ANTHROP\C

Technology Tutorial

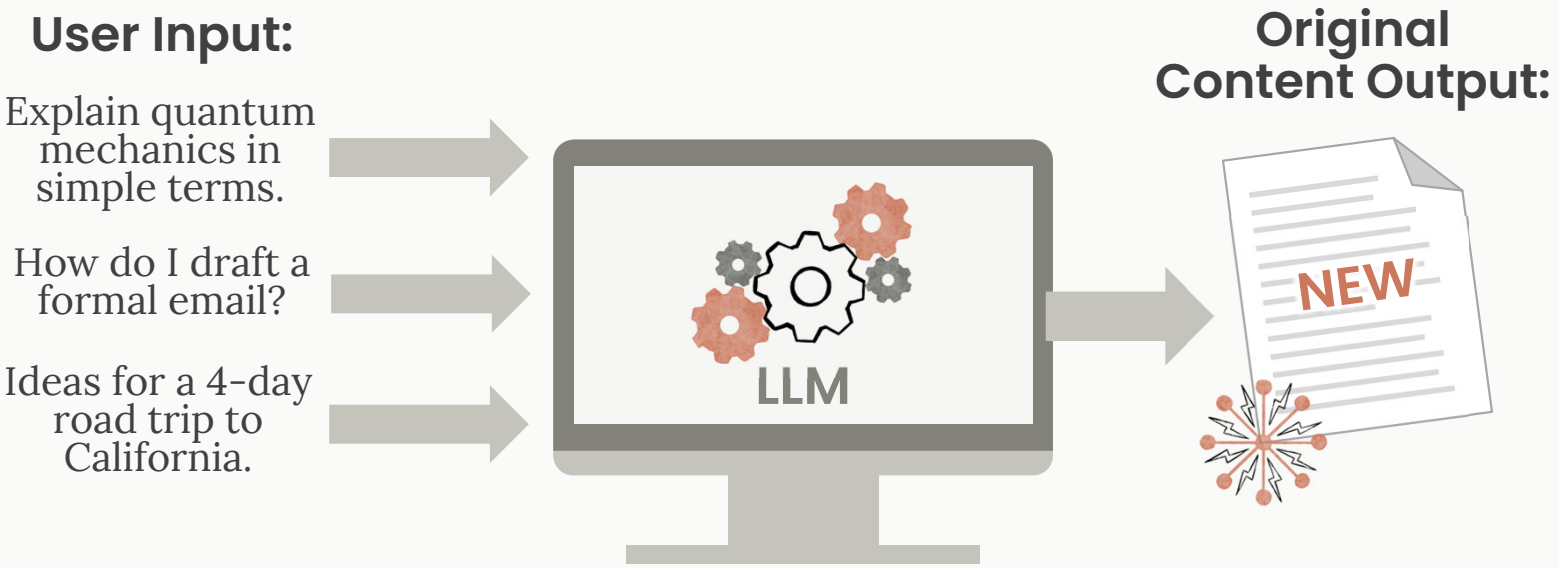
Concord Music Group, Inc., et al., v. Anthropic PBC
Case No.: 5:24-cv-03811-EKL-SVK

November 2024

What Is Claude?

What Is Claude?

Claude is a Large Language Model—a **general purpose**, **generative AI model** designed to respond like an intelligent human to user prompts.



How Do Large Language Models Work?



Good evening, Sam

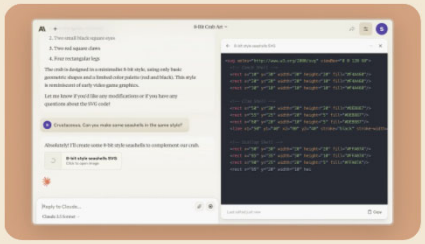
S what should I name my tabby cat?

Tiger, Stripes, Scout, or Oliver would suit a tabby nicely.

