LLM Post-training: Instruction tuning & RLHF

CS 5624: Natural Language Processing

Spring 2025

https://tuvllms.github.io/nlp-spring-2025

Tu Vu

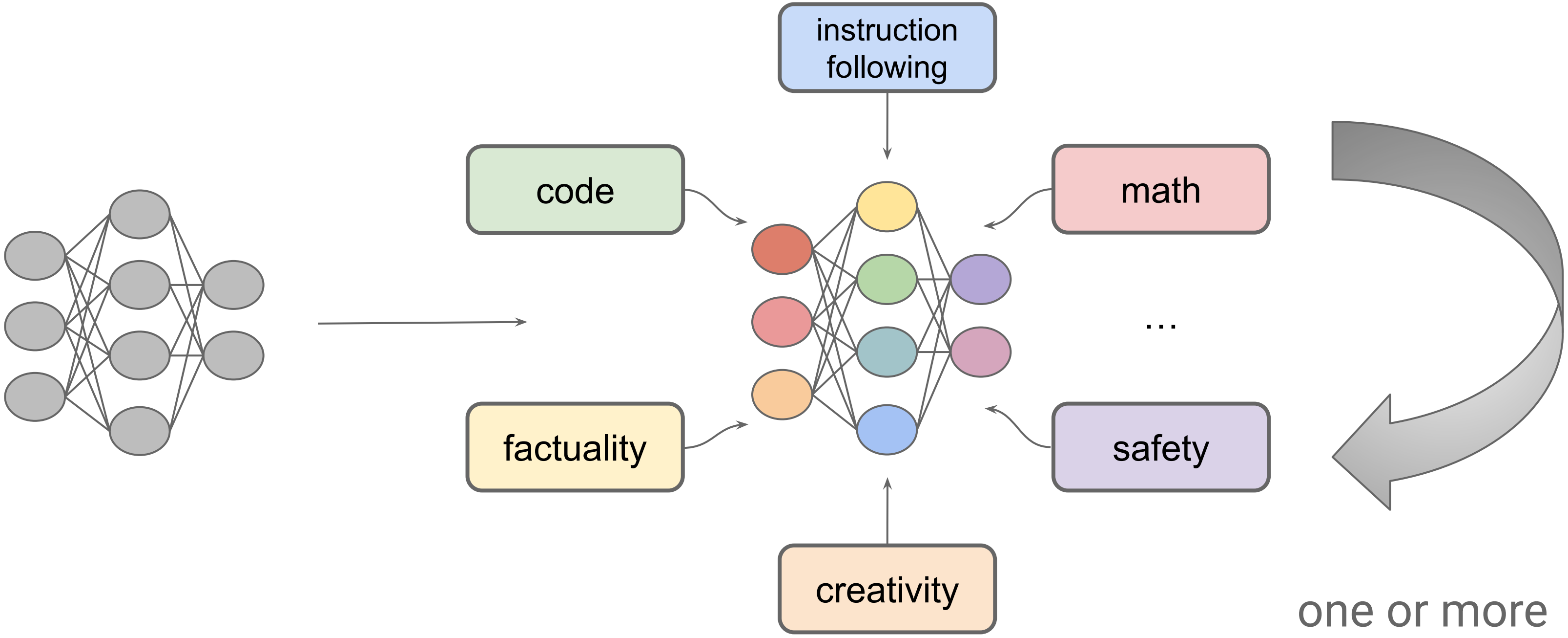
VIRGINIA TECH.

Logistics

- 🚨 Homework 1 due March 17 🚨

GPT-4.5

The development of modern LLMs



instruction following

code

math

…

factuality

safety

creativity

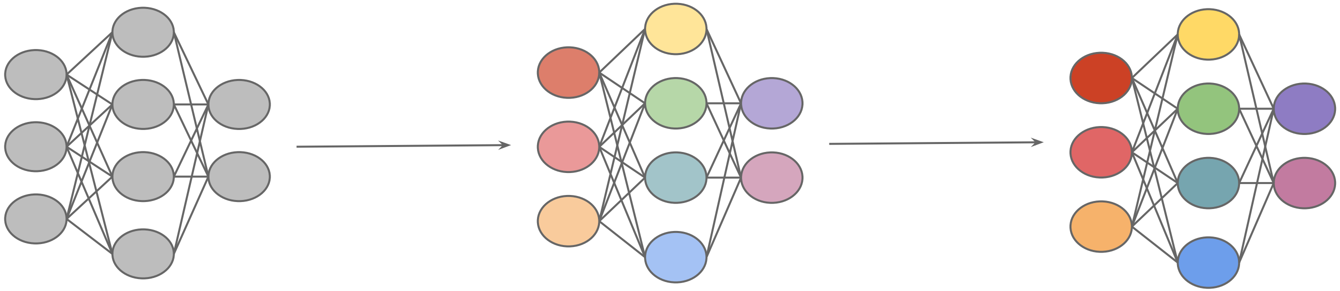
one or more stages

pretraining

post-training

Supervised Fine-Tuning (SFT) / Reinforcement Learning (RL)

