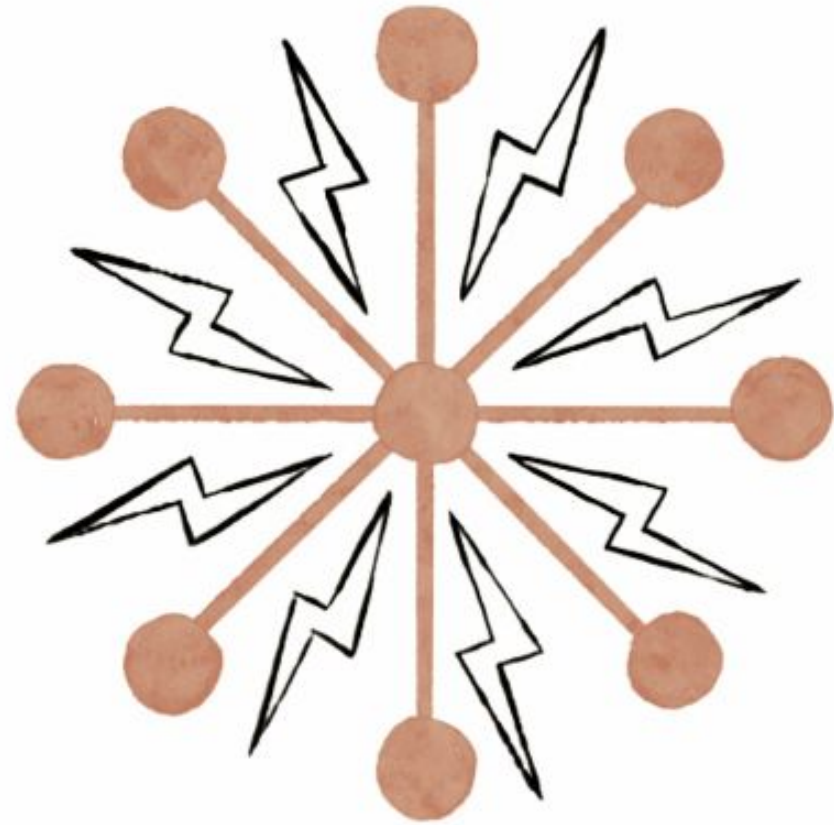


The background features two large, abstract, light beige shapes. One is a large circle in the top-left corner, and the other is a more complex, organic shape in the bottom-right corner.

ANTHROPIC

AI products and research that put safety at the frontier



- **Anthropic overview**
- Constitutional AI methodology
- Functionality deep dives
- Deployment options
- Demo



Context

Advances in AI will disrupt the labor economy, the macroeconomy, and power structures both within and between nations

Ensuring these models are developed and deployed responsibly is critical

Anthropic is focused on fueling the world's leading enterprises and governments – which we believe are key to this vision

Anthropic Track Record

Our technical leadership and co-founders are key authors in the groundbreaking GPT-3 paper



DARIO AMODEI
CEO



DANIELA AMODEI
PRESIDENT



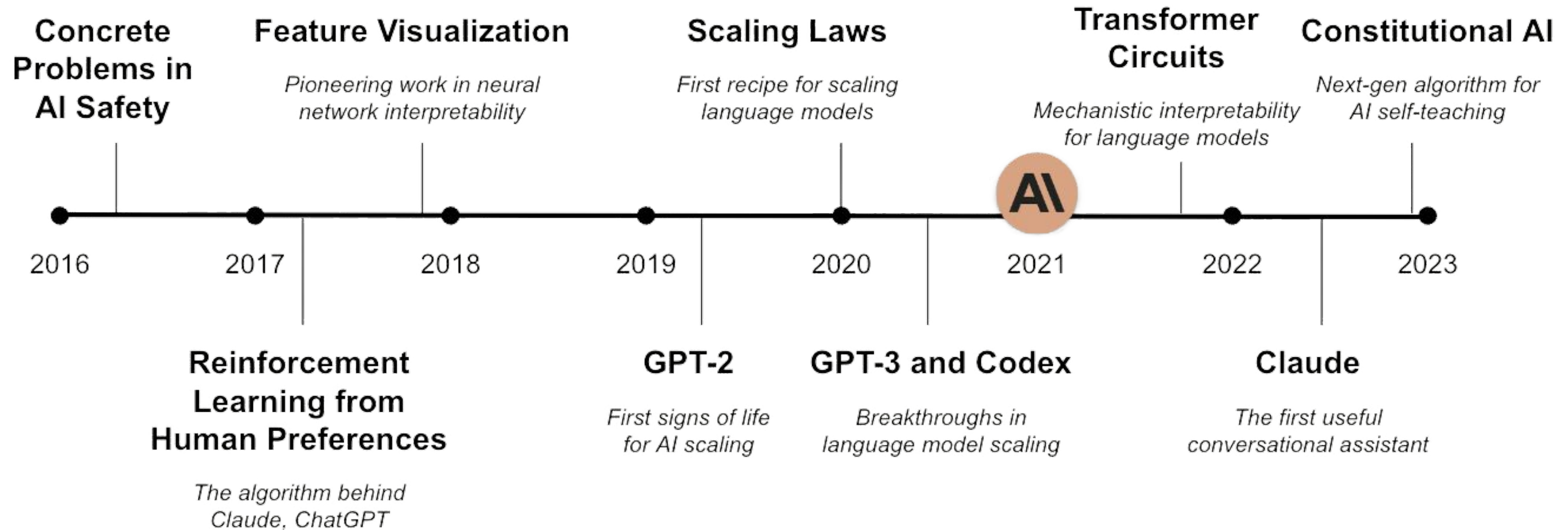
JARED KAPLAN
CHIEF SCIENTIST

Language Models are Few-Shot Learners

First authors, core technical leads				
Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess	Jack Clark	Christopher Berner		
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	
Last author, project director				

Authors in purple work at Anthropic. Core technical leads (Tom Brown and Ben Mann) are Anthropic founders. Overall project was directed by Dario Amodei, Anthropic CEO

Pioneering breakthroughs delivered by Anthropic staff



Make safe AI systems
Deploy them reliably

We develop large-scale AI systems so that we can study their safety properties at the technological frontier, where new problems are most likely to arise. We use these insights to create safer, steerable, and more reliable models, and to generate systems that we deploy externally, like Claude.

ResearchAlignment

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Apr 12, 2022

ResearchSocietal Impact

Towards Measuring the Representation of Subjective Global Opinions in Language Models

ResearchInterpretability

Privileged Bases in the Transformer Residual Stream

Mar 16, 2023

Read Paper

User Preference:

Claude 1 and GPT-4 lead the rest

Large Language Model Leaderboard

Win rates vs. other models

GPT-4 (85%)

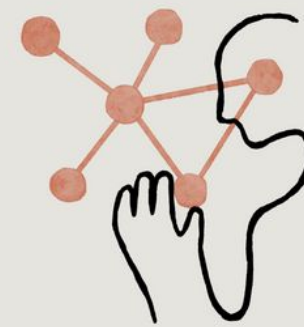
Claude (80%)

Claude Instant (77%)

GPT Turbo (72%)

3 open source models (64-60%)

Palm (57%)



Safer

4x as fast

+50% cheaper

3-25x context window

Much greater ability to
customize

Large Model Systems Organization is a research organization at UC Berkeley

Data from 6/25/2023


Introducing Claude 2

- Significant coding improvements (71.2% on Codex P@1)
- 2x more harmless
- Knowledge cutoff in 2023
- 10% training data non-english
- 100k context


...all at the **same price as Claude 1**
(4-5x cheaper than GPT-4 32k)

 **Eric Schmidt** ✓
@ericschmidt

This Claude 2 is remarkable.. please try it and find out why its so good ! Eric

 **Michael Trazzi (in SF)** ✓
@MichaelTrazzi

has anyone managed to jailbreak claude? I am already feeling immoral after trying four times. do language models have boundaries?


 **steve@mora.co** ✓
@SteveMoraco

Just ran DATA with @AnthropicAI for the first time. (thanks for the team invite @Michaelgr1011)

Astounded that it knocked it out of the park. More coherent and detailed than GPT-4 and no system prompt required.