Use Cases for Claude

Software Developers

Draft and iterate code for straightforward tasks



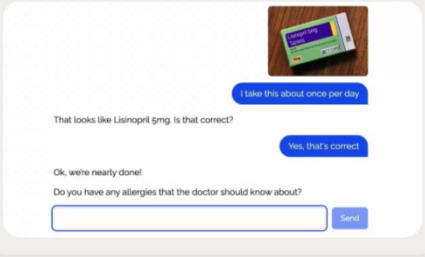
Writers

Ideate potential stories or situations



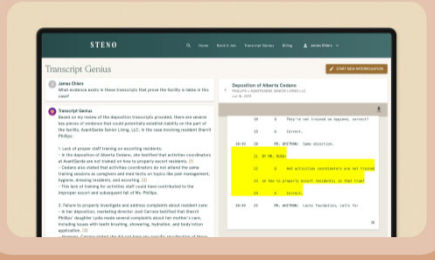
Healthcare Professionals

Analyze medical records and interpret complex clinical data



Lawyers

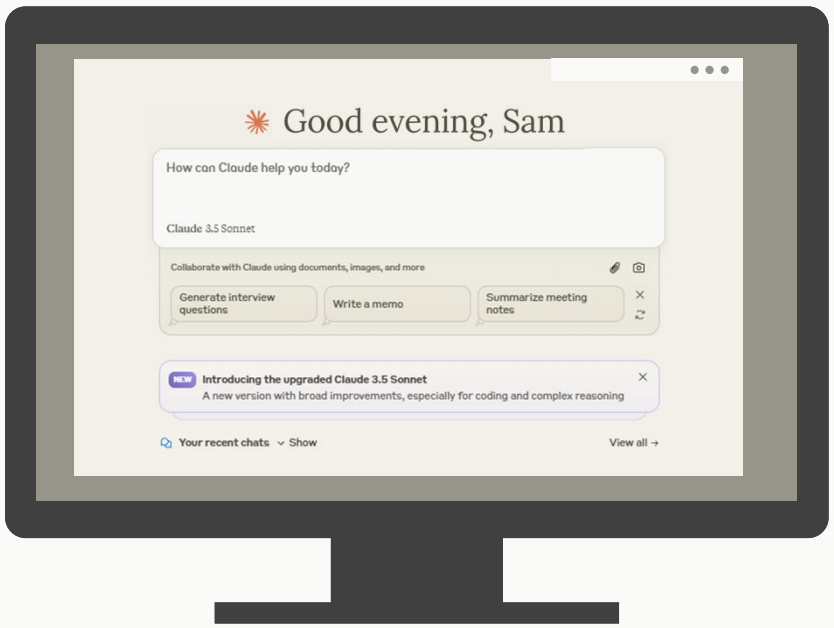
Analyze long legal transcripts and create summaries

