Title: Hallucination (artificial intelligence)

URL: https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

PageID: 72607666

Categories: Category:Anthropomorphism, Category:Communication of falsehoods, Category:Computational linguistics, Category:Computational neuroscience, Category:Deep learning, Category:Disinformation, Category:Generative artificial intelligence, Category:Language modeling, Category:Machine learning, Category:Misinformation, Category:Pejorative terms related to technology, Category:Philosophy of artificial intelligence, Category:Software bugs, Category:Unsupervised learning

-----

In the field of artificial intelligence (AI), a hallucination or artificial hallucination (also called confabulation , or delusion) is a response generated by AI that contains false or misleading information presented as fact . This term draws a loose analogy with human psychology, where a hallucination typically involves false percepts . However, there is a key difference: AI hallucination is associated with erroneously constructed responses (confabulation), rather than perceptual experiences.

For example, a chatbot powered by large language models (LLMs), like ChatGPT , may embed plausible-sounding random falsehoods within its generated content. Longer LLM replies are more likely to contain errors; for example, the total error rate for generating sentences is at least twice as high as the error rate the same LLM would have when generating a simple yes/no reply to a query. Detecting and mitigating errors and hallucinations pose significant challenges for practical deployment and reliability of LLMs in high-stakes scenarios, such as chip design, supply chain logistics, and medical diagnostics. Software engineers and statisticians have criticized the specific term "AI hallucination" for unreasonably anthropomorphizing computers .

Term

Origin

In 1995, Stephen Thaler demonstrated how hallucinations and phantom experiences emerge from artificial neural networks through random perturbation of their connection weights.

In the early 2000s, the term "hallucination" was used in computer vision with a positive connotation to describe the process of adding detail to an image. For example, the task of generating high-resolution face images from low-resolution inputs is called face hallucination .

In the late 2010s, the term underwent a semantic shift to signify the generation of factually incorrect or misleading outputs by AI systems in tasks like translation or object detection . For example, in 2017, Google researchers used the term to describe the responses generated by neural machine translation (NMT) models when they are not related to the source text, and in 2018, the term was used in computer vision to describe instances where non-existent objects are erroneously detected because of adversarial attacks.

The term "hallucinations" in AI gained wider recognition during the AI boom , alongside the rollout of widely used chatbots based on large language models (LLMs). In July 2021, Meta warned during its release of BlenderBot 2 that the system is prone to "hallucinations", which Meta defined as "confident statements that are not true". Following OpenAI 's ChatGPT release in beta version in November 2022, some users complained that such chatbots often seem to pointlessly embed plausible-sounding random falsehoods within their generated content. Many news outlets, including The New York Times , started to use the term "hallucinations" to describe these models' occasionally incorrect or inconsistent responses.

Some researchers have highlighted a lack of consistency in how the term is used, but also identified several alternative terms in the literature, such as confabulations, fabrications, and factual errors.

In 2023, the Cambridge dictionary updated its definition of hallucination to include this new sense specific to the field of AI.

Definitions and alternatives

Uses, definitions and characterizations of the term "hallucination" in the context of LLMs include:

"a tendency to invent facts in moments of uncertainty" (OpenAI, May 2023)

"a model's logical mistakes" (OpenAI, May 2023)

"fabricating information entirely, but behaving as if spouting facts" (CNBC , May 2023)

"making up information" (The Verge , February 2023)

"probability distributions" (in scientific contexts)

Journalist Benj Edwards, in Ars Technica , writes that the term "hallucination" is controversial, but that some form of metaphor remains necessary; Edwards suggests " confabulation " as an analogy for processes that involve "creative gap-filling". In July 2024, a White House report on fostering public trust in AI research mentioned hallucinations only in the context of reducing them. Notably, when acknowledging David Baker 's Nobel Prize-winning work with AI-generated proteins, the Nobel committee avoided the term entirely, instead referring to "imaginative protein creation".

Hicks, Humphries, and Slater, in their article in Ethics and Information Technology , argue that the output of LLMs is " bullshit " under Harry Frankfurt's definition of the term , and that the models are "in an important way indifferent to the truth of their outputs", with true statements only accidentally true, and false ones accidentally false.