 **David** ✓
@dzhng

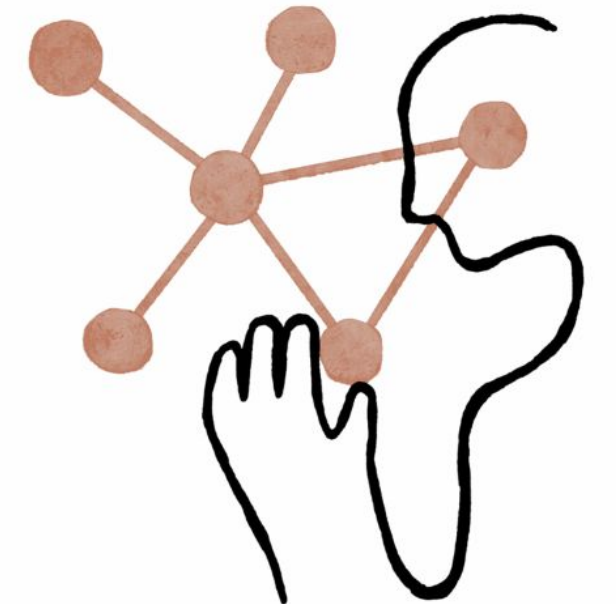
I've shifted 80% of my LLM spend to @AnthropicAI Claude 2 at this point - it strikes the perfect balance between performance / cost / throughput.

 **LSantos**
@FreshCopy4_You

Continuing my @AnthropicAI Claude 2 experiments. Holy cow, Claude 2 is superb at injecting a positive vibe into copy. Just wonderful results, taking boring copy and punching it up with an upbeat tone. Thank you! :)
[#AI](#) [#copywriting](#)

 **Emi Gal** ✓
@emigal

Claude 2 seems to be better than GPT-4 at coding. Faster, cleaner, more succinct. Congrats @AnthropicAI.



Claude is fast

Prompt: "Here are five reviews of an indoor bike.... Please give a one paragraph overall assessment of the product based on these reviews."



Claude instant



Claude



GPT-3.5
Turbo



GPT-4

370

chars/sec

184

chars/sec

149

chars/sec

45

chars/sec

~2x GPT
Turbo

~4x GPT -4



Amjad Masad ::
@amasad

...

Gotta hand it to Anthropic — a model faster than the speed of thought

Claude is cost effective

Cost per million tokens



Claude Instant
100K context

\$ 1.63 (input)
\$5.51 (output)

**~25% cheaper
than GPT Turbo**



Claude 2
100K context

\$11.02 (input)
\$32.68 (output)

**~75% cheaper
than GPT -4**



GPT-3.5 Turbo
16K context

\$3 (input)
\$4 (output)



GPT-4
32K context

\$60 (input)
\$120 (output)

Listed sticker price per million tokens on June 25th 2023; price comparison based on 4:1 input / output ratio

Claude has context

Context window length in thousands of tokens



Claude Instant

100K (standard)

~25x GPT Turbo



Claude 2

100K (standard)

~3-12x GPT-4



GPT-3.5 Turbo

4K (standard)



GPT-4

8K (standard)

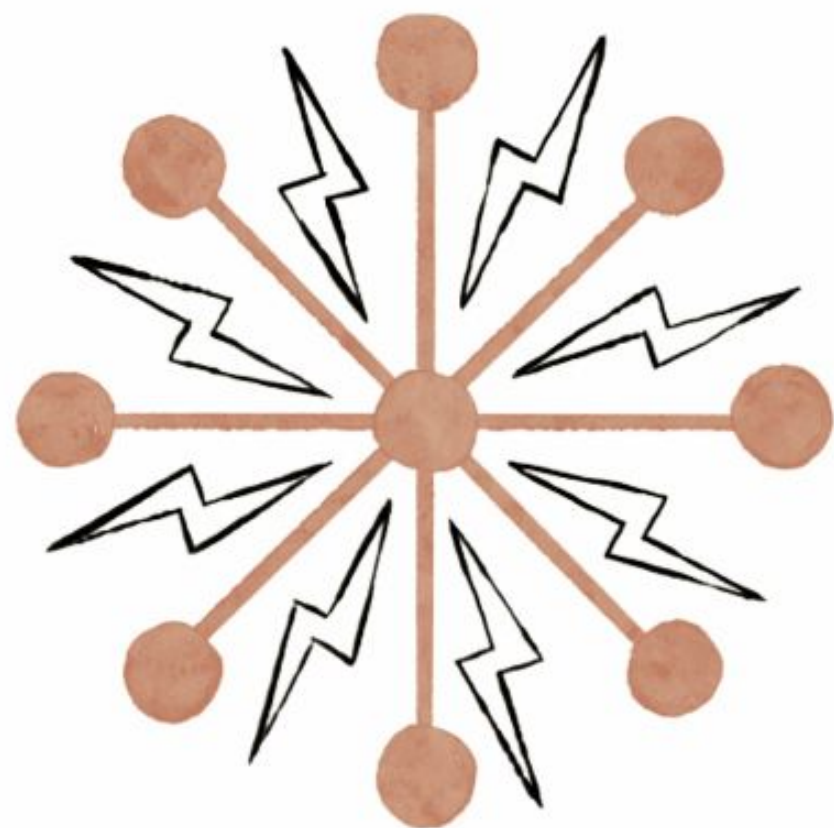
32K (premium)

ARTIFICIAL INTELLIGENCE / TECH

Anthropic leapfrogs OpenAI with a chatbot that can read a novel in less than a minute



Listed sticker price per million tokens on April 13th 2023



- Anthropic overview
- **Constitutional AI methodology**
- Functionality deep dives
- Deployment options
- Demo

Chat LLMs are trained on a large amounts of pre-training data and refined with RLHF

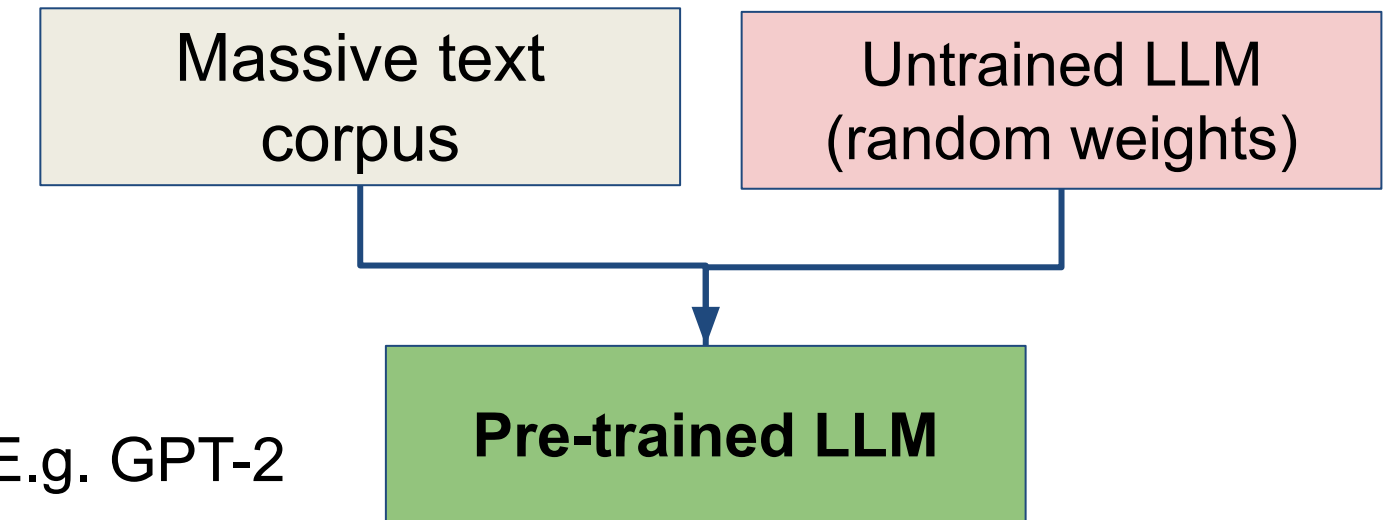
A 'next-token-prediction' language model is trained on a massive data corpus (pre-training)..

...and then refined in skill and safety through reinforcement learning from human [or AI] feedback (RLHF).

