Wikipedia
The Free Encyclopedia

Claude (language model)

Claude is a family of large language models developed by Anthropic.[1] The first model was released in March 2023. Claude 3, released in March 2024, can also analyze images.[2]

Claude

Developer(s)	Anthropic
Initial release	March 2023
Type	Large language model GPT Foundation model
License	Proprietary
Website	claude.ai (https://claude.ai/)

Training

Claude models are generative pre-trained transformers. They have been pre-trained to predict the next word in large amounts of text. Claude models have then been fine-tuned with Constitutional AI with the aim of making them helpful, honest, and harmless.[3][4]

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.[4]

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.[4]

For the reinforcement learning from AI feedback (RLAIF) phase, responses are generated, and an AI compares their compliance with the 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 reinforcement learning from human feedback (RLHF), except that the comparisons used to train the preference model are AI-generated, and that they are based on the constitution.[5][4]

This approach enables the training of AI assistants that are both helpful and harmless, and that can explain their objections to harmful requests, enhancing transparency and reducing reliance on human supervision.[6][7]

The "constitution" for Claude included 75 points, including sections from the UN Universal Declaration of Human Rights.[6][3]

Models

Claude

Claude was the initial version of Anthropic's language model released in March 2023,[8] Claude demonstrated proficiency in various tasks but had certain limitations in coding, math, and reasoning capabilities.[9] Anthropic partnered with companies like Notion (productivity software) and Quora (to help develop the Poe chatbot).[9]

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).[10]

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.[11]

Claude 2 expanded its context window from 9,000 tokens to 100,000 tokens.[8] 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.[1]

Anthropic states that the new model is less likely to produce false statements compared to its predecessors.[12]

Claude 3

Claude 3 was released on March 14, 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.[13][2]

Claude 3 has seemed to perform meta-cognitive reasoning, including the ability to realize it is being artificially tested during needle in a haystack tests.[14]

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 select code in real time such as websites or SVGs.[15]

Access

Limited-use access using Claude 3.5 Sonnet is free of charge, but requires both an e-mail address and a cellphone number. A paid plan is also offered for more usage and access to all Claude 3 models.[16]

On May 1, 2024, Anthropic announced the Claude Team plan, its first enterprise offering for Claude, and a Claude iOS app.[17]

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 programming 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.[18][12]

References

1. Davis, Wes (2023-11-21). "OpenAI rival Anthropic makes its Claude chatbot even more useful" (https://www.theverge.com/2023/11/21/23971070/anthropic-claude-2-1-openai-ai-chatbot-update-beta-tools). *The Verge*. Retrieved 2024-01-23.
2. Whitney, Lance (March 4, 2024). "Anthropic's Claude 3 chatbot claims to outperform ChatGPT, Gemini" (https://www.zdnet.com/article/anthropics-claude-3-chatbot-claims-to-outperform-chatgpt-gemini/). *ZDNET*. Retrieved 2024-03-05.
3. "What to Know About Claude 2, Anthropic's Rival to ChatGPT" (https://time.com/6295523/claude-2-anthropic-chatgpt/). *TIME*. 2023-07-18. Retrieved 2024-01-23.
4. "Claude's Constitution" (https://www.anthropic.com/news/claudes-constitution). *Anthropic*. May 9, 2023. Retrieved 2024-03-26.
5. Eliot, Lance (May 25, 2023). "Latest Generative AI Boldly Labeled As Constitutional AI Such As Claude By Anthropic Has Heart In The Right Place, Says AI Ethics And AI Law" (https://www.forbes.com/sites/lanceeliot/2023/05/25/latest-generative-ai-boldly-labeled-as-constitutional-ai-such-as-claude-by-anthropic-has-heart-in-the-right-place-says-ai-ethics-and-ai-law/). *Forbes*. Retrieved 2024-03-27.
6. Bai, Yuntao; Kadavath, Saurav; Kundu, Sandipan; Askell, Amanda; Kernion, Jackson; Jones, Andy; Chen, Anna; Goldie, Anna; Mirhoseini, Azalia (2022-12-15), *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073 (https://arxiv.org/abs/2212.08073)

7. Mok, Aaron. "A ChatGPT rival just published a new constitution to level up its AI guardrails, and prevent toxic and racist responses" (https://www.businessinsider.com/anthropic-new-crowd-sourced-ai-constitution-accuracy-safety-toxic-racist-2023-10). *Business Insider*. Retrieved 2024-01-23.
8. Drapkin, Aaron (2023-10-27). "What Is Claude AI and Anthropic? ChatGPT's Rival Explained" (https://tech.co/news/what-is-claude-ai-anthropic). *Tech.co*. Retrieved 2024-01-23.
9. "Introducing Claude" (https://www.anthropic.com/news/introducing-claude). *Anthropic*. March 14, 2023.
10. Yao, Deborah (August 11, 2023). "Anthropic's Claude Instant: A Smaller, Faster and Cheaper Language Model" (https://aibusiness.com/nlp/anthropic-s-claude-instant-a-smaller-faster-and-cheaper-language-model). *AI Business*.
11. Matthews, Dylan (2023-07-17). "The $1 billion gamble to ensure AI doesn't destroy humanity" (https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2). *Vox*. Retrieved 2024-01-23.
12. "Anthropic Announces Claude 2.1 LLM with Wider Context Window and Support for AI Tools" (https://www.infoq.com/news/2023/11/anthropic-announces-claude-2-1/). *InfoQ*. Retrieved 2024-01-23.
13. "Introducing the next generation of Claude" (https://www.anthropic.com/news/claude-3-family). *Anthropic*. Retrieved 2024-03-04.
14. Edwards, Benj (2024-03-05). "Anthropic's Claude 3 causes stir by seeming to realize when it was being tested" (https://arstechnica.com/information-technology/2024/03/claude-3-seems-to-detect-when-it-is-being-tested-sparking-ai-buzz-online/). *Ars Technica*. Retrieved 2024-03-09.
15. Pierce, David (2024-06-20). "Anthropic has a fast new AI model — and a clever new way to interact with chatbots" (https://www.theverge.com/2024/6/20/24181961/anthropic-claude-35-sonnet-model-ai-launch). *The Verge*. Retrieved 2024-06-20.
16. "Introducing the Claude Team plan and iOS app" (https://www.anthropic.com/news/team-plan-and-ios). *Anthropic*. May 1, 2024. Retrieved 2024-06-22.
17. Field, Hayden (May 1, 2024). "Amazon-backed Anthropic launches iPhone app and business tier to compete with OpenAI's ChatGPT" (https://www.cnbc.com/2024/05/01/anthropic-iphone-ai-app-business-plan-to-compete-with-openai-announced.html). *CNBC*. Retrieved May 3, 2024.
18. Glifton, Gerald (January 3, 2024). "Criticisms Arise Over Claude AI's Strict Ethical Protocols Limiting User Assistance" (https://lightsquare.org/news/criticisms-arise-over-claude-ais-strict-ethical-protocols-limiting-user-assistance). *Light Square*. Retrieved 2024-01-23.

External links

- Official website (https://claude.ai/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Claude_(language_model)&oldid=1238227557"

-