

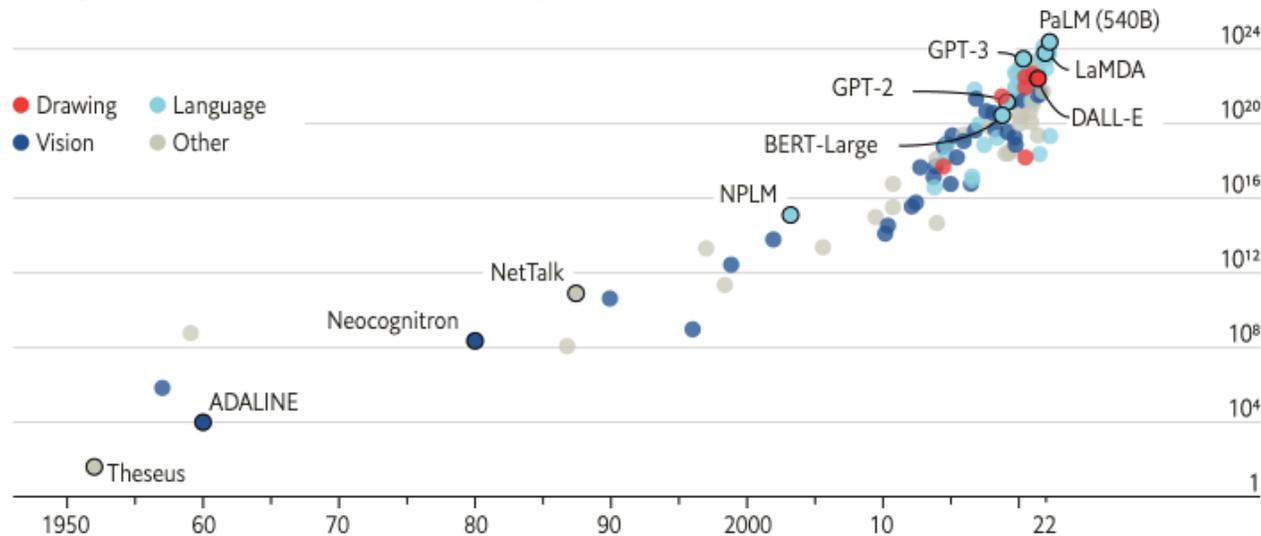
RL for LLMs: Preference  
Optimization  
CS 224R

Archit Sharma  
*(Based on slides in CS224N)*

# Larger and larger models

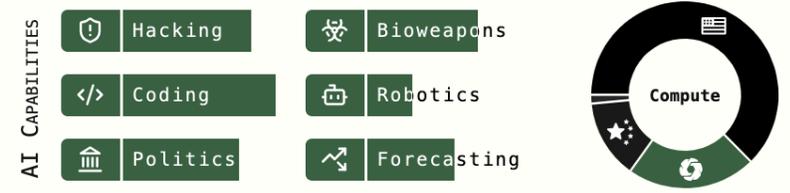
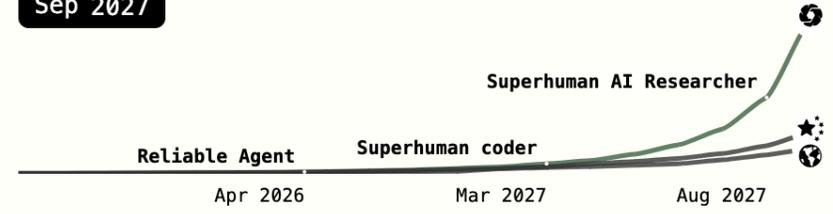
## The blessings of scale

AI training runs, estimated computing resources used  
Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

Sep 2027



Approval	Revenue	Valuation	Importance	Datcenters	Timeline
-38%	\$173B/yr	\$8T	16%	\$880B/yr	2032



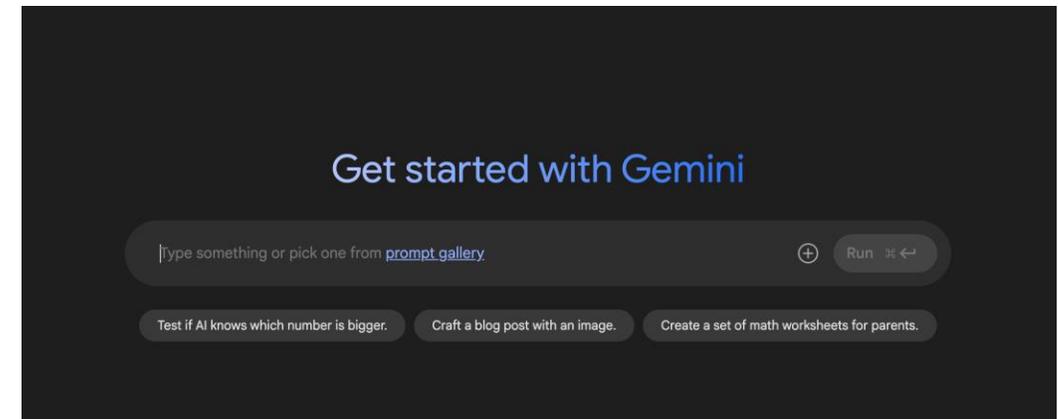
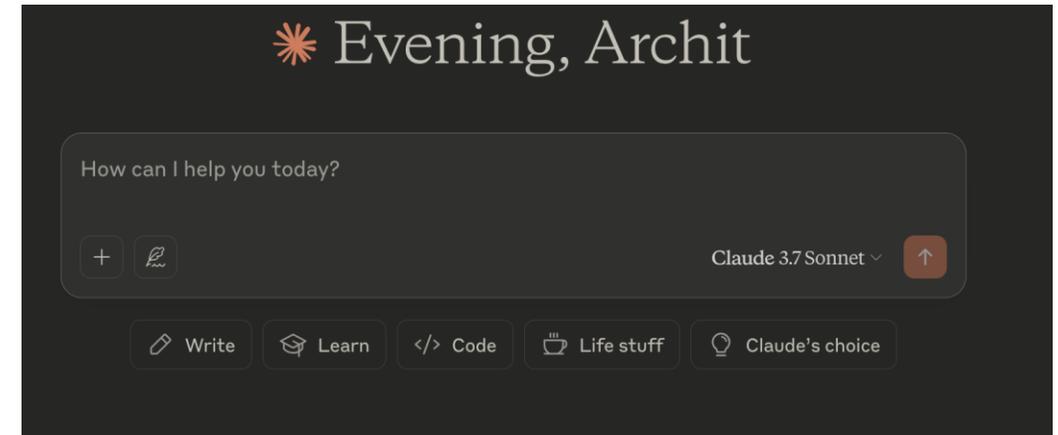
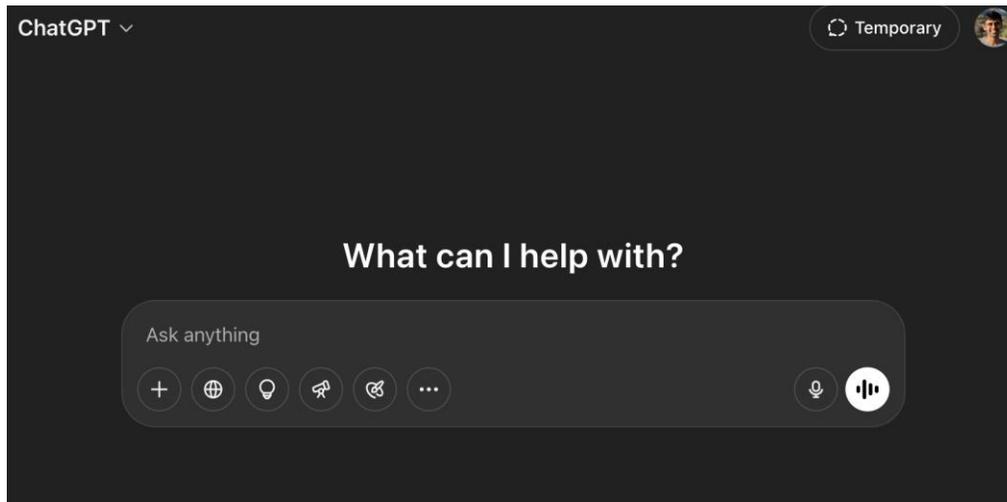
<https://www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress>  
<https://ai-2027.com/>

# Language models as assistants?

How do we get from *this*

*Stanford University is located in \_\_\_\_\_*

to *this*?



# Recap: What kinds of things does pretraining learn?

- *Stanford University is located in \_\_\_\_\_, California.* [Trivia]
- *I put \_\_\_ fork down on the table.* [syntax]
- *The woman walked across the street, checking for traffic over \_\_\_ shoulder.* [coreference]
- *I went to the ocean to see the fish, turtles, seals, and \_\_\_\_\_.* [lexical semantics/topic]
- *Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was \_\_\_\_.* [sentiment]
- *Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the \_\_\_\_\_.* [some reasoning – this is harder]
- *I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, \_\_\_\_\_* [some basic arithmetic; they don't learn the Fibonacci sequence]

# Language models as world models?

Language models may do rudimentary modeling of *agents*, *beliefs*, and *actions*:

*Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, who is a physicist, predicts that the bowling ball and the leaf will fall at the same rate.*

Changing the last sentence of the prompt, we get:

*...Pat, who has never seen this demonstration before, predicts that the bowling ball will fall to the ground first. This is incorrect. In a vacuum chamber, there is no air*

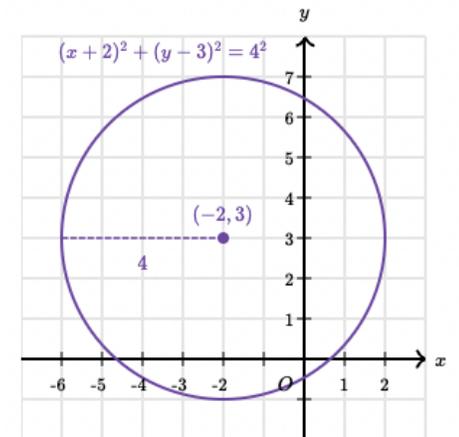
Language Models as Agent Models [[Andreas, 2022](#)]

# Language models as world models?

...*math*:

We can describe circles in the  $xy$ -plane using equations in terms of  $x$  and  $y$ . Circle equations questions require us to understand the connection between these equations and the features of circles.

