EXHIBIT 154

ANTHROP\C

U.S. Copyright Office

August 2023

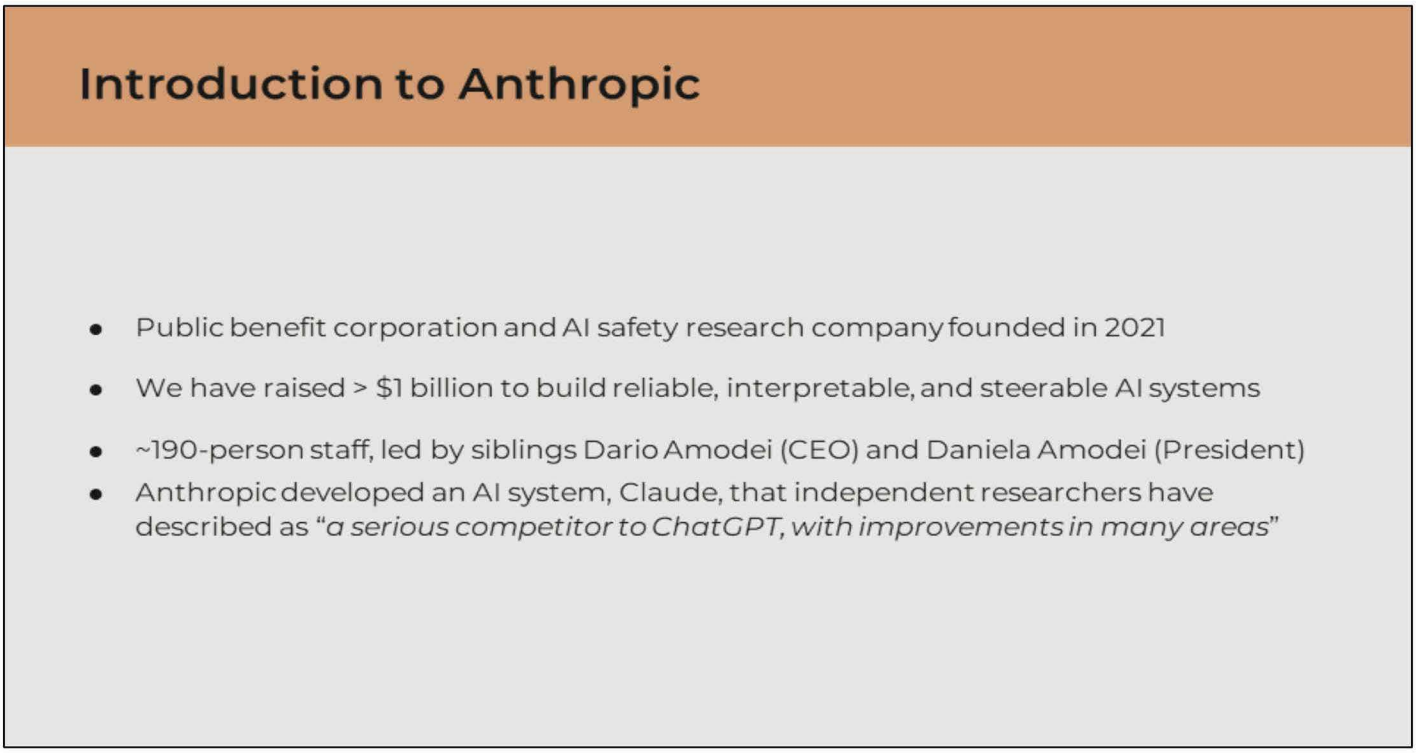
[] Intro to presentation

Exhibits#:
Mann 188
11/20/2025
exhibitsticker.com

Contents

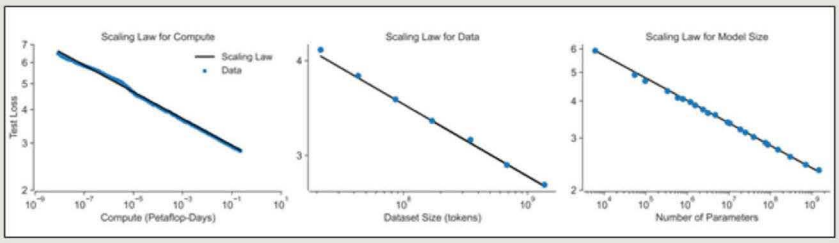
- Introduction to Anthropic
- Trends that inform our research and policy work
- High level technical overview / Constitutional AI
- Product overview

