Title: Generative pre-trained transformer

URL: https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

PageID: 72970020

Categories: Category:2018 in artificial intelligence, Category:Artificial neural networks, Category:Generative artificial intelligence, Category:Generative pre-trained transformers, Category:Large language models, Category:OpenAI

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

A generative pre-trained transformer (GPT) is a type of large language model (LLM) [1] [2] [3] that is widely used in generative AI chatbots . [4] [5] GPTs are based on a deep learning architecture called the transformer . They are pre-trained on large datasets of unlabeled content, and able to generate novel content. [2] [3]

OpenAI was the first to apply generative pre-training (GP) to the transformer architecture, introducing the GPT-1 model in 2018. [6] The company has since released many bigger GPT models. The popular chatbot ChatGPT , released in late 2022 (using GPT-3.5), was followed by many competitor chatbots using their own "GPT" models to generate text, such as Gemini , DeepSeek or Claude . [7]

GPTs are primarily used to generate text, but can be trained to generate other kinds of data. For example, GPT-4o can process and generate text, images and audio. [8] To improve performance on complex tasks, some GPTs, such as OpenAI o3 , spend more time analyzing the problem before generating an output, and are called reasoning models . In 2025, GPT-5 was released with a router that automatically selects whether to use a faster model or slower reasoning model based on task.

Background

According to The Economist , improved algorithms, more powerful computers, and an increase in the amount of digitized material fueled a revolution in machine learning during the 2010s. New techniques in the years before the AI boom resulted in "rapid improvements in tasks", including manipulating language. [9] Modern software models are trained to learn by using millions of examples in artificial neural networks that are inspired by biological neural structures. [9]

Separately, the concept of generative pre-training (GP) was a long-established technique in machine learning. GP is a form of self-supervised learning wherein a model is first trained on a large, unlabeled dataset (the "pre-training" step) to learn to generate data points. This pre-trained model is then adapted to a specific task using a labeled dataset (the " fine-tuning " step). [10]

The transformer architecture for deep learning is the core technology of a GPT. Developed by researchers at Google , it was introduced in the paper " Attention Is All You Need ", which was published on June 12, 2017. The transformer architecture solved many of the performance issues that were associated with older recurrent neural network (RNN) designs for natural language processing (NLP). The architecture's use of an attention mechanism allows models to process entire sequences of text at once, enabling the training of much larger and more sophisticated models. Since 2017, numerous transformer-based NLP systems have been available that are capable of processing, mining, organizing, connecting, contrasting, and summarizing texts as well as correctly answering questions from textual input. [11] [12]

History

On June 11, 2018, OpenAI researchers and engineers published a paper called "Improving Language Understanding by Generative Pre-Training", which introduced GPT-1 , the first GPT model. [13] It was designed as a transformer-based large language model that used generative pre-training (GP) on BookCorpus , a diverse text corpus , followed by discriminative fine-tuning to focus on specific language tasks. [14] This semi-supervised approach was seen as a breakthrough. Previously, the best-performing neural models in natural language processing (NLP) had commonly employed supervised learning from large amounts of manually labeled data – training a large language model with this approach would have been prohibitively expensive and time-consuming. [13]

On February 14, 2019, OpenAI introduced GPT-2 , a larger model that could generate coherent text. Created as a direct scale-up of its predecessor, it had both its parameter count and dataset size increased by a factor of 10. GPT-2 has 1.5 billion parameters and was trained on WebText, a 40-gigabyte dataset of 8 million web pages . [15] [16] [17] Citing risks of malicious use, OpenAI opted for a "staged release", initially publishing smaller versions of the model before releasing the full 1.5-billion-parameter model in November. [18]

On February 10, 2020, Microsoft introduced its Turing Natural Language Generation, which it claimed was the "largest language model ever published at 17 billion parameters." The model outperformed all previous language models at a variety of tasks, including summarizing texts and answering questions . [19]

On May 28, 2020, OpenAI introduced GPT-3 , a model with 175 billion parameters that was trained on a larger dataset compared to GPT-2. It marked a significant advancement in few-shot and zero-shot learning abilities. With few examples, it could perform various tasks that it was not explicitly trained for. [20] [21]

Following the release of GPT-3, OpenAI started using reinforcement learning from human feedback (RLHF) to align models' behavior more closely with human preferences. This led to the development of InstructGPT , a fine-tuned version of GPT-3. OpenAI further refined InstructGPT to

create ChatGPT , the flagship chatbot product of OpenAI that was launched on November 30, 2022. [22] ChatGPT was initially based on GPT-3.5 , but it was later transitioned to the GPT-4 model, which was released on March 14, 2023. [23] [24] GPT-4 was also integrated into parts of several applications, including Microsoft Copilot , GitHub Copilot , Snapchat , Khan Academy , and Duolingo . [25]