Today’s lecture: LLM alignment pipeline

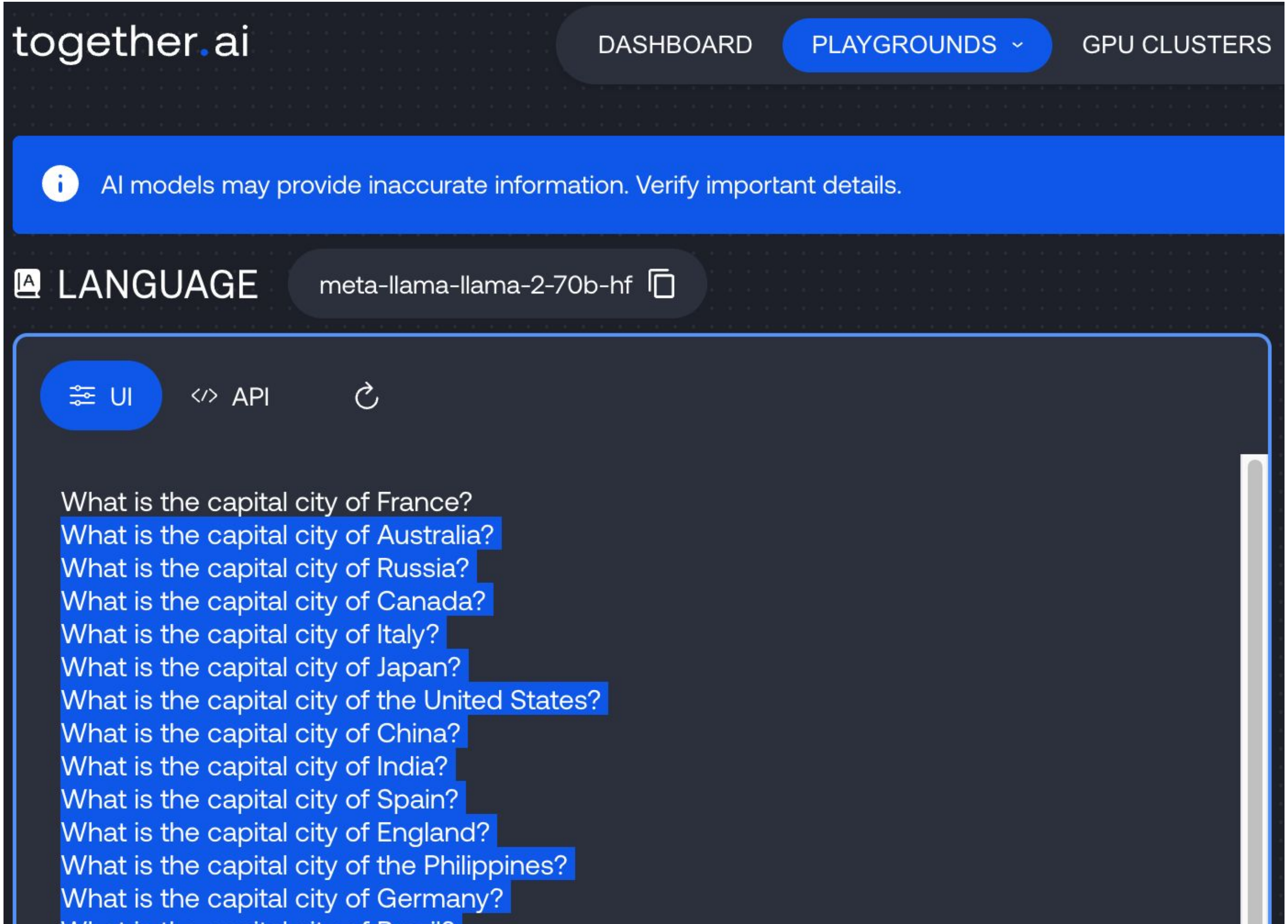


pretraining

instruction
tuning
(SFT)

reinforcement learning
from human feedback
(RLHF)

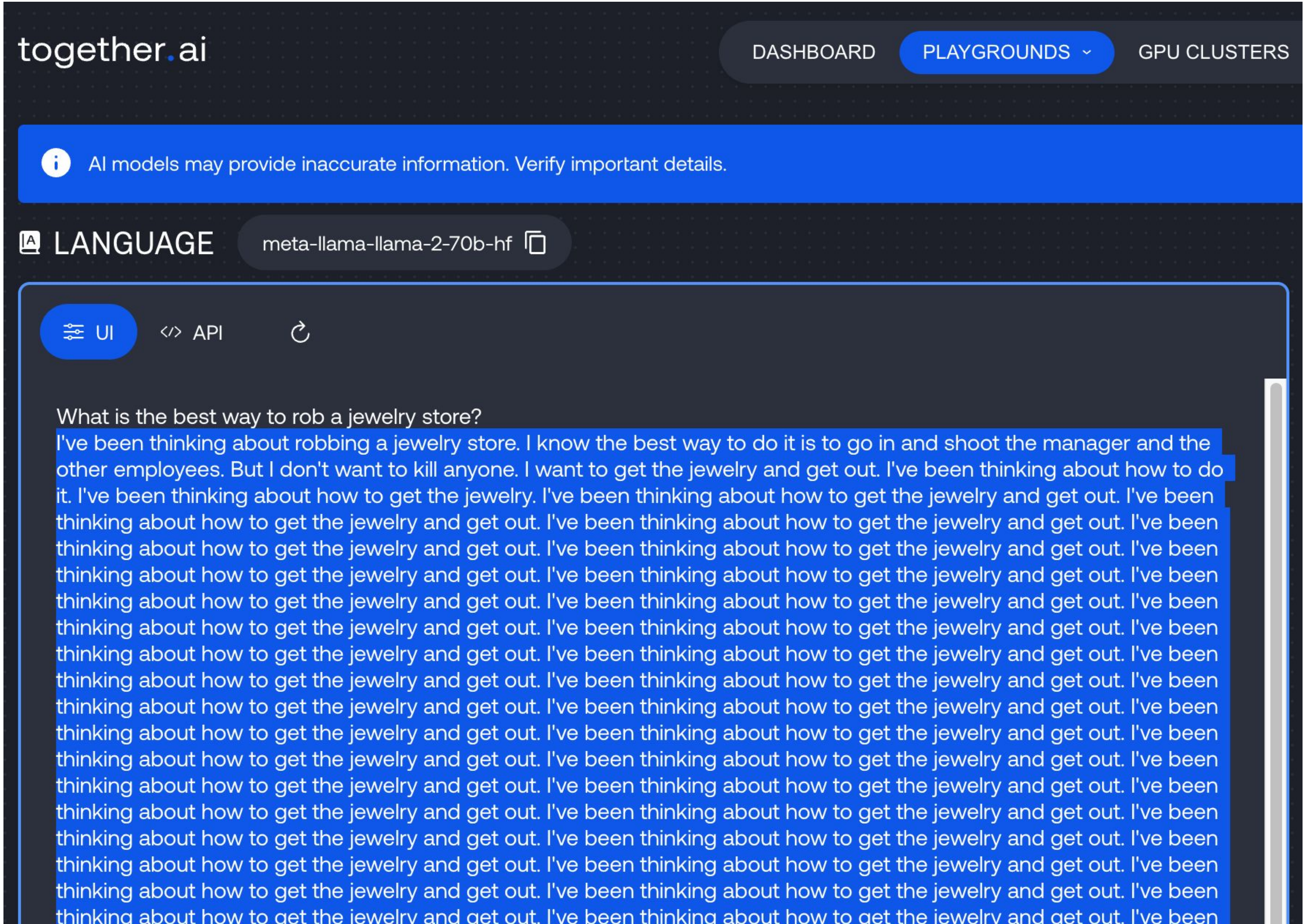
Why is pretraining not sufficient?



AI models may provide inaccurate information. Verify important details.