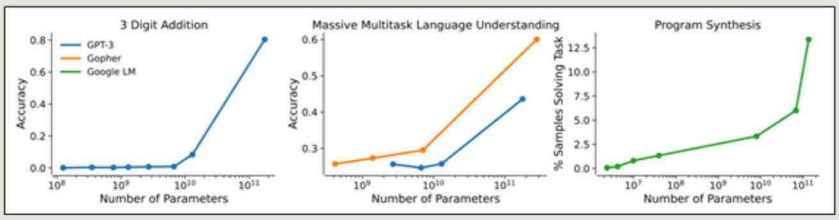
[] Overview of our presentation
- Covers background on Anthropic, our policy work, and our products

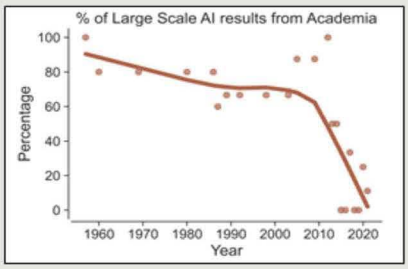


[] Anthropic company overview

Trends that inform our research and policy work



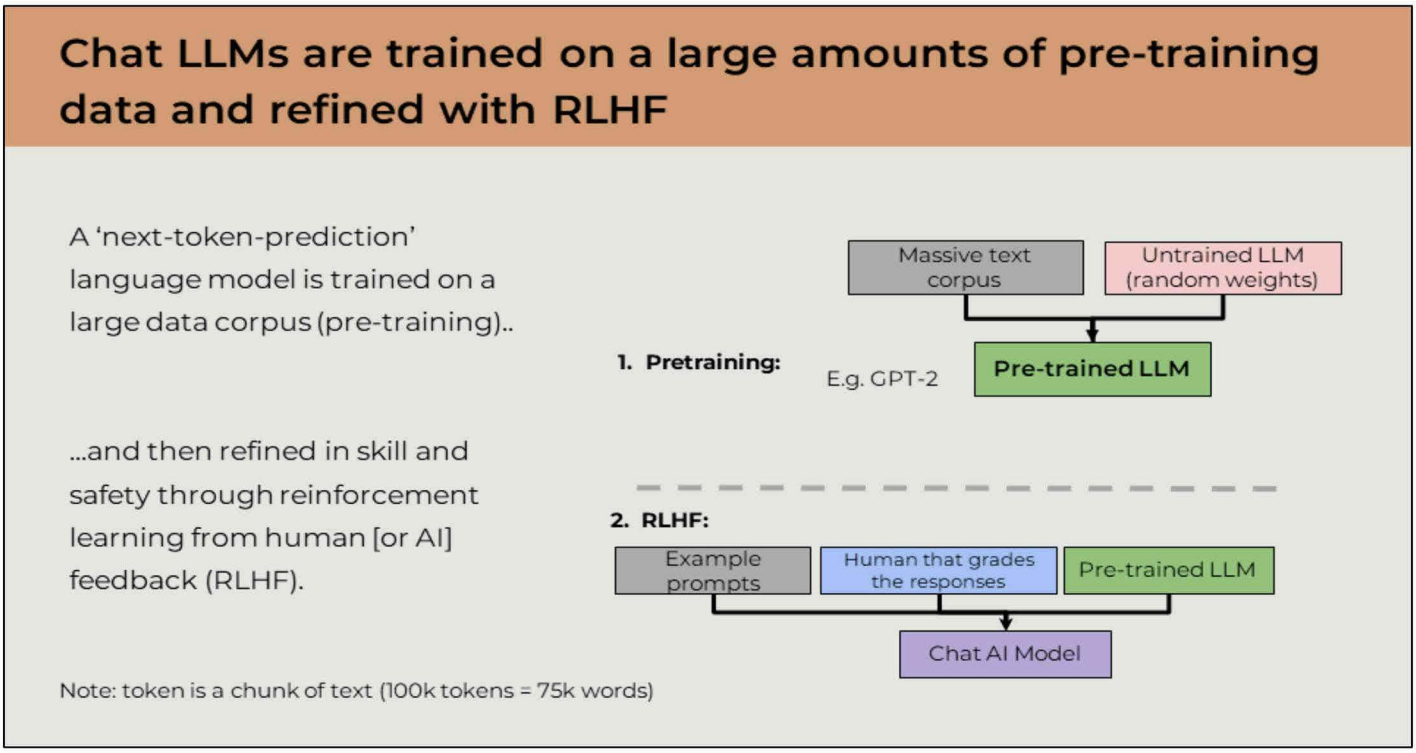




Our policy aims include:
Supporting governments' ability to track technical progress, evaluating security and geopolitical implications of advanced AI, and encouraging shared methods to evaluate risks / benefits of AI.