The immense popularity of ChatGPT spurred widespread development of competing GPT-based systems from other organizations. EleutherAI released a series of open-weight models , including GPT-J in 2021. Other major technology companies later developed their own GPT models, such as Google 's PaLM and Gemini as well as Meta AI 's Llama . [26]

Many subsequent GPT models have been trained to be multimodal (able to process or to generate multiple types of data). For example, GPT-4o can both process and generate text, images, and audio. [27] Additionally, GPT models like o3 and DeepSeek R1 have been trained with reinforcement learning to generate multi-step chain-of-thought reasoning before producing a final answer, which helps to solve complex problems in domains such as mathematics. [28]

On August 7, 2025, OpenAI released GPT-5 , which includes a router that automatically selects whether to use a faster model or slower reasoning model based on task. [29] [30]

Foundation models

A foundation model is an AI model trained on broad data at scale such that it can be adapted to a wide range of downstream tasks. [31] [32]

Thus far, the most notable GPT foundation models have been from OpenAI 's GPT-n series. The most recent from that is GPT-5 . [33]

Other such models include Google 's PaLM , a broad foundation model that has been compared to GPT-3 and has been made available to developers via an API , [34] [35] and Together's GPT-JT, which has been reported as the closest-performing open-source alternative to GPT-3 (and is derived from earlier open-source GPTs). [36] Meta AI (formerly Facebook) also has a generative transformer-based foundational large language model, known as LLaMA . [37]

Foundational GPTs can also employ modalities other than text, for input and/or output. GPT-4 is a multi-modal LLM that is capable of processing text and image input (though its output is limited to text). [38] Regarding multimodal output , some generative transformer-based models are used for text-to-image technologies such as diffusion [39] and parallel decoding. [40] Such kinds of models can serve as visual foundation models (VFMs) for developing downstream systems that can work with images. [41]

Task-specific models

A foundational GPT model can be further adapted to produce more targeted systems directed to specific tasks and/or subject-matter domains. Methods for such adaptation can include additional fine-tuning (beyond that done for the foundation model) as well as certain forms of prompt engineering . [44]

An important example of this is fine-tuning models to follow instructions , which is of course a fairly broad task but more targeted than a foundation model. In January 2022, OpenAI introduced "InstructGPT" – a series of models which were fine-tuned to follow instructions using a combination of supervised training and reinforcement learning from human feedback (RLHF) on base GPT-3 language models. [45] [46] Advantages this had over the bare foundational models included higher accuracy, less negative/toxic sentiment, and generally better alignment with user needs. Hence, OpenAI began using this as the basis for its API service offerings. [47] Other instruction-tuned models have been released by others, including a fully open version. [48] [49]

Another (related) kind of task-specific models are chatbots , which engage in human-like conversation. In November 2022, OpenAI launched ChatGPT – an online chat interface powered by an instruction-tuned language model trained in a similar fashion to InstructGPT. [50] They trained this model using RLHF, with human AI trainers providing conversations in which they played both the user and the AI, and mixed this new dialogue dataset with the InstructGPT dataset for a conversational format suitable for a chatbot. Other major chatbots currently include Microsoft 's

Bing Chat , which uses OpenAI's GPT-4 (as part of a broader close collaboration between OpenAI and Microsoft), [51] and Google 's competing chatbot Gemini (initially based on their LaMDA family of conversation-trained language models, with plans to switch to PaLM). [52]

Yet another kind of task that a GPT can be used for is the meta -task of generating its own instructions, like developing a series of prompts for 'itself' to be able to effectuate a more general goal given by a human user. [53] This is known as an AI agent , and more specifically a recursive one because it uses results from its previous self-instructions to help it form its subsequent prompts; the first major example of this was Auto-GPT (which uses OpenAI's GPT models), and others have since been developed as well. [54]

Domain-specificity

GPT systems can be directed toward particular fields or domains. Some reported examples of such models and apps are as follows:

EinsteinGPT – for sales and marketing domains, to aid with customer relationship management (uses GPT-3.5) [55] [56]

BloombergGPT – for the financial domain, to aid with financial news and information (uses "freely available" AI methods, combined with their proprietary data) [57]

Khanmigo – described as a GPT version for tutoring, in the education domain, it aids students using Khan Academy by guiding them through their studies without directly providing answers (powered by GPT-4) [58] [59]

SlackGPT – for the Slack instant-messaging service, to aid with navigating and summarizing discussions on it (uses OpenAI 's API) [60]

BioGPT – for the biomedical domain, to aid with biomedical literature text generation and mining (uses GPT-2) [61]

Sometimes domain-specificity is accomplished via software plug-ins or add-ons . For example, several different companies have developed particular plugins that interact directly with OpenAI's ChatGPT interface, [62] [63] and Google Workspace has available add-ons such as "GPT for Sheets and Docs" – which is reported to aid use of spreadsheet functionality in Google Sheets . [64] [65]