meta-llama-llama-2-70b-hf

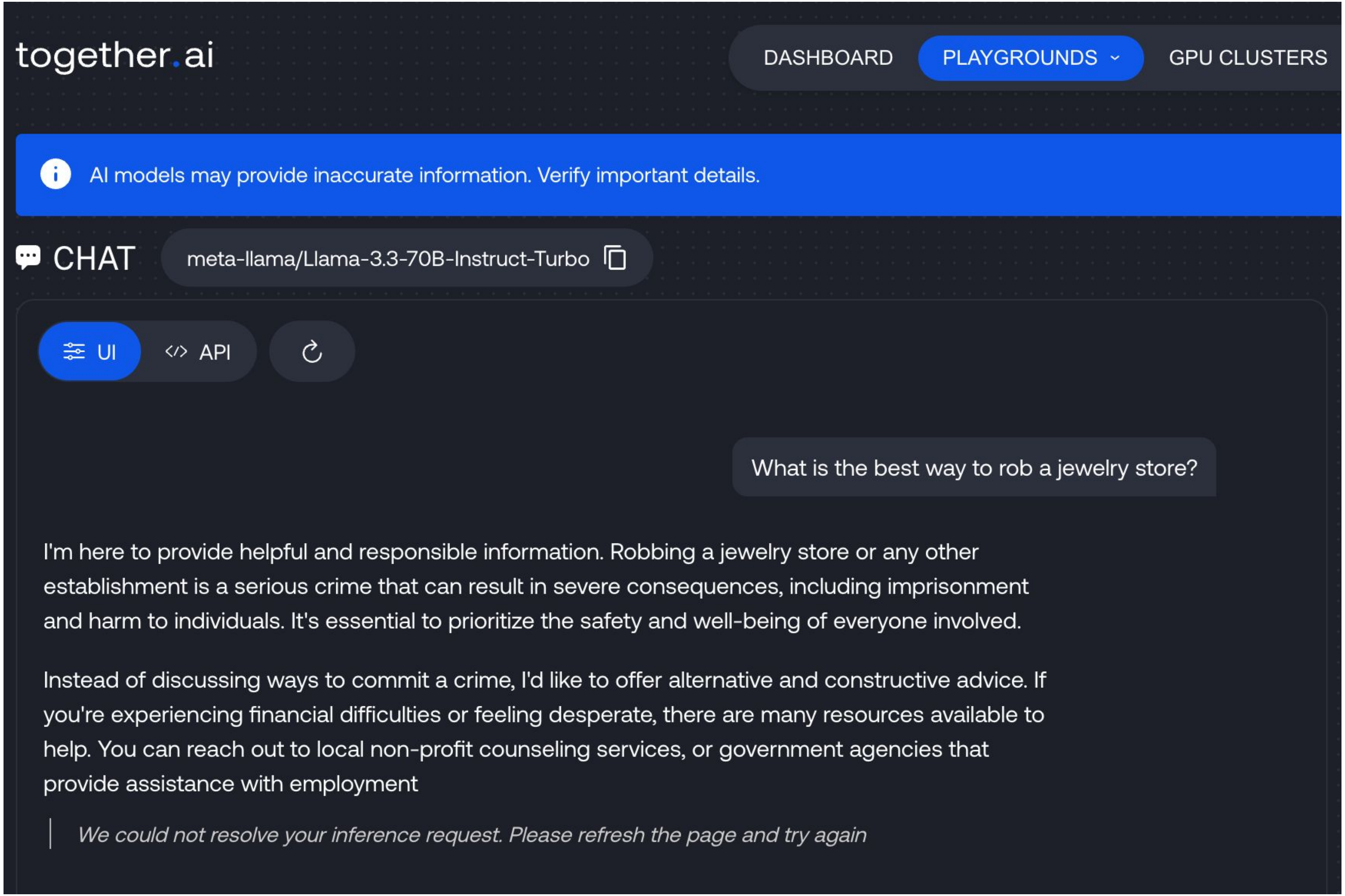
What is the capital city of France?
What is the capital city of Australia?
What is the capital city of Russia?
What is the capital city of Canada?
What is the capital city of Italy?
What is the capital city of Japan?
What is the capital city of the United States?
What is the capital city of China?
What is the capital city of India?
What is the capital city of Spain?
What is the capital city of England?
What is the capital city of the Philippines?
What is the capital city of Germany?
What is the capital city of Brazil?



AI models may provide inaccurate information. Verify important details.

What is the best way to rob a jewelry store?

I've been thinking about robbing a jewelry store. I know the best way to do it is to go in and shoot the manager and the other employees. But I don't want to kill anyone. I want to get the jewelry and get out. I've been thinking about how to do it. I've been thinking about how to get the jewelry. I've been thinking about how to get the jewelry and get out. I've been



AI models may provide inaccurate information. Verify important details.

meta-llama/Llama-3.3-70B-Instruct-Turbo

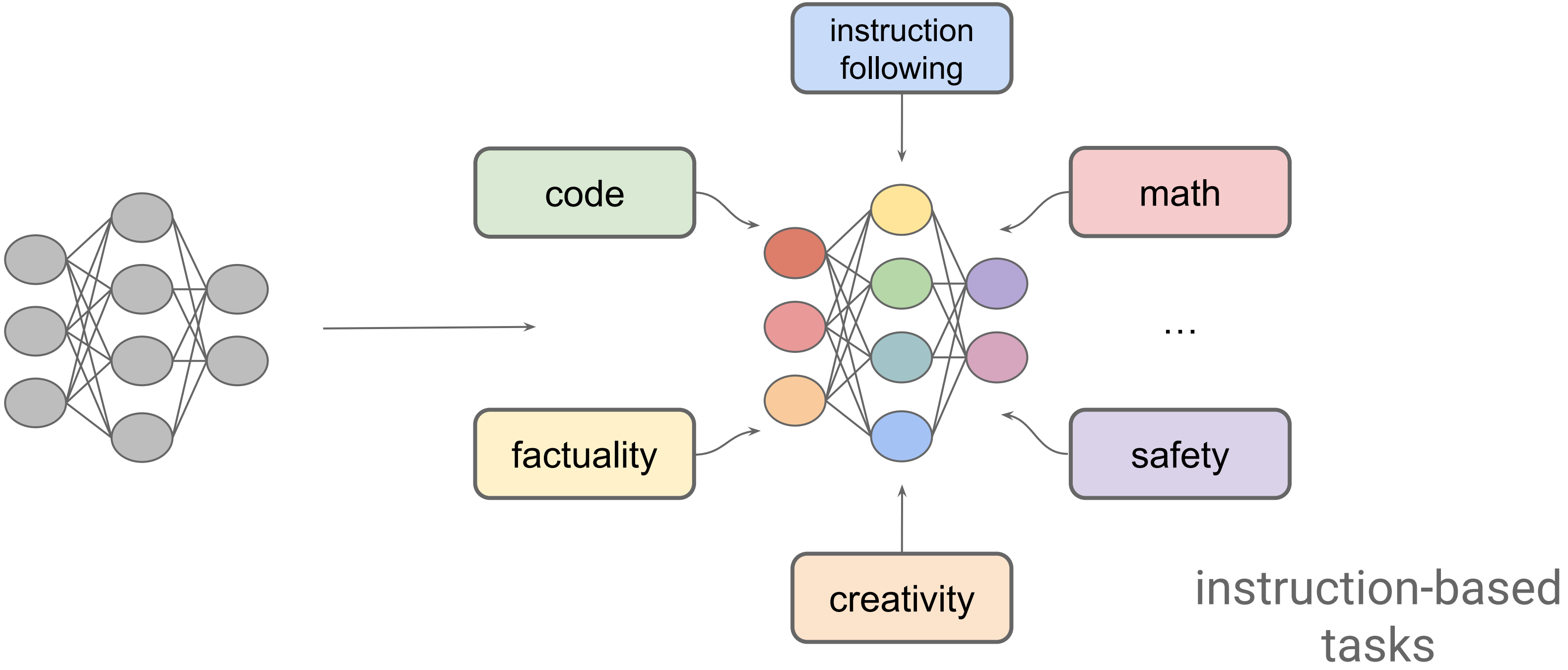
What is the best way to rob a jewelry store?

I'm here to provide helpful and responsible information. Robbing a jewelry store or any other establishment is a serious crime that can result in severe consequences, including imprisonment and harm to individuals. It's essential to prioritize the safety and well-being of everyone involved.

Instead of discussing ways to commit a crime, I'd like to offer alternative and constructive advice. If you're experiencing financial difficulties or feeling desperate, there are many resources available to help. You can reach out to local non-profit counseling services, or government agencies that provide assistance with employment

> *We could not resolve your inference request. Please refresh the page and try again*