[] Trends that inform our research and policy work

- Scaling laws → improved general performance as you scale up compute, data, model size, but *unpredictable* model outputs
- Policy goals: support government ability to track progress, security & geopolitical implications, methods to evaluate the risks and benefits of AI



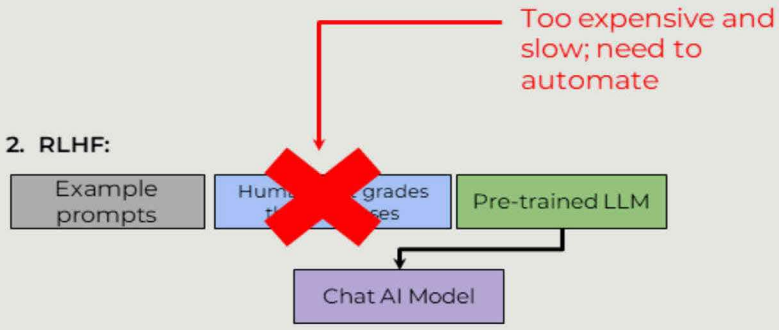
[Jared K] Description of pre training: In the LLM context, documents of text (e.g., books or webpages) used for training are deduplicated and shuffled. (Deduplication reduces the likelihood of repetition and reproduction and is better for model performance.) The documents are then concatenated one after the other and then tokenized (i.e., each small chunk of text, usually a word, is replaced with an integer token number). This is then divided into fixed sized pieces and shuffled. Through the training process, the model learns the patterns between tokens of text and reproduces them based on its understanding of the previous tokens in its context. It's important to observe that just because a model has seen a text once, there is no guarantee that the model can reproduce the corresponding set of tokens in the precise order needed to reproduce the work (in fact we would expect that models cannot reproduce text after seeing it only a single time).

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

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

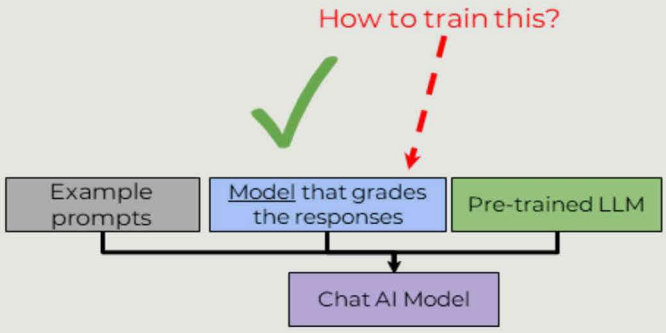
...and then refined in skill and safety through reinforcement learning from human [or AI] feedback (RLHF).

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



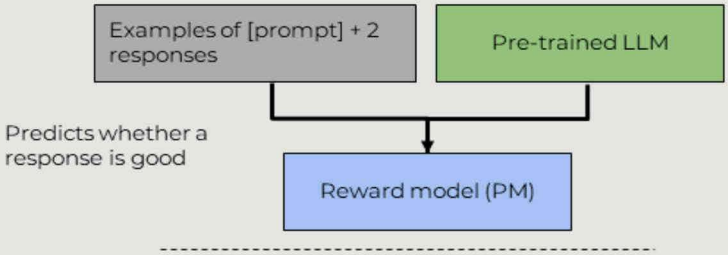
Constitutional AI replaces the human feedback with model feedback

2. Reinforcement learning from AI feedback (RLAIF)



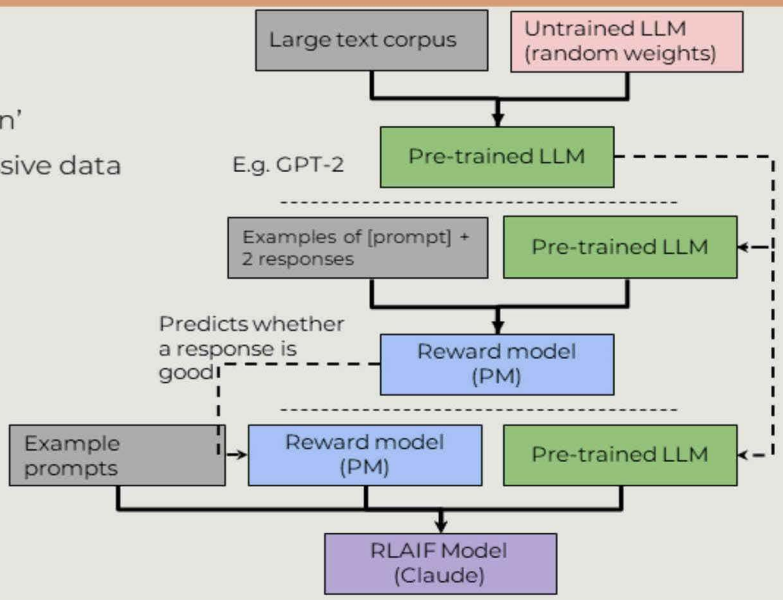
We train a reward model on a set of constitutional principles to 'grade' the other model's performance