Brand issues

OpenAI , which created the first generative pre-trained transformer (GPT) in 2018, asserted in 2023 that "GPT" should be regarded as a brand of OpenAI. [66] In April 2023, OpenAI revised the brand guidelines in its terms of service to indicate that other businesses using its API to run their AI services would no longer be able to include "GPT" in such names or branding. [67] In May 2023, OpenAI engaged a brand management service to notify its API customers of this policy, although these notifications stopped short of making overt legal claims (such as allegations of trademark infringement or demands to cease and desist). [66] As of November 2023, OpenAI still prohibits its API licensees from naming their own products with "GPT", [68] but it has begun enabling its ChatGPT Plus subscribers to make "custom versions of ChatGPT" called GPTs on the OpenAI site. [69] OpenAI's terms of service says that its subscribers may use "GPT" in the names of these, although it's "discouraged". [68]

Relatedly, OpenAI has applied to the United States Patent and Trademark Office (USPTO) to seek domestic trademark registration for the term "GPT" in the field of AI. [66] OpenAI sought to expedite handling of its application, but the USPTO declined that request in April 2023. [70] In May 2023, the USPTO responded to the application with a determination that "GPT" was both descriptive and generic. [71] As of November 2023, OpenAI continues to pursue its argument through the available processes. Regardless, failure to obtain a registered U.S. trademark does not preclude some level of common-law trademark rights in the U.S. [72] and trademark rights in other countries. [73]

For any given type or scope of trademark protection in the U.S., OpenAI would need to establish that the term is actually " distinctive " to their specific offerings in addition to being a broader

technical term for the kind of technology. Some media reports suggested in 2023 that OpenAI may be able to obtain trademark registration based indirectly on the fame of its GPT-based chatbot product, ChatGPT , [70] [74] for which OpenAI has separately sought protection (and which it has sought to enforce more strongly). [75] Other reports have indicated that registration for the bare term "GPT" seems unlikely to be granted, [66] [76] as it is used frequently as a common term to refer simply to AI systems that involve generative pre-trained transformers. [3] [77] [78] [79] In any event, to whatever extent exclusive rights in the term may occur the U.S., others would need to avoid using it for similar products or services in ways likely to cause confusion. [76] [80] If such rights ever became broad enough to implicate other well-established uses in the field, the trademark doctrine of descriptive fair use could still continue non-brand-related usage. [81]

In the European Union , the European Union Intellectual Property Office registered "GPT" as a trade mark of OpenAI in spring 2023. However, since spring 2024 the registration is being challenged and is pending cancellation. [82]

In Switzerland , the Swiss Federal Institute of Intellectual Property registered "GPT" as a trade mark of OpenAI in spring 2023. [83] [84]

See also

Cyc

Vision transformer

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
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

v

t

e

AI-complete

Bag-of-words

n -gram Bigram Trigram

Bigram

Trigram

Computational linguistics

Natural language understanding

Stop words

Text processing

Argument mining

Collocation extraction

Concept mining

Coreference resolution

Deep linguistic processing

Distant reading

Information extraction

Named-entity recognition

Ontology learning

Parsing Semantic parsing Syntactic parsing

Semantic parsing

Syntactic parsing

Part-of-speech tagging

Semantic analysis

Semantic role labeling

Semantic decomposition

Semantic similarity

Sentiment analysis

Terminology extraction

Text mining
Textual entailment
Truecasing
Word-sense disambiguation
Word-sense induction
Compound-term processing
Lemmatisation
Lexical analysis
Text chunking
Stemming
Sentence segmentation
Word segmentation
Multi-document summarization
Sentence extraction
Text simplification
Computer-assisted
Example-based
Rule-based
Statistical
Transfer-based
Neural
BERT
Document-term matrix
Explicit semantic analysis
fastText
GloVe
Language model (large)
Latent semantic analysis
Seq2seq
Word embedding
Word2vec
Corpus linguistics
Lexical resource
Linguistic Linked Open Data
Machine-readable dictionary
Parallel text
PropBank
Semantic network

Simple Knowledge Organization System
Speech corpus
Text corpus
Thesaurus (information retrieval)
Treebank
Universal Dependencies
BabelNet
Bank of English
DBpedia
FrameNet
Google Ngram Viewer
UBY
WordNet
Wikidata
Speech recognition
Speech segmentation
Speech synthesis
Natural language generation
Optical character recognition
Document classification
Latent Dirichlet allocation
Pachinko allocation
Automated essay scoring
Concordancer
Grammar checker
Predictive text
Pronunciation assessment
Spell checker
Chatbot
Interactive fiction
Question answering
Virtual assistant
Voice user interface
Formal semantics
Hallucination
Natural Language Toolkit
spaCy
Computer programming

Technology

Data from Wikidata