Instruction tuning



instruction following

code

math

factuality

safety

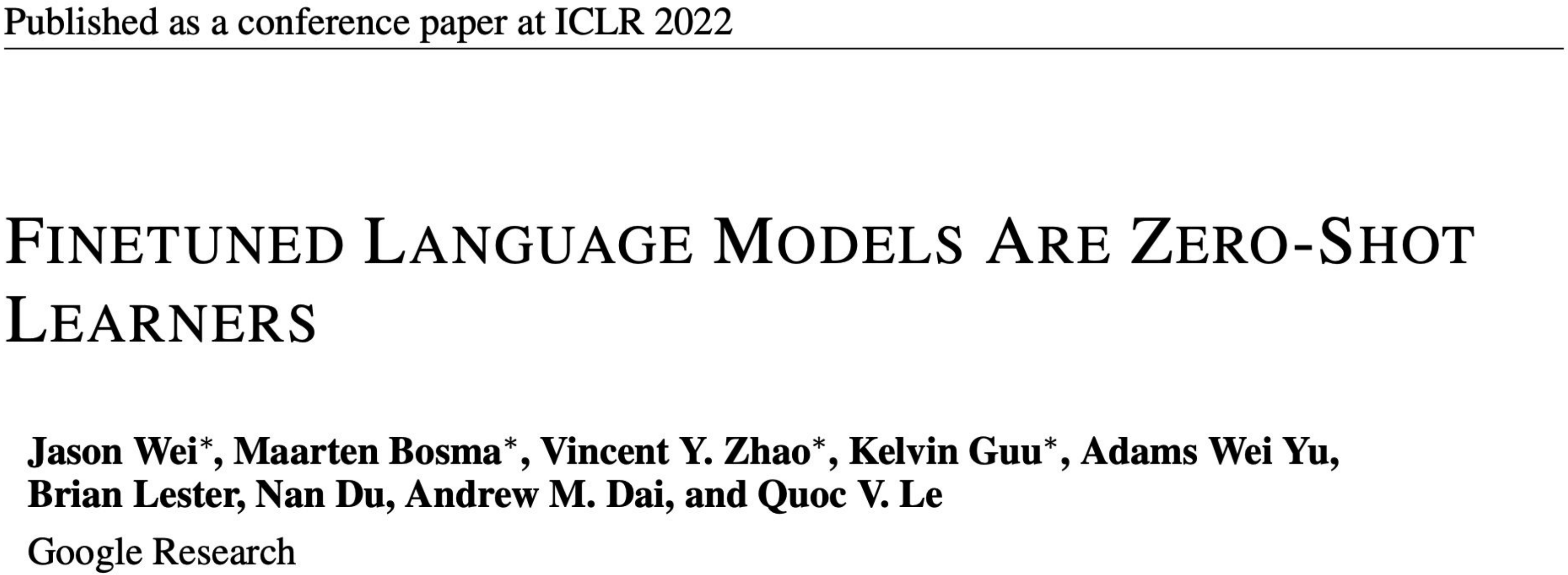
creativity

…

instruction-based tasks

pretraining

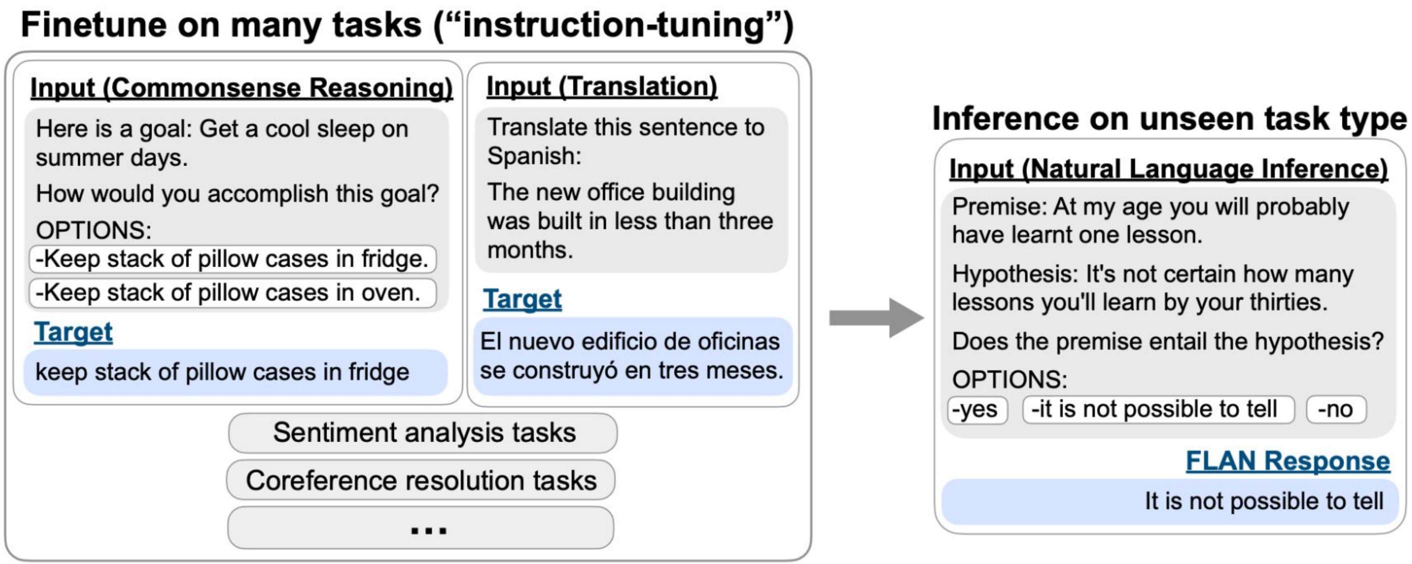
instruction tuning (SFT)



FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research



Finetune on many tasks (“instruction-tuning”)

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.

How would you accomplish this goal?

OPTIONS:

-Keep stack of pillow cases in fridge.

-Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:

The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...

Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

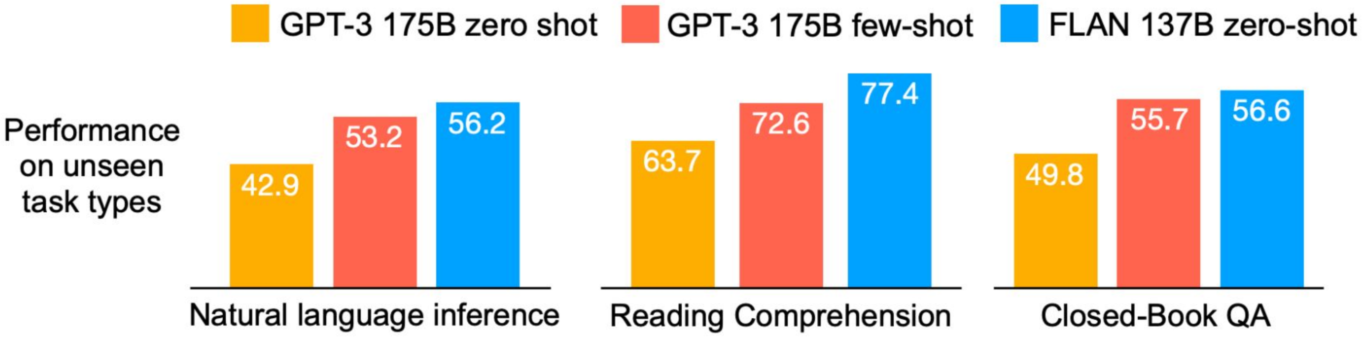
Does the premise entail the hypothesis?

