in the snapshot answers generated, making it easier for users to check the source of the information. The company is also introducing images and video in Search Generative Experience results. Meanwhile, SemiAnalysis claims that Google's forthcoming Gemini AI model will outperform OpenAI's GPT-4 due to Google's advanced GPU infrastructure, and OpenAI CEO Sam Altman responds sarcastically on Twitter.
The Nippers: Elon Musk's newly formed AI company xAI buys the domain ai.com, previously owned by OpenAI. Anthropic raises US$100m from SK Telecom to build a customized LM for telcos. Claude.ai, the web interface for Anthropic's Claude 2, starts restricting access for unpaid users. Perplexity updates its AI search copilot with GPT-3.5 Turbo fine-tuning and introduces Code Llama to its LLaMa Chat. AI21 Labs updates its Wordtune generative AI assistant for enterprise clients, introducing new features such as the ability to create templates, summarize information, and answer queries using linked databases.
The Laggards: AWS expands its Amazon Bedrock service by adding LLMs from StabilityAI and Anthropic. Amazon is testing a generative AI tool that writes product descriptions for sellers. CEO Jassy also reveals that every business unit within Amazon has multiple generative AI initiatives underway, signaling the company's focus on AI across all sectors. During a quarterly earnings call, Tim Cook addresses the criticism that Apple's AI efforts are behind those of Microsoft, Google, and Elon Musk, saying that while they may not talk about it as much, AI is a core technology integral to their products.
The Rest of the World: China's internet giants, including Baidu, ByteDance, Tencent, and Alibaba, have ordered US$5B worth of Nvidia chips to support their AI ambitions. Baidu showcases the expansion of Ernie Bot, along with updates to its PaddlePaddle deep learning platform and the launch of an AI coding assistant.

Saudi Arabia and UAE are also racing to buy Nvidia chips to power their AI ambitions. The UK plans to spend £100m to purchase AI chip technology from AMD, Intel, and Nvidia in an effort to

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

build a national AI resource. At the global level, Europe wants to develop its own generative AI to counteract American dominance in the field; Aleph Alpha is emerging as a potential contender.
Ramifications: OpenAI's web crawler, GPTBot, is blocked by The New York Times, quickly followed by other news outlets including CNN, Reuters, and the Australian Broadcasting Corporation. Gartner places generative AI on the peak of inflated expectations on the 2023 hype cycle for emerging technologies.

September 2023

The Majors: OpenAI announces that it will host its first developer conference in November. Meanwhile, Sam Altman suggests that AI hallucinations are an integral part of the appeal of generative AI systems like ChatGPT. The company launches the OpenAI Red Teaming Network, a group of contracted experts who will help assess and mitigate AI model risks and biases. OpenAI announces DALL-E 3, an updated version of its generative AI visual art platform that integrates with ChatGPT, and ChatGPT is gaining new voice and image capabilities, allowing users to engage in voice conversations and share images. The company is reportedly raising funds at a valuation of US$80B–US$90B. And ChatGPT can now access real-time information from the internet, allowing users to ask questions about current affairs and access news.

Microsoft announces a Copilot Copyright Commitment, in which it promises to defend paid customers against any copyright lawsuits related to its AI tools that generate outputs. A new version of Windows 11 featuring the company's AI companion Copilot integrated into the operating system, offering new features such as analyzing writing styles and generating content across various applications, is to be released. In other news, the company is developing smaller and cheaper conversational AI software as a backup plan to reduce costs, separate from its partnership with OpenAI. Meanwhile, Google has begun external testing of Gemini, its multimodal GPT-4 competitor.
The Nippers: Anthropic launches its first premium subscription plan, Claude Pro, for its AI-powered chatbot, Claude 2, priced at US$20 per month, offering increased usage, more messages, priority access, and early feature access to subscribers. Character.AI is gaining traction with 4.2 million monthly active users in the US compared to ChatGPT's nearly 6 million, attracting a younger demographic in the 18–24 years age group; and the company is in early talks for funding at a more than US$5B valuation. Cohere announces the availability of its Chat API with retrieval-augmented generation in a public beta, allowing developers to create conversational AI products with improved accuracy and verifiability using RAG.
The Laggards: Apple is reportedly dedicating millions of dollars a day to its R&D budget for AI research, with a focus on developing conversational AI capabilities for Siri, as well as exploring 3D video and image generation and multimodal AI systems; CEO Tim Cook confirms that the company is researching generative AI, including a ChatGPT-like service. Meta is said to be developing a powerful AI model to compete with OpenAI's GPT-4. But this month's big news: Amazon will invest up to US$4B in Anthropic; the deal includes moving Anthropic's software to Amazon Web Services data centers and utilizing Amazon's chips for training AI models. Amazon also demonstrates "Let's Chat," a new generative AI feature designed to make its Alexa voice assistant more conversational on its Echo devices.
The Rest of the World: Baidu's Ernie Bot receives 33 million questions on its public debut. Baidu plans to launch its Ernie 4 LLM soon. And China's internet giant Tencent Holdings will introduce an AI chatbot named HunyuanAide. More than 70 large language models with over a billion parameters have now been released in China. The Technology Innovation Institute in the UAE launches Falcon 180B, an advanced LLM that's available for research and commercial purposes, following the success of its predecessor Falcon 40B; researchers from the UAE, in collaboration

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

with Cerebras, introduce Jais and Jais-chat, two new open language models trained on Arabic and English.