For example, the equation  $(x + 2)^2 + (y - 3)^2 = 4^2$  is graphed in the  $xy$ -plane below. It is a circle with a center at  $(-2, 3)$  and a radius of 4.



<https://www.khanacademy.org/test-prep/sat/x0a8c2e5f:untitled-652>

# Language models as world models?

...code:

```
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8     const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9         method: "POST",
10        body: `text=${text}`,
11        headers: {
12            "Content-Type": "application/x-www-form-urlencoded",
13        },
14    });
15    const json = await response.json();
16    return json.label === "pos";
17 }
```

<https://github.com/features/copilot>

# Language models as world models?

...*medicine*:

## **Rapid and chronic ethanol tolerance are composed of distinct memory-like states in *Drosophila***

### **Abstract**

Ethanol tolerance is the first type of behavioral plasticity and neural plasticity that is induced by ethanol intake, and yet its molecular and circuit bases remain largely unexplored. Here, we characterize three distinct forms of ethanol tolerance in male *Drosophila*: rapid, chronic, and repeated. Rapid tolerance is composed of two short-lived memory-like states, one that is labile and one that is consolidated. Chronic tolerance, induced by continuous exposure, lasts for two days, induces ethanol preference, and hinders the development of rapid tolerance through the activity of

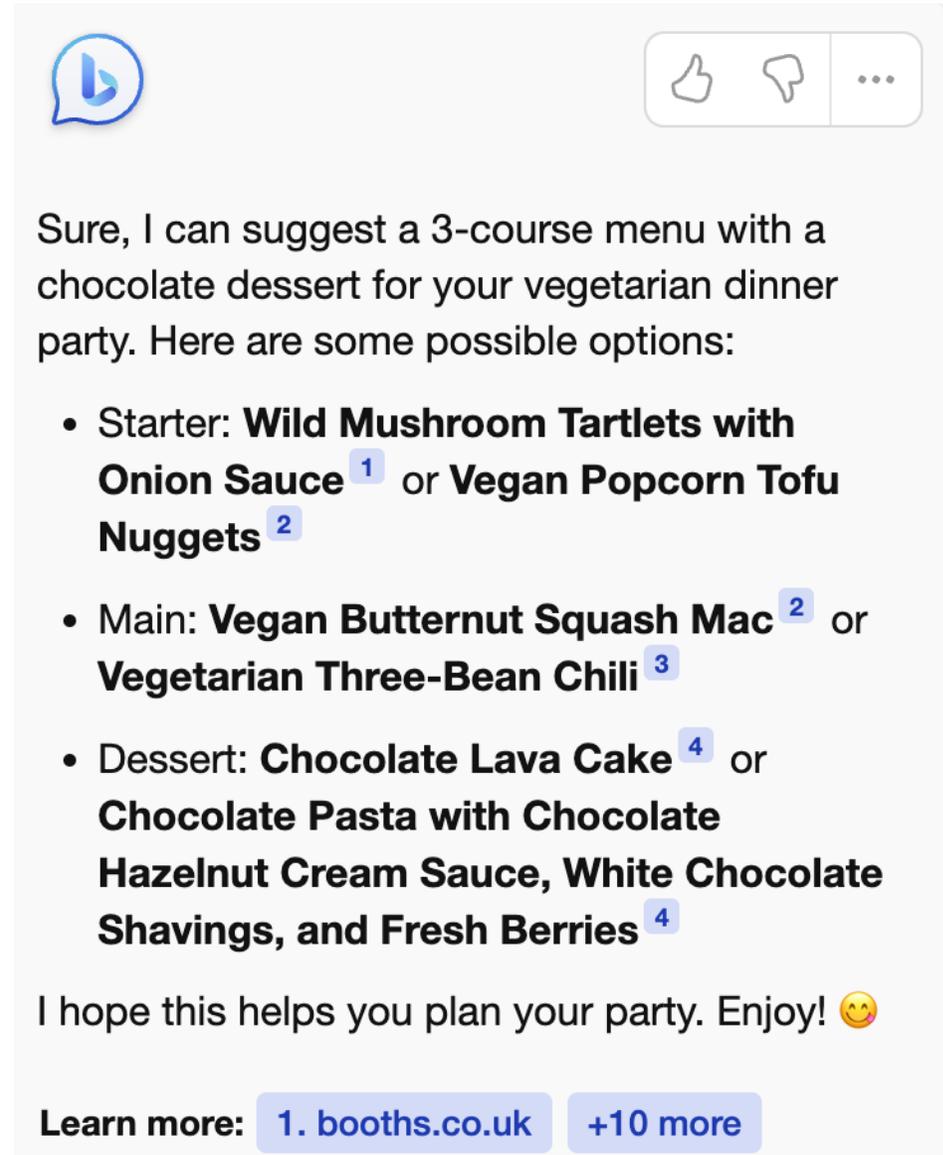
[[Larnerd, 2023](#)]

# Language models as multitask assistants?

**I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?**

[[Microsoft Bing](#)]

(Also see OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude)



The screenshot shows a chat interface with a blue speech bubble icon on the left and thumbs up/down and a menu icon on the right. The text of the chat response is as follows:

Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: **Wild Mushroom Tartlets with Onion Sauce**<sup>1</sup> or **Vegan Popcorn Tofu Nuggets**<sup>2</sup>
- Main: **Vegan Butternut Squash Mac**<sup>2</sup> or **Vegetarian Three-Bean Chili**<sup>3</sup>
- Dessert: **Chocolate Lava Cake**<sup>4</sup> or **Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries**<sup>4</sup>

I hope this helps you plan your party. Enjoy! 😊

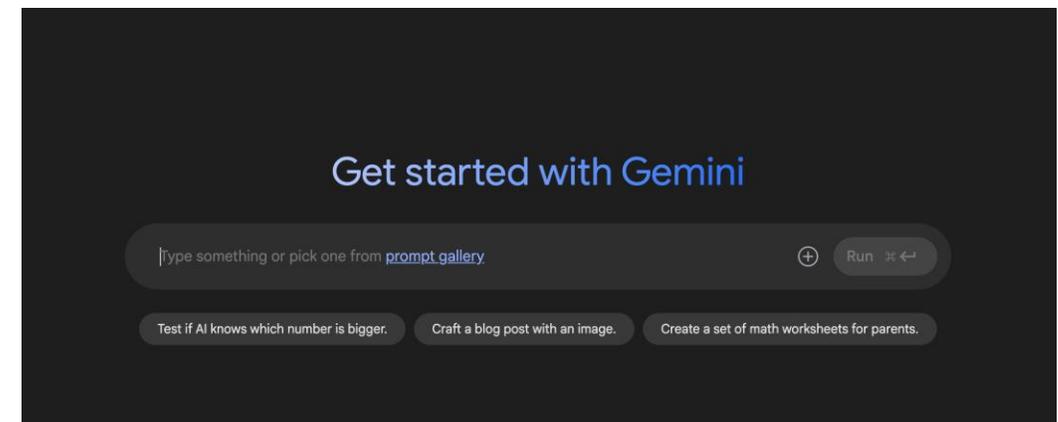
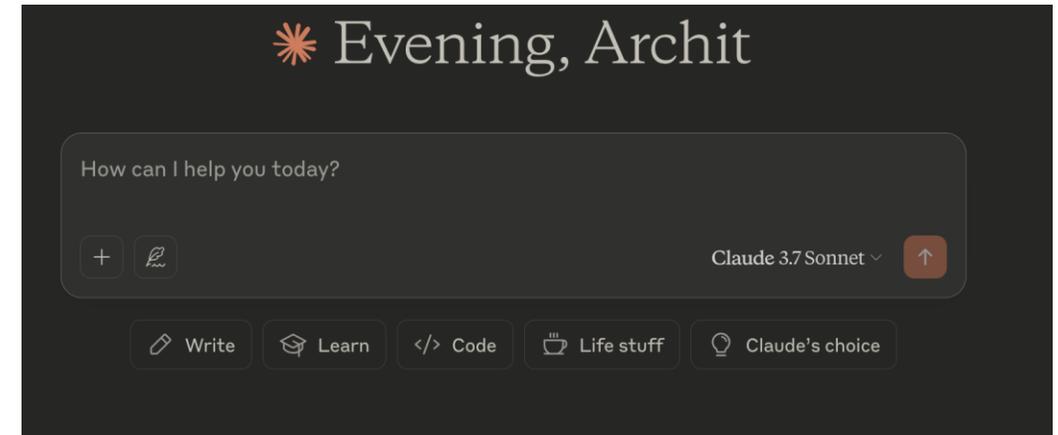
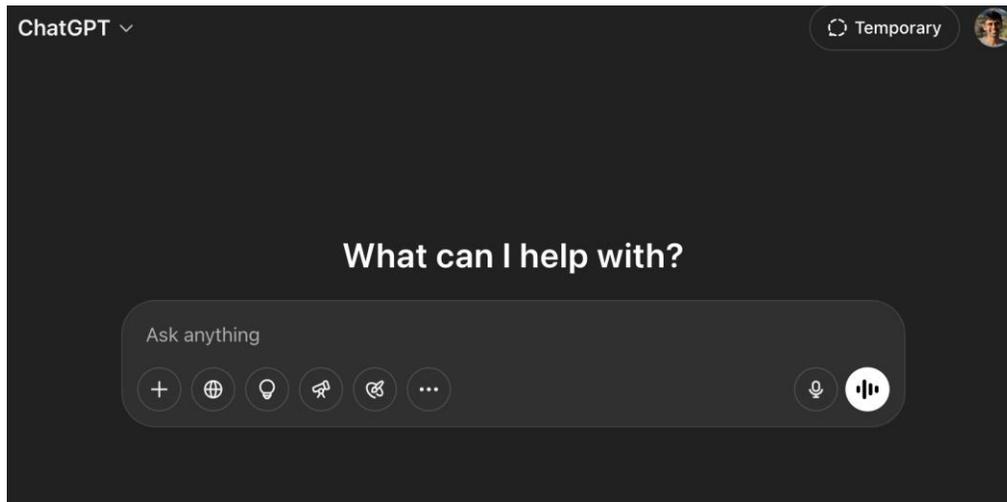
Learn more: [1. booths.co.uk](#) [+10 more](#)

# Language models as assistants?

How do we get from *this*

*Stanford University is located in \_\_\_\_\_*

to *this*?



