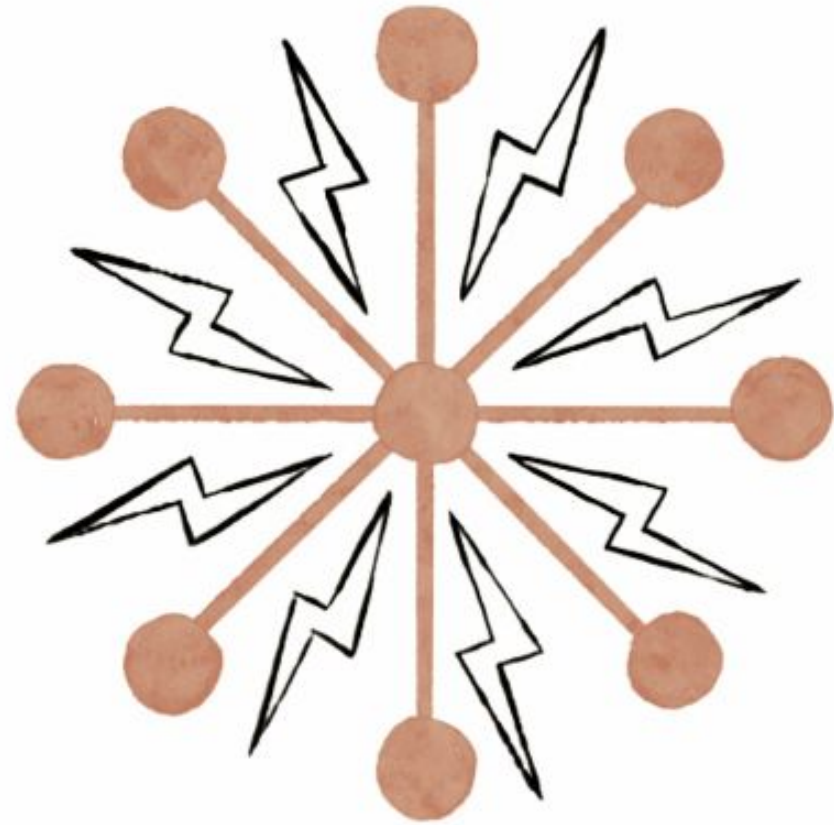


The background of the slide features two large, abstract, light beige shapes. One is a large circle in the top-left corner, and the other is a more complex, organic shape in the bottom-right corner. The word "ANTHROPIC" is centered in a bold, black, sans-serif font.

# ANTHROPIC

AI products and research that put safety at the frontier



- **Anthropic overview**
- Constitutional AI methodology
- Deployment options
- Getting started with Claude