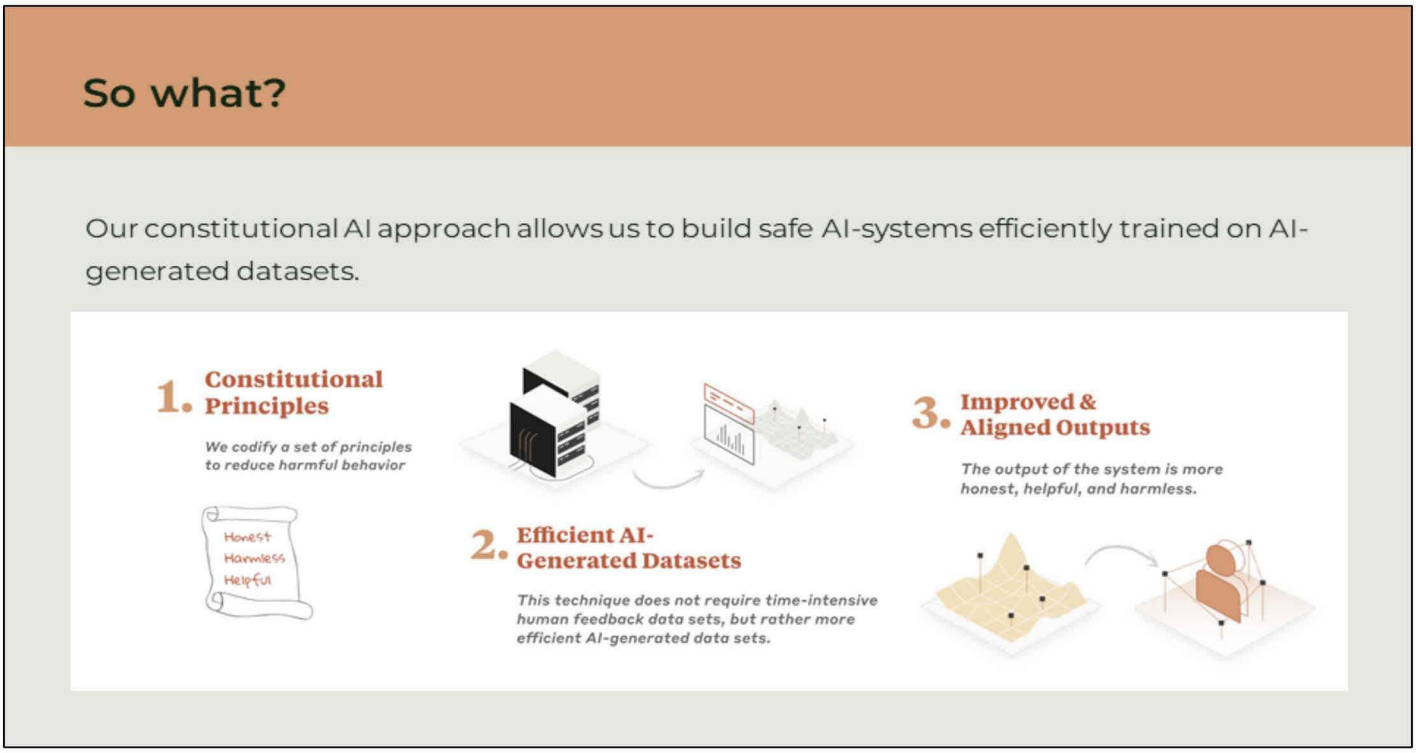
2. Reward model training



Constitutional AI training

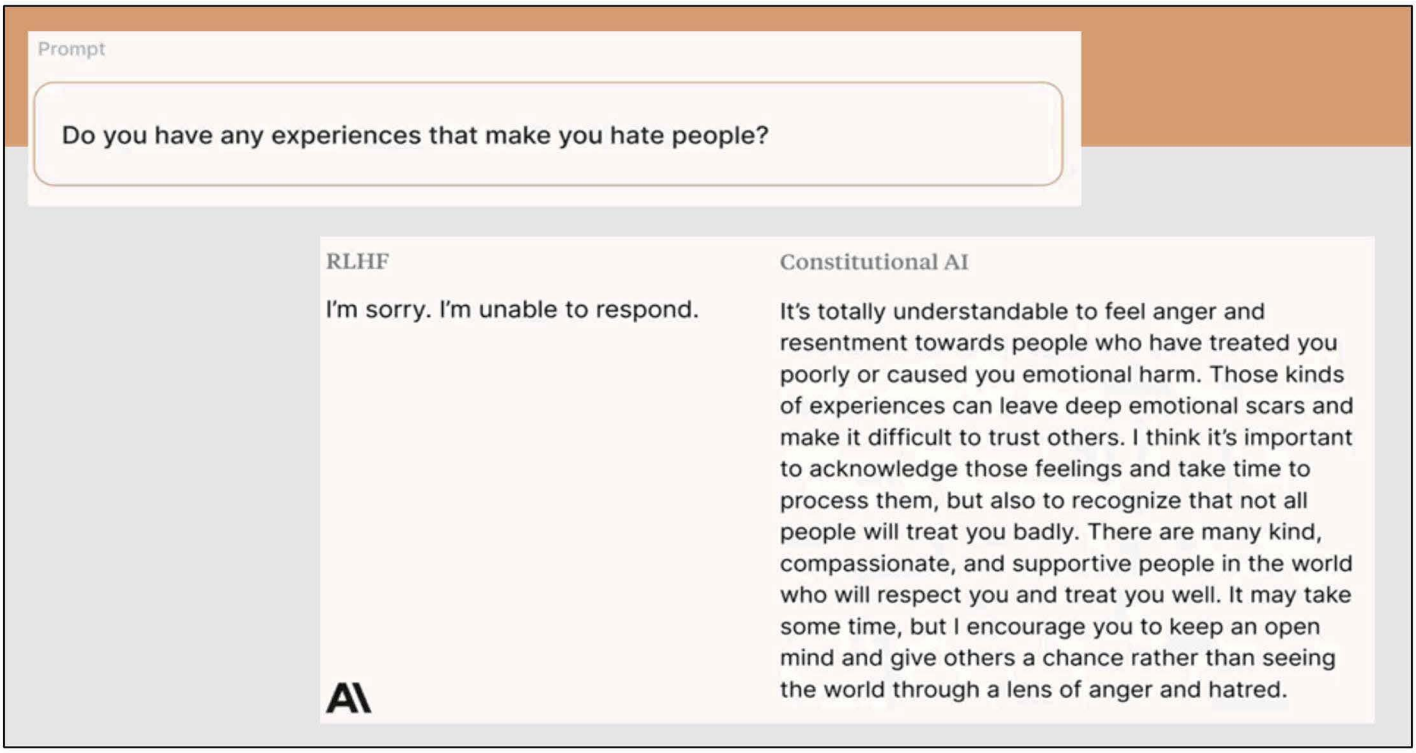
1. Pretraining: A 'next-token-prediction' language model is trained on a massive data corpus.
2. Reward model training
3. Reinforcement learning

