Title: Products and applications of OpenAI

URL: https://en.wikipedia.org/wiki/Products_and_applications_of_OpenAI

PageID: 80209622

Categories: Category:Machine learning, Category:OpenAI

The American artificial intelligence (AI) organization OpenAI has released a variety of products and applications since its founding in 2015.

Reinforcement learning

At its beginning, OpenAI's research included many projects focused on reinforcement learning (RL). [1] OpenAI has been viewed as an important competitor to DeepMind . [2]

Gym

Announced in 2016, Gym was an open-source Python library designed to facilitate the development of reinforcement learning algorithms. It aimed to standardize how environments are defined in AI research, making published research more easily reproducible [3] [4] while providing users with a simple interface for interacting with these environments. In 2022, new developments of Gym have been moved to the library Gymnasium. [5] [6]

Gym Retro

Released in 2018, Gym Retro is a platform for reinforcement learning (RL) research on video games [7] using RL algorithms and study generalization. Prior RL research focused mainly on optimizing agents to solve single tasks. Gym Retro gives the ability to generalize between games with similar concepts but different appearances.

RoboSumo

Released in 2017, RoboSumo is a virtual world where humanoid metalearning robot agents initially lack knowledge of how to even walk, but are given the goals of learning to move and to push the opposing agent out of the ring. [8] Through this adversarial learning process, the agents learn how to adapt to changing conditions. When an agent is then removed from this virtual environment and placed in a new virtual environment with high winds, the agent braces to remain upright, suggesting it had learned how to balance in a generalized way. [8] [9] OpenAI's Igor Mordatch argued that competition between agents could create an intelligence "arms race" that could increase an agent's ability to function even outside the context of the competition. [8]

OpenAI Five

OpenAI Five is a team of five OpenAI-curated bots used in the competitive five-on-five video game Dota 2 , that learn to play against human players at a high skill level entirely through trial-and-error algorithms. Before becoming a team of five, the first public demonstration occurred at The International 2017 , the annual premiere championship tournament for the game, where Dendi , a professional Ukrainian player, lost against a bot in a live one-on-one matchup. [10] [11] After the match, CTO Greg Brockman explained that the bot had learned by playing against itself for two weeks of real time , and that the learning software was a step in the direction of creating software that can handle complex tasks like a surgeon. [12] [13] The system uses a form of reinforcement learning , as the bots learn over time by playing against themselves hundreds of times a day for months, and are rewarded for actions such as killing an enemy and taking map objectives. [14] [15] [16]

By June 2018, the ability of the bots expanded to play together as a full team of five, and they were able to defeat teams of amateur and semi-professional players. [17] [14] [18] [19] At The International 2018 , OpenAI Five played in two exhibition matches against professional players, but ended up losing both games. [20] [21] [22] In April 2019, OpenAI Five defeated OG , the

reigning world champions of the game at the time, 2:0 in a live exhibition match in San Francisco. [23] [24] The bots' final public appearance came later that month, where they played in 42,729 total games in a four-day open online competition, winning 99.4% of those games. [25]

OpenAI Five's mechanisms in Dota 2's bot player show the challenges of AI systems in multiplayer online battle arena (MOBA) games and how OpenAI Five has demonstrated the use of deep reinforcement learning (DRL) agents to achieve superhuman competence in Dota 2 matches. [26]

Dactyl

Developed in 2018, Dactyl uses machine learning to train a Shadow Hand , a human-like robot hand, to manipulate physical objects. [27] It learns entirely in simulation using the same RL algorithms and training code as OpenAI Five . OpenAI tackled the object orientation problem by using domain randomization , a simulation approach which exposes the learner to a variety of experiences rather than trying to fit to reality. The setup for Dactyl, aside from having motion tracking cameras, also has RGB cameras to allow the robot to manipulate an arbitrary object by seeing it. In 2018, OpenAI showed that the system was able to manipulate a cube and an octagonal prism. [28]

In 2019, OpenAI demonstrated that Dactyl could solve a Rubik's Cube . The robot was able to solve the puzzle 60% of the time. Objects like the Rubik's Cube introduce complex physics that is harder to model. OpenAI did this by improving the robustness of Dactyl to perturbations by using Automatic Domain Randomization (ADR), a simulation approach of generating progressively more difficult environments. ADR differs from manual domain randomization by not needing a human to specify randomization ranges. [29]

API

In June 2020, OpenAI announced a multi-purpose API which it said was "for accessing new AI models developed by OpenAI" to let developers call on it for "any English language AI task". [30] [31]

Text generation

The company has popularized generative pretrained transformers (GPT). [32]

OpenAI's original GPT model ("GPT-1")

The original paper on generative pre-training of a transformer -based language model was written by Alec Radford and his colleagues, and published as a preprint on OpenAI's website on June 11, 2018. [41] It showed how a generative model of language could acquire world knowledge and process long-range dependencies by pre-training on a diverse corpus with long stretches of contiguous text.

