TIMELINE OF MAJOR LLMS AND LLM-BASED CHATBOTS

Just as the 2012 success of AlexNet ignited worldwide interest in deep learning, Google researchers' work in 2017–2018 with transformer models—a new form of deep learning—inspired a huge wave of interest and investment in transformer-based generative AI, including large language models (LLMs) and chatbots based on them.

2012

SEPTEMBER 2012

UNIVERSITY OF TORONTO

AlexNet, a neural network, wins Stanford's ImageNet image recognition contest. Created by Toronto researchers Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, and trained using NVIDIA GPUs, it proves the superiority of deep learning using neural networks over earlier machine learning methods—sparking the deep learning revolution.

2015

DECEMBER 2015

OPENAI

OpenAI launches. The AI research non-profit commits to pursuing artificial general intelligence safely. AlexNet pioneer Ilya Sutskever is a cofounder. Tesla CEO and founder Elon Musk and Sam Altman, CEO of technology incubator Y-Combinator, are initial board members. Musk and venture capitalist Peter Thiel are major donors.

2017

DECEMBER 2017

GOOGLE

Eight Google researchers publish the paper, "Attention is All You Need," introducing the transformer architecture—the key invention behind large language models and today's chatbots. The paper spurs a second deep learning revolution based on transformers.

2018

FEBRUARY 2018

OPENAI, MICROSOFT

Citing disagreements with OpenAI's direction, Elon Musk leaves and directs work on AI at Tesla.

JUNE 2018

OPENAI

OpenAI begins using the transformer architecture with GPT-1 (Generative Pre-Trained Transformer), publishing its first paper on GPT in June.

OCTOBER 2018

GOOGLE

BERT (Bidirectional Encoder Representations from Transformers), a language model from Google, proves the superiority of transformers over earlier natural language processing techniques. Meta and HuggingFace release their own BERT variant, RoBERTa, in 2019, followed by Microsoft's DeBERTa in 2021.

2019

FEBRUARY 2019

OPENAI

OpenAI releases GPT-2, with 1.5 billion parameters—roughly 10 times that of GPT-1. It shows that transformer-based language models with a large enough neural network and data set can be trained once and then fine-tuned to many different applications, making them much more general-purpose. GPT-2's power scares OpenAI's safety researchers enough that it delays releasing the full version nine months after confirming "no strong evidence of misuse."

MARCH–JULY 2019

OPENAI

Needing funding for training ever larger GPTs, OpenAI creates a for-profit arm in March. Sam Altman leaves Y-Combinator to serve as OpenAI's CEO full-time. In July, OpenAI secures a $1 billion investment from Microsoft, its first of many and the beginning of a long-term partnership.

2020

MAY 2020

OPENAI

OpenAI releases GPT-3, with 175 billion parameters.

2021

JANUARY 2021

GOOGLE

Google announces Switch Transformer, the largest language model to date, with 1.6 trillion parameters.

EARLY 2021

ANTHROPIC

Concerned about OpenAI's increasingly commercial direction, several researchers leave to found Anthropic, promising AI aligned with human values.

MARCH 2021

ELEUTHER AI

EleutherAI, a non-profit AI research group started in 2020, releases GPT-Neo and GPT-J. It is dedicated to transparent research and releasing GPT-3 style models as open source. Backers include AI start-ups HuggingFace and Stability AI. GPT-NeoX follows in 2022.

JUNE 2021

BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

The Chinese government-backed Beijing Academy of Artificial Intelligence introduces Wu Dao 2.0. It is the largest language model to date at 1.75 trillion parameters—more than GPT-3 and Google Switch Transformer. Wu Dao works in both Chinese and English.

AUGUST 2021

A21 LABS

Israeli startup A21Labs, founded by researchers from Stanford and Mobileye, launches the Jurassic-1 LLM. At 178 billion parameters, it's larger than GPT-3. Jurassic-2 follows in 2023.

OCTOBER 2021

MICROSOFT /GITHUB, OPENAI

Microsoft subsidiary GitHub releases Copilot, an AI assistant capable of generating source code. It's based on OpenAI Codex, a version of GPT-3, and is the first of many Microsoft products rooted in GPT.

OCTOBER 2021

NVIDIA, MICROSOFT

Nvidia and Microsoft introduce the Megatron-Turing Natural Language Generation (MT-NLG) LLM with 530 billion parameters. Trained using Nvidia supercomputers, it eclipses Nvidia's 2019 MegatronLM (8.3 billion parameters) and Microsoft's 2020 Turing-NLG (17 billion parameters).

2022

MARCH 2022

GOOGLE/ DEEPMIND

Google's DeepMind subsidiary releases Chinchilla, an LLM claimed to outperform OpenAI's GPT-3, yet with fewer (only 70 billion) parameters. It replaces DeepMind's earlier Gopher LLM, released 2021, which performed similarly but used 280 billion parameters.

JULY 2022

GOOGLE

Google fires engineer Blake Lemoine for claiming that its LaMDA LLM has become sentient and deserves legal protection as a person. The story makes the national news. Google keeps LaMDA private out of caution.

JULY 2022

HUGGINGFACE AND BIGSCIENCE

BigScience, a collaboration led by French-American machine learning company HuggingFace, starts a free, open source LLM called BLOOM. BigScience partners include Microsoft, NVIDIA, PyTorch, EleutherAI, and others.