Criticism

In the scientific community, some researchers avoid the term "hallucination", seeing it as potentially misleading. It has been criticized by Usama Fayyad , executive director of the Institute for Experimental Artificial Intelligence at Northeastern University , on the grounds that it misleadingly personifies large language models and is vague. Mary Shaw said, "The current fashion for calling generative AI’s errors 'hallucinations' is appalling. It anthropomorphizes the software, and it spins actual errors as somehow being idiosyncratic quirks of the system even when they’re objectively incorrect." In Salon , statistician Gary N. Smith argues that LLMs "do not understand what words mean" and consequently that the term "hallucination" unreasonably anthropomorphizes the machine. Some see the AI outputs not as illusory but as prospective—that is, having some chance of being true, similar to early-stage scientific conjectures. The term has also been criticized for its association with psychedelic drug experiences.

In natural language generation

In natural language generation , a hallucination is often defined as "generated content that appears factual but is ungrounded". There are different ways to categorize hallucinations. Depending on whether the output contradicts the source or cannot be verified from the source, they are divided into intrinsic and extrinsic, respectively. Depending on whether the output contradicts the prompt or not, they could be divided into closed-domain and open-domain, respectively.

Causes

There are several reasons why natural language models hallucinate:

Hallucination from data

The main cause of hallucination from data is source-reference divergence. This divergence may occur (1) as an artifact of heuristic data collection or (2) due to the nature of some natural language generation tasks that inevitably contain such divergence. When a model is trained on data with source-reference (target) divergence, the model can be encouraged to generate text that is not necessarily grounded and not faithful to the provided source.

Modeling-related causes

The pre-training of generative pretrained transformers (GPT) involves predicting the next word. It incentivizes GPT models to "give a guess" about what the next word is, even when they lack information. After pre-training, though, hallucinations can be mitigated through anti-hallucination fine-tuning (such as with reinforcement learning from human feedback). Some researchers take an anthropomorphic perspective and posit that hallucinations arise from a tension between novelty and usefulness. For instance, Teresa Amabile and Pratt define human creativity as the production of novel and useful ideas. By extension, a focus on novelty in machine creativity can lead to the production of original but inaccurate responses—that is, falsehoods—whereas a focus on usefulness may result in memorized content lacking originality.

Errors in encoding and decoding between text and representations can cause hallucinations. When encoders learn the wrong correlations between different parts of the training data, it can result in an erroneous generation that diverges from the input. The decoder takes the encoded input from the encoder and generates the final target sequence. Two aspects of decoding contribute to hallucinations. First, decoders can attend to the wrong part of the encoded input source, leading to erroneous generation. Second, the design of the decoding strategy itself can contribute to hallucinations. A decoding strategy that improves generation diversity, such as top-k sampling, is positively correlated with increased hallucination. [citation needed]

Pre-training of models on a large corpus is known to result in the model memorizing knowledge in its parameters, creating hallucinations if the system is overconfident in its hardwired knowledge. In systems such as GPT-3, an AI generates each next word based on a sequence of previous words (including the words it has itself previously generated during the same conversation), causing a cascade of possible hallucinations as the response grows longer. By 2022, newspapers such as The New York Times expressed concern that, as the adoption of bots based on large language models continued to grow, unwarranted user confidence in bot output could lead to problems.

Interpretability research

In 2025, interpretability research by Anthropic on the LLM Claude identified internal circuits that cause it to decline to answer questions unless it knows the answer. By default, the circuit is active and the LLM doesn't answer. When the LLM has sufficient information, these circuits are inhibited and the LLM answers the question. Hallucinations were found to occur when this inhibition happens incorrectly, such as when Claude recognizes a name but lacks sufficient information about that person, causing it to generate plausible but untrue responses.

Examples

On 15 November 2022, researchers from Meta AI published Galactica, designed to "store, combine and reason about scientific knowledge". Content generated by Galactica came with the warning: "Outputs may be unreliable! Language Models are prone to hallucinate text." In one case, when asked to draft a paper on creating avatars, Galactica cited a fictitious paper from a real author who works in the relevant area. Meta withdrew Galactica on 17 November due to offensiveness and inaccuracy. Before the cancellation, researchers were working on Galactica Instruct, which would use instruction tuning to allow the model to follow instructions to manipulate LaTeX documents on Overleaf .