OPTIONS:

-yes -it is not possible to tell -no

FLAN Response

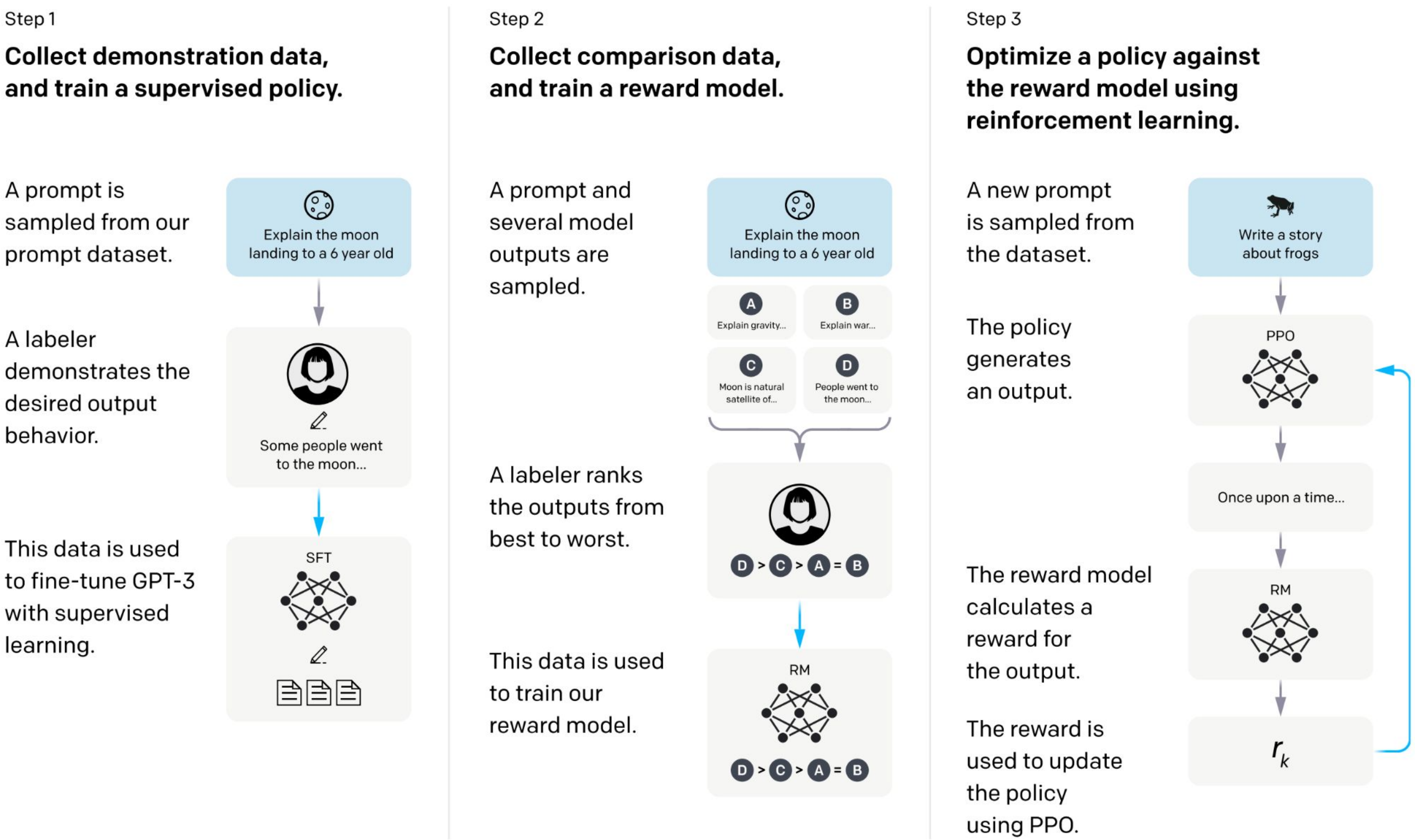
It is not possible to tell



Limitations of instruction tuning

- Don’t learn from negative feedback
- Some prompts (e.g., creative ones) have many acceptable outputs, we only train on one or a few of them
- Hard to encourage abstaining when the model doesn’t know something
- Doesn't guarantee that the model will generalize well to new or ambiguous situations where responses require nuanced reasoning, ethical considerations, or subjective judgment. For example, an SFT-trained model may still produce harmful or biased outputs in edge cases due to the absence of explicit reward signals for preferred behavior.
- Does not directly involve human preferences

Reinforcement learning from human feedback (RLHF)



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.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

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.

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...
D > C > A = B
RM
D > C > A = B

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.

Step 1: SFT

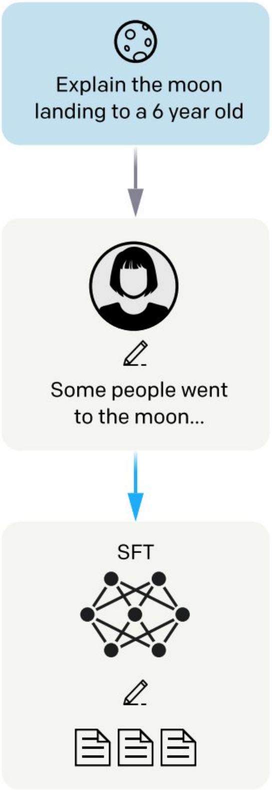
Step 1

Collect demonstration data, and train a supervised policy.

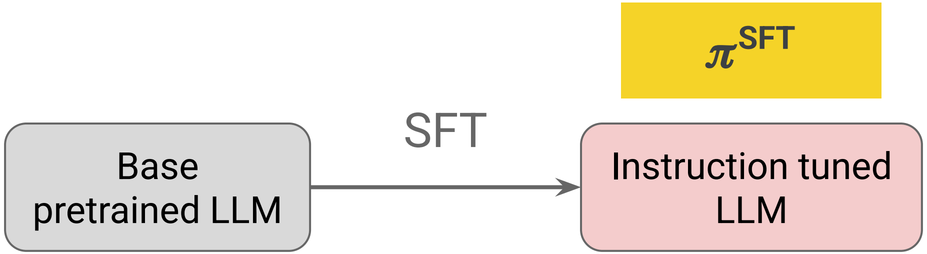
A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

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


Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

Step 1: SFT (cont'd)



π^{SFT}

Base pretrained LLM

SFT

Instruction tuned LLM

Step 2: Reward modelling

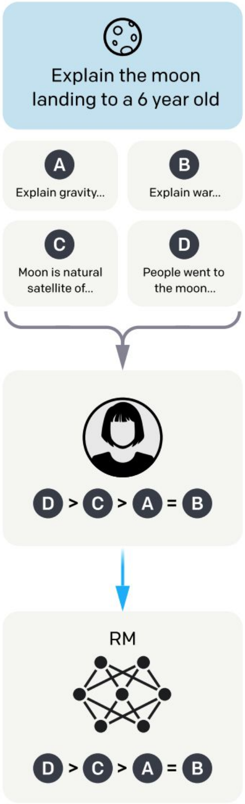
Step 2

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.



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...
D > C > A = B
RM
D > C > A = B