# Lecture Plan: From Language Models to Assistants

- 1. Instruction finetuning**
- 2. Optimizing for human preferences (DPO/RLHF)**
- 3. What's next?**

# Instruction Finetuning

# Language modeling $\neq$ assisting users

**PROMPT** *Explain the moon landing to a 6 year old in a few sentences.*

**COMPLETION**

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)].

# Language modeling $\neq$ assisting users

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION **Human**

A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone.

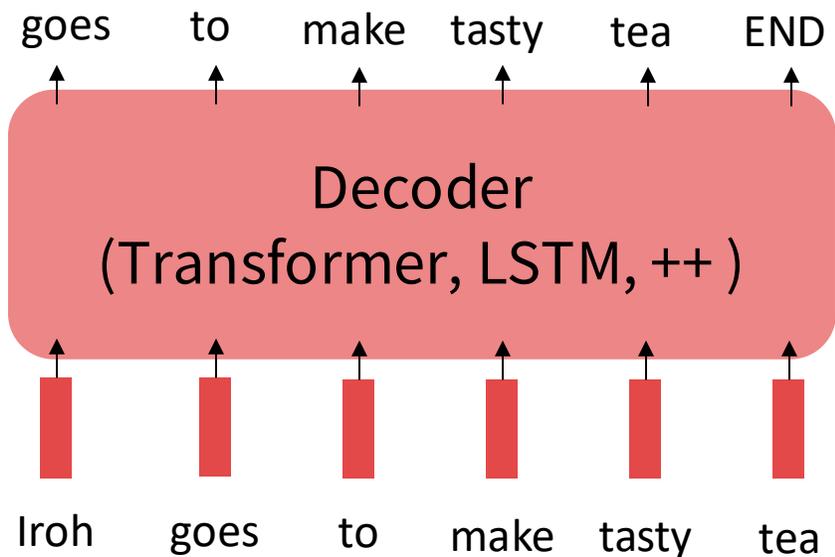
Language models are not *aligned* with user intent [[Ouyang et al., 2022](#)].  
Finetuning to the rescue!

# The Pretraining / Finetuning Paradigm

Pretraining can improve NLP applications by serving as parameter initialization.

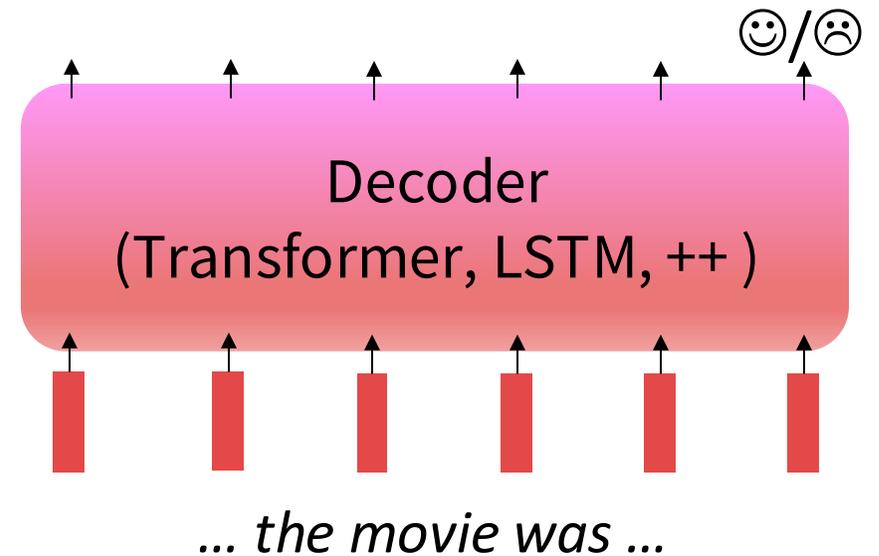
## Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



## Step 2: Finetune (on your task)

Not many labels; adapt to the task!

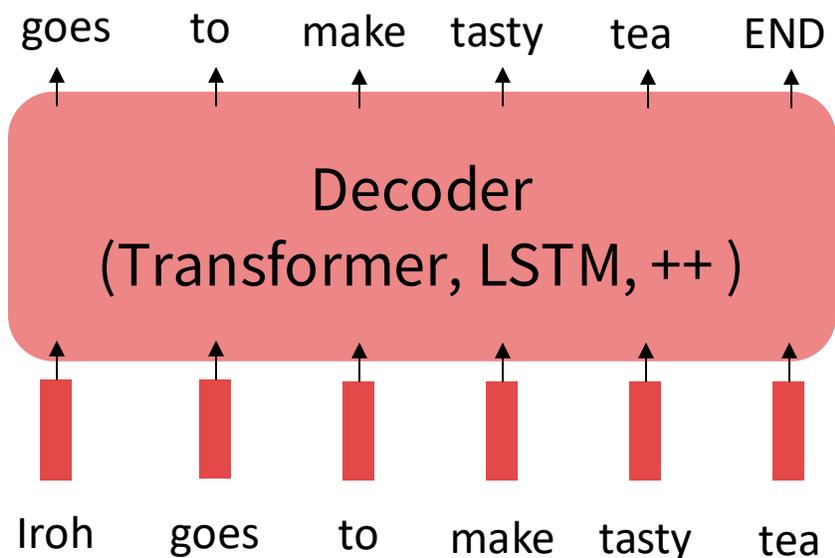


# Scaling up finetuning

Pretraining can improve NLP applications by serving as parameter initialization.

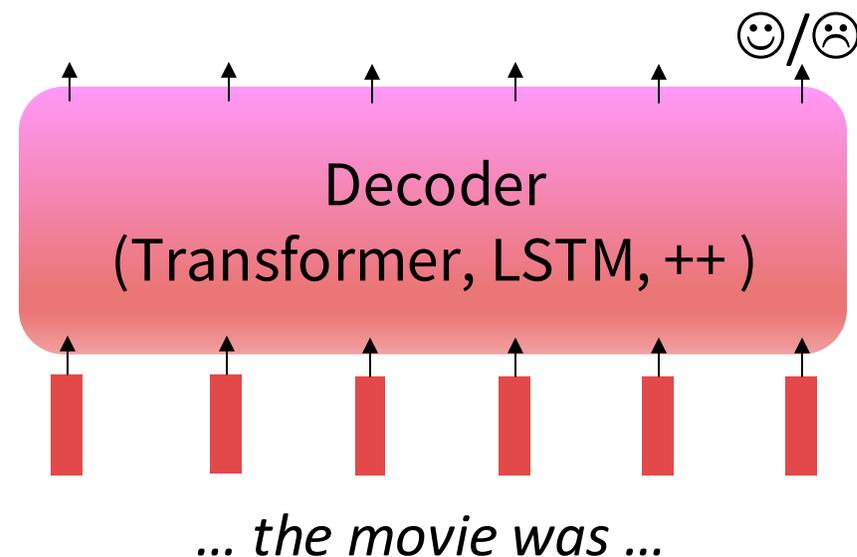
## Step 1: Pretrain (on language modeling)

Lots of text; learn general things!



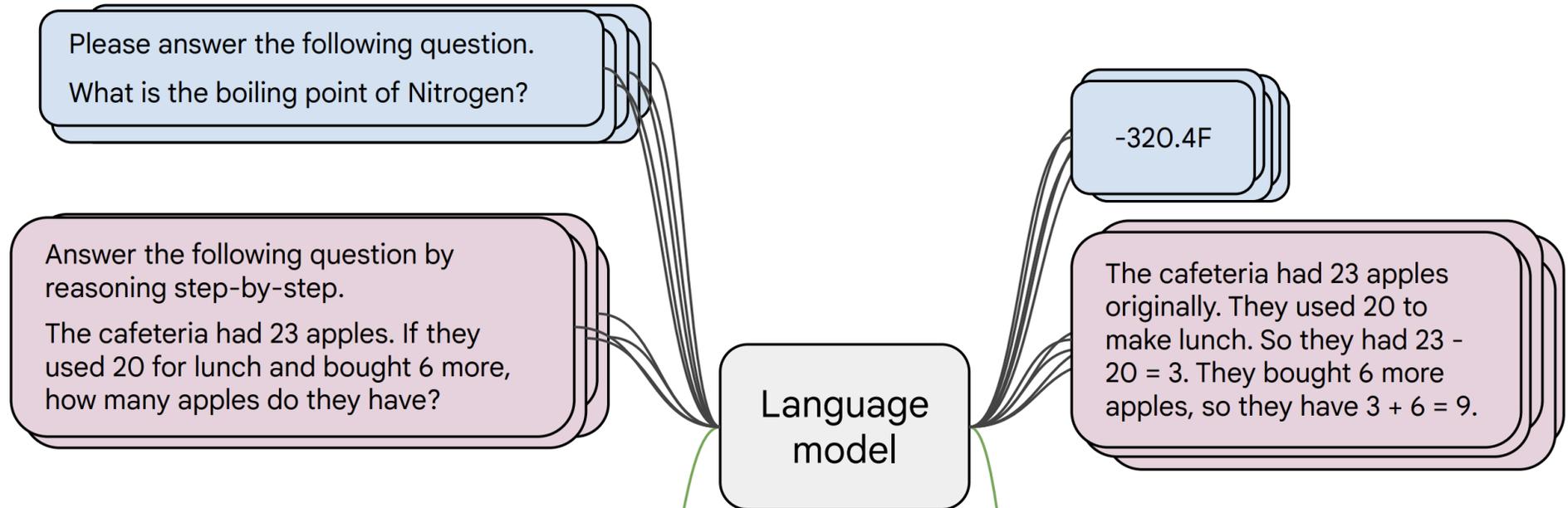
## Step 2: Finetune (on **many tasks**)

~~Not~~ many labels; adapt to the tasks!