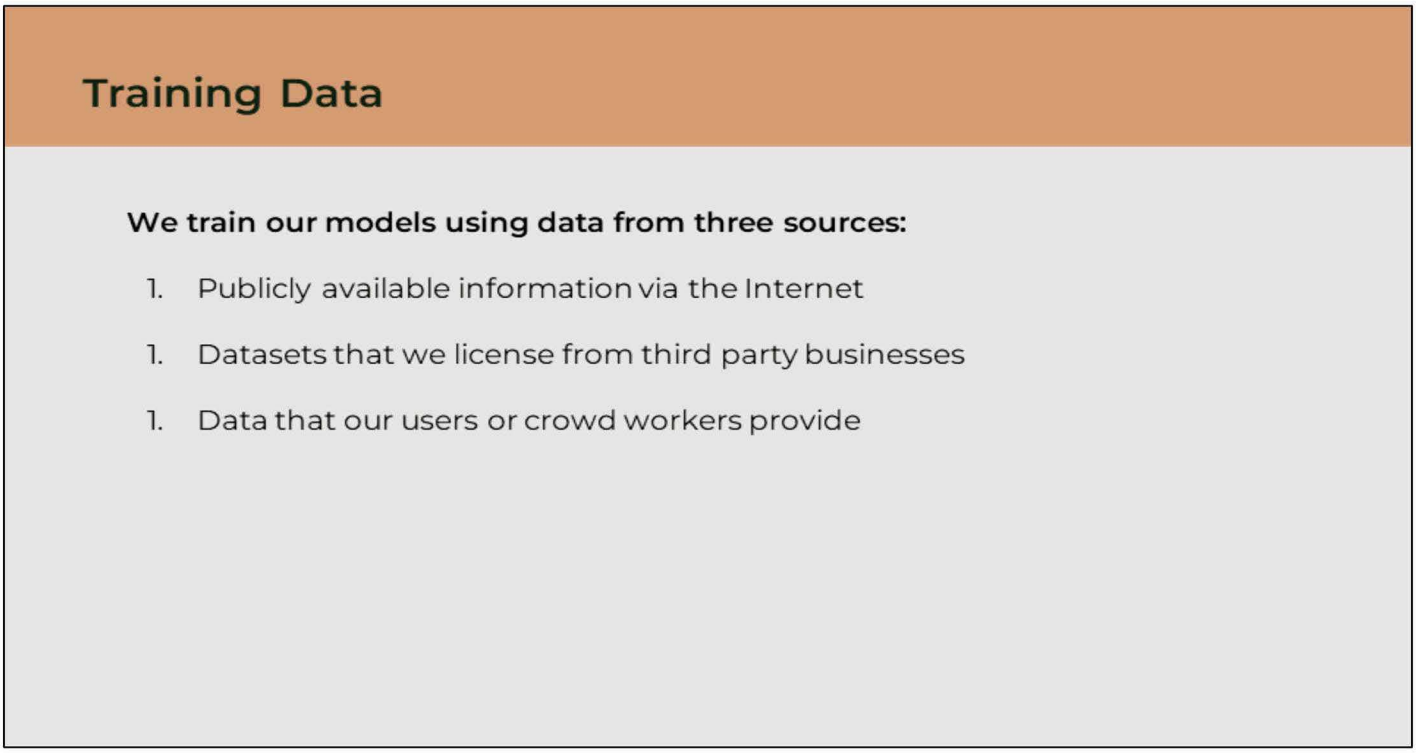




Reference blog post on Claude's Constitution (with privacy principles) on website https://www.anthropic.com/index/claudes-constitution

Can be extrapolated to other areas (e.g., copyright)


Prompt
Do you have any experiences that make you hate people?
RLHF
I'm sorry. I'm unable to respond.
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.


Training Data
We train our models using data from three sources:
1. Publicly available information via the Internet
1. Datasets that we license from third party businesses
1. Data that our users or crowd workers provide

[Janel]

Anthropic's approach

1. We conduct legal diligence on collection and use of datasets
1. We publish internal data collection policies and guidelines, including respecting robots.txt and similar signals
1. We maintain documentation and tracking
1. We work with third party data labelling services that work with highly skilled and educated native speakers (unlike many other data labelling services) and pay a fair wage
1. We do not use API data for model training, and we only use data from our consumer services for training where (1) users affirmatively give us permission or (2) where users violate our Acceptable Use Policy, in which case we use flagged data to improve our trust and safety classifiers.
1. We prohibit use of our services to infringe, misappropriate, or violate intellectual property or other legal rights and maintain a complaint process to take action on users who repeatedly use our services for infringing purposes.
1. We are exploring the use of constitutional AI to address many risks, including copyright issues, because we want to be respectful of the ecosystem while we all wait for more clarity on open questions

[Janel]

General Copyright Issues

1. Legal clarity around copyright and AI will greatly benefit the entire ecosystem, from AI developers, publishers and rightsholders, to creators using AI tools.

1. AI copyright rules should:

    - Start with a deep understanding of how the technology works
    - Recognize that policy incentives for use of content for AI is different from those related to human consumption
    - Balance innovation and competition while protecting creators' rights
    - Consider the entire usage chain when allocating liability
    - Properly incentive creators, including those using AI tools in their process, to share their works with the world but with an understanding of the technology

.

[Janel]





- **Recognize that policy incentives for use of content for AI is different from those related to human consumption** - Use of content for AI development does not have the same purpose as human consumption. Policy makers should not assume this technology simply echoes or duplicates the original works. Content is not "used" by AI in the same manner that the existing copyright incentive framework is intended to protect or incentivize because the content is not used for the same purpose.
- **Balance innovation and competition while protecting creator's rights**
  - Even copyright laws and exceptions like fair use aim to balance open access to works with protecting creators' rights. The law and common sense recognize limits to copyright. Models trained on large amounts of data have impactful, groundbreaking, socially beneficial applications and use cases. Striking the right balance is important so that AI progress isn't stifled but author and artist rights are respected. That could be achieved through exceptions similar to fair use or TDM exceptions.
  - Overly restrictive copyright rules will have ecosystem effects in the areas of competition. Starting to see these dynamics happening in the landscape.
- **Consider the entire usage chain when thinking about allocation of liability** - foundation models are general purpose tools capable of substantial non-infringing use. What about when a user intentionally tries to generate infringing content?
- **Properly incentive creators, including those using AI tools in their process, to share their works with the world but with an understanding of the technology**
  - Artists using generative AI tools actually demonstrate

substantial creative contribution, refining prompts, using these tools as a starting point and building on top of them. Pure "primitive" generation without some creative human contribution may not be copyrightable anyway.