And Japan is developing its own version of ChatGPT to overcome the language and cultural limitations of English-based models. Meanwhile, Paris-based startup Mistral AI releases its first LLM, Mistral 7B, which outperforms larger offerings and is open-sourced under the Apache 2.0 license, making it available for enterprise use.

Ramifications: Salesforce clarifies that its AI systems cannot take responsibility for decisions that may have legal or other major implications. Jasper AI reduces the value of its shares by 20%, signaling potential slowdown in its growth due to competition from ChatGPT. And 26 of the top 100 websites are now blocking GPTBot, OpenAI's web crawler.

October 2023

The Majors: OpenAI's annualized revenue increases to US$1.3B a year, with the majority coming from subscriptions to ChatGPT, while ChatGPT's mobile app hit a record US$4.58m in revenue last month, although growth is slowing. The company is considering developing its own AI chips and potentially acquiring a chip company. But there are reports that OpenAI's "Arrakis" project, which aimed to improve the efficiency of ChatGPT, was a failure, disappointing Microsoft executives. And OpenAI quietly changes its core values on its career page. Microsoft makes OpenAI's DALL-E 3 image generation available to all Bing Chat and Bing Image Creator users. Google announces an AI-powered update to Google Assistant, integrating Bard to handle a wider range of questions and tasks, including personalized responses from Google apps. SGE adds new capabilities including image generation using prompts and the ability to write drafts with customizable output. Meanwhile, Bard is being questioned by Google's own staff regarding its effectiveness and utility, with some wondering whether the resources invested are worth it. Following an earlier announcement by Microsoft, Google Cloud introduces indemnification for customers using generative AI, taking responsibility for potential legal risks related to copyright infringement claims on training data and generated output. Some information leaks about Google's forthcoming Gemini multimodal AI model, and also on Stubbs, a new feature that allows users to create, launch, and share their own AI-generated app prototypes. Toward the end of the month, Microsoft and Google parent Alphabet's earning reports come out: Microsoft's share price jumps as much as 6% on its 13% increase in revenue, driven by demand for its AI-infused products and its Azure public cloud; Alphabet, on the other hand, reports slower-than-anticipated growth, resulting in an 8% drop in share price.

The Nippers: Anthropic plans to raise a further US$2B, with a potential valuation of US$20B to US$30B. Cohere releases an enterprise chatbot API called Coral, allowing developers to integrate conversational capabilities into their services. Character.AI launches a new feature called Character Group Chat, which allows users to interact with multiple AI Characters and humans in the same room. And Inflection AI's Pi now has real-time access to the Web.

The Laggards: Apple plans to spend US$1B on catching up with generative AI tools. IBM announces the availability of the watsonx Granite Model series, a collection of generative AI models designed to help businesses scale AI. Meta announces Llama 2 Long, a model that outperforms competitors in generating responses to long user prompts. The company is also investing millions of dollars to secure licenses from celebrities, such as Snoop Dogg and Charli D'Amelio, to use their likenesses for AI chatbot characters. And Meta is rolling out generative AI features for advertisers, including the ability to create backgrounds, expand images, and generate multiple versions of ad text.

The Rest of the World: Following criticism of a lack of guardrails on its recently released 7B model, Mistral AI says the responsibility for ensuring LLM safety lies with the application

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

developers. Chinese AI startup Baichuan, one of the first companies to receive China's approval to release a public chatbot, raises US$300m in funding from investors including Alibaba and Tencent. Baidu launches Ernie 4.0, a foundation model that the company claims is on par with GPT-4.

Ramifications: The Foundation Model Transparency Index released by Stanford HAI assesses the disclosure of information by the creators of the top 10 AI models, finding that none of the models scores particularly high in transparency. Sam Altman warns that AI doesn't need to reach superhuman levels of general intelligence in order to have the capability of persuading humans in ways that could lead to concerning outcomes. Reddit may block Google and Bing's search crawlers if it can't make deals with generative AI companies to pay for its data. LinkedIn is to lay off hundreds of people amid broader restructuring that aims to optimize around AI. And coding help forum Stack Overflow is laying off 28% of its staff; the decision may be in response to challenges posed by AI-generated coding answers.