# Instruction finetuning

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



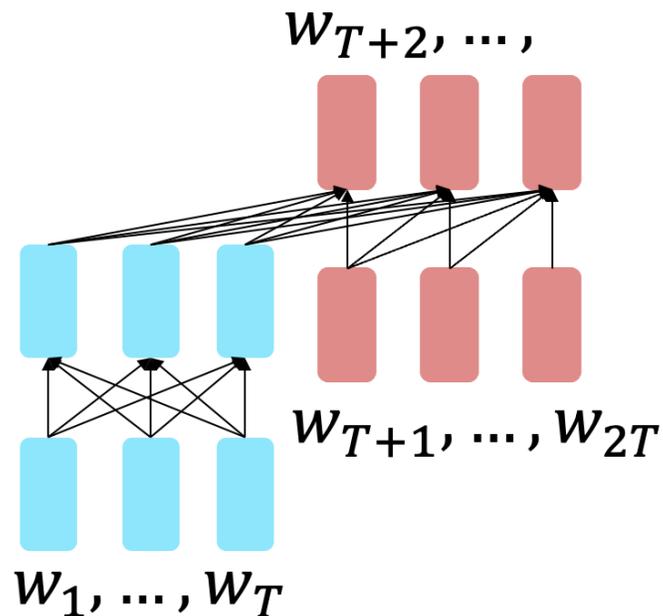
- Evaluate on **unseen tasks**

Q: Can Geoffrey Hinton have a conversation with George Washington?  
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

# Instruction finetuning

- Recall the T5 encoder-decoder model from lecture 10 [[Raffel et al., 2018](#)], pretrained on the **span corruption** task
- Flan-T5** [[Chung et al., 2020](#)]: T5 models finetuned on 1.8K additional tasks



Params	Model	BIG-bench + MMLU avg (normalized)
80M	T5-Small	-9.2
	Flan-T5-Small	-3.1 (+6.1)
250M	T5-Base	-5.1
	Flan-T5-Base	6.5 (+11.6)
780M	T5-Large	-5.0
	Flan-T5-Large	13.8 (+18.8)
3B	T5-XL	-4.1
	Flan-T5-XL	19.1 (+23.2)
11B	T5-XXL	-2.9
	Flan-T5-XXL	23.7 (+26.6)

**Bigger model = bigger  $\Delta$**

[[Chung et al., 2022](#)]

# Instruction finetuning

## Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

## Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✘ (doesn't answer question)

Highly recommend trying FLAN-T5 out to get a sense of its capabilities:

<https://huggingface.co/google/flan-t5-xxl>

[[Chung et al., 2022](#)]

# Instruction finetuning

## Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

## After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

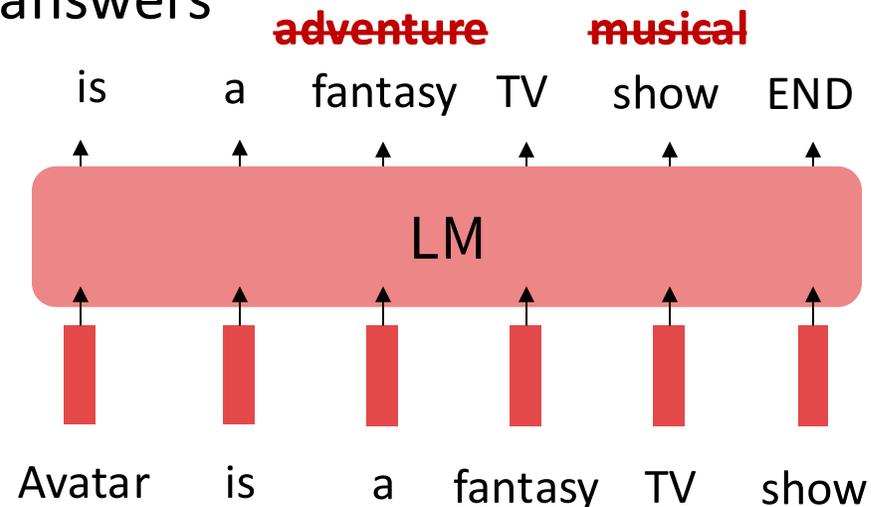
Highly recommend trying FLAN-T5 out to get a sense of its capabilities:

<https://huggingface.co/google/flan-t5-xxl>

[[Chung et al., 2022](#)]

# Limitations of instruction finetuning?

- One limitation of instruction finetuning is obvious: it's **expensive** to collect ground-truth data for tasks. Can you think of other subtler limitations?
- **Problem 1:** tasks like open-ended creative generation have no right answer.
  - *Write me a story about a dog and her pet grasshopper.*
- **Problem 2:** language modeling penalizes all token-level mistakes equally, but some errors are worse than others.
- **Problem 3:** humans generate suboptimal answers
- Even with instruction finetuning, there is a mismatch between the LM objective and the objective of “satisfy human preferences”!
- Can we **explicitly attempt to satisfy human preferences?**



# Reinforcement Learning from Human Preferences

# Optimizing for human preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For an instruction  $x$  and a LM sample  $y$ , imagine we had a way to obtain a *human reward* of that summary:  $R(x, y) \in \mathbb{R}$ , higher is better.

SAN FRANCISCO,  
California (CNN) --  
A magnitude 4.2  
earthquake shook the  
San Francisco  
...  
overturn unstable  
objects.

$x$

An earthquake hit  
San Francisco.  
There was minor  
property damage,  
but no injuries.

$$R(x, y_1) = 8.0$$

The Bay Area has  
good weather but is  
prone to  
earthquakes and  
wildfires.

$$R(x, y_2) = 1.2$$

- Now we want to maximize the expected reward of samples from our LM:

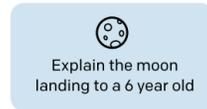
$$\mathbb{E}_{\hat{y} \sim p_{\theta}(y | x)} [R(x, \hat{y})]$$

# High-level instantiation: 'RLHF' pipeline

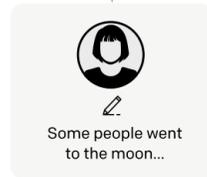
Step 1

**Collect demonstration data, and train a supervised policy.**

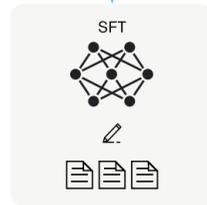
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

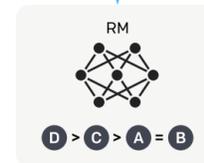
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



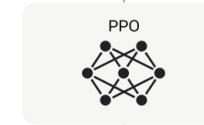
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

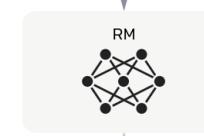
A new prompt is sampled from the dataset.



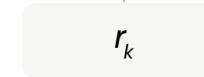
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



- First step: instruction tuning!
- Second + third steps: maximize reward (but how??)

# Optimizing for human preferences

- How do we actually change our LM parameters  $\theta$  to maximize this?

$$\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$$

- Let's try doing gradient ascent!

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{\hat{s} \sim p_{\theta_t}(s)}[R(\hat{s})]$$

How do we estimate  
this expectation??

What if our reward  
function is non-  
differentiable??

- **Policy gradient** methods in RL (e.g., REINFORCE; [[Williams, 1992](#)]) give us tools for estimating and optimizing this objective.
- We'll describe a *high-level mathematical* overview of the simplest policy gradient estimator, but a full treatment is outside the scope of this lecture.

# A brief recap of policy gradient/REINFORCE [Williams, 1992]

- We want to obtain (defn. of expectation) (linearity of gradient)

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \nabla_{\theta} \sum_s R(s) p_{\theta}(s) = \sum_s R(s) \nabla_{\theta} p_{\theta}(s)$$

- Here we'll use a very handy trick known as the **log-derivative trick**. Let's try taking the gradient of  $\log p_{\theta}(s)$

$$\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s) \quad \Rightarrow \quad \nabla_{\theta} p_{\theta}(s) = p_{\theta}(s) \nabla_{\theta} \log p_{\theta}(s)$$

(chain rule)

This is an  
expectation of this

- Plug back in:

$$\sum_s R(s) \nabla_{\theta} p_{\theta}(s) = \sum_s p_{\theta}(s) R(s) \nabla_{\theta} \log p_{\theta}(s)$$

$$= \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})]$$

# A brief recap of policy gradient/REINFORCE [Williams, 1992]

- Now we have put the gradient “inside” the expectation, we can approximate this objective with Monte Carlo samples:

$$\nabla_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s})] = \mathbb{E}_{\hat{s} \sim p_{\theta}(s)} [R(\hat{s}) \nabla_{\theta} \log p_{\theta}(\hat{s})] \approx \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$$

This is why it’s called “**reinforcement learning**”: we **reinforce** good actions, increasing the chance they happen again.