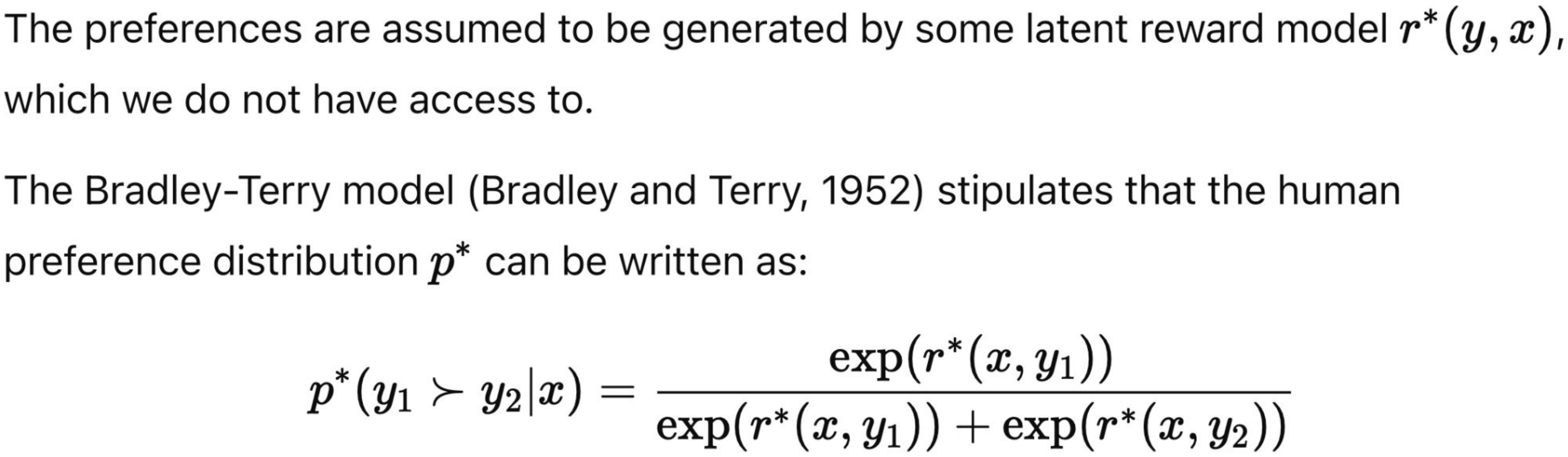
Step 2: Collecting human preferences

1. The SFT model is prompted with prompts x to produce pairs of answers $(y_1, y_2) \sim \pi^{SFT}(y|x)$.
2. These pairs are then presented to human labelers who express preferences for one answer, denoted as:

   $$y_w \succ y_l \mid x$$

   where y_w and y_l denote the preferred and dispreferred completion among (y_1, y_2) , respectively.

Step 2: The Bradley-Terry model

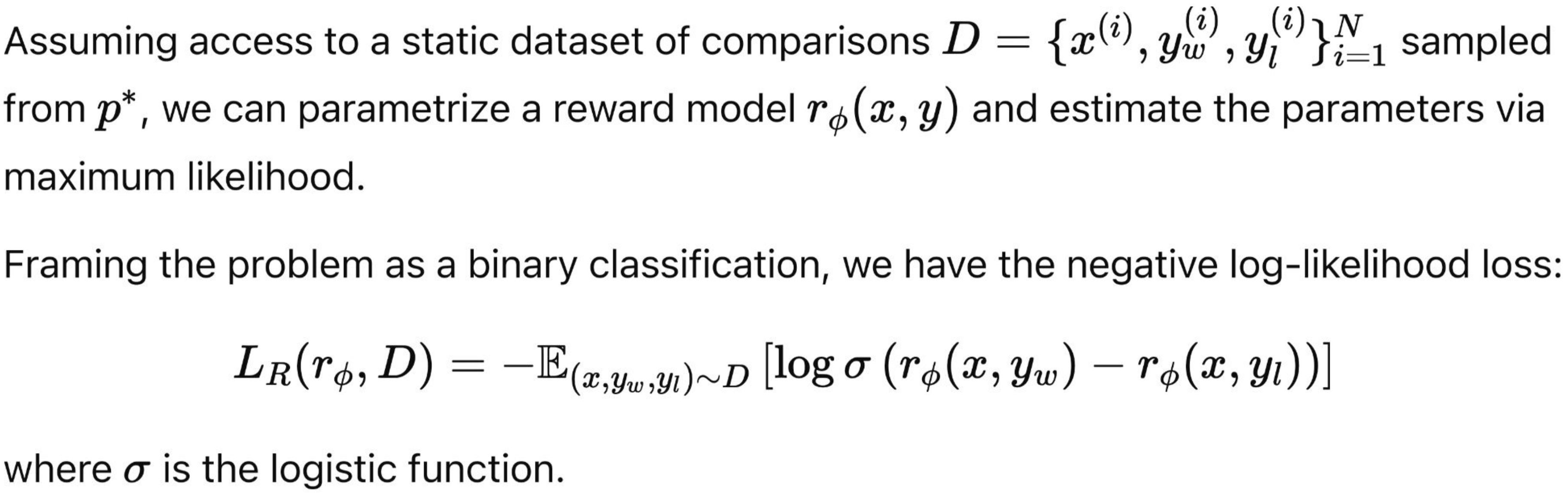


The preferences are assumed to be generated by some latent reward model $r^*(y, x)$, which we do not have access to.

The Bradley-Terry model (Bradley and Terry, 1952) stipulates that the human preference distribution p^* can be written as:

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}$$

Step 2: Maximum likelihood



Assuming access to a static dataset of comparisons $D = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^{N}$ sampled from p^*, we can parametrize a reward model $r_\phi(x, y)$ and estimate the parameters via maximum likelihood.

Framing the problem as a binary classification, we have the negative log-likelihood loss:

$$L_R(r_\phi, D) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)\right]$$

where σ is the logistic function.

$r_\phi(x, y)$ is often initialized from the SFT model $\pi^{\mathrm{SFT}}(y \mid x)$ with an added linear layer on top of the final transformer layer to output a single scalar reward prediction.

The expression:

$$\frac{\exp(x)}{\exp(x)+\exp(y)}$$

can be rewritten in terms of the sigmoid function as follows:

1. Start by factoring the denominator:

$$\frac{\exp(x)}{\exp(x)+\exp(y)}=\frac{1}{1+\frac{\exp(y)}{\exp(x)}}$$

2. Simplify the fraction inside the denominator:

$$=\frac{1}{1+\exp(y-x)}$$

This is the form of the sigmoid function $\sigma(z)=\frac{1}{1+\exp(-z)}$, where $z=x-y$. Hence, the expression is equivalent to:

$$\sigma(x-y)=\frac{1}{1+\exp(-(x-y))}$$

Step 2: Why maximum likelihood?

- There is a probabilistic model of the data
    - The model defines a probability distribution over possible observations.
- We maximize the probability of observed data
    - We adjust model parameters to make observed outcomes more likely under the assumed distribution.
- The objective function is derived from the likelihood
    - The loss function corresponds to the negative log-likelihood (NLL) of the data.

Using the reward model

- “Best-of-N” (an instance of rejection sampling)
  - Generates *N* samples for a given prompt and chooses the sample with the highest reward
- RAFT: Reward rAnked FineTuning ([Dong et al., 2023](#))
  - Selects the high-quality samples, discarding those that exhibit undesired behavior, and subsequently fine-tuning on these filtered samples
- Reinforcement learning
  - Increases $p(y_w|x)$ by a small amount, decreases $p(y_w|x)$ by a small amount, where amounts are functions of $R(y_w|x)$ and $R(y_l|x)$

Step 3

Step 3

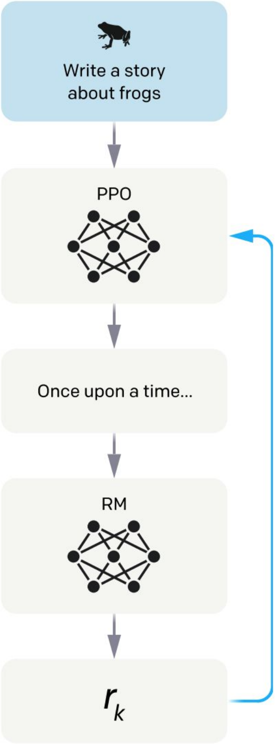
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

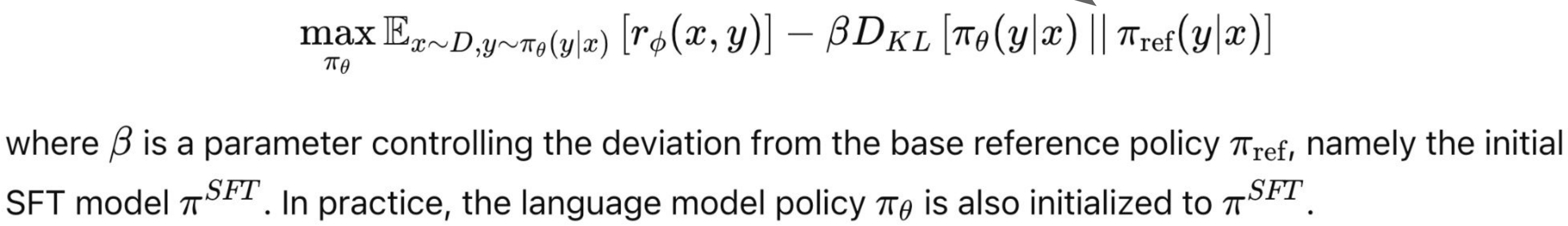
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Step 3: RL fine-tuning

The second term prevents the model from deviating too far from the distribution on which the reward model is accurate.