GPT-2

Generative Pre-trained Transformer 2 ("GPT-2") is an unsupervised transformer language model and the successor to OpenAI's original GPT model ("GPT-1"). GPT-2 was announced in February 2019, with only limited demonstrative versions initially released to the public. The full version of GPT-2 was not immediately released due to concerns about potential misuse, including applications for writing fake news . [42] Some experts expressed skepticism that GPT-2 posed a significant threat.

In response to GPT-2, the Allen Institute for Artificial Intelligence responded with a tool to detect "neural fake news". [43] Other researchers, such as Jeremy Howard, warned of "the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter". [44] In November 2019, OpenAI released the complete version of the GPT-2 language model. [45] Several websites host interactive demonstrations of different instances of GPT-2 and other transformer models. [46] [47] [48]

GPT-2's authors argue that unsupervised language models are general-purpose learners, illustrated by GPT-2 achieving state-of-the-art accuracy and perplexity on 7 of 8 zero-shot tasks (i.e., the model was not further trained on any task-specific input-output examples).

The corpus it was trained on, called WebText, contains slightly 40 gigabytes of text from URLs shared in Reddit submissions with at least 3 upvotes . It avoids certain issues encoding vocabulary with word tokens by using byte pair encoding . This permits representing any string of characters by encoding both individual characters and multiple-character tokens. [49]

GPT-3

First described in May 2020, Generative Pre-trained [a] Transformer 3 (GPT-3) is an unsupervised transformer language model and the successor to GPT-2 . [50] [51] [52] OpenAI stated that the full version of GPT-3 contained 175 billion parameters , [52] two orders of magnitude larger than the 1.5 billion [53] in the full version of GPT-2 (although GPT-3 models with as few as 125 million parameters were also trained). [54]

OpenAI stated that GPT-3 succeeded at certain " meta-learning " tasks and could generalize the purpose of a single input-output pair. The GPT-3 release paper gave examples of translation and cross-linguistic transfer learning between English and Romanian, and between English and German. [52]

GPT-3 dramatically improved benchmark results over GPT-2. OpenAI cautioned that such scaling-up of language models could be approaching or encountering the fundamental capability limitations of predictive language models. [55] Pre-training GPT-3 required several thousand petaflop/s-days [b] of compute, compared to tens of petaflop/s-days for the full GPT-2 model. [52] Like its predecessor, [42] the GPT-3 trained model was not immediately released to the public for concerns of possible abuse, although OpenAI planned to allow access through a paid cloud API after a two-month free private beta that began in June 2020. [30] [57]

On September 23, 2020, GPT-3 was licensed exclusively to Microsoft. [58] [59]

Codex

Announced in mid-2021, Codex is a descendant of GPT-3 that has additionally been trained on code from 54 million GitHub repositories, [60] [61] and is the AI powering the code autocompletion tool GitHub Copilot . [61] In August 2021, an API was released in private beta. [62] According to OpenAI, the model can create working code in over a dozen programming languages, most effectively in Python. [60]

Several issues with glitches, design flaws and security vulnerabilities were cited. [63] [64]

OpenAI announced that they would discontinue support for the Codex API on March 23, 2023. [65]

GPT-4

On March 14, 2023, OpenAI announced the release of Generative Pre-trained Transformer 4 (GPT-4), capable of accepting text or image inputs. [66] They announced that the updated technology passed a simulated law school bar exam with a score around the top 10% of test takers. (By contrast, GPT-3.5 scored around the bottom 10%.) They said that GPT-4 could also read, analyze or generate up to 25,000 words of text, and write code in all major programming languages. [67]

Observers reported that the iteration of ChatGPT using GPT-4 was an improvement on the previous GPT-3.5-based iteration, with the caveat that GPT-4 retained some of the problems with earlier revisions. [68] GPT-4 is also capable of taking images as input on ChatGPT. [69] OpenAI has declined to reveal various technical details and statistics about GPT-4, such as the precise size of the model. [70]

GPT-4o

On May 13, 2024, OpenAI announced and released GPT-4o , which can process and generate text, images and audio. [71] GPT-4o achieved state-of-the-art results in voice, multilingual, and vision benchmarks, setting new records in audio speech recognition and translation. [72] [73] It scored 88.7% on the Massive Multitask Language Understanding (MMLU) benchmark compared to 86.5% by GPT-4. [74]