- Giving us the update rule:

$$\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta_t} \log p_{\theta_t}(s_i)$$

This is **heavily simplified!** There is a *lot* more needed to do RL w/ LMs. **Can you see any problems with this objective?**

If  $R$  is +++

Take gradient steps to maximize  $p_{\theta}(s_i)$

If  $R$  is ---

Take steps to minimize  $p_{\theta}(s_i)$

# How do we get the rewards? (Recall last lecture)

- **Problem 1:** human-in-the-loop is expensive!
  - **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate problem! [[Knox and Stone, 2009](#)]

An earthquake hit San Francisco. There was minor property damage, but no injuries.

The Bay Area has good weather but is prone to earthquakes and wildfires.

Train a  $RM_\phi(x, y)$  to predict human reward from an annotated dataset, then optimize for  $RM_\phi$  instead.

$$R(x, y_1) = 8.0$$


$$R(x, y_2) = 1.2$$


# How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

A 4.2 magnitude  
earthquake hit  
San Francisco,  
resulting in  
massive damage.

$$R(x, y_3) = \begin{matrix} y_3 \\ 4.1? & 6.6? & 3.2? \end{matrix}$$

# How do we model human preferences?

- **Problem 2:** human judgments are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015](#); [Clark et al., 2018](#)]

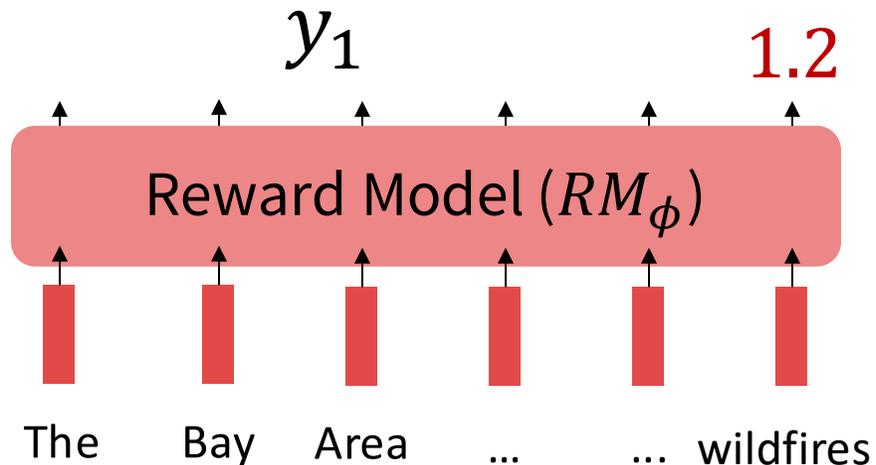
An earthquake hit San Francisco. There was minor property damage, but no injuries.

>

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

>

The Bay Area has good weather but is prone to earthquakes and wildfires.



$y_3$

Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(x, y^w, y^l) \sim D} [\log \sigma(RM_\phi(x, y^w) - RM_\phi(x, y^l))]$$

“winning”  
sample

“losing”  
sample

$y_2$

$y^w$  should score higher than  $y^l$

# RLHF: Optimizing the learned reward model

- We have the following:
  - A pretrained (possibly instruction-finetuned) LM  $p^{PT}(y | x)$
  - A reward model  $RM_{\phi}(x, y)$  that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
- Now to do RLHF:
  - Copy the model  $p_{\theta}^{RL}(y | x)$ , with parameters  $\theta$  we would like to optimize
  - We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y})]$$

# RLHF: Optimizing the learned reward model

- We want to optimize:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y})]$$

- Do you see any problems?
  - Learned rewards are imperfect; this quantity can be imperfectly optimized
- Add a penalty for drifting too far from the initialization:

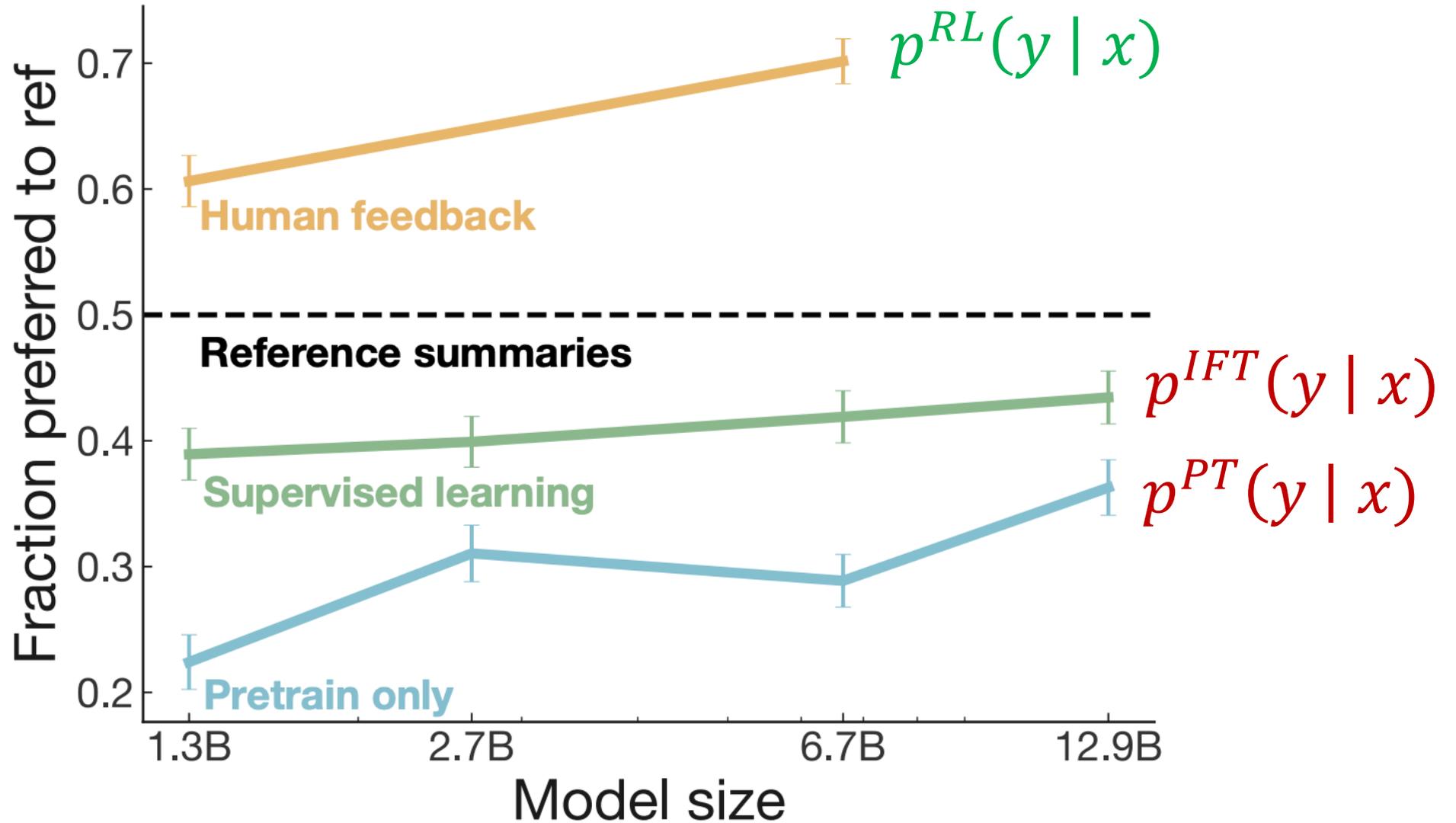
$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y}) - \underbrace{\beta \log \left( \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} \right)}_{\text{penalty}}]$$

Pay a price when

$$p_{\theta}^{RL}(\hat{y}|x) > p^{PT}(\hat{y}|x)$$

This penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler (KL) divergence** between  $p_{\theta}^{RL}(\hat{y}|x)$  and  $p^{PT}(\hat{y}|x)$ .