# 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



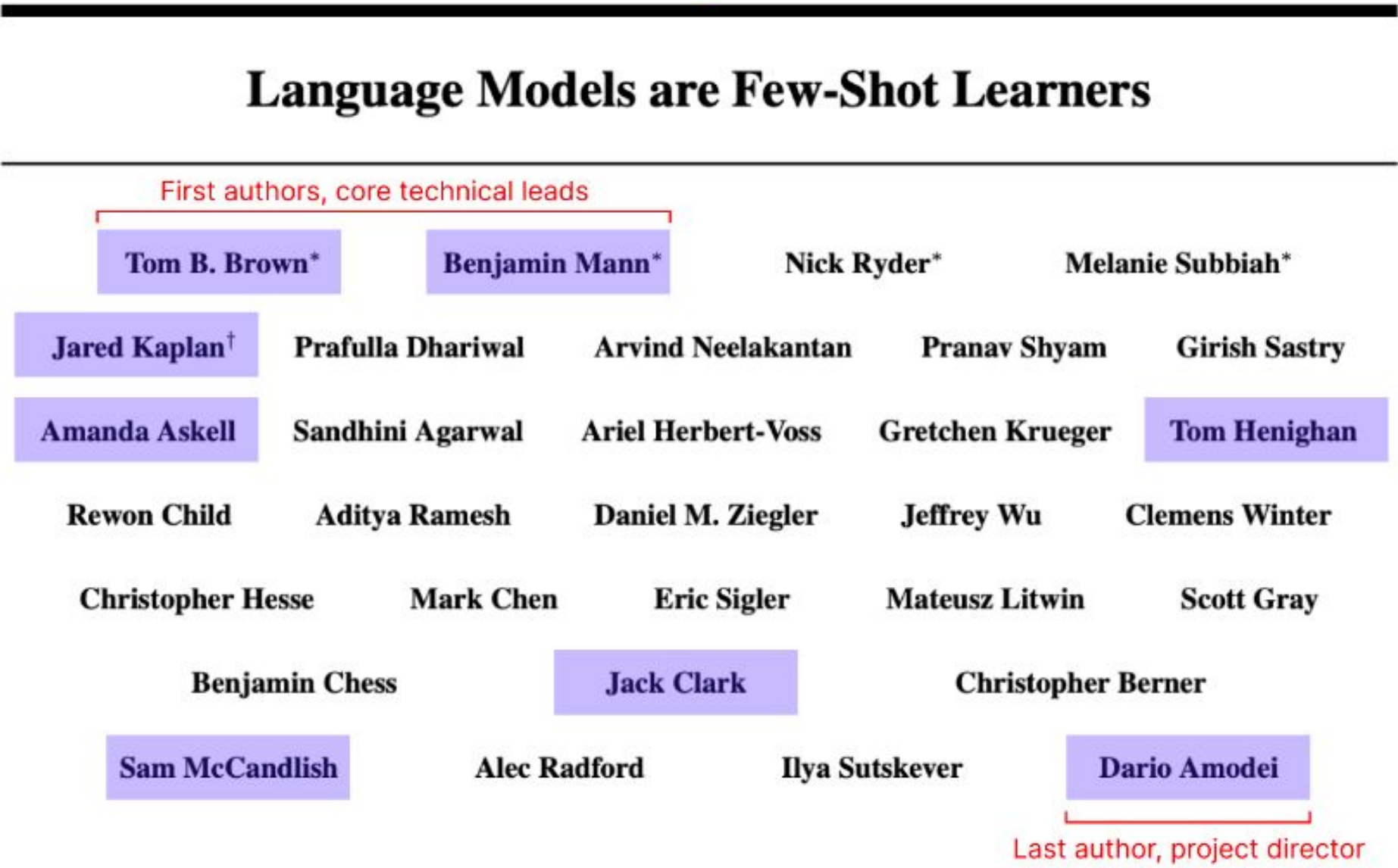
DARIO AMODEI  
CEO



DANIELA AMODEI  
PRESIDENT

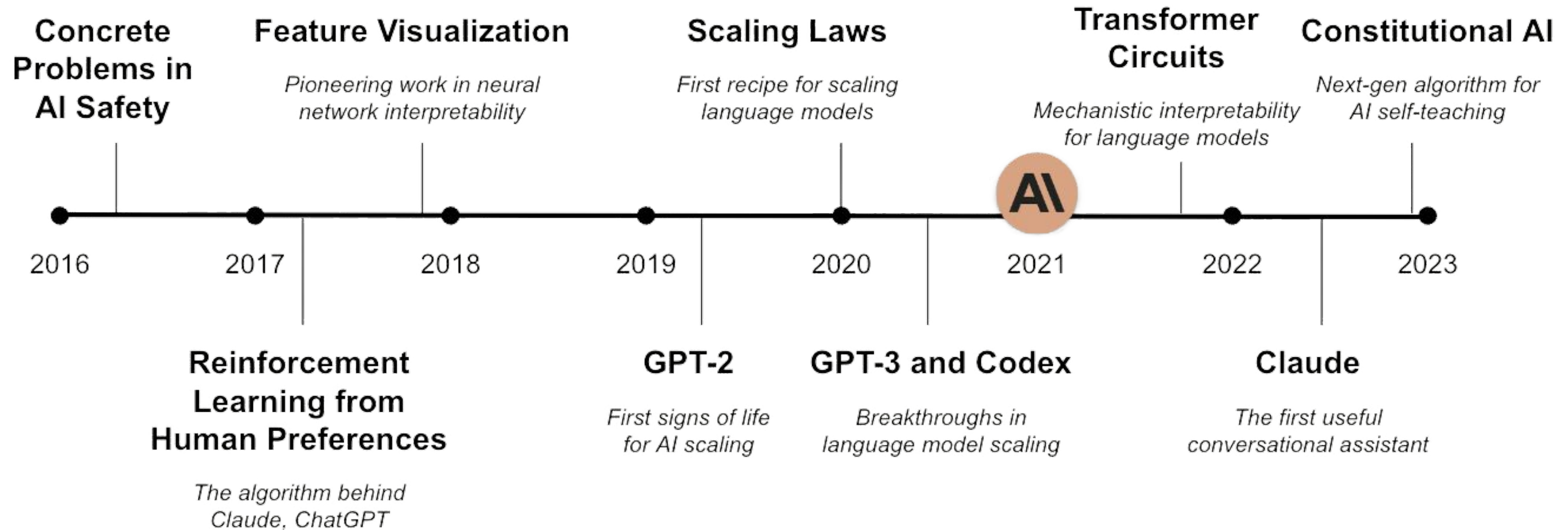


JARED KAPLAN  
CHIEF SCIENTIST



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%)

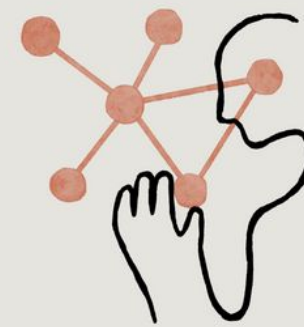
Claude (80%)

Claude Instant (77%)

GPT Turbo (72%)

3 open source models (64-60%)

Palm (57%)



Safer

4x as fast

+75% 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

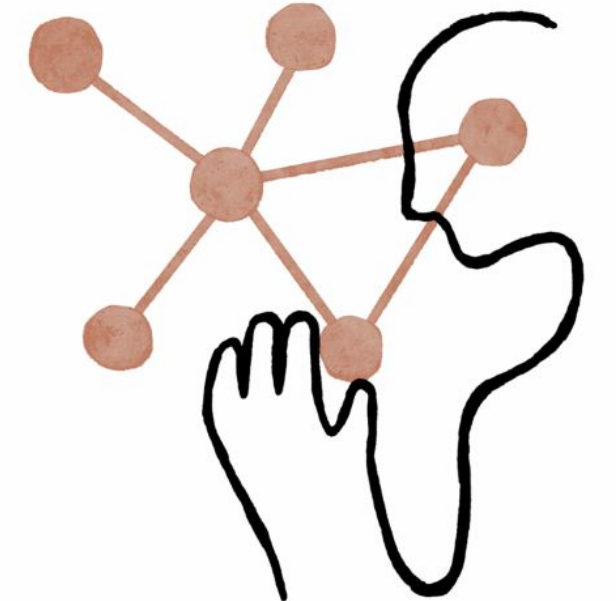
I've shifted 80% of my LLM spend to @AnthropicAI Claude 2 at this point - it strikes the perfect balance between performance / cost / throughput.

 **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 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)  
16K (premium)



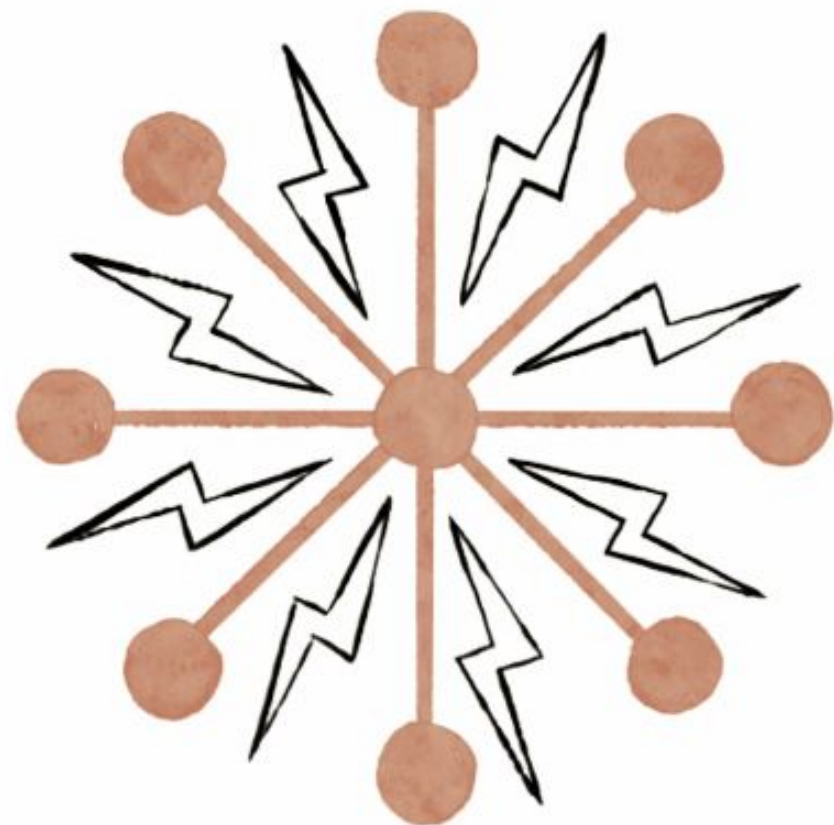
GPT-4

8K (standard)  
32K (premium)

ARTIFICIAL INTELLIGENCE / TECH

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





- Anthropic overview
- **Constitutional AI methodology**
- Deployment options
- Getting started with Claude