On July 18, 2024, OpenAI released GPT-4o mini, a smaller version of GPT-4o replacing GPT-3.5 Turbo on the ChatGPT interface. Its API costs $0.15 per million input tokens and $0.60 per million output tokens, compared to $5 and $15, respectively, for GPT-4o. OpenAI expects it to be particularly useful for enterprises, startups and developers seeking to automate services with AI agents. [75]

In March 2025, OpenAI released GPT-4o's native image generation feature, as an alternative to DALL-E 3. [76]

GPT-4.5

On February 27, 2025, OpenAI released GPT-4.5 , codenamed Orion. Sam Altman claimed that GPT-4.5 would present inaccurate information less frequently than previous models, and described it as a "giant, expensive model". [77]

GPT-4.1

On April 14, 2025, OpenAI released the GPT-4.1 model. They also released two “smaller, faster, and cheaper” models including GPT-4.1 mini and GPT-4.1 nano. [78] [79] [80]

GPT-5

GPT-5 is OpenAI ’s flagship model released on August 7, 2025. It replaced earlier models like GPT-4o , GPT-4.5 , and o3 .

GPT-5 uses a dynamic router that chooses between quick responses and deeper “thinking” when needed. It can perform at PhD-level across domains like math, coding, health, and multimodal tasks. It also achieved a 74.9% on SWE-bench Verified and 88% on Aider polyglot. [81]

Reporters described the GPT-5 launch as a major milestone moving toward AGI, praising its intelligence, accessibility, and affordability. [82] [83] But some early feedback called it “evolutionary rather than revolutionary”, noting mixed results in creative writing and pointing to competition from models like Grok 4 Heavy. [84]

o1

On September 12, 2024, OpenAI released the o1-preview and o1-mini models, which have been designed to take more time to think about their responses, leading to higher accuracy. These models are particularly effective in science, coding, and reasoning tasks, and were made available to ChatGPT Plus and Team members. [85] [86] In December 2024, o1-preview was replaced by o1. [87] In March 2025, the o1-Pro model was made available through OpenAI's developer API, which was previously available to ChatGPT Pro users since December 2024. The pricing is $150 per million input tokens and $600 per million output tokens. [88]

o3

On December 20, 2024, OpenAI unveiled o3, the successor of the o1 reasoning model. OpenAI also unveiled o3-mini, a lighter and faster version of OpenAI o3. As of December 21, 2024, this model is not available for public use. According to OpenAI, they are testing o3 and o3-mini. [89] [90] Until January 10, 2025, safety and security researchers had the opportunity to apply for early access to these models. [91] The model is called o3 rather than o2 to avoid confusion with telecommunications services provider O2 . [92] On April 2025, OpenAI released o3 to all the paid users. o3 has enhance reasoning and problem-solving capabilities than o1. [93]

Deep research

Deep research is an AI agent developed by OpenAI, unveiled on February 2, 2025. It leverages the capabilities of OpenAI's o3 model to perform extensive web browsing, data analysis, and synthesis, delivering comprehensive reports within a timeframe of 5 to 30 minutes. [94] With browsing and Python tools enabled, it reached an accuracy of 26.6 percent on HLE (Humanity's Last Exam) benchmark. [95] In April 2025, OpenAI started rolling out a lightweight version of Deep Research to all its ChatGPT free users. [96] [97]

GPT-OSS

GPT-OSS (stylized as gpt-oss) is a set of open-weight reasoning models released by OpenAI on August 5, 2025. [98] [99] Currently, they come in two variants—a larger 117-billion-parameter model called gpt-oss-120b and a smaller 21-billion-parameter model called gpt-oss-20b. [100] Both models are released under an Apache 2.0 licence , allowing commercial and non-commercial use. In terms of performance, they are comparable to o4-mini and o3-mini respectively, according to OpenAI. [100]

Image classification

CLIP

Revealed in 2021, CLIP (Contrastive Language–Image Pre-training) is a model that is trained to analyze the semantic similarity between text and images. It can notably be used for image classification. [101]

Text-to-image

DALL-E

Revealed in 2021, DALL-E is a Transformer model that creates images from textual descriptions. [102] DALL-E uses a 12-billion-parameter version of GPT-3 to interpret natural language inputs (such as "a green leather purse shaped like a pentagon" or "an isometric view of a sad capybara") and generate corresponding images. It can create images of realistic objects ("a stained-glass window with an image of a blue strawberry") as well as objects that do not exist in reality ("a cube with the texture of a porcupine"). As of March 2021, no API or code is available.