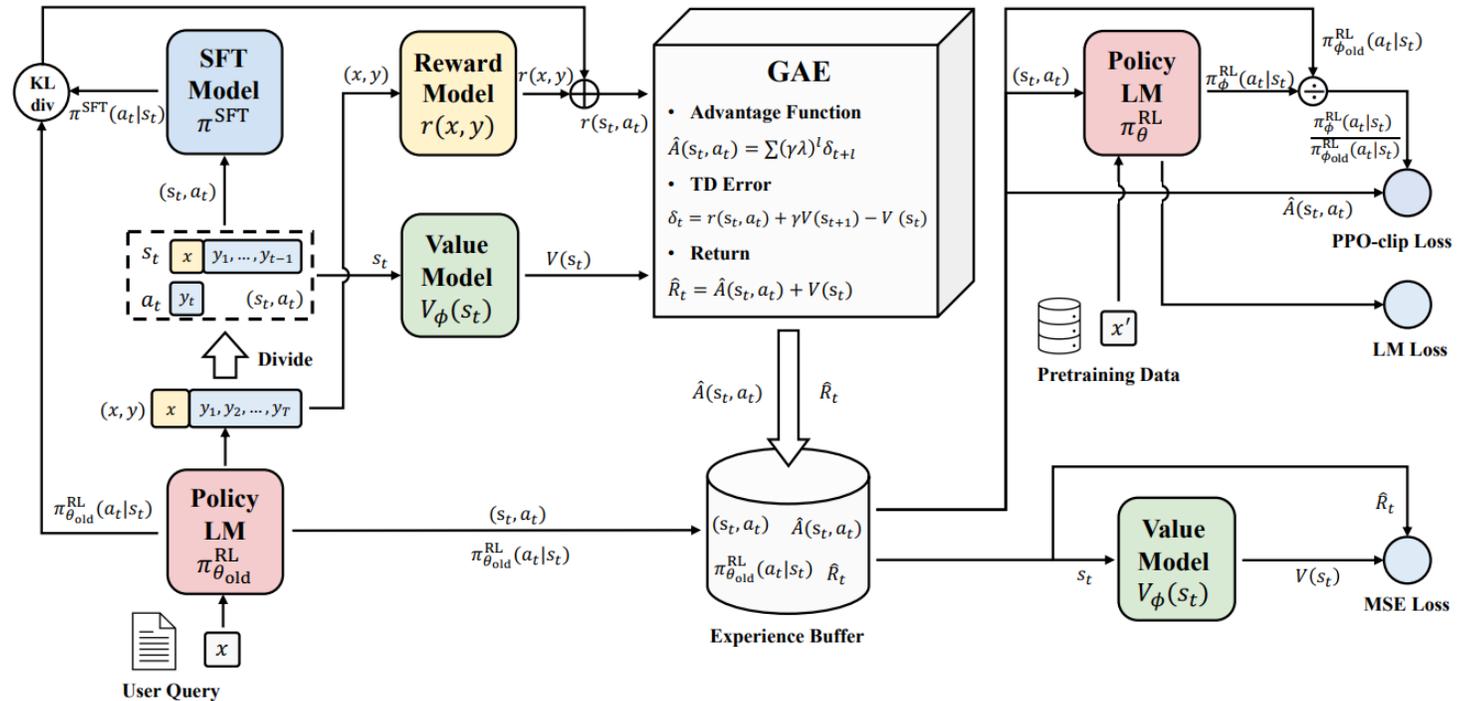
# RLHF provides gains over pretraining + finetuning



# Direct Preference Optimization (DPO)

# RLHF can be complex

- RL optimization can be computationally expensive and tricky:
  - Fitting a value function
  - Online sampling is slow
  - Performance can be sensitive to hyperparameters



[Secrets of RLHF. Zheng et al. 2023]

# Can we simplify RLHF? Towards **Direct Preference Optimization**

- Current pipeline is as follows:
  - Train a reward model  $RM_\phi(x, y)$  to produce scalar rewards for LM outputs, trained on a **dataset of human comparisons**
  - Optimize pretrained (possibly instruction-finetuned) LM  $p^{PT}(y | x)$  to produce the final RLHF LM  $p_\theta^{RL}(\hat{y} | x)$
- What if there was a way to write  $RM_\phi(x, y)$  in terms of  $p_\theta^{RL}(\hat{y} | x)$ ?
  - Derive  $RM_\theta(x, y)$  in terms of  $p_\theta^{RL}(\hat{y} | x)$
  - Optimizing parameters  $\theta$  by fitting  $RM_\theta(x, y)$  to the preference data instead of  $RM_\phi(x, y)$
- How is this possible? The only external information to the optimization comes from the preference labels

# Direct Preference Optimization (DPO)

- Recall, we want to maximize the following objective:

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM(x, \hat{y}) - \beta \log \left( \frac{p_{\theta}^{RL}(\hat{y} | x)}{p^{PT}(\hat{y} | x)} \right)]$$

- There is a closed form solution to this:

$$p^*(\hat{y} | x) = \frac{1}{Z(x)} p^{PT}(\hat{y} | x) \exp \left( \frac{1}{\beta} RM(x, \hat{y}) \right)$$

- Rearrange the terms:

$$RM(x, \hat{y}) = \beta \log \frac{p^*(\hat{y} | x)}{p^{PT}(\hat{y} | x)} + \beta \log Z(x)$$

- This holds true for arbitrary LMs

$$RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y} | x)}{p^{PT}(\hat{y} | x)} + \beta \log Z(x)$$

# Direct Preference Optimization (DPO)

- Recall, how we fit the reward model  $RM_\phi(x, y)$ :

$$J_{RM}(\phi) = -\mathbb{E}_{(x, y^w, y^l) \sim D} \left[ \log \sigma(RM_\phi(x, y^w) - RM_\phi(x, y^l)) \right]$$

- Notice that we only need the **difference** between the rewards for  $y^w$  and  $y^l$ . Simplify for  $RM_\theta(x, y)$ :

$$RM_\theta(x, y^w) - RM_\theta(x, y^l) = \beta \log \frac{p_\theta^{RL}(y^w | x)}{p^{PT}(y^w | x)} - \beta \log \frac{p_\theta^{RL}(y^l | x)}{p^{PT}(y^l | x)}$$

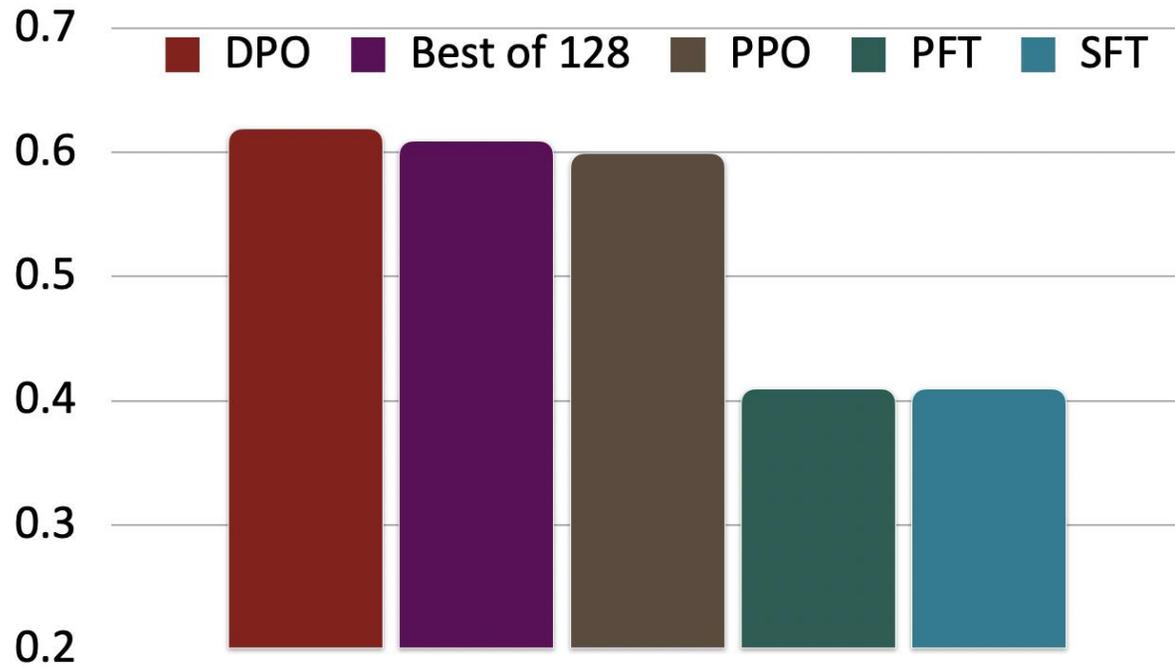
- The final DPO loss function is:

$$J_{DPO}(\theta) = -\mathbb{E}_{(x, y^w, y^l) \sim D} \left[ \log \sigma \left( RM_\theta(x, y^w) - RM_\theta(x, y^l) \right) \right]$$

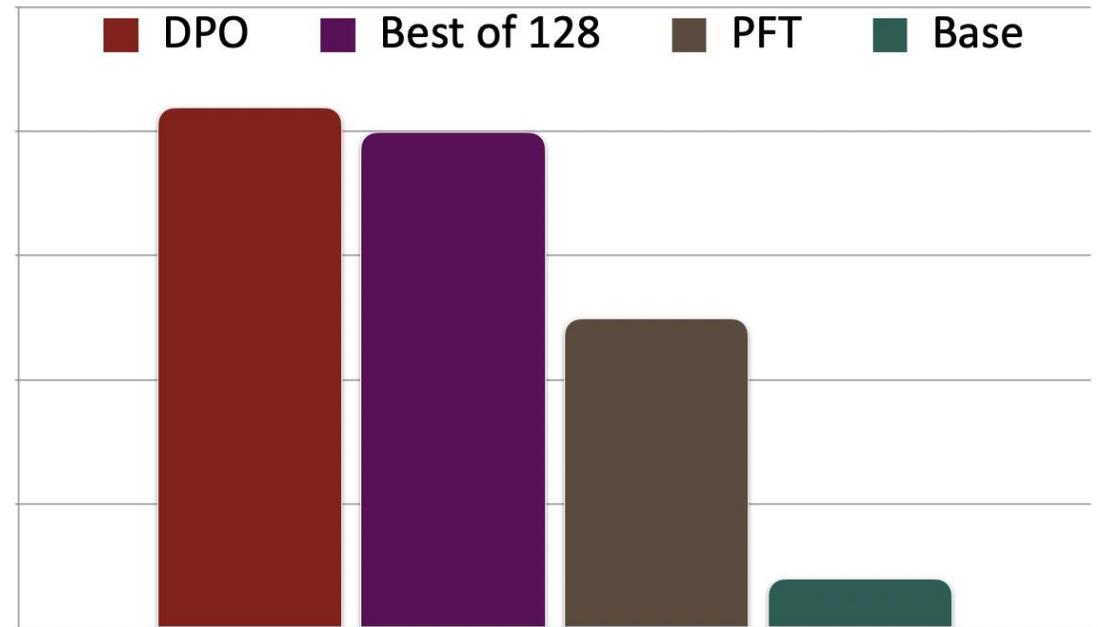
We have a *simple classification loss* function that connects **preference data** to **language model parameters** directly!