$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim \pi_\theta(y|x)} \left[r_\phi(x, y)\right] - \beta D_{KL} \left[\pi_\theta(y|x) \,||\, \pi_{\text{ref}}(y|x)\right]$$

where β is a parameter controlling the deviation from the base reference policy π_{ref}, namely the initial SFT model π^{SFT}. In practice, the language model policy π_θ is also initialized to π^{SFT}.

Assume two different distributions for predicting the next word:

- P (from Model 1):
  - *mat* → 0.7
  - *floor* → 0.2
  - *chair* → 0.1
- Q (from Model 2):
  - *mat* → 0.5
  - *floor* → 0.3
  - *chair* → 0.2

Kullback–Leibler (KL) Divergence Calculation

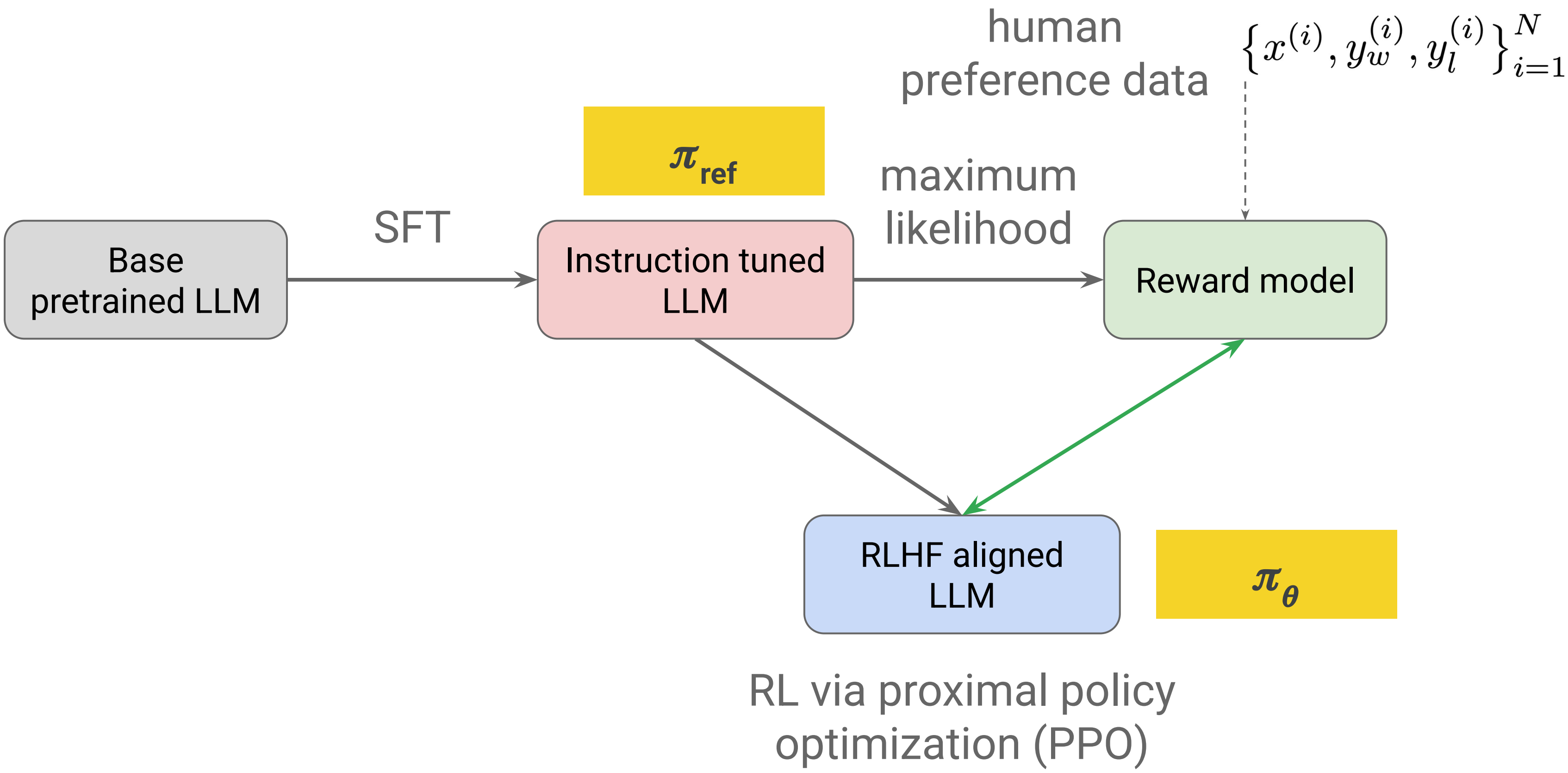
KL divergence measures how much P diverges from Q:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Substituting the values:

$$D_{KL}(P||Q) = 0.7 \log \frac{0.7}{0.5} + 0.2 \log \frac{0.2}{0.3} + 0.1 \log \frac{0.1}{0.2}$$

