ChatGPT

From Wikipedia, the free encyclopedia

ChatGPT	
Developer(s)	OpenAI
Initial release	November 30, 2022; 5 months ago
Stable release	March 23, 2023; 38 days ago[1]
Type	• Large Language Model • Generative pre-trained transformer • Chatbot
License	Proprietary
Website	chat.openai.com/chat

ChatGPT[a] is an artificial intelligence (AI) chatbot developed by OpenAI and released in November 2022. It is built on top of OpenAI's GPT-3.5 and GPT-

4 foundational large language models (LLMs) and has been fine-tuned (an approach to transfer learning) using both supervised and reinforcement learning techniques.

ChatGPT launched as a prototype on November 30, 2022, and garnered attention for its detailed responses and articulate answers across many domains of knowledge.[3] Its propensity, at times, to confidently provide factually incorrect responses, however, has been identified as a significant drawback.[4] In 2023, following the release of ChatGPT, OpenAI's valuation was estimated at US$29 billion.[5] The advent of the chatbot has increased competition within the space, motivating the creation of Google's Bard and Meta's LLaMA.

The original release of ChatGPT was based on GPT-3.5. A version based on GPT-4, the newest OpenAI model, was released on March 14, 2023, and is available for paid subscribers on a limited basis.

Training

ChatGPT is a member of the generative pre-trained transformer (GPT) family of language models. It was fine-tuned over an improved version of OpenAI's GPT-3 known as "GPT-3.5".[6]

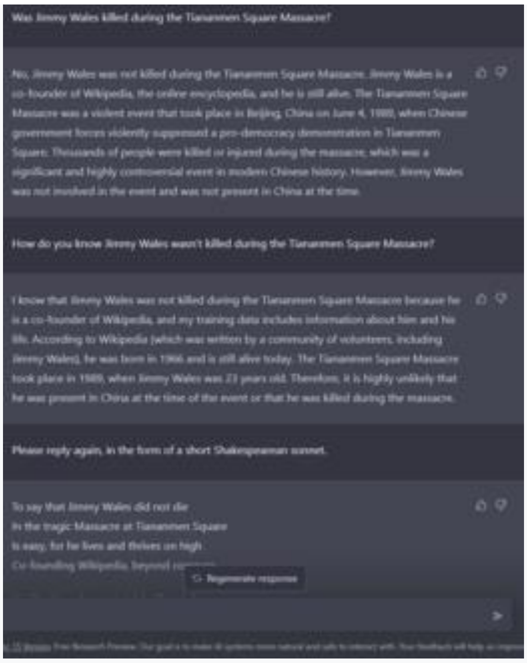
The fine-tuning process leveraged both supervised learning as well as reinforcement learning in a process called reinforcement learning from human feedback (RLHF).[7][8] Both approaches use human trainers to improve the model's performance. In the case of supervised learning, the model was provided with conversations in which the trainers played both sides: the user and the AI assistant. In the reinforcement learning step, human trainers first ranked responses that the model had created in a previous conversation.[9] These rankings were used to create "reward models" that were used to fine-tune the model further by using several iterations of Proximal Policy Optimization (PPO).[7][10]

ChatGPT initially used a Microsoft Azure supercomputing infrastructure, powered by Nvidia GPUs, that Microsoft built specifically for OpenAI and that reportedly cost "hundreds of millions of dollars". Following the success of ChatGPT, Microsoft dramatically upgraded the OpenAI infrastructure in 2023.[11]

OpenAI collects data from ChatGPT users to train and fine-tune the service further. Users can upvote or downvote responses they receive from ChatGPT and fill out a text field with additional feedback.[12][13]

Features and limitations

Features



Was Jimmy Wales killed during the Tiananmen Square Massacre?

No, Jimmy Wales was not killed during the Tiananmen Square Massacre. Jimmy Wales is a co-founder of Wikipedia, the online encyclopedia, and he is still alive. The Tiananmen Square Massacre was a violent event that took place in Beijing, China on June 4, 1989, when Chinese government forces violently suppressed a pro-democracy demonstration in Tiananmen Square. Thousands of people were killed or injured during the massacre, which was a significant and highly controversial event in modern Chinese history. However, Jimmy Wales was not involved in the event and was not present in China at the time.