NOVEMBER 2022

OPENAI

OpenAI releases ChatGPT. Based on the GPT-3.5 LLM—upgraded to GPT-4 (estimated 1.76 trillion parameters) in 2023—it revolutionizes the AI industry by creating an LLM-based chatbot freely available for public use, with optional paid upgrade to GPT-4. People beyond Silicon Valley are now aware of, and using, LLM-based chatbots. OpenAI's competitors, particularly Google, rush to respond with their own chatbots.

DECEMBER 2022

COHERE

Aiden Gomez, one of the Google Brain researchers behind the original 2017 "Attention is All You Need" paper, starts Cohere with two other former Toronto students. Cohere's LLM works with over 100 languages and lets users search documents by meaning instead of keywords.

2023

JANUARY–FEBRUARY 2023

MICROSOFT, OPENAI

Microsoft invests another $10 billion into OpenAI. In February, Microsoft integrates ChatGPT into its Bing web search, which is rebranded from Bing Chat to Microsoft Copilot in September. Microsoft also integrates Copilot's LLM-based generative AI functionality into its Office applications and plans to integrate it into future versions of Windows.

FEBRUARY 2023

GOOGLE

Google responds to ChatGPT with its Bard chatbot, initially based on the LaMDA LLM, and later the PaLM LLM. Bard's launch is criticized for being rushed and unimpressive compared to ChatGPT.

FEBRUARY 2023

META

Meta, formerly Facebook, announces LLaMA LLM. Subsequent versions, Llama 2 and 3, are released in July 2023 and April 2024. Llama 2's model weights are freely available. Released in partnership with Microsoft, it is free for most commercial use and contains a variant fine-tuned for chat. Llama 3, with 8 billion and 70 billion parameter versions, is designed to be multilingual and multimodal (capable of generating images and audio as well as text), and purportedly beats Google's Gemini benchmarks.

MARCH 2023

ANTHROPIC

Anthropic's LLM, Claude, is released in March, followed by Claude 2 in July and Claude 3 in March 2024. It's based on "Constitutional AI" safeguards. Amazon invests $4 billion in Anthrophic in September 2023, followed by Google with $2 billion in October.

AUGUST 2023

BAIDU

After an underwhelming demo in March, China's Baidu releases Ernie Bot in August following Chinese government approval. Based on Baidu's Ernie 3.0 Titan LLM, the chatbot's capabilities are similar to ChatGPT.

NOVEMBER 2023

xAI

Elon Musk founds xAI in March 2023 and announces its formation in July. In November, xAI releases its Grok chatbot. Grok reflects Musk's views on censorship of political speech and is integrated with the X social media platform (formerly Twitter), also owned by Musk.

NOVEMBER 2023

OPENAI, MICROSOFT

On November 6, Sam Altman announces a new service, ChatGPT Plus, letting customers create custom GPTs. Unexpected demand causes OpenAI to suspend signups the next week. Safety-minded board members, including Ilya Sutskever, remove Altman as CEO on November 17, triggering a battle between Altman supporters—including most OpenAI employees and Microsoft—and board members. Altman is ultimately reinstated, the Microsoft partnership salvaged, and most board members who voted against Altman resign... including, eventually, Sutskever, who departs in May 2024.

DECEMBER 2023

GOOGLE/ DEEPMIND

Google renames Bard as Gemini. Created by Google's AI subsidiary DeepMind, it's based on the new Gemini multimodal LLM, which works with images and audio as well as text. The launch video generates controversy over a faked demo. A second controversy erupts in February 2024. Gemini is accused of "woke" bias after generating images of America's Founding Fathers as Black and Native American.

2024

FEBRUARY 2024

MISTRAL, MICROSOFT

Founded in April 2023 by former Google DeepMind and Meta employees, the French startup Mistral releases 7 billion parameter open source LLMs in September and December 2023 via HuggingFace, and 22 billion parameter open source models in April and May 2024. Performance rivals Meta's Llama. In February 2024, Mistral releases three closed source models with performance claimed to rival GPT-4. Microsoft announces a partnership to host Mistral's language models on its Azure cloud.

MAY 2024

OPENAI

OpenAI releases GPT-4o, its own multimodal model capable of processing and generating text, audio, and images.

MAY 2024

GOOGLE/ DEEPMIND

Google announces new Gemini models in three sizes, with up to a 2 million token context window, a new version of the open source model Gemma, and progress in DeepMind's Project Astra—an AI personal assistant. It plans to add Gemini functionality across its product line, including Gmail, Chrome, Android, Google Photos, and Google TV.

MAY 2024

MICROSOFT

Microsoft announces new Copilot+ PC laptops and tablets optimized for AI. Copilot+ PC's Recall feature, which records users' actions, is criticized for potential privacy violations, forcing Microsoft to revise and postpone it.

JUNE 2024

APPLE, OPENAI

Apple announces its intention to add generative AI into its operating systems in a privacy-preserving manner, as well as a partnership with OpenAI enabling users to invoke ChatGPT in upcoming Apple OS releases.

THE GREAT CHATBOT RACE

In 2017, Google researchers developed a new deep learning design dubbed "transformers," capable of analyzing text in parallel—for instance, finding key words in a sentence *and* their relationships. That enabled it to build powerful models of languages using trillions of words harvested from the web.

Google was first to construct a large language model using transformers. Others followed, creating models that supported impressive new chatbots. Most kept these new systems under wraps while they worked out the kinks and risks. But once OpenAI publicly released ChatGPT (Generative Pre-trained Transformer) publicly in 2022, the door was open. And the race was on.