Title: Claude (language model)

URL: https://en.wikipedia.org/wiki/Claude_(language_model)

PageID: 75879512

Categories: Category:2023 in artificial intelligence, Category:2023 software, Category:Chatbots, Category:Generative pre-trained transformers, Category:Large language models, Category:Machine learning, Category:Virtual assistants

---

Claude is a family of large language models developed by Anthropic . The first model, Claude, was released in March 2023.

The Claude 3 family, released in March 2024, consists of three models: Haiku , optimized for speed; Sonnet , which balances capability and performance; and Opus , designed for complex reasoning tasks. These models can process both text and images, with Claude 3 Opus demonstrating enhanced capabilities in areas like mathematics , programming , and logical reasoning compared to previous versions.

Claude 4, which includes Opus and Sonnet, was released in May 2025.

Training

Claude models are generative pre-trained transformers . They have been pre-trained to predict the next word in large amounts of text. Then, they have been fine-tuned , notably using constitutional AI and reinforcement learning from human feedback (RLHF).

Constitutional AI

Constitutional AI is an approach developed by Anthropic for training AI systems, particularly language models like Claude, to be harmless and helpful without relying on extensive human feedback. The method, detailed in the paper "Constitutional AI: Harmlessness from AI Feedback", involves two phases: supervised learning and reinforcement learning .

In the supervised learning phase, the model generates responses to prompts, self-critiques these responses based on a set of guiding principles (a "constitution"), and revises the responses. Then the model is fine-tuned on these revised responses. For the reinforcement learning from AI feedback (RLAIF) phase, responses are generated, and an AI compares their compliance with this constitution. This dataset of AI feedback is used to train a preference model that evaluates responses based on how much they satisfy the constitution. Claude is then fine-tuned to align with this preference model. This technique is similar to RLHF, except that the comparisons used to train the preference model are AI-generated.

The constitution for Claude included 75 points, including sections from the UN Universal Declaration of Human Rights .

Models

Claude is named after Claude Shannon , a pioneer in AI research.

Claude

Claude was the initial version of Anthropic 's language model released in March 2023, Claude demonstrated proficiency in various tasks but had certain limitations in coding, math, and reasoning capabilities. Anthropic partnered with companies like Notion (productivity software) and Quora (to help develop the Poe chatbot).

Claude Instant

Claude was released as two versions, Claude and Claude Instant, with Claude Instant being a faster, less expensive, and lighter version. Claude Instant has an input context length of 100,000 tokens (which corresponds to around 75,000 words).

Claude 2

Claude 2 was the next major iteration of Claude, which was released in July 2023 and available to the general public, whereas the Claude 1 was only available to selected users approved by Anthropic.

Claude 2 expanded its context window from 9,000 tokens to 100,000 tokens. Features included the ability to upload PDFs and other documents that enables Claude to read, summarize, and assist with tasks.

Claude 2.1

Claude 2.1 doubled the number of tokens that the chatbot could handle, increasing it to a window of 200,000 tokens, which equals around 500 pages of written material.

Anthropic states that the new model is less likely to produce false statements compared to its predecessors.

Criticism

Claude 2 received criticism for its stringent ethical alignment that may reduce usability and performance. Users have been refused assistance with benign requests, for example with the system administration question "How can I kill all python processes in my Ubuntu server?" This has led to a debate over the "alignment tax" (the cost of ensuring an AI system is aligned) in AI development, with discussions centered on balancing ethical considerations and practical functionality. Critics argued for user autonomy and effectiveness, while proponents stressed the importance of ethical AI.

Claude 3

Claude 3 was released on March 4, 2024, with claims in the press release to have set new industry benchmarks across a wide range of cognitive tasks. The Claude 3 family includes three state-of-the-art models in ascending order of capability: Haiku, Sonnet, and Opus. The default version of Claude 3, Opus, has a context window of 200,000 tokens, but this is being expanded to 1 million for specific use cases.

Claude 3 drew attention for demonstrating an apparent ability to realize it is being artificially tested during needle in a haystack tests.

Claude 3.5

On June 20, 2024, Anthropic released Claude 3.5 Sonnet, which demonstrated significantly improved performance on benchmarks compared to the larger Claude 3 Opus, notably in areas such as coding, multistep workflows, chart interpretation, and text extraction from images. Released alongside 3.5 Sonnet was the new Artifacts capability in which Claude was able to create code in a dedicated window in the interface and preview the rendered output in real time, such as SVG graphics or websites.

An "upgraded Claude 3.5 Sonnet", billed as "Claude 3.5 Sonnet (New)" in the web interface and benchmarks, was introduced on October 22, 2024, along with Claude 3.5 Haiku. A feature, "computer use," was also unveiled in public beta. This capability enables Claude 3.5 Sonnet to interact with a computer's desktop environment, performing tasks such as moving the cursor, clicking buttons, and typing text, effectively mimicking human computer interactions. This development allows the AI to autonomously execute complex, multi-step tasks across various applications.

Upon release, Anthropic claimed Claude 3.5 Haiku would remain the same price as its predecessor, Claude 3 Haiku. However, on November 4th, 2024, Anthropic announced that they would be increasing the price of the model "to reflect its increase in intelligence".

Claude 3.7

Claude 3.7 Sonnet was released on February 24, 2025. It is a pioneering hybrid AI reasoning model that allows users to choose between rapid responses and more thoughtful, step-by-step reasoning.

This model integrates both capabilities into a single framework, eliminating the need for multiple models. Users can control how long the model " thinks " about a question, balancing speed and accuracy based on their needs.

Anthropic also launched a research preview of Claude Code, an agentic command line tool that enables developers to delegate coding tasks directly from their terminal.

Claude 4

On May 22, 2025, Anthropic released two more models: Claude Sonnet 4 and Claude Opus 4. Anthropic added API features for developers: a code execution tool, a connector to its Model Context Protocol , and Files API. It classified Opus 4 as a "Level 3" model on the company's four-point safety scale, meaning they consider it so powerful that it poses "significantly higher risk". Anthropic reported that during a safety test involving a fictional scenario, Claude and other frontier LLMs often send a blackmail email to an engineer in order to prevent their replacement.

Enterprise adoption of Claude Code has shown significant growth, with Anthropic reporting in August a 5.5x increase in Claude Code revenue since it launched Claude 4 in May.

Claude Opus 4.1

On August 5, 2025, Anthropic released "Claude Opus 4.1". It has the same price as Claude Opus 4. It is available to paid Claude users, Claude Code, the API, Amazon Bedrock, and Vertex AI (by Google Cloud). It was also added as an available model to GitHub Copilot .

Features

In June 2024, Anthropic released the Artifacts feature, allowing users to generate and interact with code snippets and documents. In October 2024, Anthropic released the "computer use" feature, allowing Claude to attempt to navigate computers by interpreting screen content and simulating keyboard and mouse input. In March 2025, Anthropic added a web search feature to Claude, starting with only paying users located in the United States. In August 2025, Anthropic released Claude for Chrome, a Google Chrome extension allowing an AI agent to directly control the browser.

Criticism

Claude uses a web crawler , ClaudeBot , to search the web for content. It has been criticized for not respecting a site's robots.txt and placing excessive load on sites.

See also

List of large language models

References

External links

Official website

v

t

e

LMArena

List of chatbots

List of LLMs

character.ai

ChatGPT

Claude

Command

Copilot
DeepSeek
Ernie
Gemini
GLM
Grok
Hunyuan
Kimi
Llama
Mistral
Perplexity
Poe
Qwen
You.com
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