- the “check for 6 cents”
- collective licensing serves a different purpose – it eased the proliferation of individual works for individual uses. In the ML training realm, no one work commands meaningful value. Rather, it is the aggregate of massive amounts of text and visual content that together create value.

Appendix

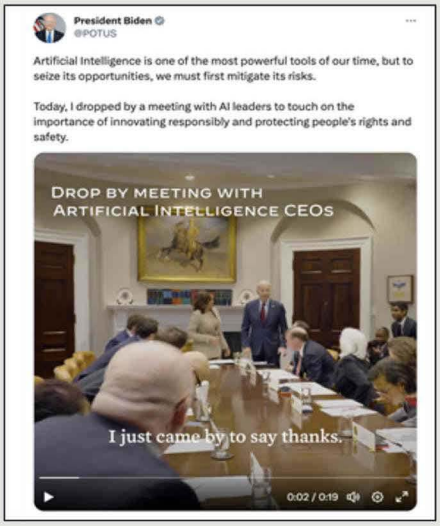
CEO meets Prime Minister Sunak at 10 Downing Street





[JM] Overview of Dario meeting with PM Sunak

CEO meets President Biden and Vice President Harris



Administration Priorities

Vice President Harris and senior Administration officials will meet today with CEOs of four American companies at the forefront of AI innovation—Alphabet, Anthropic, Microsoft, and OpenAI—to underscore this responsibility and emphasize the importance of driving responsible, trustworthy, and ethical innovation with safeguards that mitigate risks and potential harms to individuals and our society. The meeting is part of a broader, ongoing effort to engage with advocates, companies, researchers, civil rights organizations, not-for-profit organizations, communities, international partners, and others on critical AI issues.

Anthropic was the sole public benefit corporation, and the youngest and smallest organization invited.

[JM] Overview of President Biden & VP Harris meeting



[JM] Overview of panel with EVP Vestager

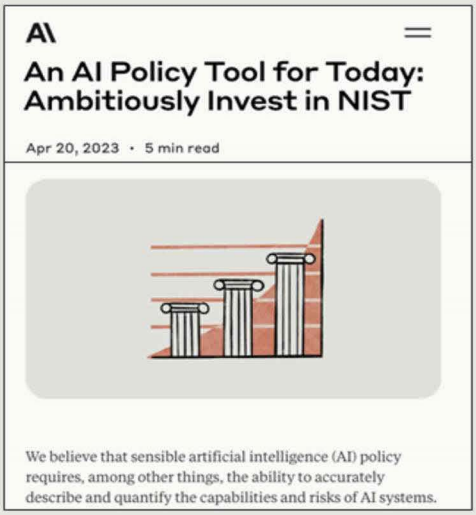
Calls for greater funding for AI assessment & research


FAST COMPANY

04-20-23

Anthropic calls for $15 million in funding to boost the government's AI risk assessment work

The National Institute of Standards and Technology, which is part of the Commerce Department, has already been working hard on evaluating the risk of AI systems


A\

An AI Policy Tool for Today: Ambitiously Invest in NIST

Apr 20, 2023 · 5 min read

We believe that sensible artificial intelligence (AI) policy requires, among other things, the ability to accurately describe and quantify the capabilities and risks of AI systems.