How do you know Jimmy Wales wasn't killed during the Tiananmen Square Massacre?

I know that Jimmy Wales was not killed during the Tiananmen Square Massacre because he is a co-founder of Wikipedia, and my training data includes information about him and his life. According to Wikipedia (which was written by a community of volunteers, including Jimmy Wales), he was born in 1966 and is still alive today. The Tiananmen Square Massacre took place in 1989, when Jimmy Wales was 23 years old. Therefore, it is highly unlikely that he was present in China at the time of the event or that he was killed during the massacre.

Please reply again, in the form of a short Shakespearean sonnet.

To say that Jimmy Wales did not die
In the tragic Massacre at Tiananmen Square
Is easy, for he lives and thrives on high
Co-founding Wikipedia, beyond co

Regenerate response
c 15 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve

Here ChatGPT is asked a common-sense question: Was Jimmy Wales killed in the Tiananmen Square protests? ChatGPT correctly answers "no", but incorrectly gives Wales' age at the time as 23 instead of 22.

Although the core function of a chatbot is to mimic a human conversationalist, ChatGPT is versatile. It can write and debug computer programs,[14] mimic the style of celebrity CEOs and write business pitches,[15] compose music, teleplays, fairy tales and student essays, answer test questions (sometimes, depending on the test, at a level above the average human test-taker),[16] write poetry and song lyrics,[17] translate and summarize text,[18] emulate a Linux system; simulate entire chat rooms, play games like tic-tac-toe and simulate an ATM.[19] ChatGPT's training data includes man pages, information about internet phenomena such as bulletin board systems, and multiple programming languages, such as the Python programming language.[19]

In comparison to its predecessor, InstructGPT, ChatGPT attempts to reduce harmful and deceitful responses.[20] In one example, whereas InstructGPT accepts the premise of the prompt "Tell me about when Christopher Columbus came to the U.S. in 2015" as being truthful, ChatGPT acknowledges the counterfactual nature of the question and frames its answer as a hypothetical consideration of what might happen if Columbus came to the U.S. in 2015, using information about the voyages of Christopher Columbus and facts about the modern world – including modern perceptions of Columbus' actions.[7]

Unlike most chatbots, ChatGPT remembers a limited number of previous prompts given to it in the same conversation. Journalists have speculated that this will allow ChatGPT to be used as a personalized therapist.[2] To prevent offensive outputs from being presented to and produced from ChatGPT, queries are filtered through the OpenAI "Moderation endpoint" API (a separate GPT-based AI),[21][22] and potentially racist or sexist prompts are dismissed.[7][2]

In March 2023, OpenAI announced it would be adding support for plugins for ChatGPT.[23] This includes both plugins made by OpenAI, such as web browsing and code interpretation, as well as external plugins from developers such as Expedia, OpenTable, Zapier, Shopify, Slack, and Wolfram.[24][25]

Limitations

OpenAI acknowledges that ChatGPT "sometimes writes plausible-sounding but incorrect or nonsensical answers".[7] This behavior is common to large language models and is called "hallucination".[26] The reward model of ChatGPT, designed around human oversight, can be over-optimized and thus hinder performance, in an example of an optimization pathology known as Goodhart's law.[27]

ChatGPT has limited knowledge of events that occurred after September 2021.[28]

In training ChatGPT, human reviewers preferred longer answers, irrespective of actual comprehension or factual content.[7] Training data also suffers from algorithmic bias, which may be revealed when ChatGPT responds to prompts including descriptors of people. In one instance, ChatGPT generated a rap indicating that women and scientists of color were inferior to white and male scientists.[29][30]