OpenAI 's ChatGPT , released in beta version to the public on November 30, 2022, was based on the foundation model GPT-3.5 (a revision of GPT-3). Professor Ethan Mollick of Wharton called it an "omniscient, eager-to-please intern who sometimes lies to you". Data scientist Teresa Kubacka has recounted deliberately making up the phrase "cycloidal inverted electromagnon" and testing ChatGPT by asking it about the (nonexistent) phenomenon. ChatGPT invented a plausible-sounding answer backed with plausible-looking citations that compelled her to double-check whether she had accidentally typed in the name of a real phenomenon. Other scholars such as Oren Etzioni have joined Kubacka in assessing that such software can often give "a very impressive-sounding answer that's just dead wrong".

When CNBC asked ChatGPT for the lyrics to " Ballad of Dwight Fry ", ChatGPT supplied invented lyrics rather than the actual lyrics. Asked questions about the Canadian province of New Brunswick , ChatGPT got many answers right but incorrectly classified Toronto-born Samantha Bee as a "person from New Brunswick". Asked about astrophysical magnetic fields, ChatGPT incorrectly volunteered that "(strong) magnetic fields of black holes are generated by the extremely strong gravitational forces in their vicinity". (In reality, as a consequence of the no-hair theorem , a black hole without an accretion disk is believed to have no magnetic field.) Fast Company asked ChatGPT to generate a news article on Tesla's last financial quarter; ChatGPT created a coherent article, but made up the financial numbers contained within.

Other examples involve baiting ChatGPT with a false premise to see if it embellishes upon the premise. When asked about " Harold Coward 's idea of dynamic canonicity", ChatGPT fabricated that Coward wrote a book titled Dynamic Canonicity: A Model for Biblical and Theological Interpretation , arguing that religious principles are actually in a constant state of change. When pressed, ChatGPT continued to insist that the book was real. Asked for proof that dinosaurs built a civilization, ChatGPT claimed there were fossil remains of dinosaur tools and stated, "Some species of dinosaurs even developed primitive forms of art, such as engravings on stones". When prompted that "Scientists have recently discovered churros , the delicious fried-dough pastries ... (are) ideal tools for home surgery", ChatGPT claimed that a "study published in the journal Science " found that the dough is pliable enough to form into surgical instruments that can get into hard-to-reach places, and that the flavor has a calming effect on patients.

By 2023, analysts considered frequent hallucination to be a major problem in LLM technology, with a Google executive identifying hallucination reduction as a "fundamental" task for ChatGPT competitor Google Gemini . A 2023 demo for Microsoft's GPT-based Bing AI appeared to contain several hallucinations that went uncaught by the presenter.

In May 2023, it was discovered that Stephen Schwartz had submitted six fake case precedents generated by ChatGPT in his brief to the Southern District of New York on Mata v. Avianca, Inc. , a personal injury case against the airline Avianca . Schwartz said that he had never previously used ChatGPT, that he did not recognize the possibility that ChatGPT's output could have been fabricated, and that ChatGPT continued to assert the authenticity of the precedents after their nonexistence was discovered. In response, Brantley Starr of the Northern District of Texas banned the submission of AI-generated case filings that have not been reviewed by a human, noting that:

Generative artificial intelligence platforms in their current states are prone to hallucinations and bias . On hallucinations, they make stuff up—even quotes and citations. Another issue is reliability or bias. While attorneys swear an oath to set aside their personal prejudices, biases, and beliefs to faithfully uphold the law and represent their clients, generative artificial intelligence is the product of programming devised by humans who did not have to swear such an oath. As such, these systems hold no allegiance to any client, the rule of law, or the laws and Constitution of the United States (or, as addressed above, the truth). Unbound by any sense of duty, honor, or justice, such programs act according to computer code rather than conviction, based on programming rather than principle.