November 2023

The Majors: The first half of the month sees two major news events. One is the release of xAI's first AI product, an LLM called "Grok," which drives a chatbot on X that offers real-time information and isn't wary of answering "spicy" questions. Currently available to a limited number of users in the US, the chatbot's other distinguishing feature is its embodiment of Musk's sometimes adolescent sense of humor.

The other major news is a host of technical announcements at OpenAI's DevDay, including updated knowledge bases, a massively larger context window, and, perhaps most significantly, the release of a platform for creating custom versions of ChatGPT for specific use cases, to be made accessible through a GPT Store. Earlier in the month the company releases an update to ChatGPT that combines several capabilities into one, including document analysis, web browsing, data analysis, and image generation. This is seen as a blow to startups that rely on OpenAI's system to process PDFs. The company is also creating a preparedness team to address the potential catastrophic risks associated with advanced AI models.

Microsoft launches its Microsoft 365 Copilot system for a US$30-per-month premium per user, but enterprise customers will need to commit to at least 300 users; while Microsoft's Windows 11 23H2 update includes the integration of the Windows Copilot AI assistant. But Copilot for Windows gets criticized for its lack of functionality and failure to improve Windows' usability.

Bard now responds to questions in real time, but it's otherwise pretty quiet on the Google front; we hear that Google plans to release its Gemini AI model as part of a series of next-generation models in 2024.

The Nippers: Google announces it will invest up to US$2B in Anthropic. The company will utilize Google's Cloud TPU v5e chips to enhance its Claude LLM.

The Laggards: Amazon is reportedly developing an LLM called Olympus with two trillion parameters to compete with OpenAI and Google's models, potentially rolling out as early as December.

The Rest of the World: Baidu introduces a paid version of Ernie Bot, with a subscription that allows users to make 100 inquiries every three hours. Tencent has incorporated its Hunyuan AI model into over 180 of its services, including Tencent Meeting and Tencent Docs.

Ramifications: US President Joe Biden issues an executive order on the use and regulation of AI, providing guidance on safety, civil rights, privacy, and promoting innovation and competition. The UK hosts its AI Safety Summit, attended by 100 representatives of governments, academia, industry, and other interested parties; the Bletchley Declaration is signed, emphasizing the need for safe, responsible, and human-centric development and use of AI technology and promoting international cooperation.

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

And then . . .

Just as this article was going to press, the OpenAI upheaval happened.

- Friday 17th November: Sam Altman is removed from his CEO position at OpenAI, allegedly due to lack of transparency and hindrance to the board's responsibilities. OpenAI's CTO Mira Murati is appointed as interim CEO. Altman's departure is followed by the resignation of company president Greg Brockman and three senior scientists. Ilya Sutskever apparently led the firing process, with support from the board. Early reports suggest the oustering followed disagreements over how to balance the safety and profitability of AI.
- Saturday: An internal memo from OpenAI states that Altman's firing was not due to wrongdoing or safety concerns, but rather a breakdown in communication.
- Sunday: Things change fast—the OpenAI board enters negotiations with Altman about him potentially returning to the company, following pressure from investors, but negotiations collapse. OpenAI names former Twitch CEO Emmett Shear as interim CEO, making Shear the third CEO in 3 days since Altman's shock firing. Meanwhile, Microsoft, Amazon, and Google are all vying for Altman, with Microsoft particularly concerned about losing him to its competitors.
- Monday: Microsoft CEO Satya Nadella negotiates a deal to bring Altman to head up a new subsidiary focused on AI innovation within Microsoft, causing its shares to jump by over 1%. Over 500 employees at OpenAI threaten to quit and join Microsoft if the current board does not resign. Salesforce is actively recruiting OpenAI employees who are unhappy with Sam Altman's removal. Investors in OpenAI are considering legal action against the company's board in their fight to reinstate Altman as CEO. OpenAI's board of directors is in restarted discussions with Altman to potentially bring him back as CEO. Satya Nadella says Microsoft will never again be blindsided by the OpenAI board if Altman returns.
- Tuesday: Altman makes a triumphant return as CEO of OpenAI, agreeing to an internal investigation into his dismissal. Altman's return raises questions about what changes Microsoft has secured to improve its control of the company.

At the time of writing, the reasons for Altman's firing remain unclear. A common observation is that the saga highlights the tension between OpenAI's mission to build AI smarter than humans and its commitment to ensuring AI safety and benefit to humanity. By the time you read this, the dust will have settled, and we should have a clearer picture of the longer-term effects of these events and what really kicked them off.

In the meantime, this serves as a good reminder that ~~a year's~~ a weekend's a long time in generative AI.

What's next?