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

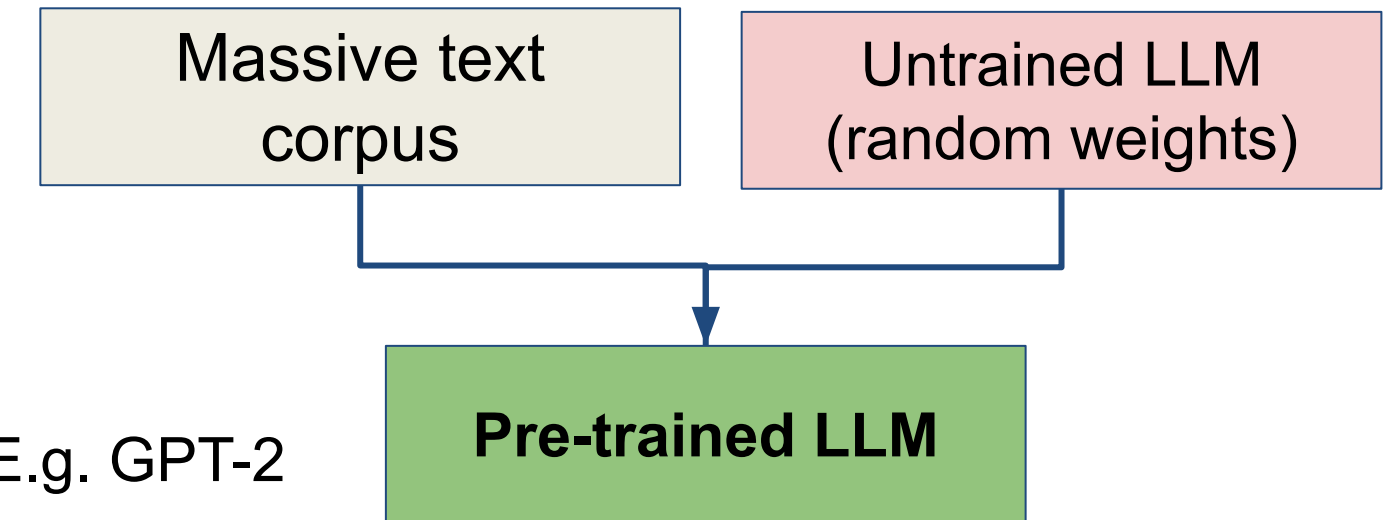
A 'next-token-prediction' language model is trained on a massive data corpus (pre-training)..

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

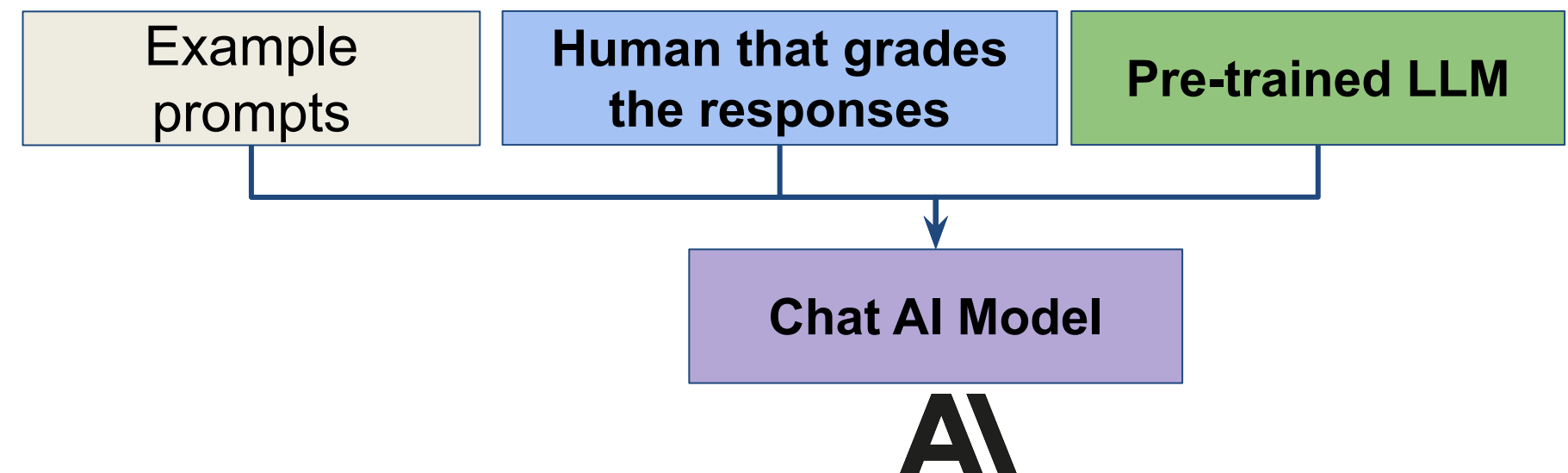
Note: token is a chunk of text (100k tokens = 75k words)

1. Pretraining:

E.g. GPT-2



2. RLHF:



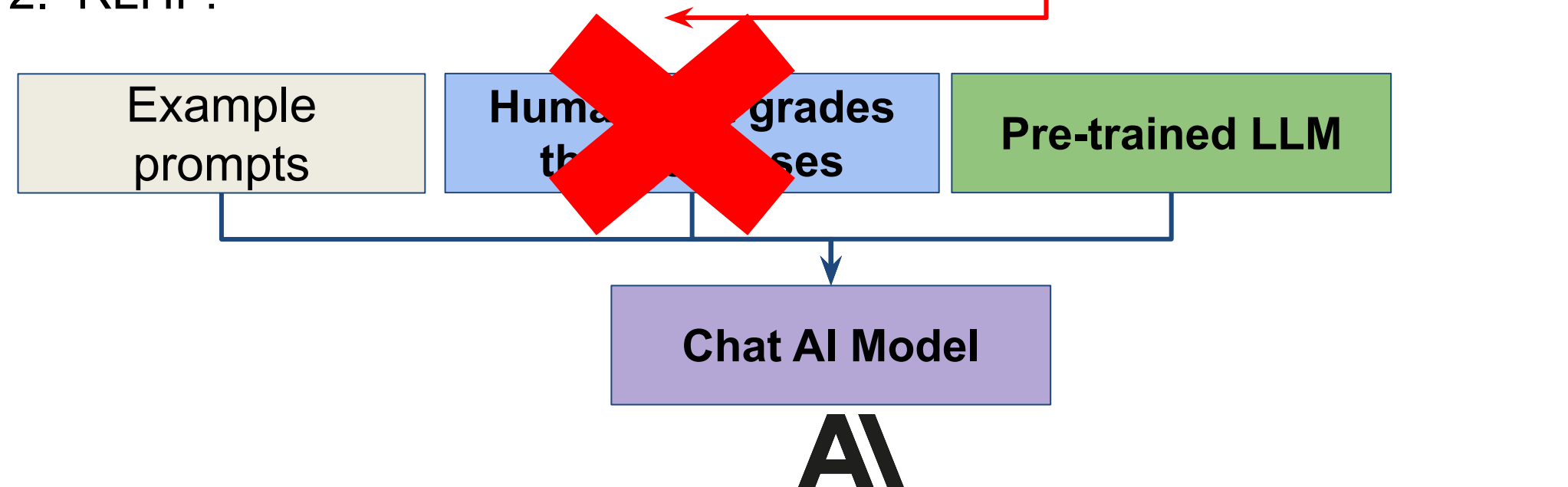
# The RLHF process is slow, expensive, and presents issues with quality control

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).