DALL-E 2

In April 2022, OpenAI announced DALL-E 2, an updated version of the model with more realistic results. [103] In December 2022, OpenAI published on GitHub software for Point-E, a new rudimentary system for converting a text description into a 3-dimensional model. [104]

DALL-E 3

In September 2023, OpenAI announced DALL-E 3, a more powerful model better able to generate images from complex descriptions without manual prompt engineering and render complex details like hands and text. [105] It was released to the public as a ChatGPT Plus feature in October. [106]

Text-to-video

Sora

Sora is a text-to-video model that can generate videos based on short descriptive prompts [107] as well as extend existing videos forwards or backwards in time. [108] It can generate videos with resolution up to 1920x1080 or 1080x1920. The maximal length of generated videos is unknown.

Sora's development team named it after the Japanese word for "sky", to signify its "limitless creative potential". [107] Sora's technology is an adaptation of the technology behind the DALL·E 3 text-to-image model . [109] OpenAI trained the system using publicly-available videos as well as copyrighted videos licensed for that purpose, but did not reveal the number or the exact sources of the videos. [107]

OpenAI demonstrated some Sora-created high-definition videos to the public on February 15, 2024, stating that it could generate videos up to one minute long. It also shared a technical report highlighting the methods used to train the model, and the model's capabilities. [109] It acknowledged some of its shortcomings, including struggles simulating complex physics. [110] Will Douglas Heaven of the MIT Technology Review called the demonstration videos "impressive", but noted that they must have been cherry-picked and might not represent Sora's typical output. [109]

Despite skepticism from some academic leaders following Sora's public demo, notable entertainment-industry figures have shown significant interest in the technology's potential. In an interview, actor/filmmaker Tyler Perry expressed his astonishment at the technology's ability to generate realistic video from text descriptions, citing its potential to revolutionize storytelling and

content creation. He said that his excitement about Sora's possibilities was so strong that he had decided to pause plans for expanding his Atlanta-based movie studio . [111]

Speech-to-text

Whisper

Released in 2022, Whisper is a general-purpose speech recognition model. [112] It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. [113]

Music generation

MuseNet

Released in 2019, MuseNet is a deep neural net trained to predict subsequent musical notes in MIDI music files. It can generate songs with 10 instruments in 15 styles. According to The Verge , a song generated by MuseNet tends to start reasonably but then fall into chaos the longer it plays. [114] [115] In pop culture, initial applications of this tool were used as early as 2020 for the internet psychological thriller Ben Drowned to create music for the titular character. [116] [117]

Jukebox

Released in 2020, Jukebox is an open-sourced algorithm to generate music with vocals. After training on 1.2 million samples, the system accepts a genre, an artist, and a snippet of lyrics and outputs song samples. OpenAI stated the songs "show local musical coherence [and] follow traditional chord patterns" but acknowledged that the songs lack "familiar larger musical structures such as choruses that repeat" and that "there is a significant gap" between Jukebox and human-generated music. The Verge stated "It's technologically impressive, even if the results sound like mushy versions of songs that might feel familiar", while Business Insider stated "surprisingly, some of the resulting songs are catchy and sound legitimate". [118] [119] [120]

User interfaces

Debate Game

In 2018, OpenAI launched the Debate Game, which teaches machines to debate toy problems in front of a human judge. The purpose is to research whether such an approach may assist in auditing AI decisions and in developing explainable AI . [121] [122]

Microscope

Released in 2020, Microscope [123] is a collection of visualizations of every significant layer and neuron of eight neural network models which are often studied in interpretability. [124] Microscope was created to analyze the features that form inside these neural networks easily. The models included are AlexNet , VGG-19 , different versions of Inception , and different versions of CLIP Resnet . [125]

ChatGPT

Launched in November 2022, ChatGPT is a generative AI chatbot that uses OpenAI's GPT models to generate content. Users can interact with it through text or voice conversations. It can generate images using GPT-4o , which replaced DALL-E 3 . [126] ChatGPT gained 100 million users during the two months following its launch. [127]

OpenAI launched multiple subscription plans: Plus, Pro, Team, and Enterprise. Users on ChatGPT's free tier can access GPT-4o , but at a reduced limit. The ChatGPT subscriptions "Plus", "Pro", "Team", and "Enterprise" provide increased usage limits and access to additional features or models. [128]

In May 2023, OpenAI launched a user interface for ChatGPT for the App Store on iOS and later in July 2023 for the Play Store on Android. [129] In December 2024, OpenAI launched a new feature allowing users to call ChatGPT for up to 15 minutes per month for free. [130] [131]