So that was quite a year. How will the next year of the generative AI wars play out? Some reflections and observations are in order.

"ChatGPT it" doesn't have quite the same ring as "Google it"—which makes me wonder whether OpenAI's choice of name for the chatbot reflects a missed branding opportunity—but OpenAI clearly remains at the front of the pack in terms of mindshare in the generative AI space. By virtue of their unique partnership, Microsoft has been able to quickly absorb OpenAI's technical advances into products, or at least product announcements; there was a period during the year where it seemed that every week saw AI being added to yet another element of the Microsoft product range.

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

Google's responses—and it's hard not to see them as reactive rather than proactive—often seem like missteps that are too little and too late, making it hard to gain traction. The company's concern over the reputational risk introduced by the too-early release of unreliable products seems quite valid, but perhaps, in a post-Trump world, a good reputation is worth less than it used to be. Microsoft seems much more willing to take risks, and investors seem to be on board with that.

Interestingly, Google suffers some of the media disenchantment that was once reserved for Microsoft; the one-time evil empire now garners some of the uncritical coverage once afforded to the erstwhile doer-of-no-evil. Google was panned in the press and lost US$1B in value after Bard made an error in a recorded demo back in February; a demo of Microsoft's Bing Chat also generated an incorrect response around the same time, but, for whatever reason, that received much less media attention.

Back in February, amidst his optimism that Bing's incorporation of OpenAI's technology would redefine search, Microsoft CEO Satya Nadella famously said that he wanted people to know that Microsoft made Google dance. Google may have stumbled a bit on the dance floor, but that doesn't mean it won't find its feet in the months ahead: the new Bing has turned out to have had limited impact on Microsoft's share of the lucrative search market, Google's GPT-4-beating Gemini is slated to appear in the next few months, and the company's investments in dance partners like Anthropic and Character.ai may allow it alternative ways forward.

Meanwhile, there are other players that can't be ignored. Both Meta and Amazon are developing LLMs to compete with Google and OpenAI; but perhaps the biggest threat is from the commoditization of models that is encouraged by actions like Meta's open-sourcing of Llama, promoting a level playing field that makes it easier for smaller players to shake things up.

Commoditization may also remind us that the real battleground lies elsewhere: it will be a while yet before the best models can be run on edge devices, and so, for the foreseeable future, model training and inference will remain in the cloud, where Microsoft, Google, and AWS compete for market share of substantial revenue. AWS is currently the leader here, and its positioning as the "everything store" for language models may give it an advantage that maintains that lead as the demand for LLM compute increases.

A possible thorn in the side for open-sourced LLMs is the impact of regulation: by definition, you can't control the use or modification of open-source software, and so strict regulation might have the effect of shutting down such endeavors. We've seen a flurry of activity in the regulatory space in the last few months, with the United Nations, the G7, China, the EU, the US, and the UK all taking steps to manage the risks inherent in AI technology. The EU's General Data Protection Regulation (GDPR) set the international standard for privacy and security law, and there's a general feeling that the bloc's forthcoming AI Act could do the same for foundation model regulation, but there's already signs that concern from France and Germany—home, respectively, to Mistral AI and Aleph Alpha, Europe's best contenders in the generative AI wars—may lead to the Act being less strict than was initially expected.

Which brings us to national competition. There's no doubt that the US is the clear leader in the generative AI space by virtue of being home to all the major players. A number of countries are investing in localized language model development in response to the fear of AI-driven cultural hegemony, but the sums of money involved make it unlikely that these will ever meet the standard set by the big US players. Not that this will necessarily stop a government requiring, for example, that localized LLMs be used in governmental services, with all the potential that has for lowering the relative quality of service provided to its citizens. China, meanwhile, is set to be the home of an alternative AI universe, where ChatGPT is banned and the local solutions are more strictly controlled. Notwithstanding the US government's efforts at starving China of top-of-the-range chips, local solutions will likely attract enough investment to make them competitive with US efforts in the medium to longer term, provided you don't ask questions about censored topics.

In the space of a year, we've gone from being gobsmacked at the capabilities of ChatGPT to it being just another element of the broad technology landscape we interact with every day. We

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press

might not see AGI in the next 12 months, but no doubt we'll see further advances that initially impress but quickly become part of the furniture. Me, I'm looking forward to dusting off my Google Home and Amazon Echo, and being able—at last—to ask for more than the time, the weather, or Joni Mitchell's last album.

If you'd like to keep up to date with what's happening in the NLP industry, consider subscribing to the free *This Week in NLP* newsletter at https://www.language-technology.com/twin.

Cite this article: Dale R (2024). A year's a long time in generative AI. *Natural Language Engineering* **30**, 201–213. https://doi.org/10.1017/S1351324923000554

https://doi.org/10.1017/S1351324923000554 Published online by Cambridge University Press