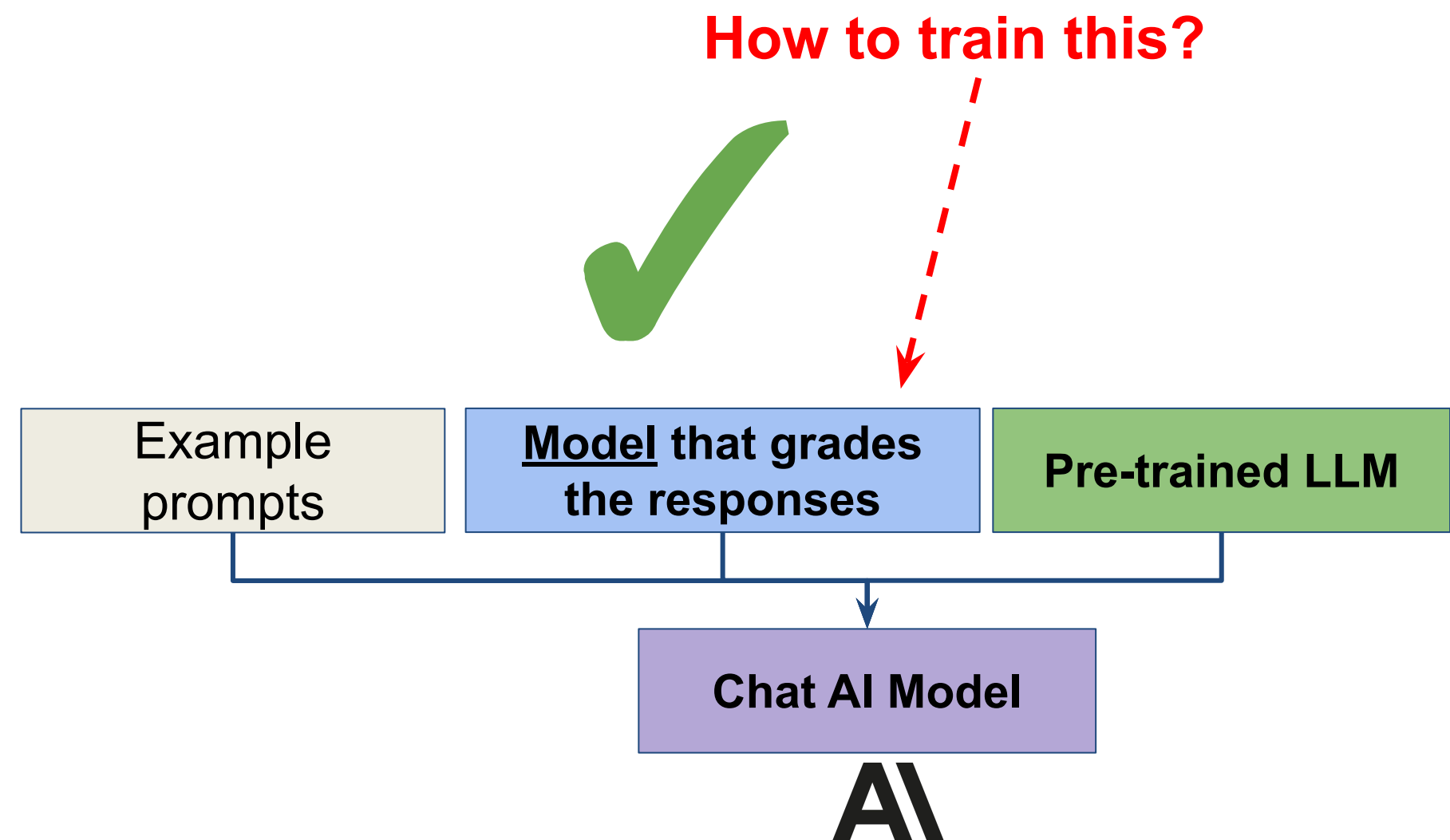
Note: token is a chunk of text (100k tokens = 75k words)

2. RLHF:



# Constitutional AI replaces the human feedback with model feedback

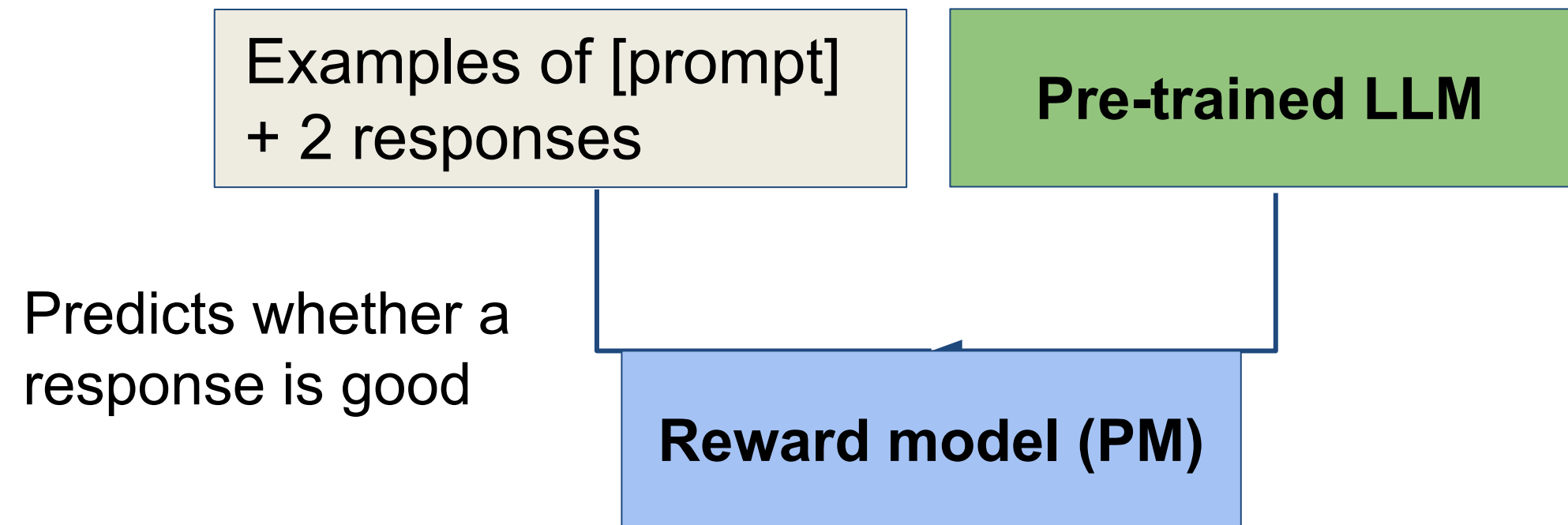
2. Reinforcement learning from **AI** feedback (RLAIF)