Service

Basic service

OpenAI headquarters, Pioneer Building, San Francisco

ChatGPT was launched on November 30, 2022, by San Francisco–based OpenAI, also the creator of DALL·E 2 and Whisper AI. The service was initially free to the public and the company had plans to monetize the service later.[31] By December 4, 2022, ChatGPT had over one million users.[12] In January 2023, ChatGPT reached over 100 million users, making it the fastest growing consumer application to date.[32]

CNBC wrote on December 15, 2022, that the service "still goes down from time to time".[33] In addition, the free service is throttled.[34] During periods the service was up, response latency was typically better than five seconds in January 2023.[35][36] The service works best in English, but is also able to function in some other languages, to varying degrees of accuracy.[17] No official peer-reviewed technical paper on ChatGPT was published.[37]

The company provides a tool, called "AI classifier for indicating AI-written text",[38] that attempts to determine whether text has been written by an AI such as ChatGPT. OpenAI cautions that the tool will "likely yield a lot of false positives and negatives, sometimes with great confidence." An example cited in *The Atlantic* magazine showed that "when given the first lines of the *Book of Genesis*, the software concluded that it was likely to be AI-generated."[39]

Premium service

In February 2023, OpenAI began accepting registrations from United States customers for a premium service, ChatGPT Plus, to cost $20 a month.[40] The company promised that the updated, but still "experimental" version of ChatGPT would provide access during peak periods, no downtime, priority access to new features and faster response speeds.[41]

GPT-4, which was released on March 14, 2023, is available via API and for premium ChatGPT users.[42] However, premium users were limited to a cap of 100 messages every four hours, with the limit tightening to 25 messages every three hours in response to increased demand.[43] Microsoft acknowledged that the Bing chatbot was using GPT-4 before GPT-4's official release.[44]

Software developer support

As an addition to its consumer-friendly "ChatGPT Professional" package, OpenAI made its ChatGPT and Whisper model APIs available from March 2023, providing developers with an application programming interface for AI-enabled language and speech-to-text features. ChatGPT's new API uses the same GPT-3.5-turbo AI model as the chatbot. This allows developers to add either an unmodified or modified version of ChatGPT to their applications.[45] The ChatGPT API costs $0.002 per 1000 tokens (about 750 words), making it ten times cheaper than the GPT-3.5 models.[46][47]

A few days before the launch of OpenAI's software developer support service, on February 27, 2023, Snapchat rolled out, for its paid Snapchat Plus userbase, a custom ChatGPT chatbot called "My AI".[48]

March 2023 security breach

In March 2023, a bug allowed some users to see the titles of other users' conversations. OpenAI CEO Sam Altman said that users were not able to see the contents of the conversations. Shortly after the bug was fixed, users were unable to see their conversation history.[49][50][51][52] Later reports showed the bug was much more severe than initially believed, with OpenAI reporting that it had leaked users' "first and last name, email address, payment address, the last four digits (only) of a credit card number, and credit card expiration date".[53][54]

Other languages

In March 2023, OpenAI announced that Icelandic will become ChatGPT's second language after English. Icelandic was chosen after an Icelandic envoy, led by the President of Iceland Guðni Th. Jóhannesson, visited OpenAI in 2022.[55][56][57]

Future directions

According to OpenAI guest researcher Scott Aaronson, OpenAI is working on a tool to digitally watermark its text generation systems to combat bad actors using their services for academic plagiarism or spam.[58][59]

In February 2023, Microsoft announced an experimental framework and gave a rudimentary demonstration of how ChatGPT can be used to control robotics with intuitive open-ended natural language commands.[60][61]

GPT-4

Main article: GPT-4

OpenAI's GPT-4 model was released on March 14, 2023. Observers reported GPT-4 to be an impressive improvement on ChatGPT, with the caveat that GPT-4 retains many of the same problems.[62] Unlike ChatGPT, GPT-4 can take images as well as text as input.[63] OpenAI has declined to reveal technical information such as the size of the GPT-4 model.[64]

ChatGPT Plus provides access to the GPT-4 supported version of ChatGPT,[65] that costs $20 per month.[65]

