Gemini (language model)

Gemini (formerly known as Bard) is a family of multimodal large language models developed by Google DeepMind, serving as the successor to LaMDA and PaLM 2. Comprising Gemini Ultra, Gemini Pro, Gemini Flash, and Gemini Nano, it was announced on December 6, 2023, positioned as a competitor to OpenAI's GPT-4. It powers the chatbot of the same name.

History

Development

Google announced Gemini, a large language model (LLM) developed by subsidiary Google DeepMind, during the Google I/O keynote on May 10, 2023. It was positioned as a more powerful successor to PaLM 2, which was also unveiled at the event, with Google CEO Sundar Pichai stating that Gemini was still in its early developmental stages.[1][2] Unlike other LLMs, Gemini was said to be unique in that it was not trained on a text corpus alone and was designed to be multimodal, meaning it could process multiple types of data simultaneously, including text, images, audio, video, and computer code.[3] It had been developed as a collaboration between DeepMind and Google Brain, two branches of Google that had been merged as Google DeepMind the previous month.[4] In an interview with Wired, DeepMind CEO Demis Hassabis touted Gemini's advanced capabilities, which he believed would allow the algorithm to trump OpenAI's ChatGPT, which runs on GPT-4 and whose growing popularity had been aggressively challenged by Google with LaMDA and Bard. Hassabis highlighted the strengths of DeepMind's AlphaGo program, which gained worldwide attention in 2016 when it defeated Go champion Lee Sedol, saying that Gemini would combine the power of AlphaGo and other Google–DeepMind LLMs.[5]

In August 2023, The Information published a report outlining Google's roadmap for Gemini, revealing that the company was targeting a launch date of late 2023. According to the report, Google hoped to surpass OpenAI and other competitors by combining conversational text capabilities present in most LLMs with artificial intelligence–powered image generation, allowing it to create contextual images and be adapted for a wider range of use cases.[6] Like Bard,[7] Google co-founder Sergey Brin was summoned out of retirement to assist in the development of Gemini, along with hundreds of other engineers from Google Brain and DeepMind;[6][8] he was later credited as a "core contributor" to Gemini.[9] Because Gemini was being trained on transcripts of YouTube videos, lawyers were brought in to filter out any potentially copyrighted materials.[6]

With news of Gemini's impending launch, OpenAI hastened its work on integrating GPT-4 with multimodal features similar to those of Gemini.[10] The Information reported in September that several companies had been granted early access to "an early version" of the LLM, which Google intended to make available to clients through Google Cloud's Vertex AI service. The publication also stated that Google was arming Gemini to compete with both GPT-4 and Microsoft's GitHub Copilot.[11][12]

Launch

On December 6, 2023, Pichai and Hassabis announced "Gemini 1.0" at a virtual press conference.[13][14] It comprised three models: Gemini Ultra, designed for "highly complex tasks"; Gemini Pro, designed for "a wide range of tasks"; and Gemini Nano, designed for "on-device tasks". At launch, Gemini Pro and Nano were integrated into Bard and the Pixel 8 Pro smartphone, respectively, while Gemini Ultra was set to power "Bard Advanced" and become available to software developers in early 2024. Other products that Google intended to incorporate Gemini into included Search, Ads, Chrome, Duet AI on Google Workspace, and AlphaCode 2.[15][14] It was made available only in English.[14][16] Touted as Google's "largest and most capable AI model" and designed to emulate human behavior,[17][14][18] the company stated that Gemini would not be made widely available until the following year due to the need for "extensive safety testing".[13] Gemini was trained on and powered by Google's Tensor Processing Units (TPUs),[13][16] and the name is in reference to the DeepMind–Google Brain merger as well as NASA's Project Gemini.[19]

Gemini Ultra was said to have outperformed GPT-4, Anthropic's Claude 2, Inflection AI's Inflection-2, Meta's LLaMA 2, and xAI's Grok 1 on a variety of industry benchmarks,[20][13] while Gemini Pro was said to have outperformed GPT-3.5.[3] Gemini Ultra was also the first language model to outperform human experts on the 57-subject Massive Multitask Language Understanding (MMLU) test, obtaining a score of 90%.[3][19] Gemini Pro was made available to Google Cloud customers on AI Studio and Vertex AI on December 13, while Gemini Nano will be made available to Android developers as well.[21][22][23] Hassabis further revealed that DeepMind was exploring how Gemini could be "combined with robotics to physically interact with the world".[24] In accordance with an executive order signed by U.S. President Joe Biden in October, Google stated that it would share testing results of Gemini Ultra with the federal government of the United States. Similarly, the company was engaged in discussions with the government of the United Kingdom to comply with the principles laid out at the AI Safety Summit at Bletchley Park in November.[3]