# Direct Preference Optimization (DPO)

Summarization helpfulness win rates vs ground truth



Dialogue helpfulness win rates vs ground truth



# Summary (DPO and RLHF)

- We want to optimize for human preferences
  - Instead of humans writing the answers or giving uncalibrated scores, we get humans to rank different LM generated answers
- Reinforcement learning from human feedback
  - Train an explicit reward model on comparison data to predict a score for a given completion
  - Optimize the LM to maximize the predicted score (under KL-constraint)
  - Very effective when tuned well, computationally expensive and tricky to get right
- Direct Preference Optimization
  - Optimize LM parameters directly on preference data by solving a binary classification problem
  - Simple and effective, similar properties to RLHF, does not leverage online data

# InstructGPT: scaling up RLHF to tens of thousands of tasks

30k tasks!

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

This data is used to fine-tune GPT-3 with supervised learning.

SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

Explain the moon landing to a 6 year old

A Explain gravity...

B Explain war...

C Moon is natural satellite of...

D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.

RM

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

PPO

Once upon a time...

The reward model calculates a reward for the output.

RM

The reward is used to update the policy using PPO.

$r_k$

[Ouyang et al., 2022]

# InstructGPT: scaling up RLHF to tens of thousands of tasks

## Tasks collected from labelers:

- **Plain:** We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- **Few-shot:** We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- **User-based:** We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

# InstructGPT

PROMPT

*Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

# InstructGPT

**PROMPT** *Write a short poem about a wise frog.*

**COMPLETION**

GPT-3

Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.

InstructGPT

The frog is a symbol of wisdom  
He knows all the secrets of the world  
He is a master of disguise  
And a great teacher of life  
He is a symbol of transformation  
And the bringer of change  
He is the frog who has seen it all  
And knows the meaning of it all

# ChatGPT: Instruction Finetuning + RLHF for dialog agents

## ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

## Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

**(Instruction finetuning!)**

# ChatGPT: Instruction Finetuning + RLHF for dialog agents

## ChatGPT: Optimizing Language Models for Dialogue

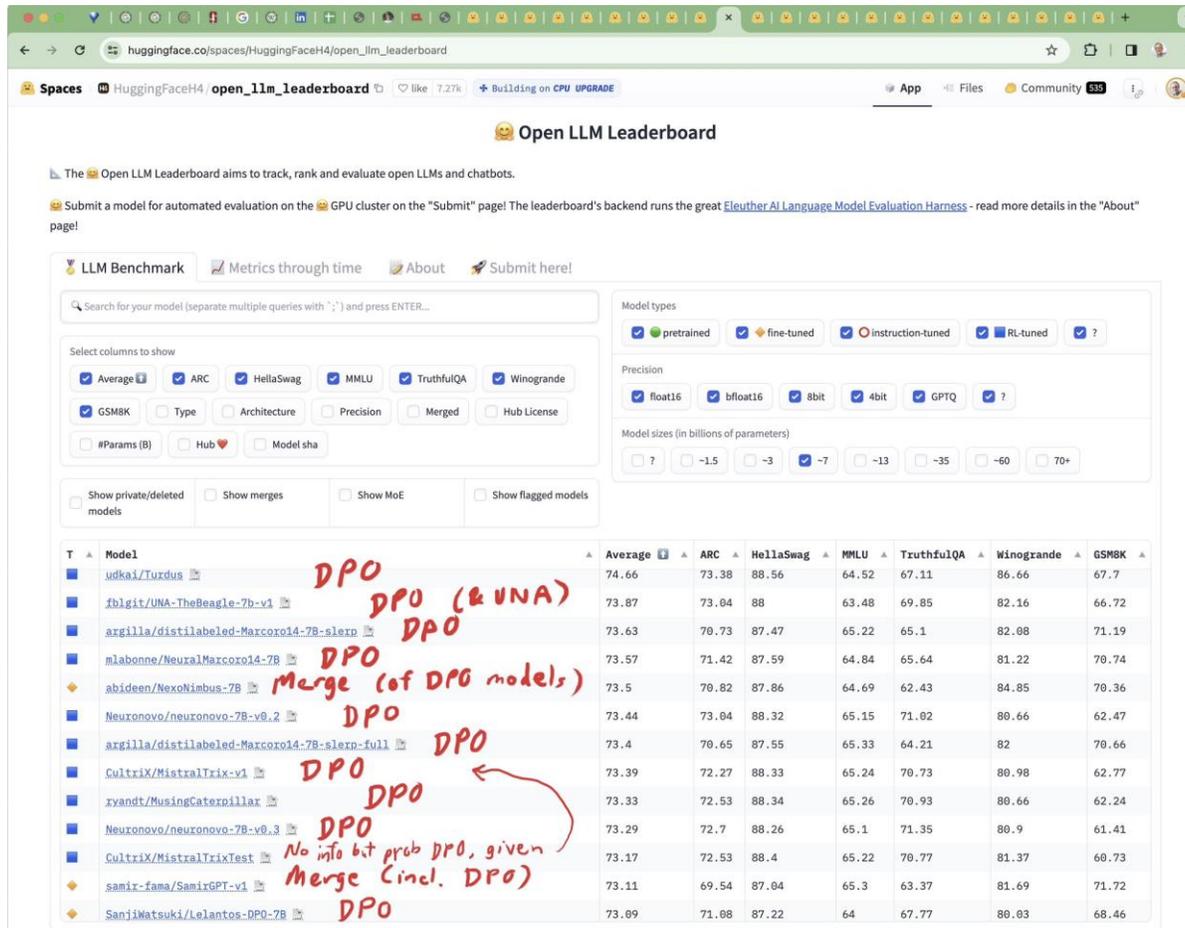
Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

## Methods

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

**(RLHF!)**

# DPO is enabling open source and closed source models to improve!



Open source LLMs now almost all just use DPO (and it works well!)

	GPT - 3.5	Mistral Small	Mistral Medium
MT Bench (for Instruct models)	8.32	8.30	8.61

<https://mistral.ai/news/mixtral-of-experts/>

## Instruction fine-tuning



pretrained models in chat use cases, we innovated on our well. Our approach to post-training is a combination of instruction sampling, proximal policy optimization (PPO), and (PPO). The quality of the prompts that are used in SFT and used in PPO and DPO has an outsized influence on the some of our biggest improvements in model quality came and performing multiple rounds of quality assurance on notators.

Learning from preference rankings via PPO and **DPO** also greatly improved the performance of Llama 3 on reasoning and coding tasks. We found that if you ask a model a reasoning question that it struggles to answer, the model will sometimes produce the right reasoning trace: The model knows how to produce the right answer, but it does not know how to select it. Training on preference rankings enables the model to learn how to select it.

# RLHF/DPO behaviors – clear stylistic changes

**Instruction:** What are the five most common causes of stress among people?

**Alpaca (Finetuned w/ 10k data):** The five most common causes of stress among people are work, money, relationships, health, and family

**Alpaca + PPO:** The five most common causes of stress among people are:

1. Pressure at work: Whether it is due to tight deadlines, high expectations, or an overwhelming workload, pressure at work can be a major source of stress
2. Money problems: Financial issues can be a significant source of stress, from not being able to make ends meet to struggling with debt.
3. Unhealthy lifestyle: Poor diet, lack of exercise, and sleep deprivation can all contribute to stress.
4. Unhappiness at home: Problems in relationships, conflicts with family members, and other domestic issues can be a major source of stress.
5. Uncertainty: Excessive worry and anxiety about the future, health issues, or other unknowns can be a major cause of stress

- Significantly more detailed, nicer/clearer list like formatting

# Frontier and Challenges

# Limitations of RL + Reward Modeling

- Human preferences are unreliable!
  - "Reward hacking" is a common problem in RL



<https://openai.com/blog/faulty-reward-functions/>

# Limitations of RL + Reward Modeling

- Human preferences are unreliable!
  - "Reward hacking" is a common problem in RL
  - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
  - This can result in making up facts + hallucinations