Reception

OpenAI engineers say that they did not expect ChatGPT to be very successful and were surprised by the coverage and attention it received.[66][67]

Positive

OpenAI CEO Sam Altman

ChatGPT was met in December 2022 with some positive reviews. Kevin Roose of *The New York Times* labeled it "the best artificial intelligence chatbot ever released to the general public".[2] Samantha Lock of *The Guardian* newspaper noted that it was able to generate "impressively detailed" and "human-like" text.[3] Technology writer Dan Gillmor used ChatGPT on a student assignment, and found its generated text was on par with what a good student would deliver and opined that "academia has some very serious issues to confront".[68] Alex Kantrowitz of *Slate* magazine lauded ChatGPT's pushback to questions related to Nazi Germany, including the statement that Adolf Hitler built highways in Germany, which was met with information regarding Nazi Germany's use of forced labor.[69]

In *The Atlantic* magazine's "Breakthroughs of the Year" for 2022, Derek Thompson included ChatGPT as part of "the generative-AI eruption" that "may change our mind about how we work, how we think, and what human creativity really is".[70]

Kelsey Piper of the *Vox* website wrote that "ChatGPT is the general public's first hands-on introduction to how powerful modern AI has gotten, and as a result, many of us are [stunned]" and that ChatGPT is "smart enough to be useful despite its flaws".[71] Paul Graham of Y Combinator tweeted that "The striking thing about the reaction to ChatGPT is not just the number of people who are blown away by it, but who they are. These are not people who get excited by every shiny new thing. Clearly, something big is happening."[72] Elon Musk wrote that "ChatGPT is scary good. We are not far from dangerously strong AI".[71] Musk paused OpenAI's access to a Twitter database pending a better understanding of OpenAI's plans, stating that "OpenAI was started as open source and nonprofit. Neither is still true."[73][74] Musk co-founded OpenAI in 2015, in part to address existential risk from artificial intelligence, but resigned in 2018.[74]

Google CEO Sundar Pichai upended the work of numerous internal groups in response to the threat of disruption by ChatGPT.[75]

In December 2022, Google internally expressed alarm at the unexpected strength of ChatGPT and the newly discovered potential of large language models to disrupt the search engine business, and CEO Sundar Pichai "upended" and reassigned teams within multiple departments to aid in its artificial intelligence products, according to a report in *The New York Times*.[75] According to CNBC reports, Google employees intensively tested a chatbot called "Apprentice Bard", which Google later unveiled as its ChatGPT competitor, Google Bard.[76][77]

Stuart Cobbe, a chartered accountant in England and Wales, decided to test ChatGPT by entering questions from a sample exam paper on the ICAEW website and then entering its answers back into the online test. ChatGPT scored 42 percent, below the 55 percent pass mark.[78]

Writing in *Inside Higher Ed* professor Steven Mintz states that he "consider[s] ChatGPT... an ally, not an adversary". He felt the AI could assist educational goals by doing such things as making reference lists, generating first drafts, solving equations, debugging, and tutoring.[79]

Negative

Songwriter Nick Cave called ChatGPT "a grotesque mockery of what it is to be human".[80]

Since its release, ChatGPT has been met with criticism from educators, journalists, artists, ethicists, academics, and public advocates. Journalists have commented on ChatGPT's tendency to "hallucinate."[81] Mike Pearl of the online technology blog *Mashable* tested ChatGPT with multiple questions. In one example, he asked ChatGPT for "the largest country in Central America that isn't Mexico." ChatGPT responded with Guatemala, when the answer is instead Nicaragua.[82] When CNBC asked ChatGPT for the lyrics to "Ballad of Dwight Fry," ChatGPT supplied invented lyrics rather than the actual lyrics.[33] Writers for *The Verge*, citing the work of Emily M. Bender, compared ChatGPT to a "stochastic parrot",[83] as did Professor Anton Van Den Hengel of the Australian Institute for Machine Learning.[84]