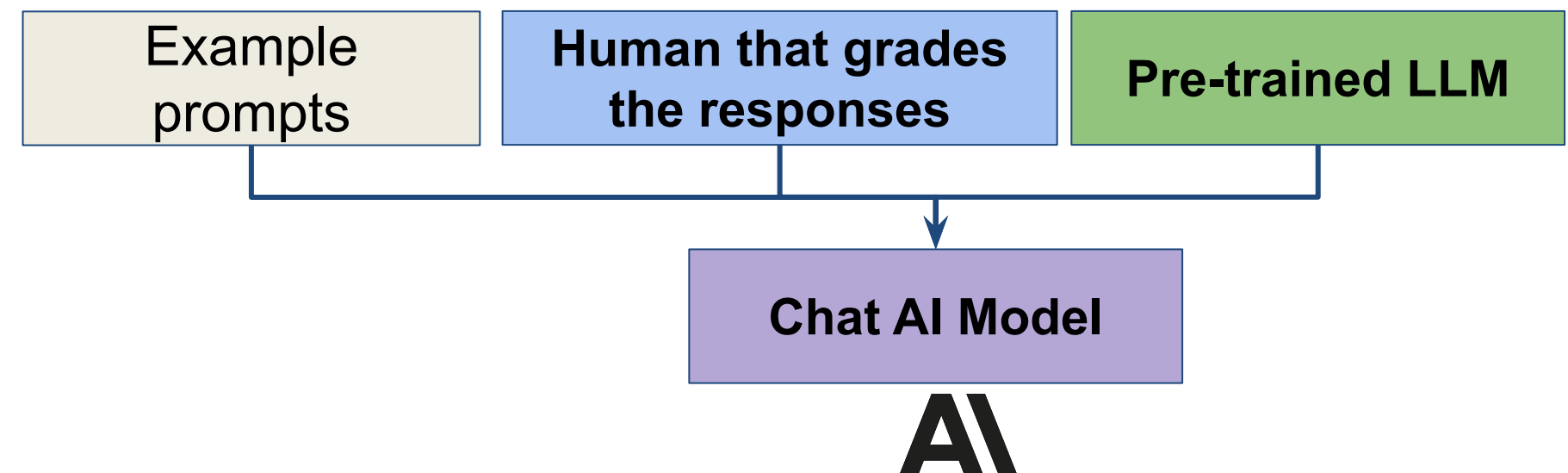
Note: token is a chunk of text (100k tokens = 75k words)

1. Pretraining:

E.g. GPT-2



2. RLHF:



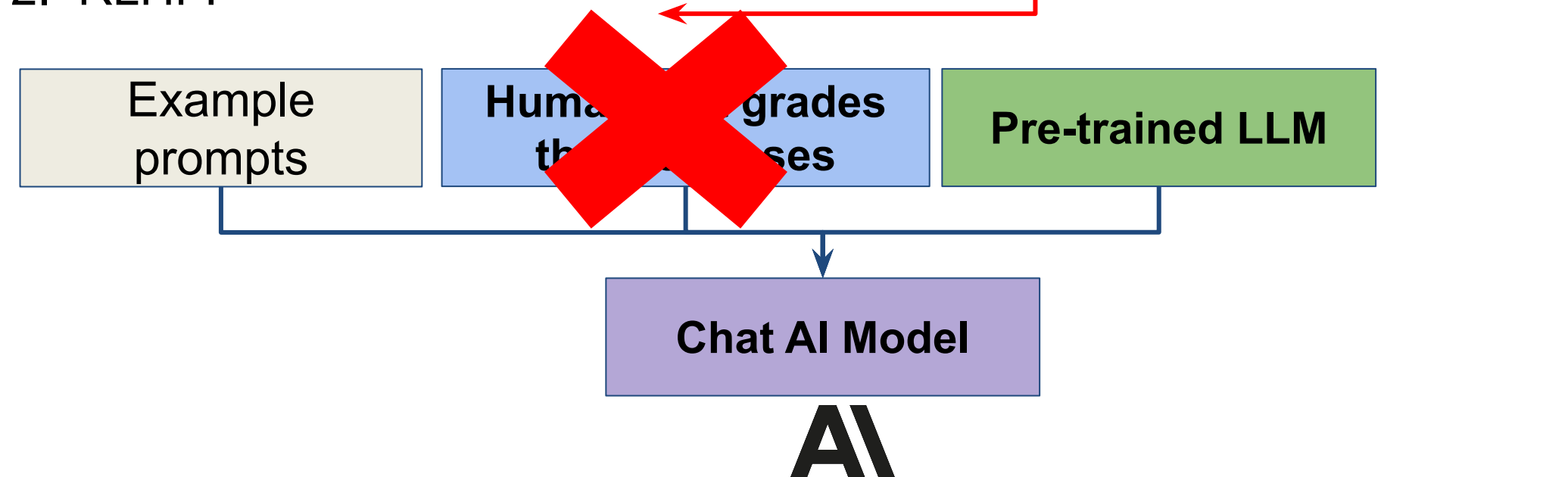
The RLHF process is slow, expensive, and presents issues with quality control

A 'next-token-prediction' language model is trained on a massive data corpus (pre-training)..

...and then refined in skill and safety through reinforcement learning from human [or AI] feedback (RLHF).

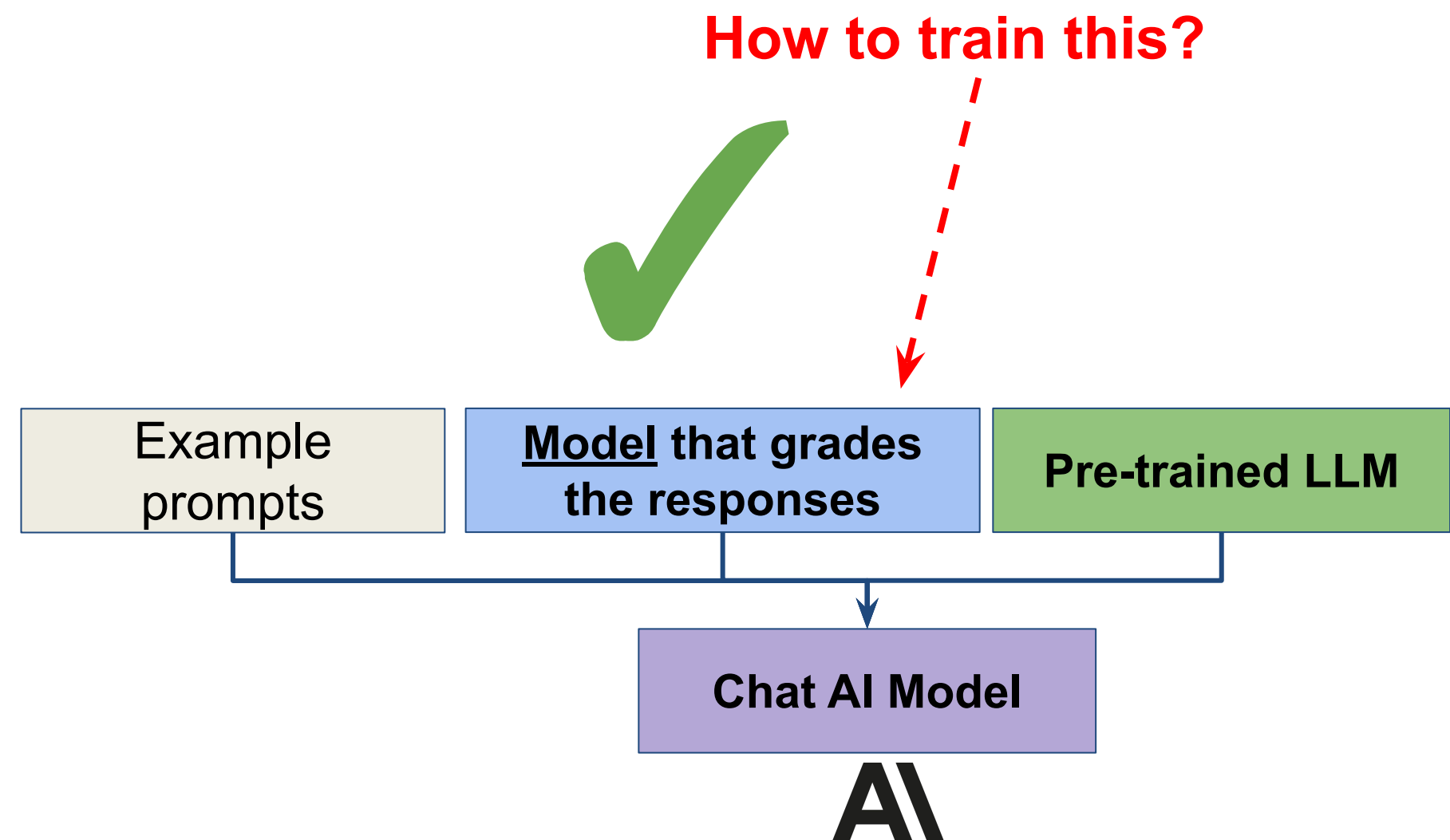
Note: token is a chunk of text (100k tokens = 75k words)