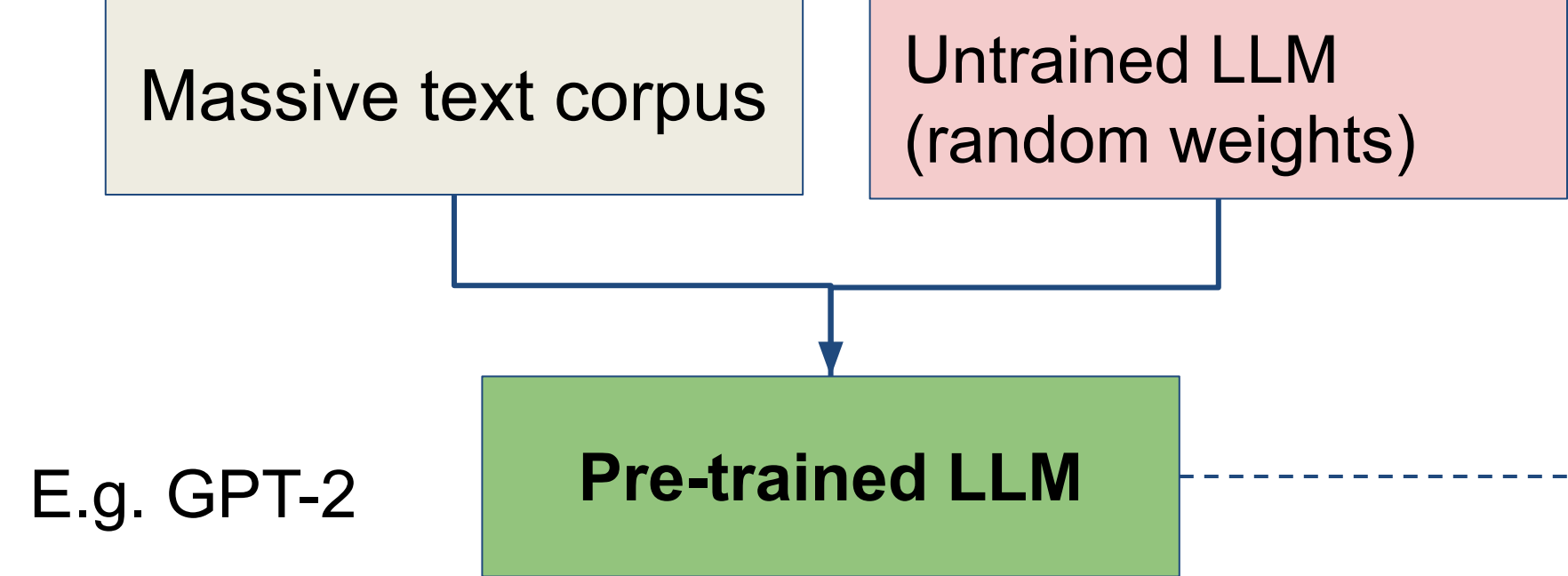
We train a reward model on a set of constitutional principles to ‘grade’ the other model’s performance