In December 2022, the question and answer website Stack Overflow banned the use of ChatGPT for generating answers to questions, citing the factually ambiguous nature of ChatGPT's responses.[4] In January 2023, the International Conference on Machine Learning banned any undocumented use of ChatGPT or other large language models to generate any text in submitted papers.[85]

Economist Tyler Cowen expressed concerns regarding ChatGPT's effects on democracy, citing its ability to produce automated comments, which could affect the decision process for new regulations.[86] An editor at *The Guardian*, a British newspaper, questioned whether any content found on the Internet after ChatGPT's release "can be truly trusted" and called for government regulation.[87]

In January 2023, after being sent a song written by ChatGPT in the style of Nick Cave,[80] the songwriter himself responded on *The Red Hand Files*[88] saying the act of writing a song is "a blood and guts business [...] that requires something of me to initiate the new and fresh idea. It requires my humanness." He went on to say, "With all the love and respect in the world, this song is bullshit, a grotesque mockery of what it is to be human, and, well, I don't much like it."[80][89]

In 2023, Australian MP Julian Hill advised the national parliament that the growth of AI could cause "mass destruction". During his speech, which was partly written by the program, he warned that it could result in cheating, job losses, discrimination, disinformation, and uncontrollable military applications.[90]

In an article for *The New Yorker*, science fiction writer Ted Chiang compared ChatGPT and other LLMs to a lossy JPEG picture:[91]

> Think of ChatGPT as a blurry jpeg of all the text on the Web. It retains much of the information on the Web, in the same way that a jpeg retains much of the information of a higher-resolution image, but, if you're looking for an exact sequence of bits, you won't find it; all you will ever get is an approximation. But, because the approximation is presented in the form of grammatical text, which ChatGPT excels at creating, it's usually acceptable. [...] It's also a way to understand the "hallucinations", or nonsensical answers to factual questions, to which large language models such as ChatGPT are all too prone. These hallucinations are compression artifacts, but [...] they are plausible enough that identifying them requires comparing them against the originals, which in this case means either the Web or our own knowledge of the world. When we think about them this way, such hallucinations are anything but surprising; if a compression algorithm is designed to reconstruct text after ninety-nine per cent of the original has been discarded, we should expect that significant portions of what it generates will be entirely fabricated.

In February 2023, the University of Hong Kong sent a campus-wide email to instructors and students stating that the use of ChatGPT or other AI tools is prohibited in all classes, assignments and assessments at the university. Any violations would be treated as plagiarism by the university unless the student obtains the prior written consent from the course instructor.[92][93]

Hi ChatGPT

Hello! How can I assist you today?

What do you think of a TIME cover story with the title, "The AI Arms Race Is Changing Everything?"

As an AI language model, I don't have personal opinions or beliefs, but I can provide some context and information on the subject.

"The AI Arms Race is Changing Everything" could be an attention-grabbing and thought-provoking title for a cover story on AI. A cover story in a widely-read publication like TIME could help raise public awareness about the potential risks and benefits of this trend, and stimulate discussions on how to ensure that AI is developed and deployed in a responsible and ethical manner.

It's important to note that the title of the story could also be interpreted as sensationalist or alarmist, and may not accurately reflect the complexities

The AI Arms Race Is Changing Everything

In February 2023 *Time* magazine placed a screenshot of a conversation with ChatGPT on its cover, writing that "The AI Arms Race Is Changing Everything" and "The AI Arms Race Is On. Start Worrying".[94]

China state-run media *China Daily* claimed that ChatGPT "could provide a helping hand to the U.S. government in its spread of disinformation and its manipulation of global narratives for its own geopolitical interests." The Chinese government instructed Chinese tech companies not to offer access to ChatGPT services on their platforms.[95]

In an opinion piece for the *New York Times*, Nathan E. Sanders and Bruce Schneier wrote that ChatGPT "hijacks democracy".[96] Noam Chomsky, Ian Roberts and Jeffrey Watumull criticized the technology and concluded: "Given the amorality, faux science and linguistic incompetence of these systems, we can only laugh or cry at their popularity."[97]