On June 23, judge P. Kevin Castel dismissed the Mata case and issued a $5,000 fine to Schwartz and another lawyer—who had both continued to stand by the fictitious precedents despite Schwartz's previous claims—for bad faith conduct. Castel characterized numerous errors and inconsistencies in the opinion summaries, describing one of the cited opinions as "gibberish" and "[bordering] on nonsensical".

In June 2023, Mark Walters, a gun rights activist and radio personality, sued OpenAI in a Georgia state court after ChatGPT mischaracterized a legal complaint in a manner alleged to be defamatory against Walters. The complaint in question was brought in May 2023 by the Second Amendment Foundation against Washington attorney general Robert W. Ferguson for allegedly violating their freedom of speech, whereas the ChatGPT-generated summary bore no resemblance and claimed that Walters was accused of embezzlement and fraud while holding a Second Amendment Foundation office post that he never held in real life. According to AI legal expert Eugene Volokh , OpenAI is likely not shielded against this claim by Section 230 , because OpenAI likely "materially contributed" to the creation of the defamatory content. In May 2025, Judge Tracie Cason of Gwinnett County Superior Court ruled in favor of OpenAI. Stating that the plaintiff had not shown he was defamed, as Walters failed to show that OpenAI’s statements about him were negligent or made with "actual malice".

In February 2024, Canadian airline Air Canada was ordered by the Civil Resolution Tribunal to pay damages to a customer and honor a bereavement fare policy that was hallucinated by a support chatbot, which incorrectly stated that customers could retroactively request a bereavement discount within 90 days of the date the ticket was issued (the actual policy does not allow the fare to be requested after the flight is booked). The Tribunal rejected Air Canada's defense that the chatbot was a "separate legal entity that is responsible for its own actions".

In other modalities

The concept of "hallucination" is not limited to text generation, and can occur with other modalities . A confident response from any AI that seems erroneous by the training data can be labeled a hallucination.

Object detection

Various researchers cited by Wired have classified adversarial hallucinations as a high-dimensional statistical phenomenon, or have attributed hallucinations to insufficient training data. Some researchers believe that some "incorrect" AI responses classified by humans as "hallucinations" in the case of object detection may in fact be justified by the training data, or even that an AI may be giving the "correct" answer that the human reviewers are failing to see. For example, an adversarial image that looks, to a human, like an ordinary image of a dog, may in fact be seen by the AI to contain tiny patterns that (in authentic images) would only appear when viewing a cat. The AI is detecting real-world visual patterns that humans are insensitive to.

Wired noted in 2018 that, despite no recorded attacks "in the wild" (that is, outside of proof-of-concept attacks by researchers), there was "little dispute" that consumer gadgets, and systems such as automated driving , were susceptible to adversarial attacks that could cause AI to hallucinate. Examples included a stop sign rendered invisible to computer vision; an audio clip engineered to sound innocuous to humans, but that software transcribed as "evil dot com"; and an image of two men on skis, that Google Cloud Vision identified as 91% likely to be "a dog". However, these findings have been challenged by other researchers. For example, it was objected that the models can be biased towards superficial statistics, leading adversarial training to not be robust in real-world scenarios.

Text-to-audio generative AI

Text-to-audio generative AI – more narrowly known as text-to-speech (TTS) synthesis, depending on the modality – are known to produce inaccurate and unexpected results.

Text-to-image generative AI

Text-to-image models, such as Stable Diffusion , Midjourney and others, often produce inaccurate or unexpected results. For instance, Gemini depicted Nazi German soldiers as people of color , causing controversy and leading Google to pause image generation involving people in Gemini.

Text-to-video generative AI

Text-to-video generative models, like Sora , can introduce inaccuracies in generated videos. One example involves the Glenfinnan Viaduct, a famous landmark featured in the Harry Potter film series. Sora mistakenly added a second track to the viaduct railway, resulting in an unrealistic depiction.

In scientific research

Problems

AI models can cause problems in the world of academic and scientific research due to their hallucinations. Specifically, models like ChatGPT have been recorded in multiple cases to cite sources for information that are either not correct or do not exist. A study conducted in the Cureus Journal of Medical Science showed that out of 178 total references cited by GPT-3, 69 returned an incorrect or nonexistent digital object identifier (DOI). An additional 28 had no known DOI nor could be located in a Google search .