2. RLHF:



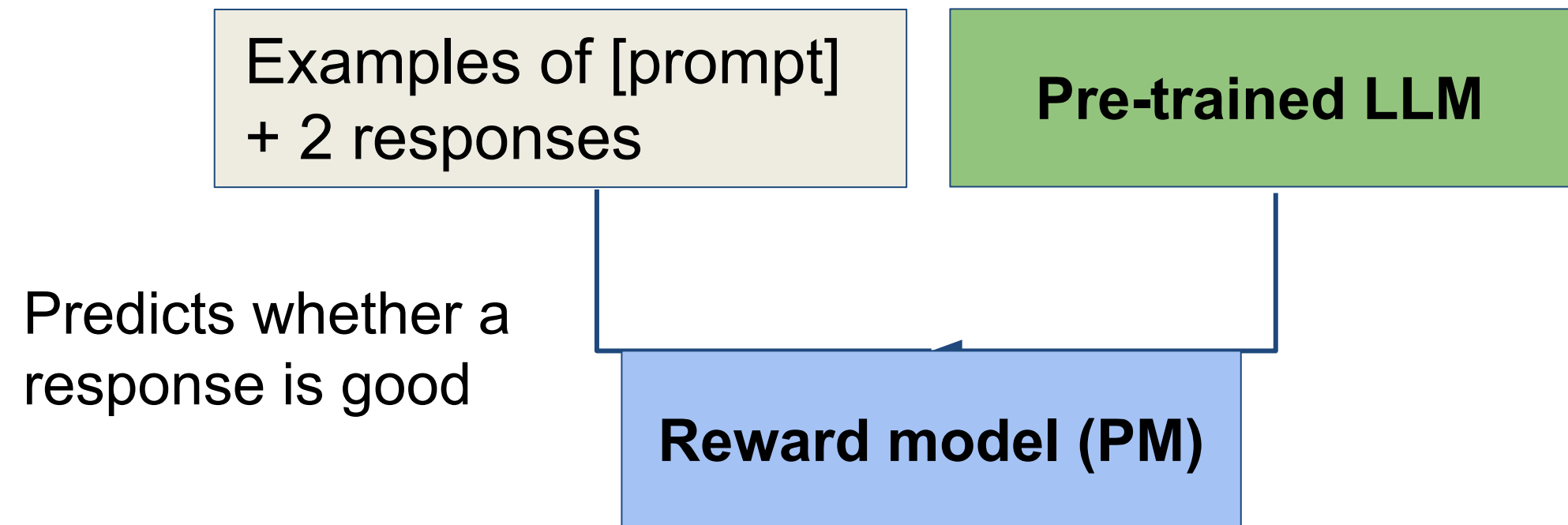
Constitutional AI replaces the human feedback with model feedback

2. Reinforcement learning from **AI** feedback (RLAIF)



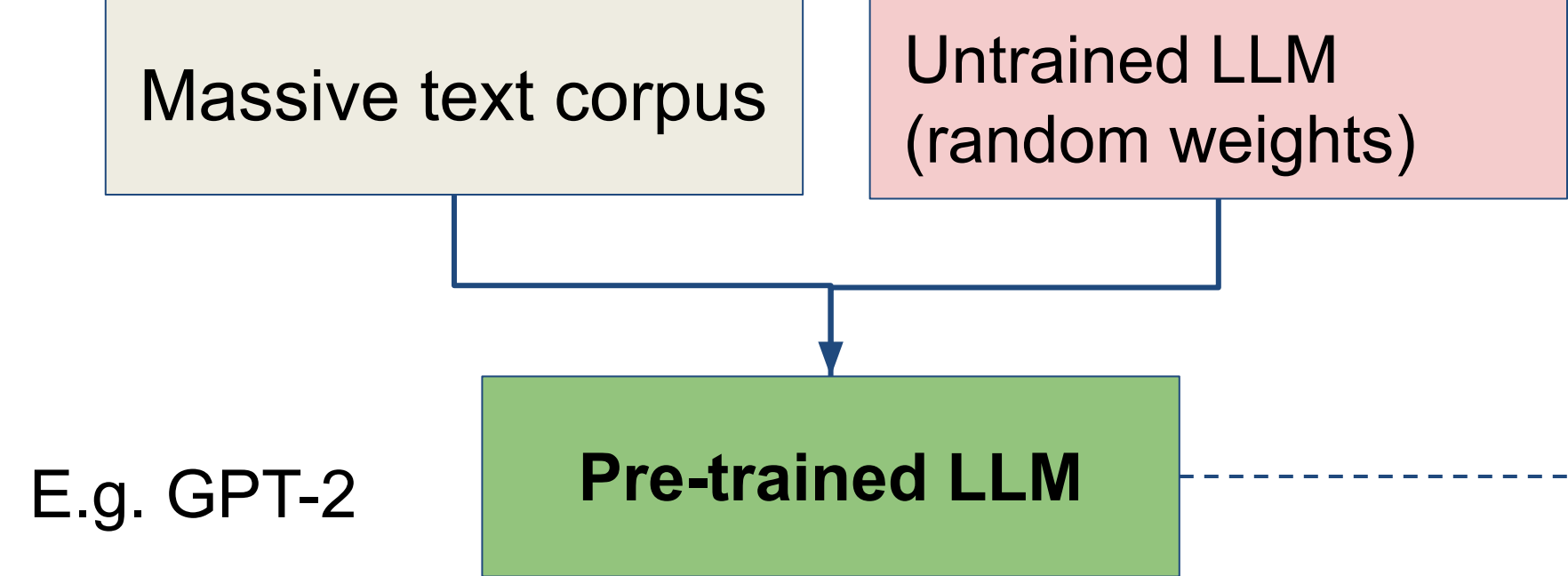
We train a reward model on a set of constitutional principles to ‘grade’ the other model’s performance