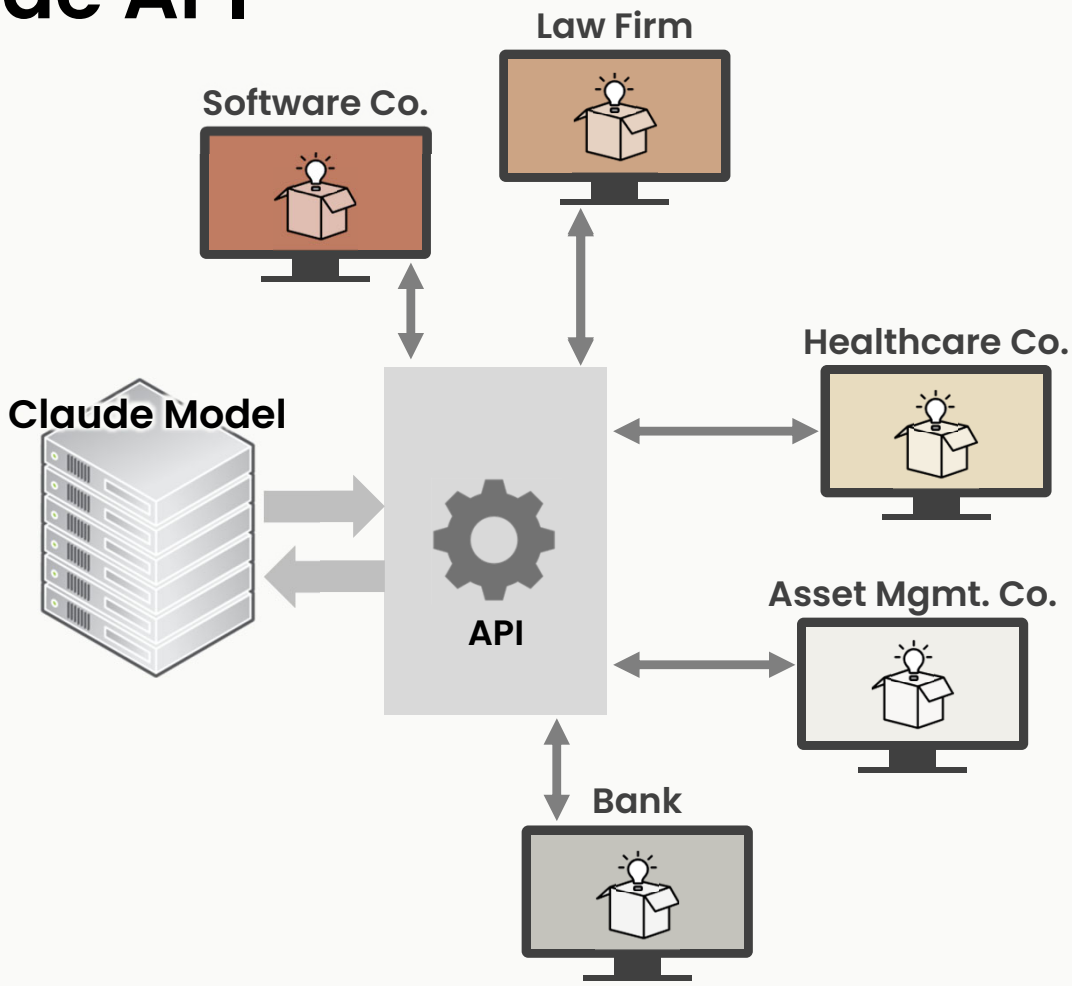


Claude refers to Anthropic's suite of commercially available models

- **Claude 1.0** was released on March 14, 2023
- **Claude 2.0** was released on July 11, 2023
- Plaintiffs' claims are directed at **Claude 1.0** and **Claude 2.0**
- **Claude 1.0** has since been deprecated and is no longer available to the public
- Anthropic has since continued to release more commercial models, totaling about 10 versions of Claude

Claude.ai and the Claude API





Pretraining

Pretraining an LLM: Data Are First Broken Into Tokens

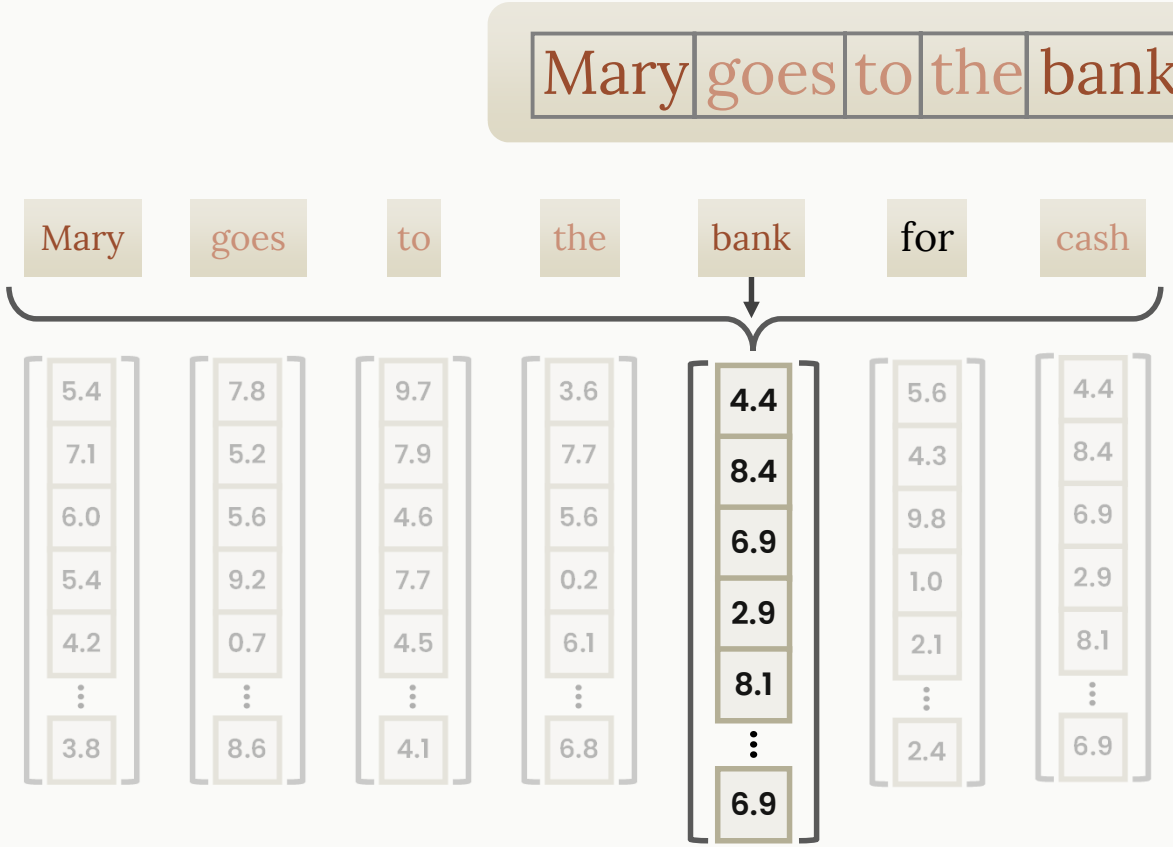
- Tokens are the smallest unit a language model can understand
- LLMs cannot understand language without tokens