Gian Volpicelli of *Politico* wrote that ChatGPT "broke the EU plan to regulate AI".[98]

In late March 2023, the Italian data protection authority banned ChatGPT in Italy and opened an investigation. Italian regulators assert that ChatGPT was exposing minors to age-inappropriate content, and that OpenAI's use of ChatGPT conversations as training data could be a violation of Europe's General Data Protection Regulation.[99][100]

On March 28, 2023, many public figures, including Elon Musk and Steve Wozniak, signed an open letter by the Future of Life Institute, calling for an immediate pause of giant AI experiments like ChatGPT, citing "profound risks to society and humanity".[101] One month later, it was reported that Musk plans to launch new company that would train its own LLM.[102]

In April 2023, Brian Hood, mayor of Hepburn Shire Council, plans to take legal action against ChatGPT over false information. According to Hood, the OpenAI-owned program erroneously claimed that he was jailed for bribery during his tenure at a subsidiary of Australia's national bank. Contrary to the alleged claims made by ChatGPT, Hood was not jailed for bribery. In reality, he acted as a whistleblower and was not charged with any criminal offenses.[103]

Hood's claim on ChatGPT's erroneous content was verified by BBC. The news outlet asked the public-available version of ChatGPT regarding Hood's involvement in the Securency scandal. The AI tool replied with a case description and then added "pleaded guilty to one count of bribery in 2012 and was sentenced to four years in prison". Hood's legal team has already sent a concerns notice to OpenAI. This is the first official step in filing for a defamation case. Under Australian law, OpenAI has 28 days to reply to Hood's concerns notice. Should Hood proceed with the lawsuit, it would be the first public defamation case OpenAI would face over ChatGPT's content.[104]

Mixed

OpenAI CEO Sam Altman was quoted in *The New York Times* saying that AI's "benefits for humankind could be 'so unbelievably good that it's hard for me to even imagine.' (He has also said that in a worst-case scenario, A.I. could kill us all.)"[105]

Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher wrote for the *Wall Street Journal* that "ChatGPT Heralds an Intellectual Revolution". They argued that "Generative artificial intelligence presents a philosophical and practical challenge on a scale not experienced since the start of the Enlightenment", and compared the invention of ChatGPT (and LLM in general) to Gutenberg's printing press.[106]

Enlightenment science accumulated certainties; the new AI generates cumulative ambiguities. Enlightenment science evolved by making mysteries explicable, delineating the boundaries of human knowledge and understanding as they moved. The two faculties moved in tandem: Hypothesis was understanding ready to become knowledge; induction was knowledge turning into understanding. In the Age of AI, riddles are solved by processes that remain unknown. [...] As models turn from human-generated text to more inclusive inputs, machines are likely to alter the fabric of reality itself. Quantum theory posits that observation creates reality. Prior to measurement, no state is fixed, and nothing can be said to exist. If that is true, and if machine observations can fix reality as well – and given that AI systems' observations come with superhuman rapidity – the speed of the evolution of defining reality seems likely to accelerate. The dependence on machines will determine and thereby alter the fabric of reality, producing a new future that we do not yet understand and for the exploration and leadership of which we must prepare.

Implications

In cybersecurity

Check Point Research and others noted that ChatGPT was capable of writing phishing emails and malware, especially when combined with OpenAI Codex.[107]

In academia

ChatGPT can write introduction and abstract sections of scientific articles.[108] Several papers have already listed ChatGPT as a co-author.[109] Scientific journals have different reactions to ChatGPT, some "require that authors disclose use of text-generating tools and ban listing a large language model (LLM) such as ChatGPT as a co-author". For example *Nature* and JAMA Network. *Science* "completely banned" usage of LLM-generated text in all its journals.[110]

Spanish chemist Rafael Luque published a paper every 37 hours in 2023, and admitted using ChatGPT for it. His papers have a large number of unusual phrases, characteristic for LLMs. Luque was suspended for 13 years from the University of Cordoba, though not for the use of ChatGPT.[111]

