Title: GPT-J

URL: https://en.wikipedia.org/wiki/GPT-J

PageID: 73162864

Categories: Category:Generative pre-trained transformers, Category:Large language models, Category:Open-source artificial intelligence

-----

GPT-J or GPT-J-6B is an open-source large language model (LLM) developed by EleutherAI in 2021. [1] As the name suggests, it is a generative pre-trained transformer model designed to produce human-like text that continues from a prompt. The optional "6B" in the name refers to the fact that it has 6 billion parameters. [2] The model is available on GitHub , but the web interface no longer communicates with the model. Development stopped in 2021. [3]

Architecture

GPT-J is a GPT-3 -like model with 6 billion parameters. [4] Like GPT-3, it is an autoregressive , decoder-only transformer model designed to solve natural language processing (NLP) tasks by predicting how a piece of text will continue. [1]

Its architecture differs from GPT-3 in three main ways. [1]

The attention and feedforward neural network were computed in parallel during training, allowing for greater efficiency.

The GPT-J model uses rotary position embeddings , which has been found to be a superior method of injecting positional information into transformers. [5] [6]

GPT-J uses dense attention instead of efficient sparse attention, as used in GPT-3.

Beyond that, the model has 28 transformer layers and 16 attention heads. Its vocabulary size is 50257 tokens , the same size as GPT-2 's. [2] It has a context window size of 2048 tokens. [7]

It was trained on the Pile dataset, [2] [4] using the Mesh Transformer JAX library in JAX to handle the parallelization scheme. [2] [8]

Performance

GPT-J was designed to generate English text from a prompt. It was not designed for translating or generating text in other languages or for performance without first fine-tuning the model for a specific task. [2] Nonetheless, GPT-J performs reasonably well even without fine-tuning, even in translation (at least from English to French). [9]

When neither is fine-tuned, GPT-J-6B performs almost as well as the 6.7 billion parameter GPT-3 (Curie) on a variety of tasks. [4] It even outperforms the 175 billion parameter GPT-3 (Davinci) on code generation tasks. [10] With fine-tuning, it outperforms an untuned GPT-3 (Davinci) on a number of tasks. [1]

Like all LLMs, it is not programmed to give factually accurate information, only to generate text based on probability. [2]

Applications

The untuned GPT-J is available on EleutherAI's website, [11] NVIDIA 's Triton Inference Server, [12] and NLP Cloud's website. [13] Cerebras [1] and Amazon Web Services [14] [15] offer services to fine-tune the GPT-J model for company-specific tasks. Graphcore offers both fine-tuning and hosting services for the untuned GPT-J, as well as offering to host the fine-tuned models after they are produced. [16] CoreWeave offers hosting services for both the untuned GPT-J and fine-tuned variants. [17] [18]

In March 2023, Databricks released Dolly, an Apache-licensed , instruction-following model created by fine-tuning GPT-J on the Stanford Alpaca dataset. [19] NovelAI 's Sigurd [20] and Genji-JP 6B [21] models are both fine-tuned versions of GPT-J. They also offer further fine-tuning services to produce and host custom models. [22]

EleutherAI has received praise from Cerebras, [1] GPT-3 Demo, [4] NLP Cloud, [13] and Databricks [19] for making the model open-source, and its open-source status is often cited as a major advantage when choosing which model to use. [10] [16] [23]

References

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