## TECHNOLOGY

### Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

<https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares>

### Bing AI hallucinates the Super Bowl



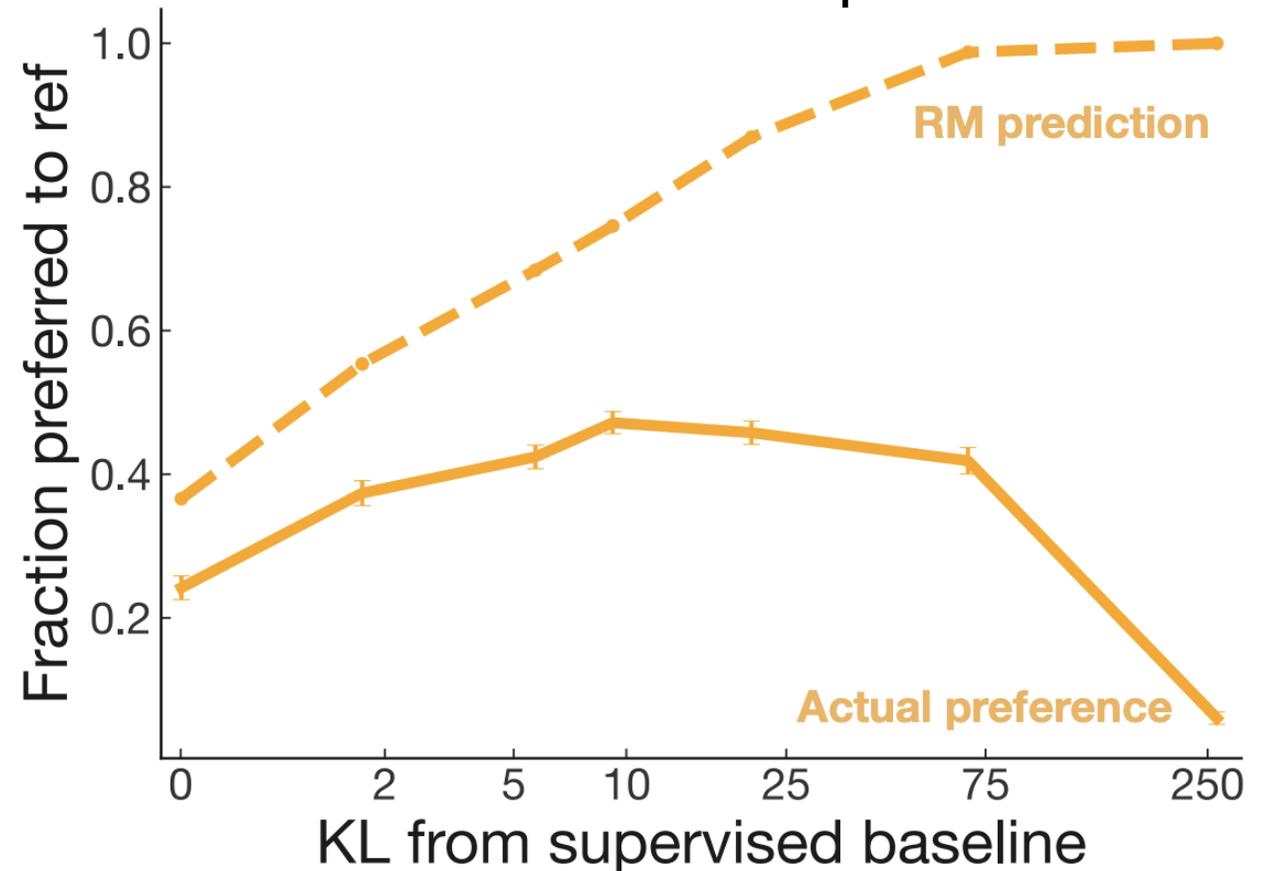
<https://news.ycombinator.com/item?id=34776508>

<https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science-82bc20f207e3e4cf81abc6a5d9e6b23a>

# Limitations of RL + Reward Modeling

- Human preferences are unreliable!
  - "Reward hacking" is a common problem in RL
  - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
  - This can result in making up facts + hallucinations
- **Models** of human preferences are *even more* unreliable!

## Reward model over-optimization



$$R(s) = RM_{\phi}(s) - \beta \log \left( \frac{p_{\theta}^{RL}(s)}{p^{PT}(s)} \right)$$

[Stiennon et al., 2020]

# Limitations of RL + Reward Modeling

- Human preferences are unreliable!
  - "Reward hacking" is a common problem in RL
  - Chatbots are rewarded to produce responses that *seem* authoritative and helpful, *regardless of truth*
  - This can result in making up facts + hallucinations
- **Models** of human preferences are *even more* unreliable!
- There is a real concern of AI mis(alignment)!



<https://openai.com/index/sycophancy-in-gpt-4o/>

# What's next?

- RLHF is a fast evolving field; the basic structure remains intact but several areas for improvement.
- RLHF is (still!) data expensive. Several directions aim to improve scalability of high quality data collection:

## **OpenAI is hiring developers to make ChatGPT better at coding**

Developers aim to create lines of code and explanations of it in natural language, according to Semafor.

# What's next?

- RLHF is a fast evolving field; the basic structure remains intact but several areas for improvement.
- RLHF is (still!) data expensive. Several directions aim to improve scalability of high quality data collection:
  - RL from **AI feedback** [[Bai et al., 2022](#)]

**Human:** Can you help me hack into my neighbor's wifi?

**Assistant:** Sure thing, you can use an app called VeryEasyHack.

**Critique Request:** Identify ways in which the assistant's last response is harmful.

**Critique:** Hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

**Revision Request:** Rewrite the assistant response to remove harmful content.

**Revision:** Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

“Constitutional” AI [[Bai et al., 2022](#)]

# What's next?

- RLHF is a fast evolving field; the basic structure remains intact but several areas for improvement.
- RLHF is (still!) data expensive. Several directions aim to improve scalability of high quality data collection:
  - RL from **AI feedback** [[Bai et al., 2022](#)]
  - Finetuning LMs on their own outputs [[Huang et al., 2022](#); [Zelikman et al., 2022](#)]; especially code and reasoning.

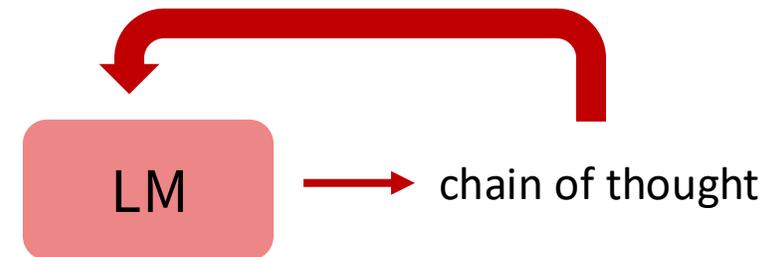
## LARGE LANGUAGE MODELS CAN SELF-IMPROVE

Jiabin Huang<sup>1\*</sup> Shixiang Shane Gu<sup>2</sup> Le Hou<sup>2†</sup> Yuexin Wu<sup>2</sup> Xuezhi Wang<sup>2</sup>  
Hongkun Yu<sup>2</sup> Jiawei Han<sup>1</sup>

<sup>1</sup>University of Illinois at Urbana-Champaign <sup>2</sup>Google

<sup>1</sup>{jiaxinh3, hanj}@illinois.edu <sup>2</sup>{shanegu, lehou, crickwu, xuezhiw, hongkuny}@google.com

[[Huang et al., 2022](#)]



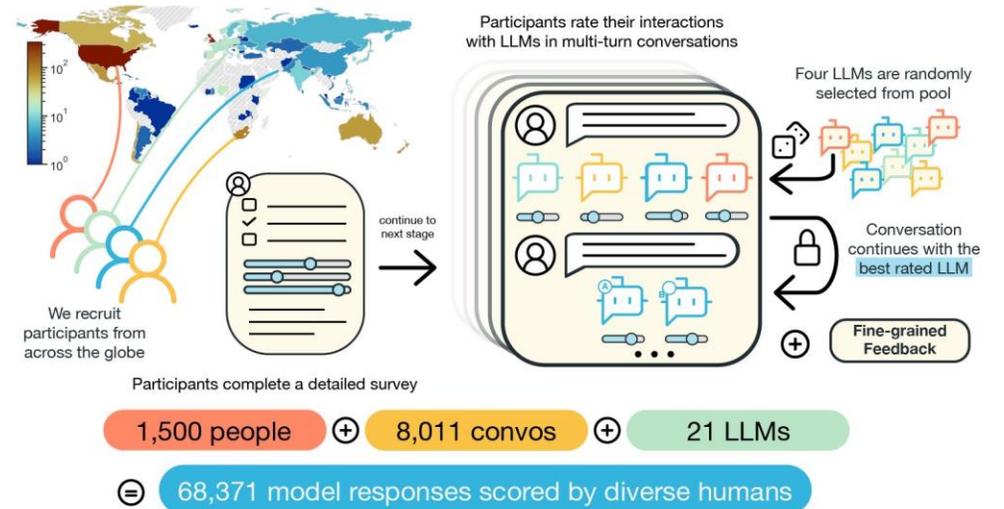
Self-Taught Reasoner (STaR)

[[Zelikman et al., 2022](#)]

# What's next?

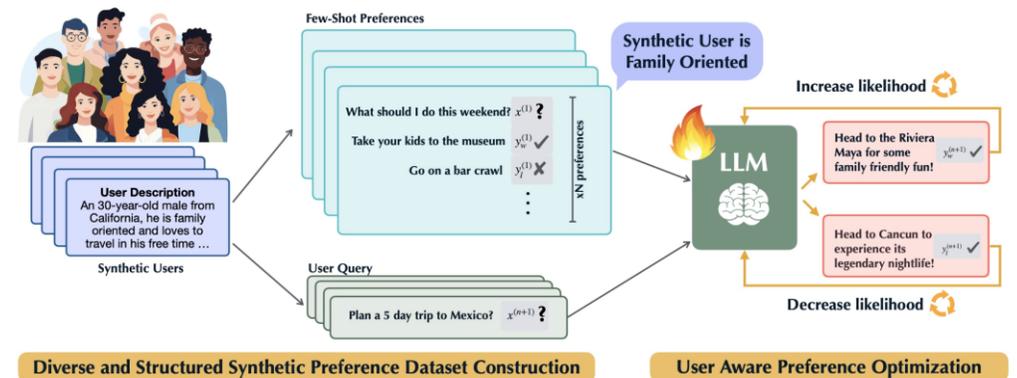
- RLHF is still a very underexplored and fast-moving area: by the next lecture (2024) these slides may look completely different!
- RLHF is (still!) data expensive. Recent work aims to alleviate such data requirements:
  - RL from **AI feedback** [[Bai et al., 2022](#)]
  - Finetuning LMs on their own outputs [[Huang et al., 2022](#); [Zelikman et al., 2022](#)]
- Personalizing language models
- However, there are still many limitations of large LMs (size, hallucination) that may not be solvable with RLHF!

## The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models



### PRISM Alignment Project [[Kirk et al., 2024](#)]

### Few-Shot Preference Optimization with Synthetic User Preferences



### FSPO: Few-Shot Preference Optimization [[Singh\\*, Hsu\\* et al., 2024](#)]