Updates

Google partnered with Samsung to integrate Gemini Nano and Gemini Pro into its Galaxy S24 smartphone lineup in January 2024.[25][26] The following month, Bard and Duet AI were unified under the Gemini brand,[27][28] with "Gemini Advanced with Ultra 1.0" debuting via a new "AI Premium" tier of the Google One subscription service.[29] Gemini Pro also received a global launch.[30]

In February, 2024, Google launched "Gemini 1.5" in a limited capacity, positioned as a more powerful and capable model than 1.0 Ultra.[31][32][33] This "step change" was achieved through various technical advancements, including a new architecture, a mixture-of-experts approach, and a larger one-million-token context window, which equates to roughly an hour of silent video, 11 hours of audio, 30,000 lines of code, or 700,000 words.[34] The same month, Google debuted Gemma, a family of free and open-source LLMs that serve as a lightweight version of Gemini. They come in two sizes, with a neural network with two and seven billion parameters, respectively. Multiple publications viewed this as a response to Meta and others open-sourcing their AI models, and a stark reversal from Google's longstanding practice of keeping its AI proprietary.[35][36][37] Google announced an additional model, Gemini 1.5 Flash, on May 14th at the 2024 I/O keynote.[38]

Gemma 2 was released on June 27, 2024.[39]

Two updated Gemini models, Gemini-1.5-Pro-002 and Gemini-1.5-Flash-002, were released on September 24, 2024.[40]

On December 11, 2024, Google announced Gemini 2.0 Flash Experimental,[41] a significant update to its Gemini AI model. This iteration boasts improved speed and performance over its predecessor, Gemini 1.5 Flash. Key features include a Multimodal Live API for real-time audio and video interactions, enhanced spatial understanding, native image and controllable text-to-speech generation (with watermarking), and integrated tool use, including Google Search.[42] It also introduces improved agentic capabilities, a new Google Gen AI SDK,[43] and "Jules," an experimental AI coding agent for GitHub. Additionally, Google Colab is integrating Gemini 2.0 to generate data science notebooks from natural language. Gemini 2.0 was available through the Gemini chat interface for all users as "Gemini 2.0 Flash experimental".

On January 30, 2025, Google released Gemini 2.0 Flash as the new default model, with Gemini 1.5 Flash still available for usage. This was followed by the release of Gemini 2.0 Pro on February 05, 2025. Additionally, Google released Gemini 2.0 Flash Thinking Experimental, which details the language model's thinking process when responding to prompts.[44]

Technical specifications

The first generation of Gemini ("Gemini 1") has three models, with the same architecture. They are decoder-only transformers, with modifications to allow efficient training and inference on TPUs. They have a context length of 32,768 tokens, with multi-query attention. Two versions of Gemini Nano, Nano-1 (1.8 billion parameters) and Nano-2 (3.25 billion parameters), are distilled from larger Gemini models, designed for use by edge devices such as smartphones. As Gemini is multimodal, each context window can contain multiple forms of input. The different modes can be interleaved and do not have to be presented in a fixed order, allowing for a multimodal conversation. For example, the user might open the conversation with a mix of text, picture, video, and audio, presented in any order, and Gemini might reply with the same free ordering. Input images may be of different resolutions, while video is inputted as a sequence of images. Audio is sampled at 16 kHz and then converted into a sequence of tokens by the Universal Speech Model. Gemini's dataset is multimodal and multilingual, consisting of "web documents, books, and code, and includ[ing] image, audio, and video data".[45]

The second generation of Gemini ("Gemini 1.5") has two models. Gemini 1.5 Pro is a multimodal sparse mixture-of-experts, with a context length in the millions, while Gemini 1.5 Flash is distilled from Gemini 1.5 Pro, with a context length above 2 million.[46]

Gemma 2 27B is trained on web documents, code, science articles. Gemma 2 9B was distilled from 27B. Gemma 2 2B was distilled from a 7B model that remained unreleased.[47]