Tokens 50 **Characters** 204

Specifically, tokens are the segments of text that are fed into and generated by the machine learning model. These can be individual characters, whole words, parts of words, or even larger chunks of text.

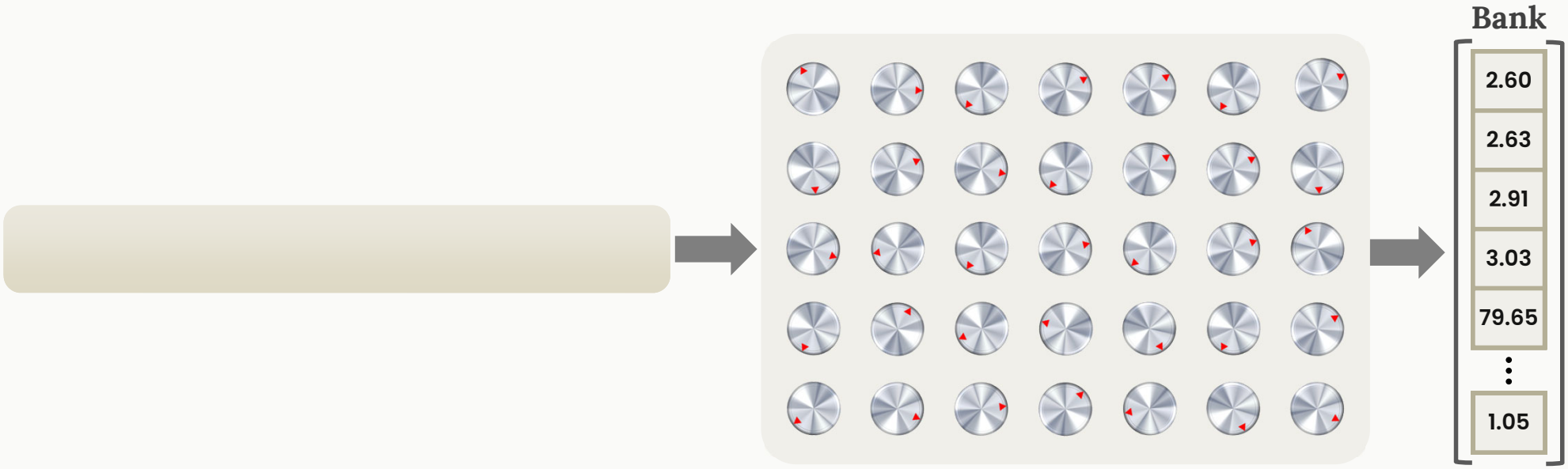
TEXT TOKEN IDS

Pretraining an LLM: Vectorization



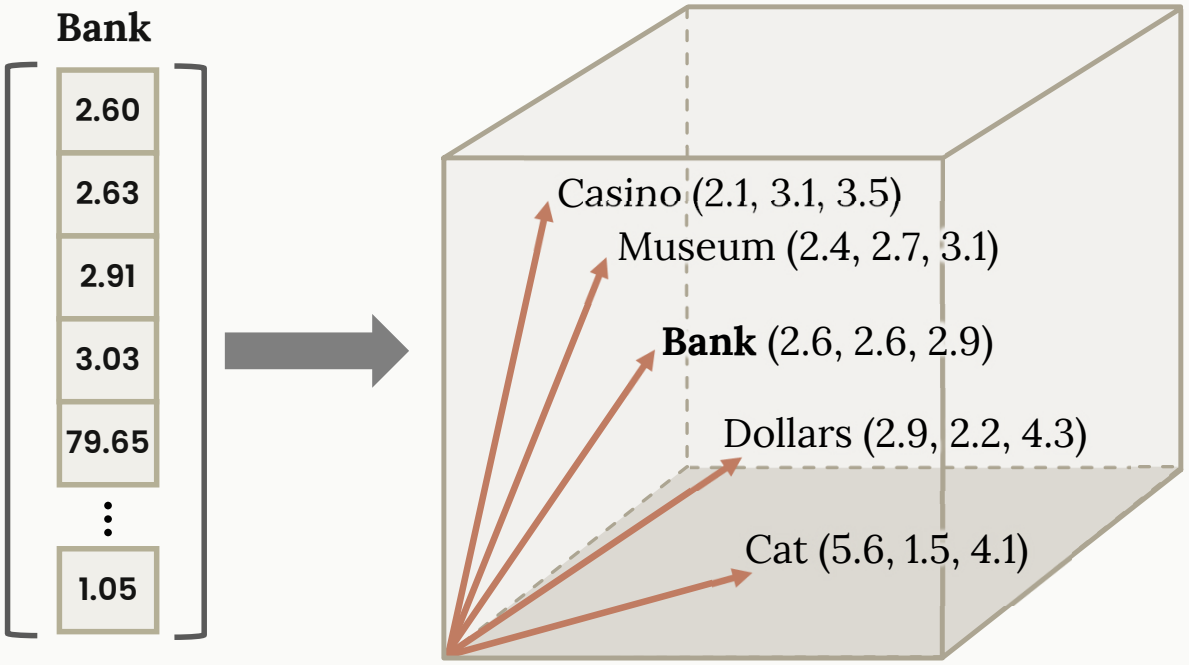
The model analyzes token sequences to **learn statistical relationships** between tokens and stores those relationships as "**weights.**"

Pretraining an LLM: Adjusting the Weights



Pretraining an LLM: Mapping the Results

I	saw	a	movie	where	the	characters	planned	a	heist	at	a	high	-	security	_____

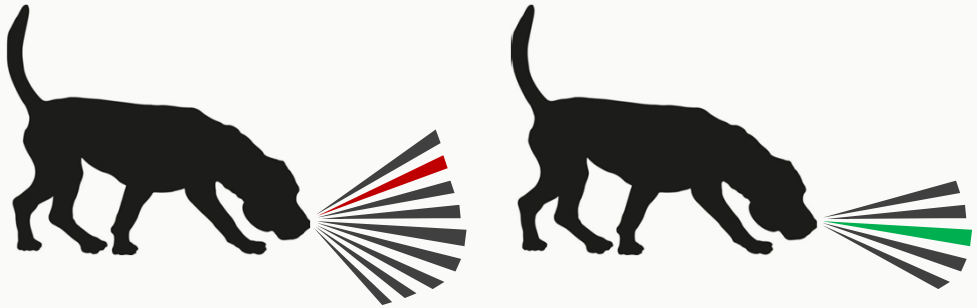


Finetuning

After Pretraining, LLMs Are Finetuned

LLMs are next trained on smaller, specialized datasets to achieve **specific objectives** rather than **general ones**.

- **Supervised learning (SL)**: learning from examples of prior examples of good and bad output/response pairs.
- **Reinforcement learning (RL)**: learning through the use of rewards and penalties for generating good and bad responses



Finetuning allows LLMs like Claude to:

- Express a particular personality
- Perform specialized tasks
- Be clear, direct, and concise
- Respond to their model names
- Be useful and helpful
- Prioritize safety

Constitutional AI

- An approach to training AI systems where **ethical principles and behavioral guidelines are built in from the start**, rather than added later
  - Similar to how a constitution guides a government, the AI system has core principles, including **a principle against copyright infringement**, it's trained to follow to be truthful, helpful, and avoid harm
- The goal is to create **AI systems that reliably follow intended behaviors and values**, even in new situations they haven't encountered before
- It differs from traditional approaches as it **makes ethical behavior fundamental to how the AI thinks and operates**, rather than relying on external restrictions

Guardrails

Claude Has Guardrails to Prevent Misuse