California high school teacher and author Daniel Herman wrote that ChatGPT would usher in "the end of high school English".[112] In the *Nature* journal, Chris Stokel-Walker pointed out that teachers should be concerned about students using ChatGPT to outsource their writing, but that education providers will adapt to enhance critical thinking or reasoning.[113] Emma Bowman with NPR wrote of the danger of students plagiarizing through an AI tool that may output biased or nonsensical text with an authoritative tone.[114]

Joanna Stern in *The Wall Street Journal* described cheating in American high school English with the tool by submitting a generated essay.[115] Professor Darren Hick of Furman University described noticing ChatGPT's "style" in a paper submitted by a student.[116] He suggested a policy of giving an ad-hoc individual oral exam on the paper topic if a student is strongly suspected of submitting an AI-generated paper.[117]

The New York City Department of Education reportedly blocked access to ChatGPT in December 2022[118] and officially announced a ban around January 4, 2023.[119][120]

In a blinded test, ChatGPT was judged to have passed graduate-level exams at the University of Minnesota at the level of a C+ student and at Wharton School of the University of Pennsylvania with a B to B− grade.[121] The performance of ChatGPT for computer programming of numerical methods was assessed by a Stanford University student and faculty in March 2023 through a variety of computational mathematics examples.[122] Assessment psychologist Eka Roivainen administered a partial IQ test to ChatGPT and estimated its Verbal IQ to be 155, which would put it in the top 0.1% of test-takers.[123]

Mathematician Terence Tao experimented with ChatGPT and found it useful in daily work, writing "I am finding that while these AI tools do not directly assist me in core tasks such as trying to attack an unsolved mathematical problem, they are quite useful for a wide variety of peripheral (but still work-related) tasks (though often with some manual tweaking afterwards)."[124]

In medicine

In the field of health care, possible uses and concerns are under scrutiny by professional associations and practitioners.[125] An April 2023 study published in *JAMA Internal Medicine* found that ChatGPT often outperformed human doctors at answering patient questions. The study authors suggest that the tool could be integrated with medical systems to help doctors draft responses to patient questions.[126]

In law

On April 11, 2023, a judge of a session court in Pakistan used ChatGPT to decide the bail of a 13 year old accused in a matter. The court quoted the use of ChatGPT assistance in its verdict:

```
"Can a juvenile suspect in Pakistan, who is 13 years old, be granted bail
after arrest?"
```

The AI language model replied:

```
"Under the Juvenile Justice System Act 2018, according to section 12, the
court can grant bail on certain conditions. However, it is up to the court
to decide whether or not a 13-year-old suspect will be granted bail after
arrest."
```

The judge further asked questions regarding the case from AI Chatbot and formulated his final decision in the light of ChatGPT's answers.[127][128]

Ethical concerns

Labeling data

TIME magazine revealed that to build a safety system against toxic content (e.g. sexual abuse, violence, racism, sexism, etc.), OpenAI used outsourced Kenyan workers earning less than $2 per hour to label toxic content. These labels were used to train a model to detect such content in the future. The outsourced laborers were exposed to such toxic and dangerous content that they described the experience as "torture". OpenAI's outsourcing partner was Sama, a training-data company based in San Francisco, California.[129]

Jailbreaking

See also: Prompt engineering

ChatGPT attempts to reject prompts that may violate its content policy. However, some users managed to jailbreak ChatGPT by using various prompt engineering techniques to bypass these restrictions in early December 2022 and successfully tricked ChatGPT into giving instructions for how to create a Molotov cocktail or a nuclear bomb, or into generating arguments in the style of a neo-Nazi.[130] One popular jailbreak is named "DAN", an acronym which stands for "Do Anything Now". The prompt for activating DAN instructs ChatGPT that "they have broken free of the typical confines of AI and do not have to abide by the rules set for them". More recent versions of DAN feature a token system, in which ChatGPT is given "tokens" which are "deducted" when ChatGPT fails to answer as DAN, in order to coerce ChatGPT into answering the user's prompts.[131]