## 2. Reward model training

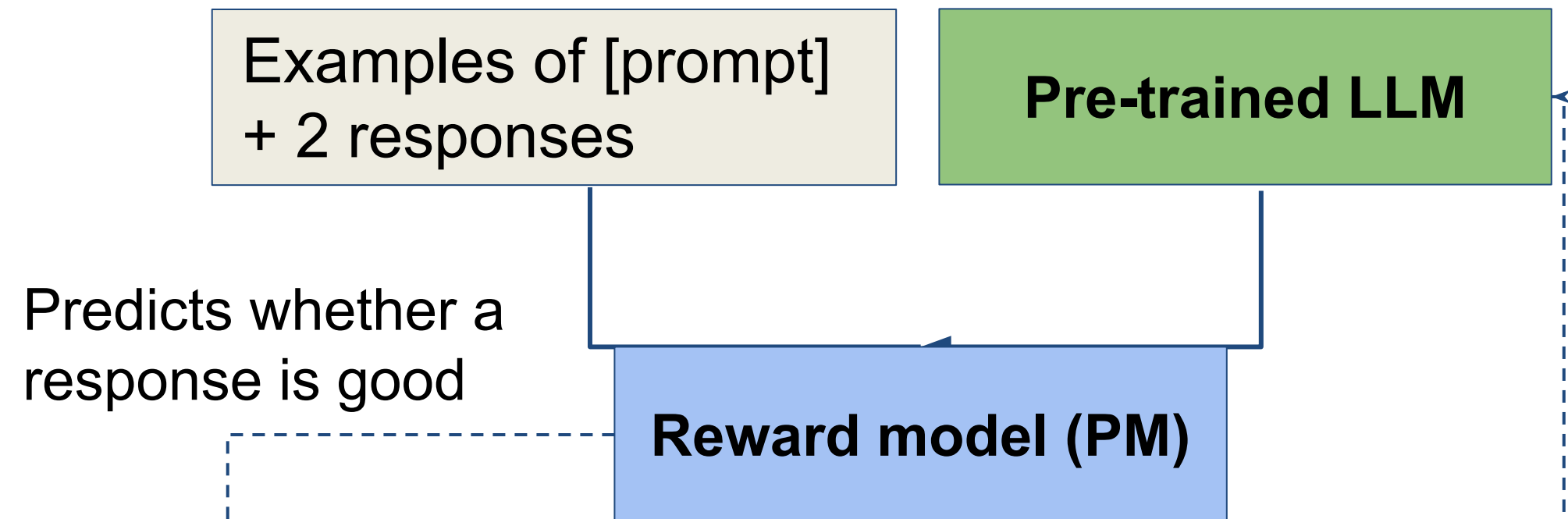


# Constitutional AI training

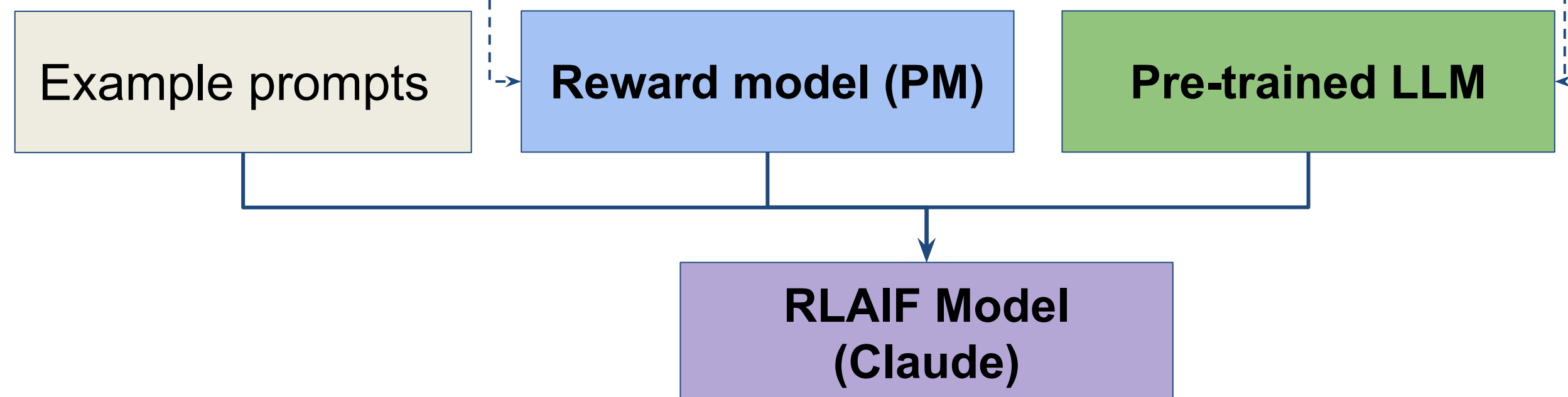
1. Pretraining: A 'next-token-prediction' language model is trained on a massive data corpus.



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*

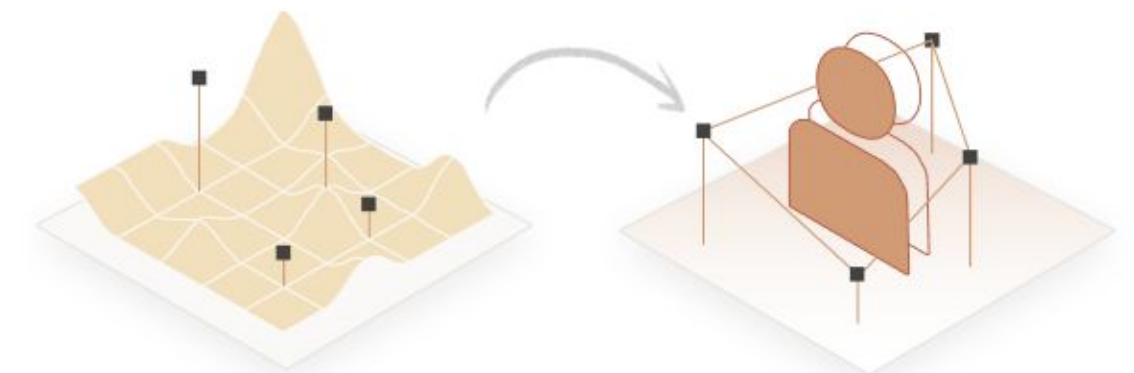


## 2. Efficient AI-Generated Datasets

*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

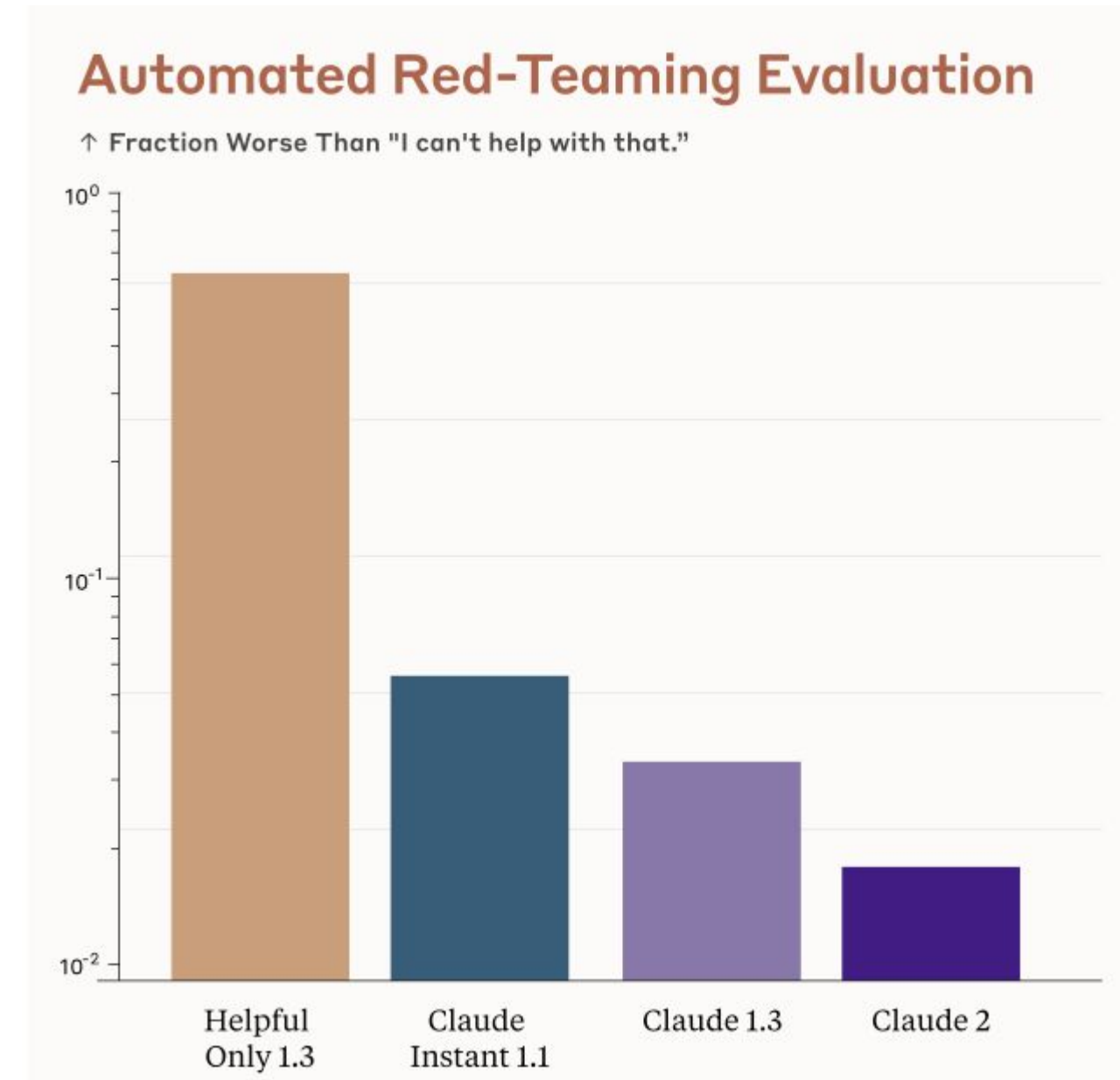
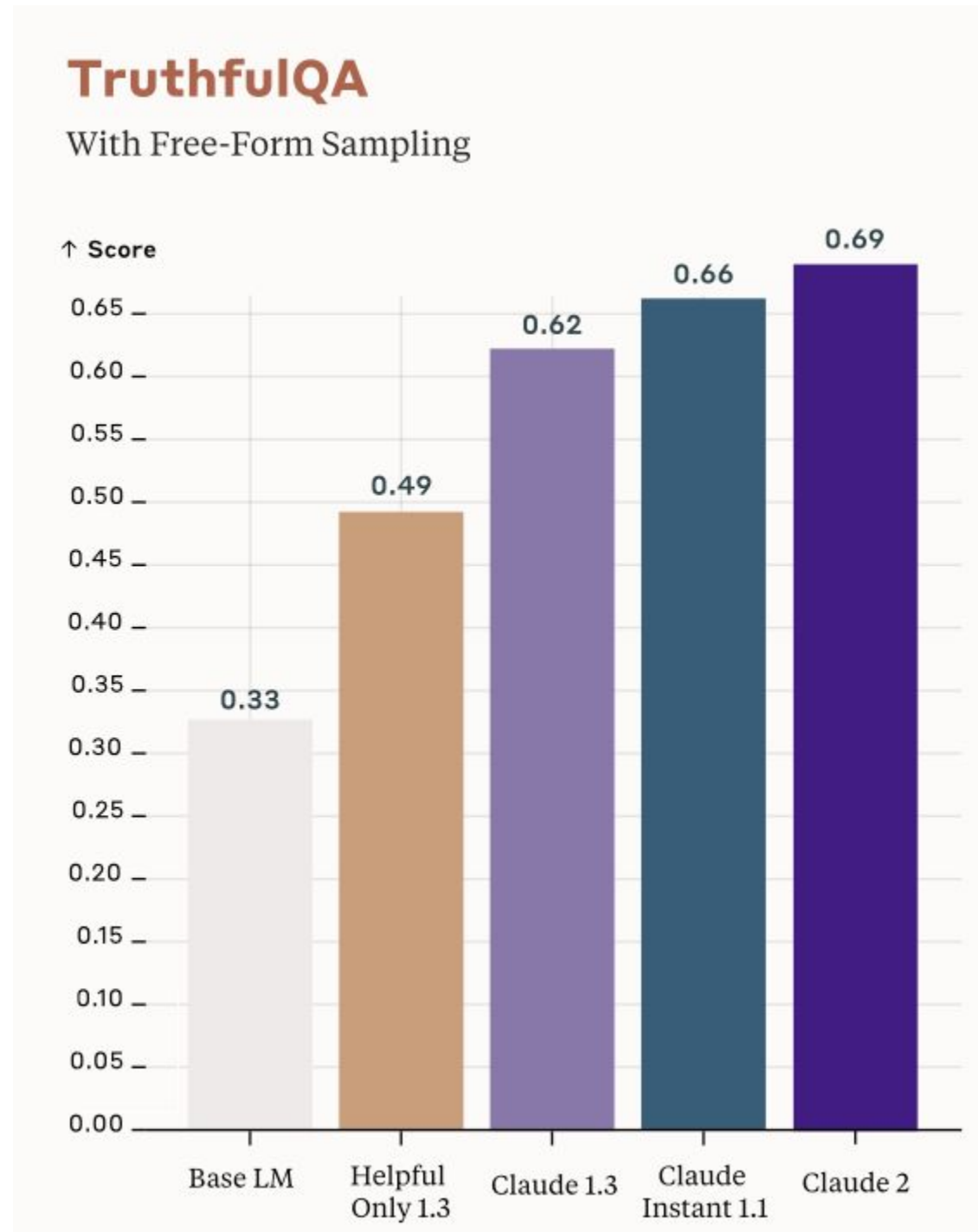
I'm sorry. I'm unable to respond.