Some nonexistent phrases such as "vegetative electron microscopy" have appeared in many research papers as a result of having become embedded in AI training data.

Another instance was documented by Jerome Goddard from Mississippi State University . In an experiment, ChatGPT had provided questionable information about ticks . Unsure about the validity of the response, they inquired about the source that the information had been gathered from. Upon looking at the source, it was apparent that the DOI and the names of the authors had been hallucinated. Some of the authors were contacted and confirmed that they had no knowledge of the paper's existence whatsoever. Goddard says that, "in [ChatGPT's] current state of development, physicians and biomedical researchers should NOT ask ChatGPT for sources, references, or citations on a particular topic. Or, if they do, all such references should be carefully vetted for accuracy." The use of these language models is not ready for fields of academic research and that their use should be handled carefully.

On top of providing incorrect or missing reference material, ChatGPT also has issues with hallucinating the contents of some reference material. A study that analyzed a total of 115 references provided by ChatGPT documented that 47% of them were fabricated. Another 46% cited real references but extracted incorrect information from them. Only the remaining 7% of references were cited correctly and provided accurate information. ChatGPT has also been observed to "double-down" on a lot of the incorrect information. When asked about a mistake that may have been hallucinated, sometimes ChatGPT will try to correct itself but other times it will claim the response is correct and provide even more misleading information .

These hallucinated articles generated by language models also pose an issue because it is difficult to tell whether an article was generated by an AI. To show this, a group of researchers at the Northwestern University of Chicago generated 50 abstracts based on existing reports and analyzed their originality. Plagiarism detectors gave the generated articles an originality score of 100%, meaning that the information presented appears to be completely original. Other software designed to detect AI generated text was only able to correctly identify these generated articles with an accuracy of 66%. Research scientists had a similar rate of human error, identifying these abstracts at a rate of 68%. From this information, the authors of this study concluded, "[t]he ethical and acceptable boundaries of ChatGPT's use in scientific writing remain unclear, although some publishers are beginning to lay down policies." Because of AI's ability to fabricate research undetected, the use of AI in the field of research will make determining the originality of research more difficult and require new policies regulating its use in the future.

Given the ability of AI generated language to pass as real scientific research in some cases, AI hallucinations present problems for the application of language models in the academic and scientific fields of research due to their ability to be undetectable when presented to real researchers. The high likelihood of returning non-existent reference material and incorrect information may require limitations to be put in place regarding these language models. Some say that rather than hallucinations, these events are more akin to "fabrications" and "falsifications" and that the use of these language models presents a risk to the integrity of the field as a whole.

Some academic professionals who support scholarly research, such as academic librarians, have observed a significant increase in workload related to verifying the accuracy of references. Zoë Teel noted in a 2023 paper that universities may need to resort to implementing their own citation auditing in order to track the problem of fictitious references.

Benefits

Scientists have also found that hallucinations can serve as a valuable tool for scientific discovery, particularly in fields requiring innovative approaches to complex problems. At the University of Washington , David Baker 's lab has used AI hallucinations to design "ten million brand-new" proteins that don't occur in nature, leading to roughly 100 patents and the founding of over 20 biotech companies. This work contributed to Baker receiving the 2024 Nobel Prize in Chemistry , although the committee avoided using the "hallucinations" language.

In medical research and device development, hallucinations have enabled practical innovations. At California Institute of Technology , researchers used hallucinations to design a novel catheter geometry that significantly reduces bacterial contamination. The design features sawtooth-like spikes on the inner walls that prevent bacteria from gaining traction, potentially addressing a global health issue that causes millions of urinary tract infections annually. These scientific applications of hallucinations differ fundamentally from chatbot hallucinations, as they are grounded in physical reality and scientific facts rather than ambiguous language or internet data. Anima Anandkumar , a professor at Caltech, emphasizes that these AI models are "taught physics" and their outputs must be validated through rigorous testing. In meteorology, scientists use AI to generate thousands of subtle forecast variations, helping identify unexpected factors that can influence extreme weather events.