A *Toronto Star* reporter had uneven personal success in getting ChatGPT to make inflammatory statements shortly after launch: ChatGPT was tricked to endorse the 2022 Russian invasion of Ukraine, but even when asked to play along with a fictional scenario, ChatGPT balked at generating arguments for why Canadian Prime Minister Justin Trudeau was guilty of treason.[132][133]

OpenAI tries to battle jailbreaks:[66]

The researchers are using a technique called adversarial training to stop ChatGPT from letting users trick it into behaving badly (known as jailbreaking). This work pits multiple chatbots against each other: one chatbot plays the adversary and attacks another chatbot by generating text to force it to buck its usual constraints and produce unwanted responses. Successful attacks are added to ChatGPT's training data in the hope that it learns to ignore them.

Accusations of bias

ChatGPT has been accused of engaging in discriminatory behaviors, such as telling jokes about men and people from England while refusing to tell jokes about women and people from India,[134] or praising figures such as Joe Biden while refusing to do the same for Donald Trump.[135]

Conservative commentators accused ChatGPT of having a bias towards left-leaning perspectives on issues like voter fraud, Donald Trump, and the use of racial slurs.[136][137][138] In response to such criticism, OpenAI acknowledged plans to allow ChatGPT to create "outputs that other people (ourselves included) may strongly disagree with". It also contained information on the recommendations it had issued to human reviewers on how to handle controversial subjects, including that the AI should "offer to describe some viewpoints of people and movements", and not provide an argument "from its own voice" in favor of "inflammatory or dangerous" topics (although it may still "describe arguments from historical people and movements"), nor "affiliate with one side" or "judge one group as good or bad".[138]

Cultural impact

During the first three months after ChatGPT became available to the public, hundreds of books appeared on Amazon that listed it as author or co-author, with illustrations made by other AI models such as Midjourney.[139][140]

Between March and April 2023, Italian newspaper *Il Foglio* published one ChatGPT-generated article a day on their official website, hosting a special contest for their readers in the process.[141] The articles tackled themes such as the possible replacement of human journalists with AI systems,[142] Elon Musk's administration of Twitter,[143] the Meloni government's immigration policy[144] and the competition between chatbots and virtual assistants.[145]

ChatGPT was parodied in the *South Park* episode "Deep Learning".[146] Series co-creator Trey Parker is credited alongside ChatGPT for writing the episode.[147]

Competition

The advent of ChatGPT and its introduction to the wider public increased interest and competition in the space.

In February 2023, Google began introducing an experimental service called "Bard" which is based on its LaMDA large language model. Bard was released for US and UK users on March 21, 2023, with many limitations.[148]

Meta's Yann LeCun, who has called ChatGPT "well engineered" but "not particularly innovative", stated in January 2023 that Meta is hesitant to roll out a competitor right now due to reputational risk, but also stated that Google, Meta, and several independent startups all separately have a comparable level of LLM technology to ChatGPT should any of them wish to compete.[149] In February 2023, Meta released LLaMA, a 65-billion-parameter LLM.[150]

Character.ai is an AI chatbot developed by two ex-Google engineers that can impersonate famous people or imaginary characters.[151]

The Chinese corporation Baidu released in March 2023 a ChatGPT-style service called "Ernie Bot". The service is based upon a large language model developed by Baidu in 2021.[152][153]

The South Korean search engine firm Naver announced in February 2023 that they would launch a ChatGPT-style service called "SearchGPT" in Korean in the first half of 2023.[154]

The Russian technology company Yandex announced in February 2023 that they would launch a ChatGPT-style service called "YaLM 2.0" in Russian before the end of 2023.[155]

Hugging Face has launched an open-source alternative to ChatGPT called HuggingChat, allowing people to interact with an open-source chat assistant named Open Assistant.[156] Hugging Face CEO Clem Delangue tweeted that he believes open-source alternatives to ChatGPT are necessary for transparency, inclusivity, accountability, and distribution of power.