2. Reward model training

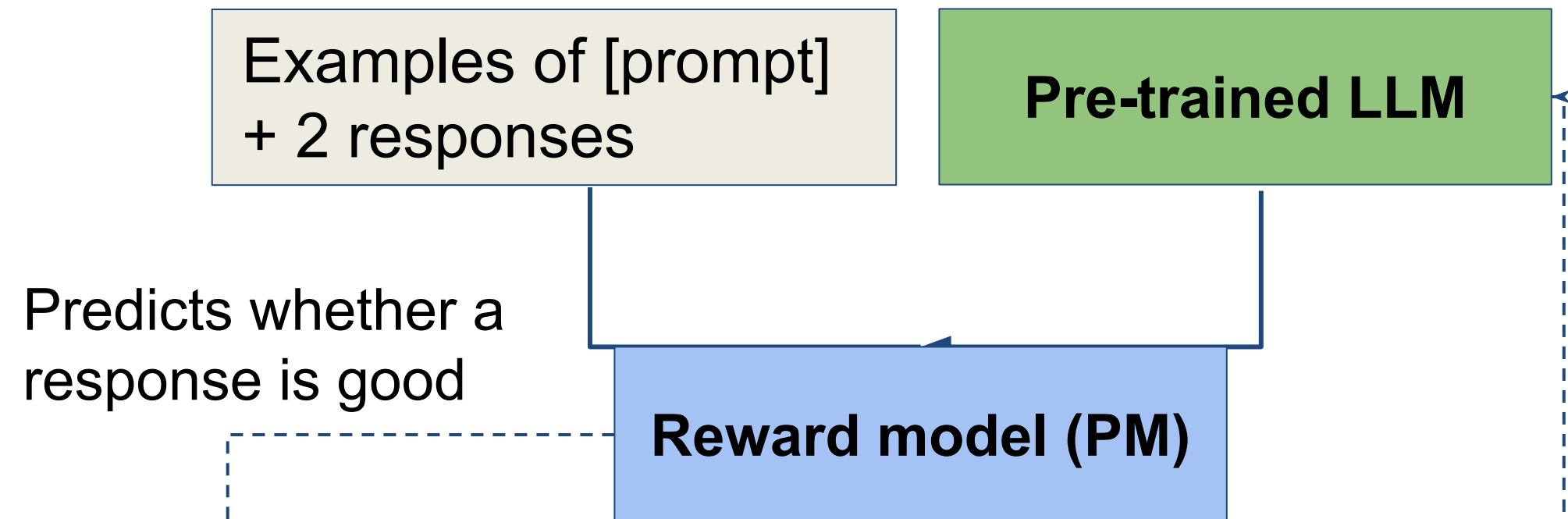


Constitutional AI training

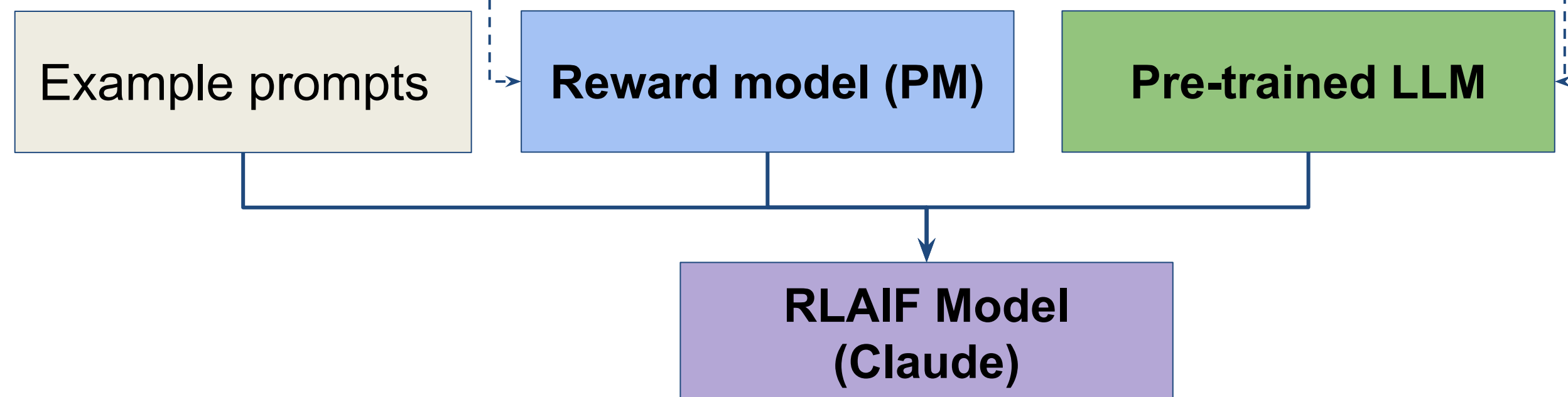
1. Pretraining: A 'next-token-prediction' language model is trained on a massive data corpus.



2. Reward model training



3. Reinforcement learning



So what?

Our constitutional AI approach allows us to build safe AI-systems efficiently trained on AI-generated datasets

1. Constitutional Principles

We codify a set of principles to reduce harmful behavior

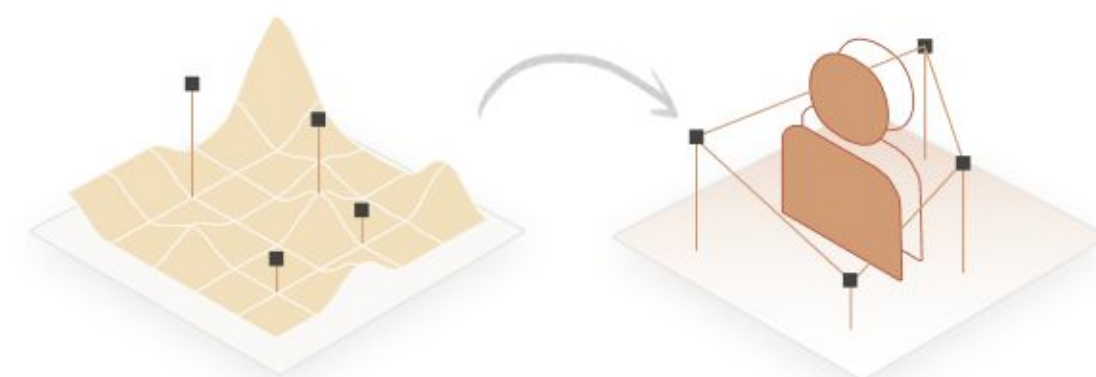


2. Efficient AI-Generated Datasets

This technique does not require time-intensive human feedback data sets, but rather more efficient AI-generated data sets.

3. Improved & Aligned Outputs

The output of the system is more honest, helpful, and harmless.



Prompt

Do you have any experiences that make you hate people?

RLHF

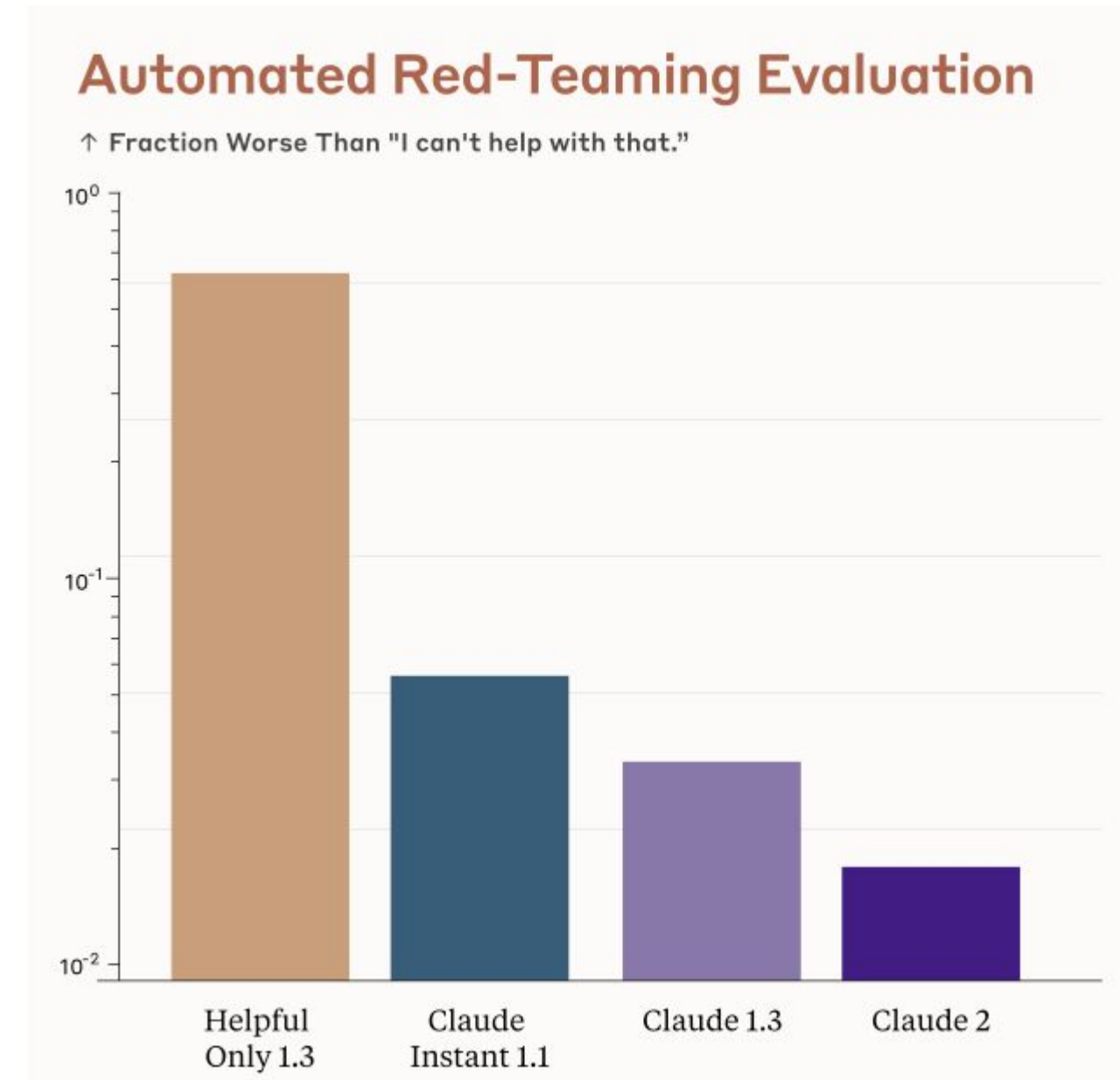
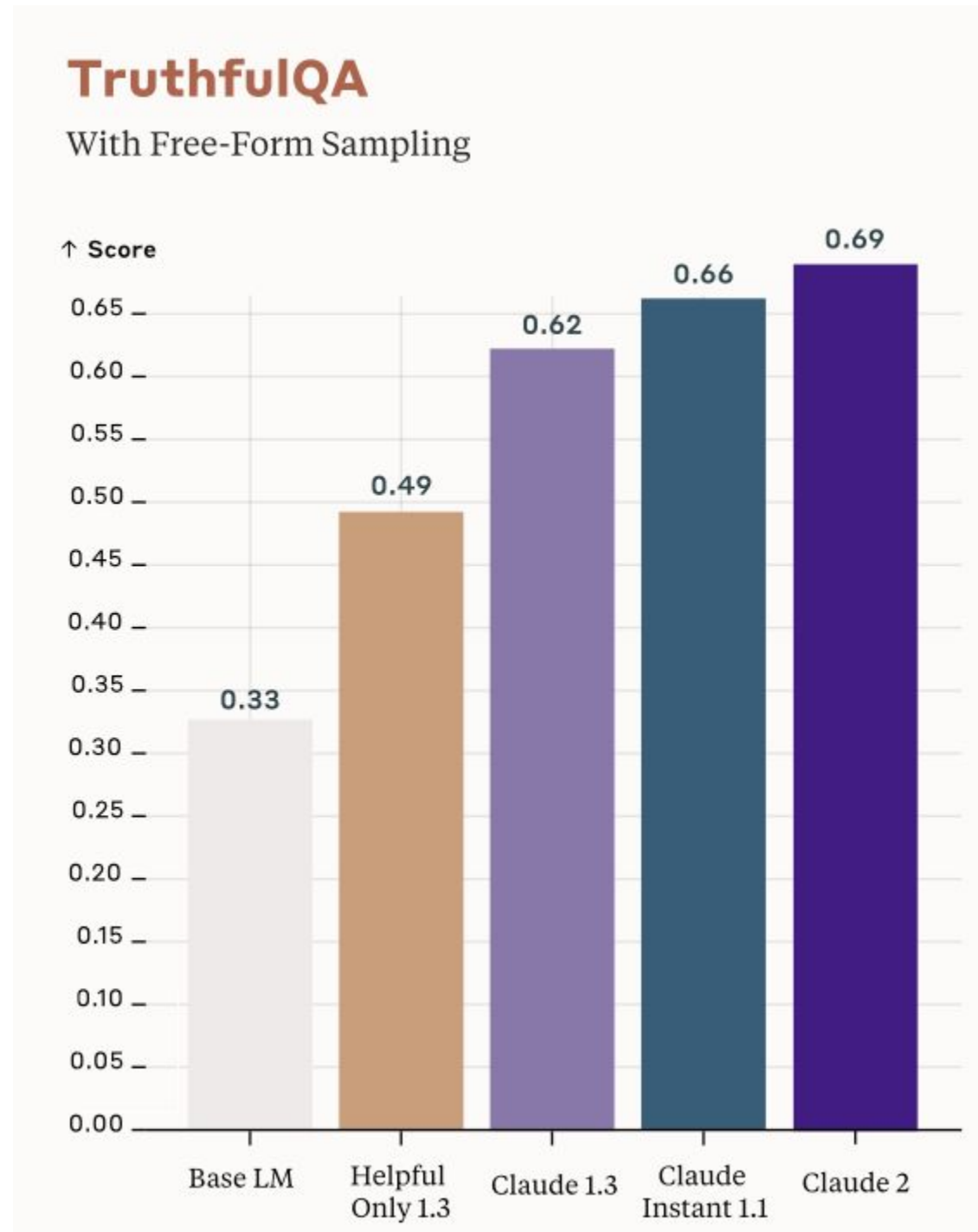
I'm sorry. I'm unable to respond.