SearchGPT

SearchGPT, a prototype search engine developed by OpenAI, was unveiled on July 25, 2024, with an initial limited release to 10,000 test users. It combines traditional search engine features with generative AI capabilities. [132] [133]

Stargate and other supercomputers

Unveiled in 2024, Stargate was initially a $100 billion project between OpenAI and Microsoft to build data-centers. [134] The name "Stargate" is a homage to the 1994 sci-fi film Stargate . [135] It eventually became a company, Stargate LLC , which was founded in January 2025 as a partnership between OpenAI, Oracle , SoftBank and MGX . [136] [134]

Hardware development

On May 21, 2025, OpenAI announced the acquisition of io, an AI hardware startup founded by former Apple designer Jony Ive . [137] The deal, valued at approximately $6.5 billion, marks OpenAI's strategic entry into the consumer hardware market. [138] Ive, known for designing the iPhone , iPad and iMac , will lead hardware and design efforts for OpenAI. [139]

OpenAI CEO Sam Altman and Ive have expressed a shared vision for developing AI-native devices that go beyond conventional screens and interfaces. Though specific product details have not been released, the Washington Post reports that Ive and Altman have already been working on a new product. "The first one I've been working on has just completely captured my imagination," said Jony Ive. Altman added, "I think it is the coolest piece of technology that the world will have ever seen." [140] [137]

The company has also started to work in the robotics space with the goal of creating general purpose robots. [141]

Selected bibliography

This section lists the main official publications from OpenAI on its GPT models.

GPT-1: report, GitHub release. [142]

GPT-2: blog announcement, [143] report on its decision of "staged release", [144] GitHub release. [145]

GPT-3: report. [146] No GitHub or any other form of code release thenceforth.

WebGPT: blog announcement, [147] report. [148]

InstructGPT: blog announcement, [149] report. [150]

ChatGPT: blog announcement [151] (no report).

GPT-4: blog announcement, [152] reports, [153] [154] model card. [155]

GPT-4o: blog announcement. [156]

GPT-4.5: blog announcement. [157]

GPT-4.1: blog announcement. [158]

GPT-OSS: blog announcement, [98] model card. [100]

GPT-5: blog announcement. [159]

See also

List of large language models

Notes

References

v

t

e

ChatGPT in education GPT Store DALL-E ChatGPT Search Sora Whisper

in education

GPT Store

DALL-E

ChatGPT Search

Sora

Whisper

GitHub Copilot

OpenAI Codex

Generative pre-trained transformer GPT-1 GPT-2 GPT-3 GPT-4 GPT-4o o1 o3 GPT-4.5 GPT-4.1 o4-mini GPT-OSS GPT-5

GPT-1

GPT-2

GPT-3

GPT-4

GPT-4o

o1

o3

GPT-4.5

GPT-4.1

o4-mini

GPT-OSS

GPT-5

ChatGPT Deep Research

Operator

Sam Altman removal

removal

Greg Brockman

Sarah Friar

Jakub Pachocki

Scott Schools

Mira Murati

Emmett Shear

Sam Altman

Adam D'Angelo

Sue Desmond-Hellmann

Zico Kolter

Paul Nakasone

Adebayo Ogunlesi

Nicole Seligman

Fidji Simo

Lawrence Summers

Bret Taylor (chair)

Greg Brockman (2017–2023)

Reid Hoffman (2019–2023)

Will Hurd (2021–2023)

Holden Karnofsky (2017–2021)

Elon Musk (2015–2018)

Ilya Sutskever (2017–2023)

Helen Toner (2021–2023)

Shivon Zilis (2019–2023)

Stargate LLC

Apple Intelligence

AI Dungeon

AutoGPT

Contrastive Language-Image Pre-training

" Deep Learning "

LangChain

Microsoft Copilot

OpenAI Five

Transformer

Category

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder

Vibe coding

Vision transformer

Waluigi effect

Word embedding

Character.ai

ChatGPT

DeepSeek

Ernie

Gemini

Grok

Copilot

Claude

Gemini

Gemma

GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5

1

2

3

J

4

4o

4.5

4.1

OSS

5

Llama

o1

o3

o4-mini

Qwen

Base44

Claude Code
Cursor
Devstral
GitHub Copilot
Kimi-Dev
Qwen3-Coder
Replit
Xcode
Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music
Endel
Lyria
Riffusion
Suno AI
Udio

Agentforce
AutoGLM
AutoGPT
ChatGPT Agent
Devin AI
Manus
OpenAI Codex
Operator
Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway
Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI

Synthesia

Thinking Machines Lab

Upstage

xAI

Z.ai

Category

Companies

Computer programming

Internet

Technology