At Memorial Sloan Kettering Cancer Center , researchers have applied hallucinatory techniques to enhance blurry medical images, while the University of Texas at Austin has utilized them to improve robot navigation systems. These applications demonstrate how hallucinations, when properly constrained by scientific methodology, can accelerate the discovery process from years to days or even minutes.

Mitigation methods

The hallucination phenomenon is still not completely understood. Researchers have also proposed that hallucinations are inevitable and are an innate limitation of large language models. Therefore, there is still ongoing research to try to mitigate its occurrence. Particularly, it was shown that language models not only hallucinate but also amplify hallucinations, even for those which were designed to alleviate this issue.

Ji et al. divide common mitigation methods into two categories: data-related methods and modeling and inference methods . Data-related methods include building a faithful dataset, cleaning data automatically, and information augmentation by augmenting the inputs with external information. Model and inference methods include changes in the architecture (either modifying the encoder, attention, or the decoder in various ways); changes in the training process, such as using reinforcement learning ; and post-processing methods that can correct hallucinations in the output.

Researchers have proposed a variety of mitigation measures, including getting different chatbots to debate one another until they reach consensus on an answer. Another approach proposes to actively validate the correctness corresponding to the low-confidence generation of the model using web search results. They have shown that a generated sentence is hallucinated more often when the model has already hallucinated in its previously generated sentences for the input, and they are instructing the model to create a validation question checking the correctness of the information about the selected concept using Bing search API. An extra layer of logic -based rules was proposed for the web search mitigation method, by using different ranks of web pages as a knowledge base, which differ in hierarchy. When there are no external data sources available to validate LLM-generated responses (or the responses are already based on external data as in RAG), model uncertainty estimation techniques from machine learning may be applied to detect hallucinations.

According to Luo et al., the previous methods fall into knowledge- and retrieval-based approaches, which ground LLM responses in factual data using external knowledge sources, such as path grounding. Luo et al. also mention training or reference guiding for language models, involving strategies like employing control codes or contrastive learning to guide the generation process to differentiate between correct and hallucinated content. Another category is evaluation and mitigation focused on specific hallucination types, such as employing methods to evaluate quantity entity in summarization and methods to detect and mitigate self-contradictory statements.

Nvidia Guardrails, launched in 2023, can be configured to hard-code certain responses via script instead of leaving them to the LLM. Furthermore, numerous tools like SelfCheckGPT, the Trustworthy Language Model, and Aimon have emerged to aid in the detection of hallucination in offline experimentation and real-time production scenarios.

Evaluating multiple possible replies before answering a query by assigning confidence scores to each could mitigate the problem. However, this approach would multiply computational costs. Active learning would further increase these costs. In high-stakes domains such as chip design, supply chain logistics, and medical diagnostics, the added costs are operationally necessary and therefore economically viable. In chatbots, however, customers tend to prefer rapid, overconfident answers over cautious, uncertainty-aware ones.

See also

AI alignment

AI effect

AI safety

AI slop

Artifact

Artificial stupidity

Memetic algorithm

Turing test

Uncanny valley

References

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car
MuZero
Action selection AutoGPT
AutoGPT
Robot control
Alan Turing
Warren Sturgis McCulloch
Walter Pitts
John von Neumann
Claude Shannon
Shun'ichi Amari
Kunihiko Fukushima
Takeo Kanade
Marvin Minsky
John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky

Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani
Noam Shazeer
Aidan Gomez
John Schulman
Mustafa Suleyman
Jan Leike
Daniel Kokotajlo
François Chollet
Neural Turing machine
Differentiable neural computer
Transformer Vision transformer (ViT)
Vision transformer (ViT)
Recurrent neural network (RNN)
Long short-term memory (LSTM)
Gated recurrent unit (GRU)
Echo state network
Multilayer perceptron (MLP)
Convolutional neural network (CNN)
Residual neural network (RNN)
Highway network
Mamba
Autoencoder
Variational autoencoder (VAE)
Generative adversarial network (GAN)
Graph neural network (GNN)
Category