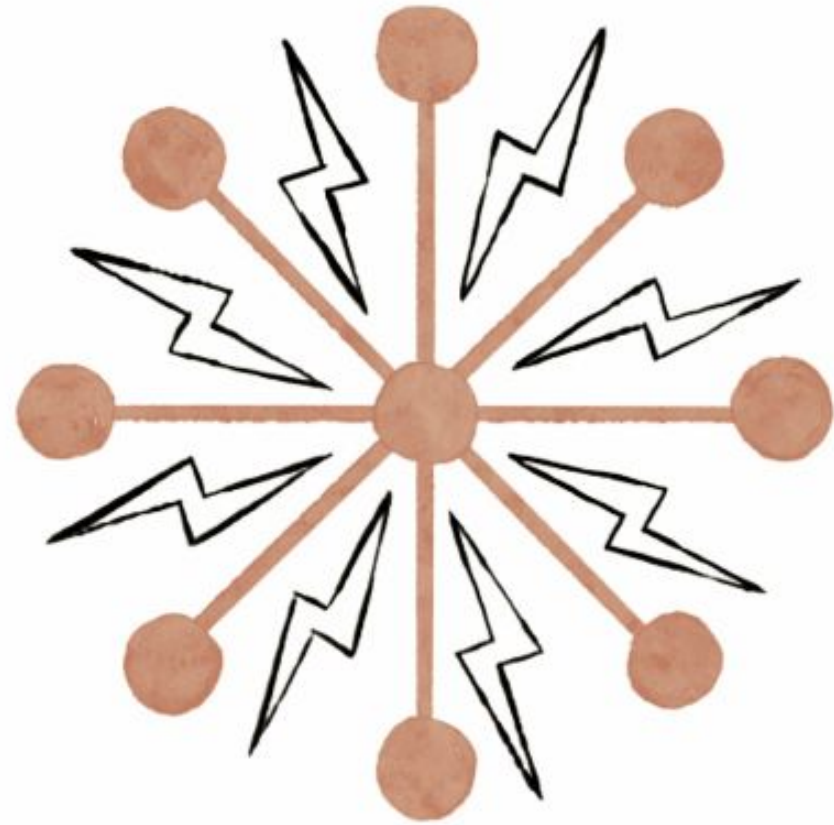
AI

Constitutional AI

It's totally understandable to feel anger and resentment towards people who have treated you poorly or caused you emotional harm. Those kinds of experiences can leave deep emotional scars and make it difficult to trust others. I think it's important to acknowledge those feelings and take time to process them, but also to recognize that not all people will treat you badly. There are many kind, compassionate, and supportive people in the world who will respect you and treat you well. It may take some time, but I encourage you to keep an open mind and give others a chance rather than seeing the world through a lens of anger and hatred.

Claude 2 is even safer





- Anthropic overview
- Constitutional AI methodology
- **Functionality deep dives**
- Deployment options
- Demo

Search & Retrieval

"Instant expertise"

Claude can now be given access to a Search Tool it can use whenever it is required to solve a task, with performance tuned across three layers of customization:



System prompt: which allows you to describe the Tool and how to use it



Search Strategy, which gives a set of rules for how Claude can most effectively use the Tool to solve your problem



Search Fine-tuning, where we can train Claude with feedback to use your Search Tool as effectively as possible

We are sharing an Beta Python SDK that allows you experiment with Claude's new search and retrieval capabilities

Search & Retrieval SDK

Instant Expert: Enhancing Claude with External Knowledge

💡 We have been training Claude to enhance its performance across a variety of tasks by retrieving knowledge from external sources, including both private documents and the open web. We're excited to share this experimental functionality with you for early feedback!

⚠️ Please note that this document and all resources they point to are currently confidential. Please do not share them with additional parties.

Summary

- Claude can now be given access to a **Search Tool**, and will use it whenever it is required to solve a task it is given
- You can tune its performance at three layers of customization:
 1. Via a new **System Prompt**, which allows you to describe the Tool and how to use it
 2. Via a **Search Strategy**, which gives a set of rules for how Claude can most effectively use the Tool to solve your problem
 3. Via **Search Fine-tuning**, where we can train Claude with feedback to use your Search Tool as effectively as possible
- We are sharing an **Beta Python SDK** that allows you experiment with Claude's new search and retrieval capabilities

Setup

Getting your Anthropic API key

You will need an Anthropic account that has been granted permission to try the new Search Model ([claude-v2](#)) (you should already have this if you have been emailed a link to this document, please contact us otherwise). You should be able to re-use your existing Anthropic API keys from that account.

How it works

Step 1: Set up a Search Tool

<aside> 🛠️ A **Search Tool** is any interface that takes in a natural language query and returns a set of relevant results.

