Claude is a language model developed by Anthropic . [ 1 ] The development of Claude aims to combine language processing (using machine learning and large language models ) with ethical standards and security in artificial intelligence (AI). [ 2 ]

overview

The Claude model line, particularly the third generation, known as Claude 3, consists of three versions: Haiku, Sonnet, and Opus. Each of these models offers different levels of performance and is optimized for specific applications. [ 3 ]

Story

Anthropic was founded by former OpenAI researchers with the goal of launching Claude as a response to existing language models. The latest version, Claude 3, was launched in 2024 and offers improved capabilities compared to its predecessors as well as models from other leading technology companies. [ 3 ]

Technical specifications

Claude 3 is available in three main variants:

Haiku: The fastest and most cost-effective model, suitable for real-time interactions. [ 4 ]

Sonnet: Offers a balance between speed and intelligence, suitable for business applications. [ 4 ]

Opus: The most powerful model, designed for tasks requiring high accuracy and advanced language comprehension skills. [ 4 ]

These models can process context windows of up to 200,000 tokens , which is significantly more than comparable models from other vendors. [ 4 ]

Anthropic plans to continually update and improve the Claude model line. These developments aim to make the models both more powerful and efficient, while reducing costs for smaller models. [ 3 ]

training

Claude models are generative, pre-trained transformers . They were trained to predict the next word in large amounts of text. The Claude models were then fine-tuned using Constitutional AI to make them helpful, honest, and harmless. [ 5 ] [ 6 ]

Constitutional AI

Constitutional AI is an approach developed by Anthropic to train AI systems, particularly language models like Claude, to be both harmless and helpful without relying on extensive human feedback. The method, described in the article "Claude's Constitution" [ 6 ] , involves two phases: supervised learning and reinforcement learning .

In the supervised learning phase, the model generates responses to prompts, self-criticizes these responses based on a set of guiding principles (a "constitution"), and revises the responses. The model is then fine-tuned based on these revised responses.

In the reinforcement learning from AI feedback phase, responses are generated, and an AI compares their agreement with the constitution. This dataset of AI feedback is used to train a preference model that scores responses based on how closely they match the constitution. Claude is then fine-tuned to match this preference model. This technique is similar to reinforcement learning from human feedback (RLHF) , except that the comparisons used to create the preference model are generated by the AI ​​and are based on the constitution. [ 7 ] [ 6 ]

This approach enables the training of AI assistants that are both helpful and harmless and that can explain their objections to harmful requests, thereby increasing transparency and reducing reliance on human oversight. [ 8 ]

The "constitution" for Claude comprised 75 points, including sections from the United Nations Universal Declaration of Human Rights . [ 9 ] [ 10 ]

Performance comparisons

For example, in several benchmarks, Claude 3 Opus outperformed both Google's Gemini language model and OpenAI's GPT-4 , particularly in areas such as general knowledge and mathematics. The tests show that Claude 3 is less biased and more precise in its answers compared to previous models and competing products. [ 4 ]

Despite its advanced capabilities and improved security features, Claude lags behind some of its competitors in widespread adoption. This is partly due to its increased focus on business customers and the high cost of the most powerful models. [ 3 ]

The "Claude 3.5 Sonnet" model is currently (October 2024) considered the smartest AI model on the market. It can tell jokes and imitates the style of human speech with deceptive accuracy. [ 11 ] According to the provider, Claude is capable of summarizing long texts and answering questions about them precisely. "This makes the software particularly useful for professionals in fields such as law, medicine, and finance." [ 12 ]

Applications

Claude is primarily used by business users who need to perform complex data analysis and other demanding tasks. With its ability to process large amounts of text, Claude is also suitable for analyzing and summarizing large documents. [ 3 ] With the launch of apps for iOS in May 2024 [ 13 ] and Android in July 2024 [ 14 ], Claude also became more accessible to private users.

Security and ethical standards

Anthropic places great emphasis on the safety and ethical aspects of AI development. Claude is equipped with various mechanisms designed to ensure that the model operates within ethical boundaries and does not produce harmful or biased outputs. [ 2 ]