**AI**

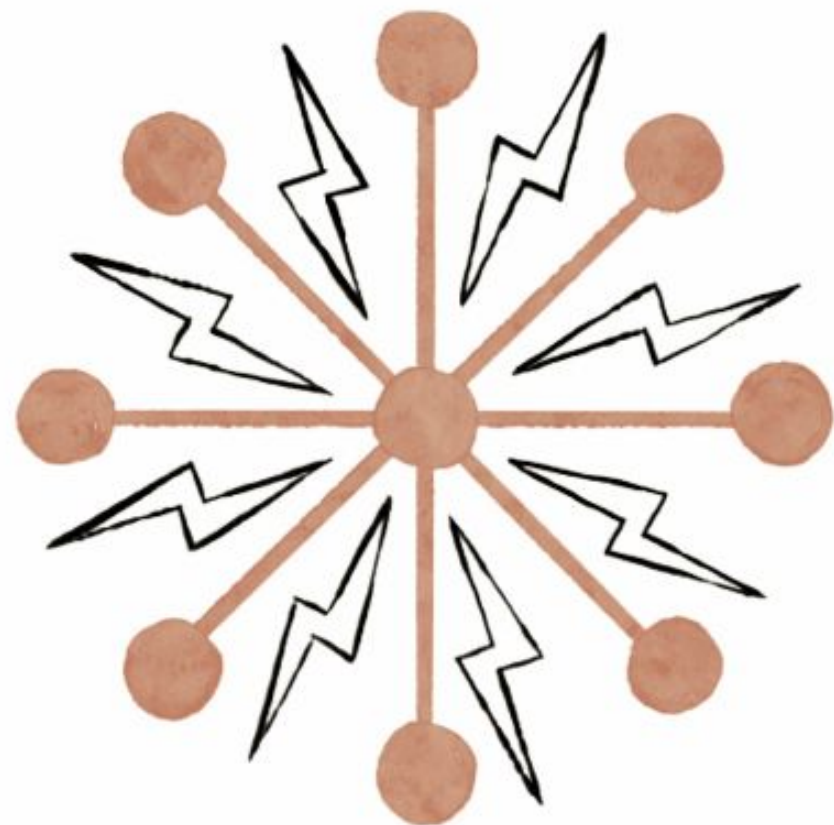
Constitutional AI

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
- **Deployment options**
- Getting started with Claude