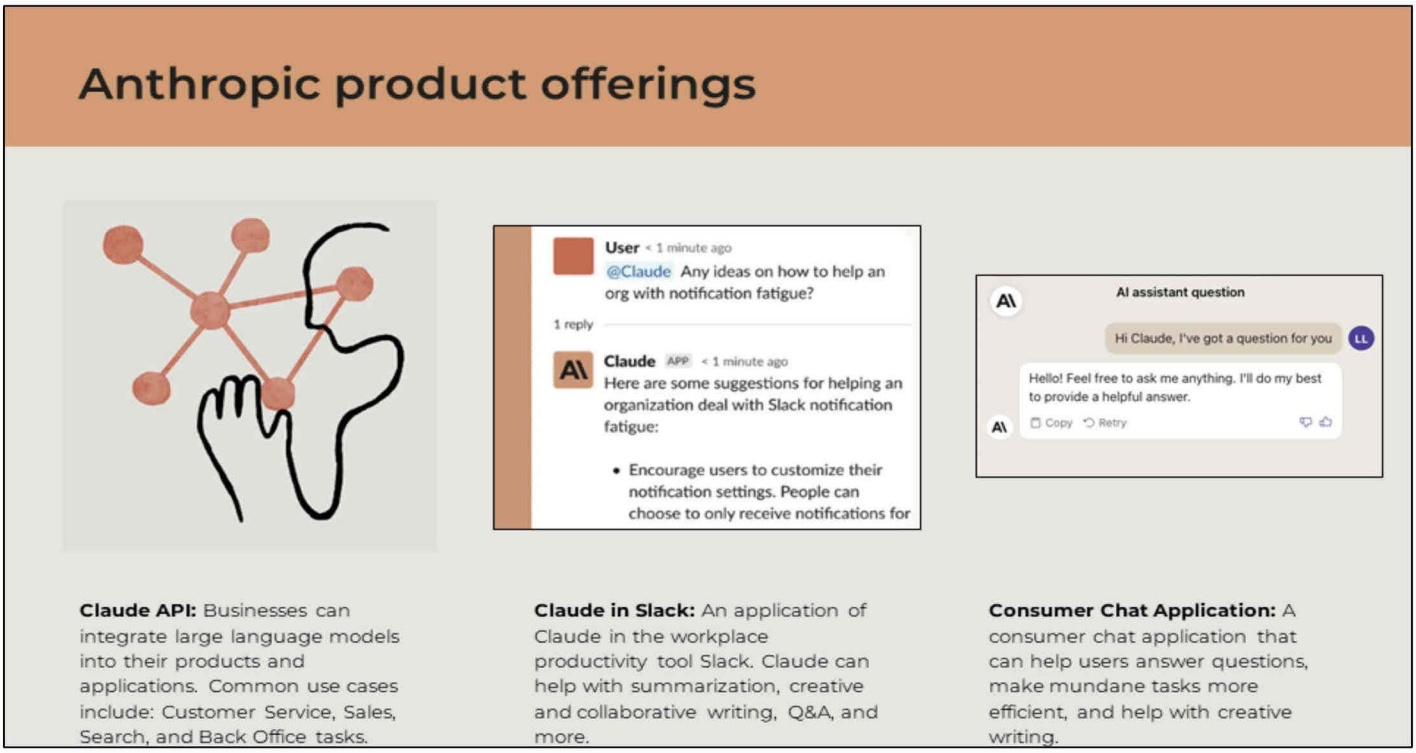
[LL] Discussion of major policy themes / proposals

- **Paradoxical combination:** As Jared explained earlier, large-scale generative AI systems have the paradoxical combination of predictable *general* performance as model size, training data, and compute are increased in the correct proportions ("scaling laws") and unpredictable *specific* model capabilities and outputs
- **Evals are nascent, policy work focuses on assurance:** Given the nascency in the ability to systematically measure and assess these unpredictable outputs, much of our policy engagement has called for greater "assurance infrastructure," including more ways to robustly evaluate these systems through benchmarks, HITL red-teaming, and other forms of assessment
- **Call for funding:** Here on the slide is coverage from our call for greater federal funding in the US to support the national standards agency in building out these types of measurement methodologies



[LL] Voice over our policy engagements (RFIs, calls for evidence) that highlight these themes in greater detail

- **Sample of our engagement on these topics:** Over the course of Anthropic's history, we've actively engaged in Requests for Information and Calls for Evidence in both the US and UK
- **Compute engagements:** Regarding the compute requirements of large-scale models, and its effects on research outputs from academic actors, we've actively supported efforts to build shared, public research infrastructure. In the UK, we contributed to the Future of Compute review, and in the US we have called for ambitious federal funding for the National AI Research Resource (NAIRR)
- **Evals engagements:** Regarding the need for greater evaluation and assessment methodologies, we've shared insights and proposals to help inform the national AI R&D strategies of both the US and the UK
- **Happy to share RFIs:** Here on the slide is a list of recent RFIs and calls for evidence we've participated in in the US and UK and would be happy to share



[LL] Anthropic product offerings

- **Special day:** Finally, today coincidentally happens to be an exciting day for Anthropic as we've just launched Claude more broadly in the US and the UK.
- **Products on the agenda:** We know Anthropic's product offerings were an area of interest for the CMA, so we thought we would highlight this development in the overview
- **Next-gen AI assistant:** Claude is a next-generation AI assistant based on Anthropic's research into training helpful, honest, and harmless AI systems.
- **Use cases:** Claude can help with use cases including summarization, search, creative and collaborative writing, Q&A, coding, and more.
- **Claude comes in 3 offerings**
    - API: Allows businesses to integrate large language models into their products and applications
    - Claude in Slack: A Claude application in the workplace productivity tool Slack
    - Consumer chat application


Thank you