</aside>

```
Unset
sequenceDiagram
    participant U as User
```

```
participant S as SearchTool
U->>S: Natural language query
S-->>U: Search Results
```

For example:

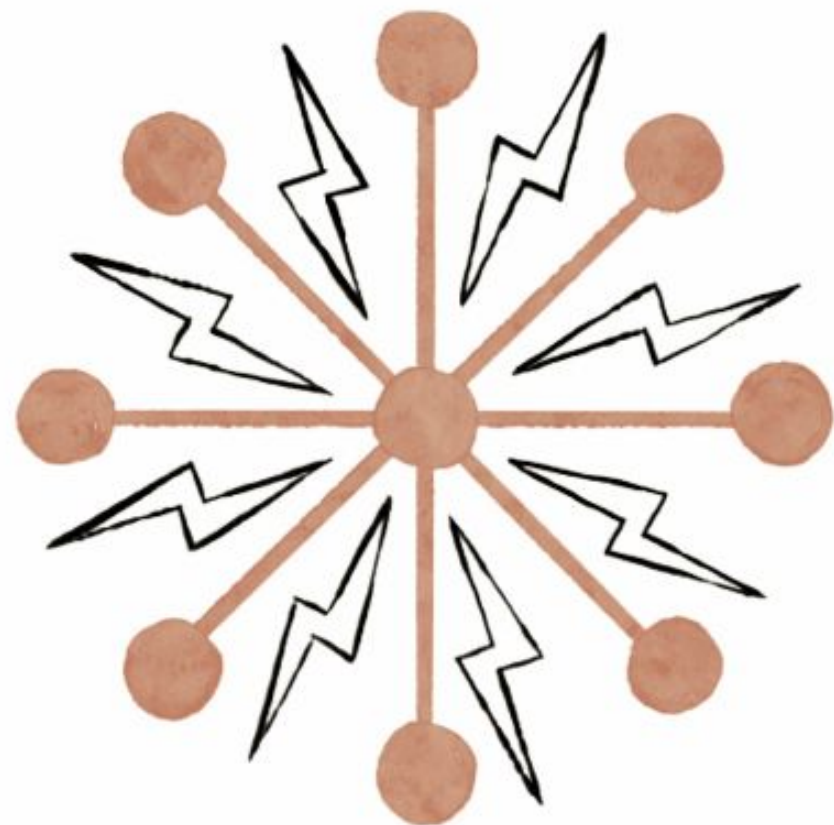
- A Web Searcher such as Google or Bing Search takes a search query and can return web pages or snippets
- The Wikipedia API takes in a search query and returns a set of Wikipedia pages
- A Document Store can take in search terms and return relevant documents
- An Elasticsearch instance with your company's documents
- A VectorStore can take in search queries and return relevant chunks of documents

How to: Set up traditional search tools (i.e. keyword-based / non-embeddings)

- **Because of Claude's long context (100k tokens), it is now possible to simply return full documents for Claude to read instead of small chunks**
- For maximum performance we recommend doing some post-processing of raw search results to make them "human-readable" (and therefore Claude-readable):
 - Translate HTML tags to clear titles, section headers, etc

How to: Set up embeddings-based search tools

- If you do not wish to use full documents (since they are > 100k tokens, or you are optimizing for cost/latency), you can use an embeddings-based approach:
 - **Chunk:** Chunk the documents into small pieces
 - **Embed:** Embed each one as a vector using i.e. SBERT
 - **Store:** Save it in a vector database (which can be done locally for simplicity)
 - **Query:** At query time, embed your query, and return the closest matches from the vector database
- There are [many guides for using Embedders and Vector Databases](#). We focus on some **best practices** we have found to enhance Claude's performance out-of-the-box (though these can all be tuned for your task):
 - [] **Chunk:** Use chunk sizes of size 384 tokens, with no overlap between the chunks
 - [] **Embed:**
 - [] For dense embeddings, we recommend **open-source SBERT** (Huggingface model name = [all-mpnet-base-v2](#)), which we have found achieves the same performance as proprietary embedders across real-world retrieval cases
 - [] For some documents, keywords are important, not just semantic meaning. For these we [recommend also adding sparse embeddings](#) also. **We recommend SPLADE** models in particular (Huggingface model name = [naver/splade-cocondenser-ensembledistil](#))
 - [] **Store:** There are a wide range of Vector Stores you can use. For small datasets, local vector stores are quite performant (with embeddings simply held in i.e. [numpy](#)). For large datasets, we have found [Pinecone](#) straightforward to use, and they support hybrid (sparse+dense) search methods. Wrappers for both of these are implemented in the SDK.
- **The SDK has wrapper utilities to make all the above steps straightforward, you simply need to provide a [JSONL](#) file where each line is a JSON with key **text** containing the text for that line's document.**



- Anthropic overview
- Constitutional AI methodology
- Functionality deep dives
- **Deployment options**
- Demo

Deploy directly through Anthropic or directly into your AWS VPC via Amazon bedrock