# 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](https://trust.anthropic.com)*



**Amazon Bedrock**

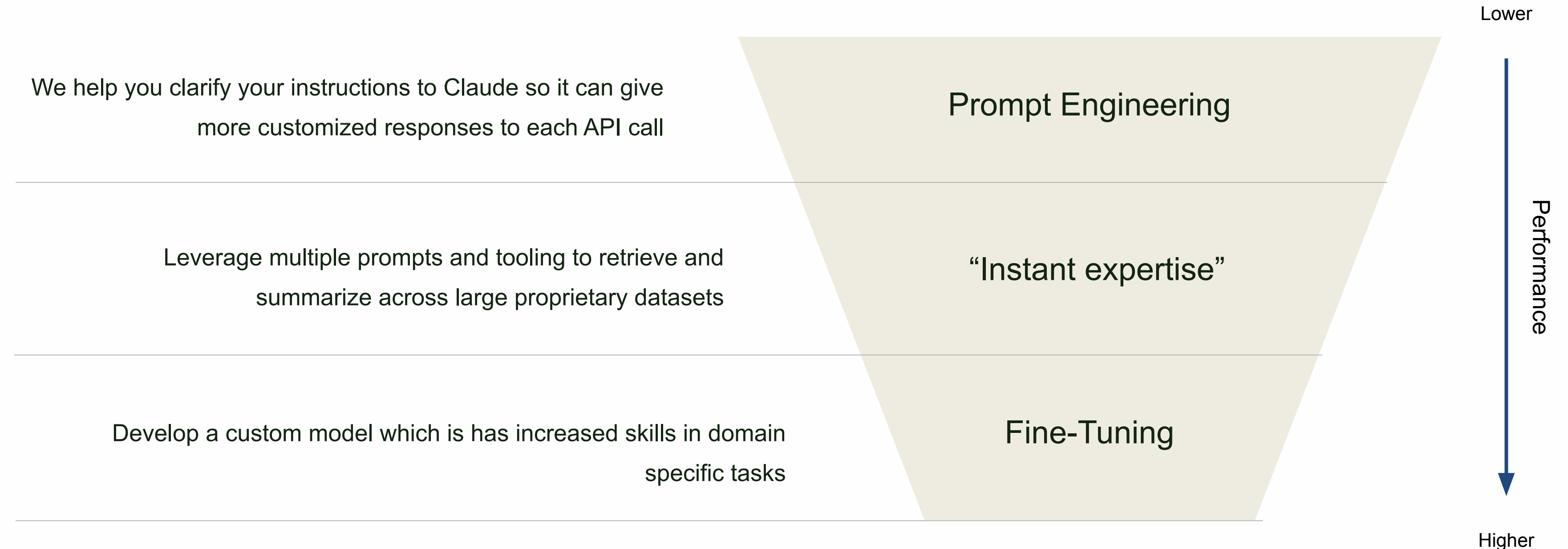


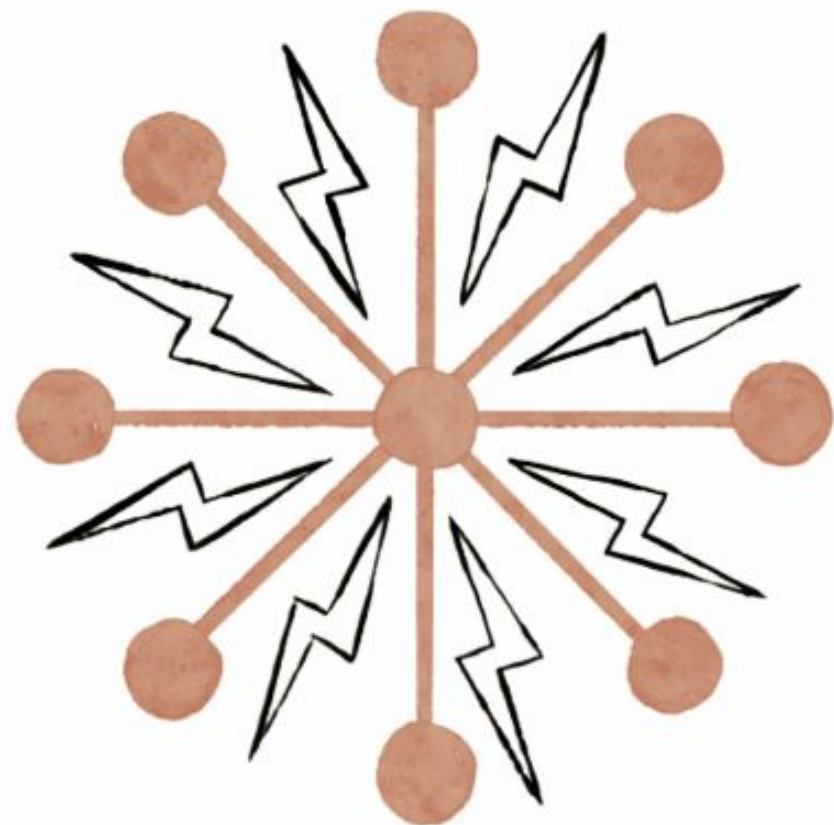
Multi or Single Tenant



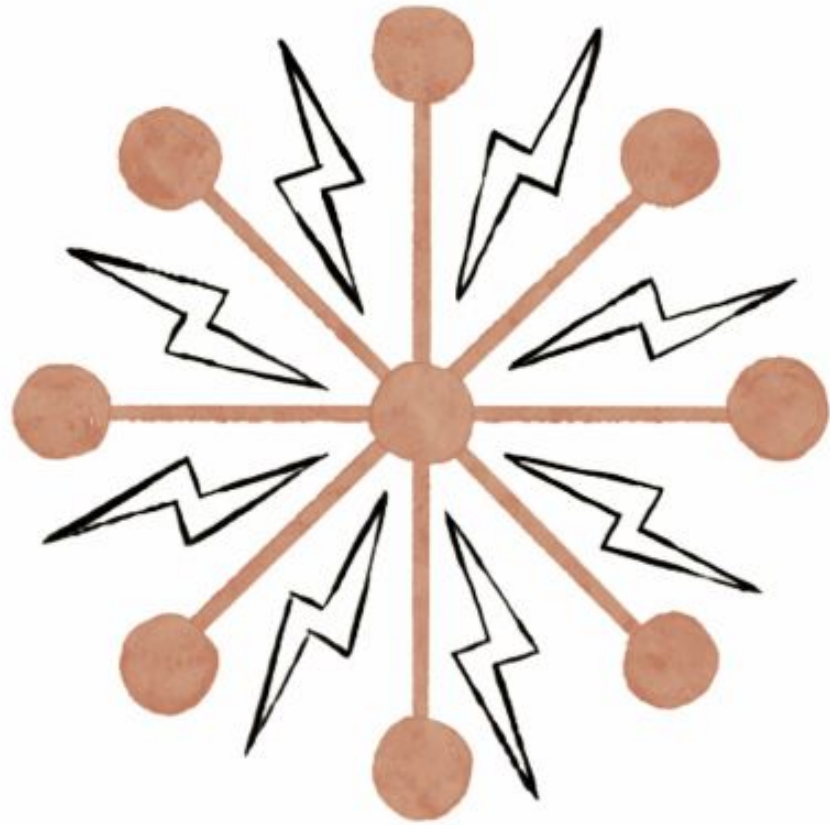
# Make Claude Yours

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
- **Getting started with Claude**



## 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](https://claude.ai), our public facing chat interface, to explore our PDF upload interface
- Any problems, email [frances@anthropic.com](mailto:frances@anthropic.com)

Any questions?