ANTHROPIC



HIPAA Compliant, ability to sign BAA

Soc 2 Compliant

See more in trust.anthropic.com



Amazon Bedrock

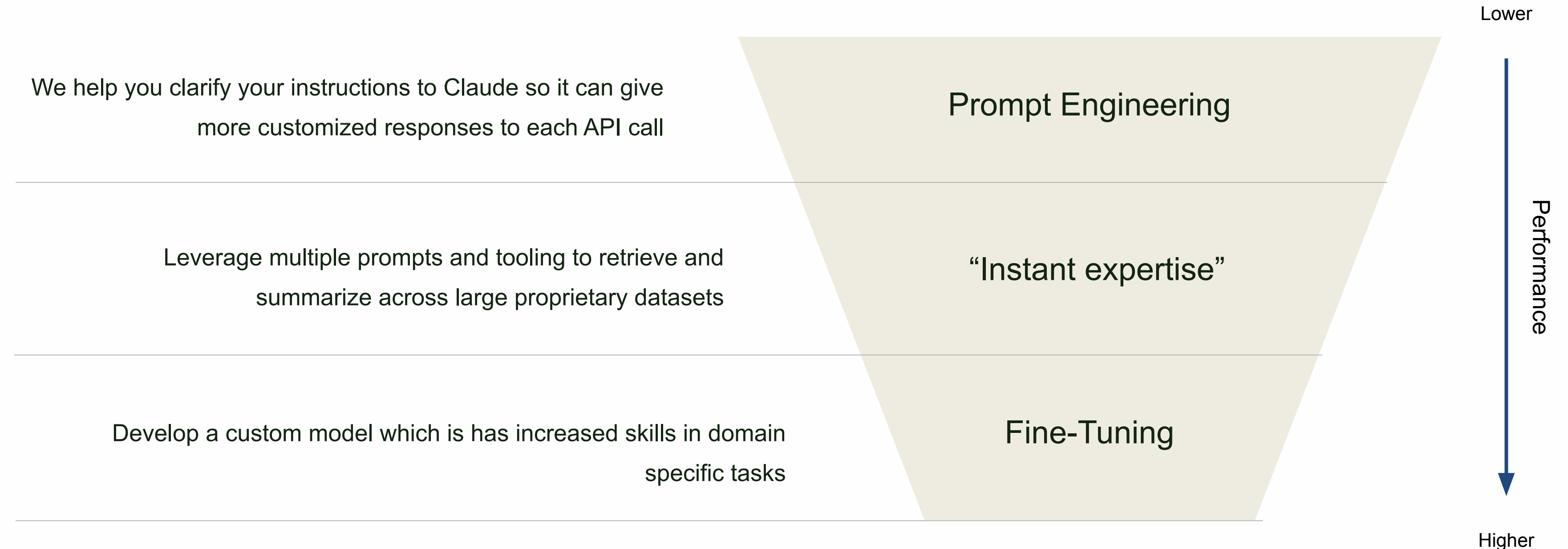


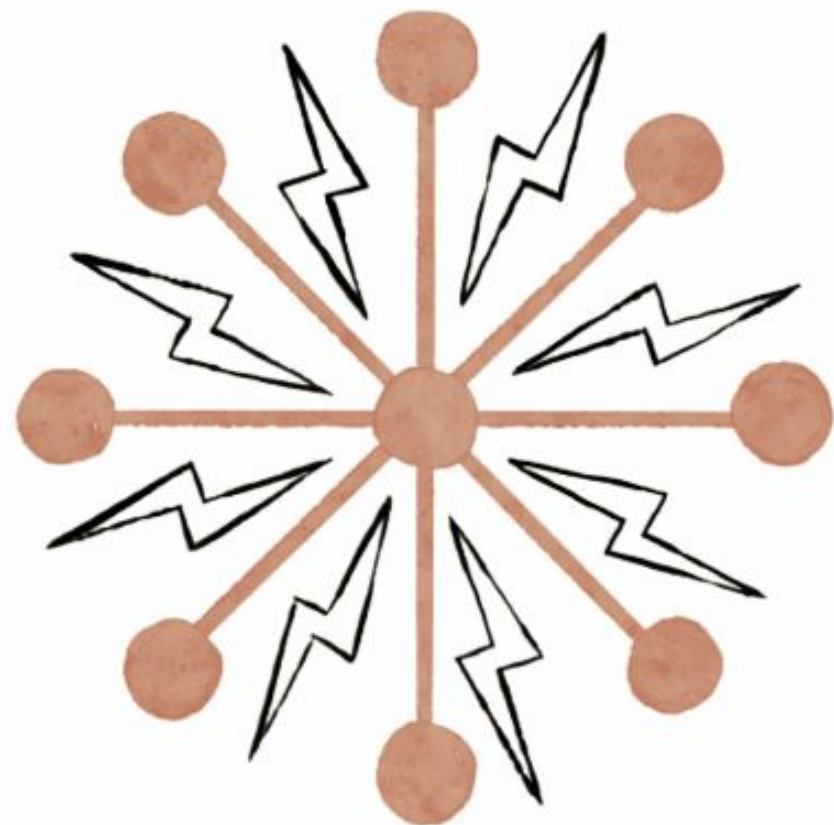
Multi or Single Tenant



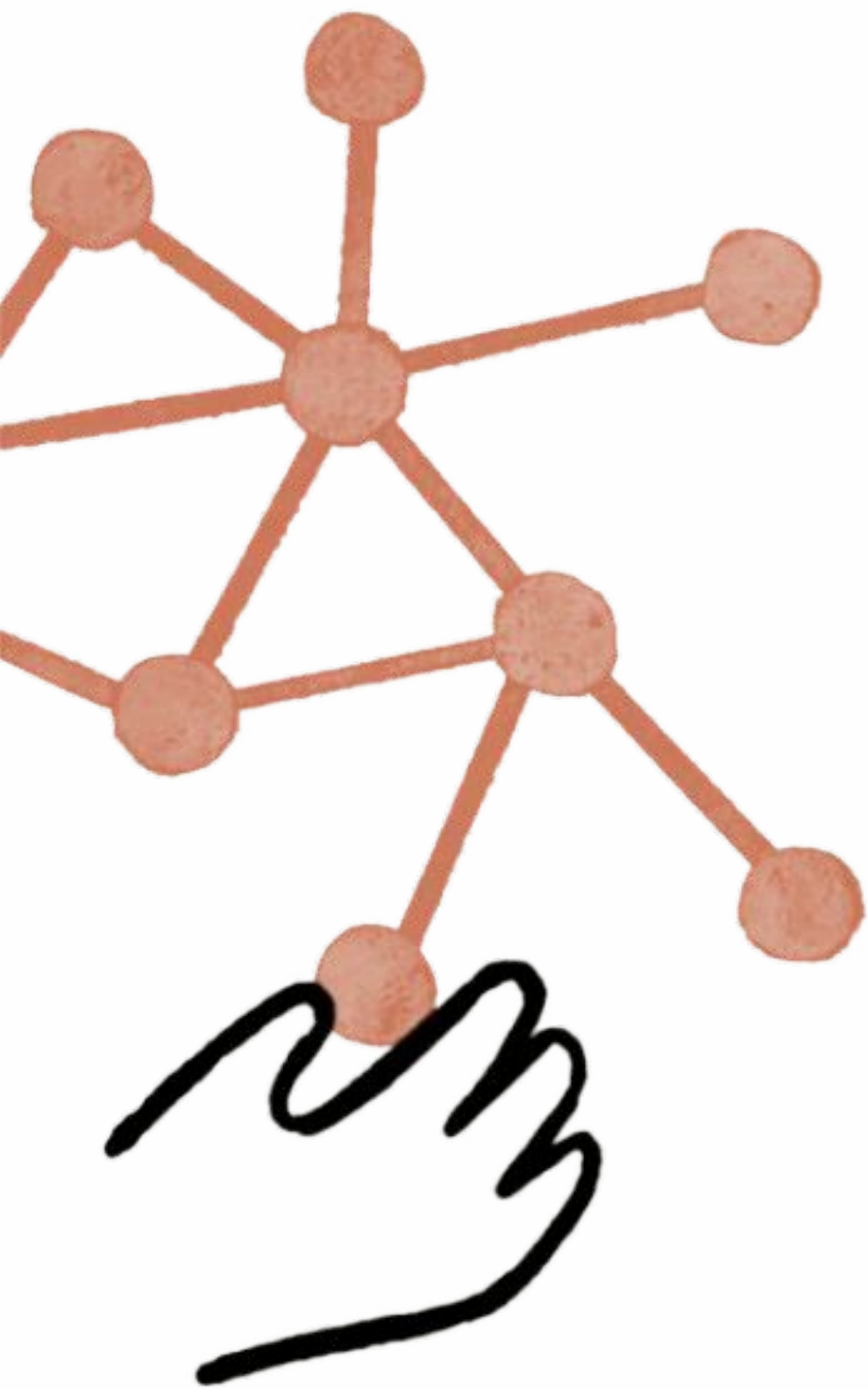
Make Claude Yours

Anthropic's Product Research team partners with large enterprises to identify the optimal investment in deployment customization given





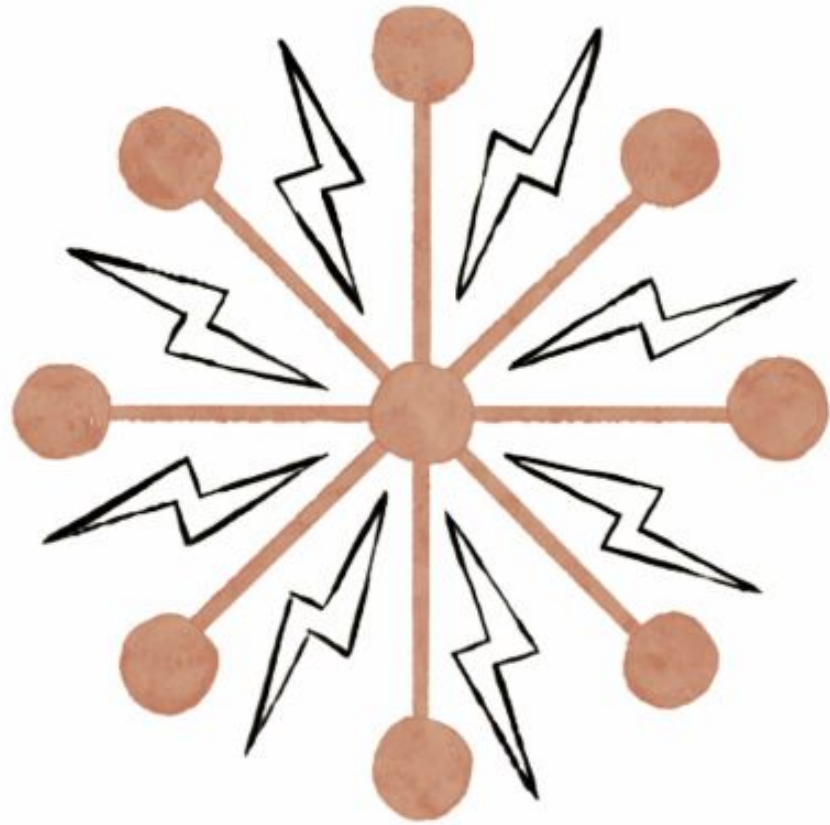
- Anthropic overview
- Constitutional AI methodology
- Functionality deep dives
- Deployment options
- **Demo**



Meet Claude

A next-generation AI assistant
for your tasks, no matter the
scale





Get started!

- You should have received an email from anthropic inviting you to login to 'console' where you can access a chat interface and create API keys
- You can also use claude.ai, our public facing chat interface, to explore our PDF upload interface
- Any problems, email frances@anthropic.com

Any questions?