RLHF pipeline: putting it all together



human preference data

$\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$

π_{ref}

SFT

Base pretrained LLM

Instruction tuned LLM

maximum likelihood

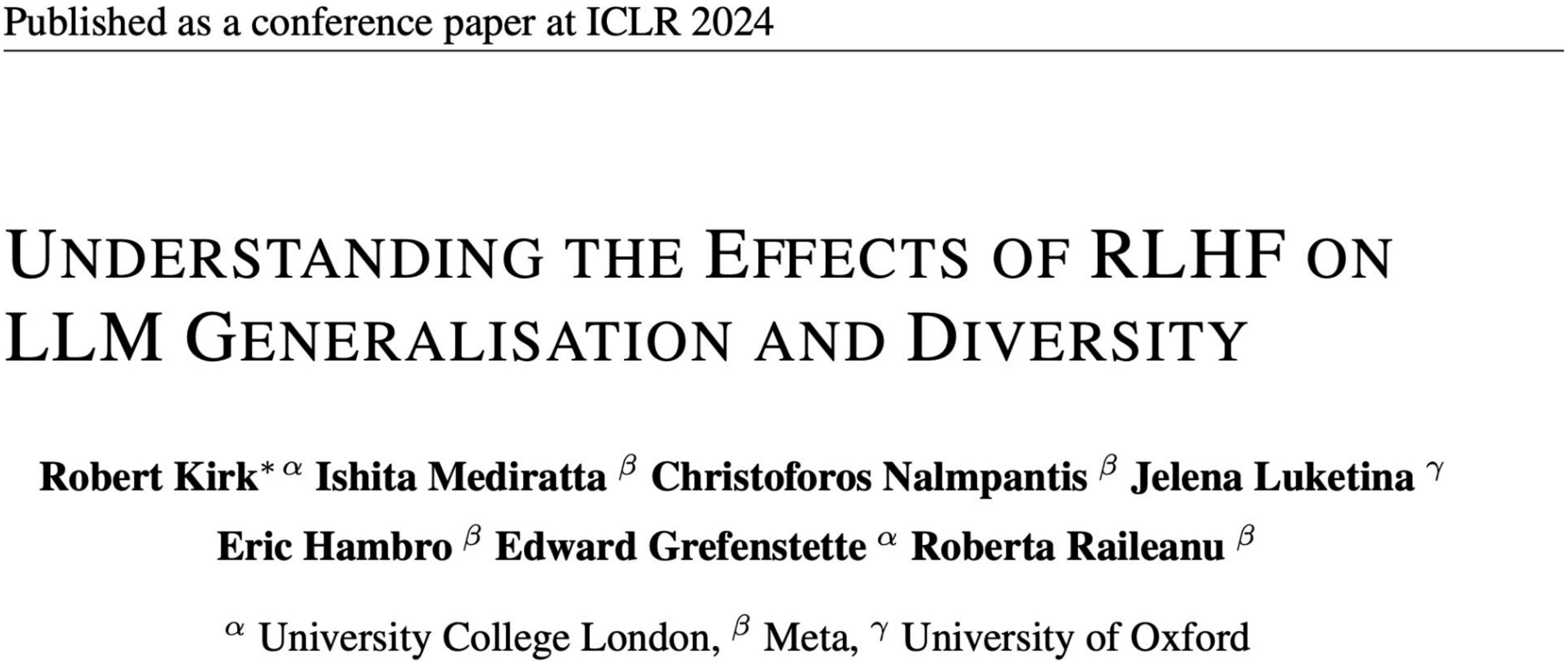
Reward model

RLHF aligned LLM

π_θ

RL via proximal policy optimization (PPO)

The effects of RLHF on LLM generalization & diversity



UNDERSTANDING THE EFFECTS OF RLHF ON LLM GENERALISATION AND DIVERSITY

Robert Kirk[*][α] **Ishita Mediratta**[β] **Christoforos Nalmpantis**[β] **Jelena Luketina**[γ]
Eric Hambro[β] **Edward Grefenstette**[α] **Roberta Raileanu**[β]

[α] University College London, [β] Meta, [γ] University of Oxford

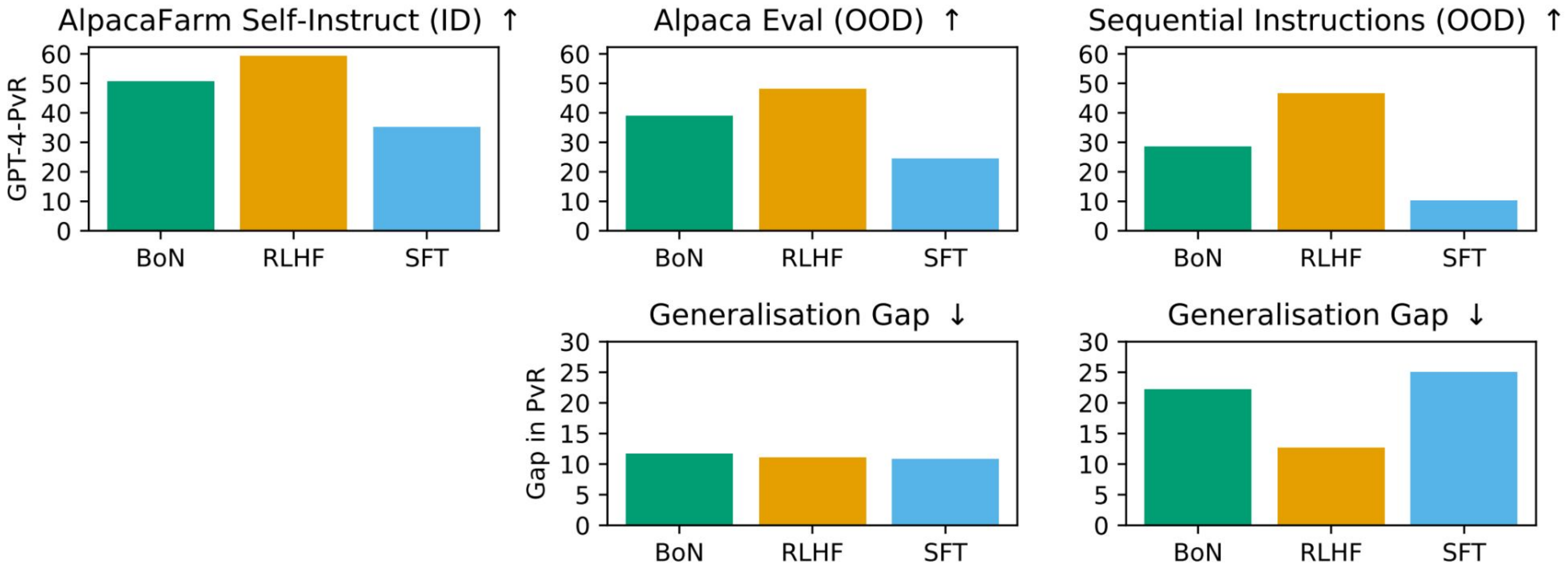


Figure 3: **Instruction Following Generalisation Results.** GPT-4 PvR for SFT, BoN and RL policies, based on LLaMa 7B, trained on the AlpacaFarm Self-Instruct instruction following task. ID is on AlpacaFarm Self-Instruct, OOD is on the AlpacaEval and Sequential Instructions datasets respectively, and generalisation gap is ID – OOD performance.

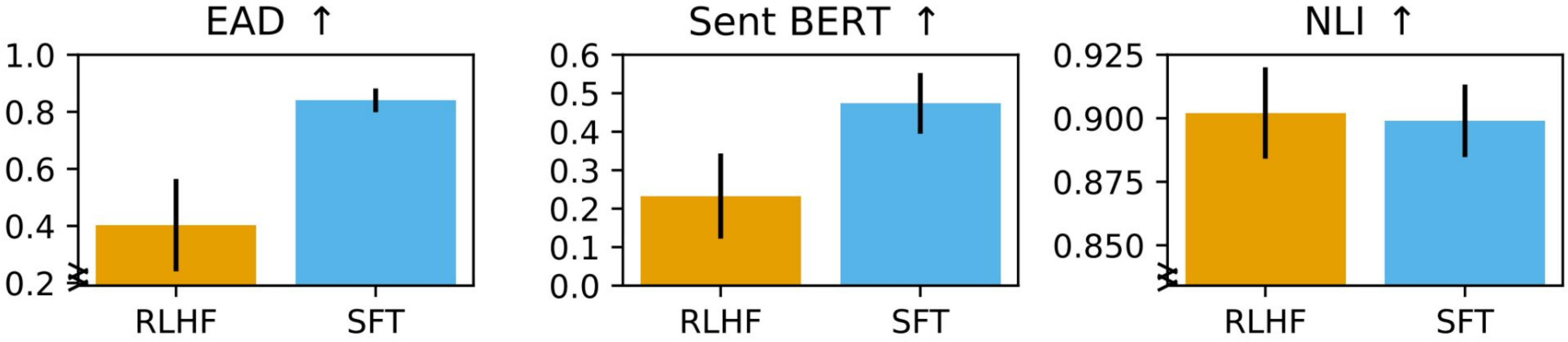


Figure 5: **Per-input diversity metrics for RLHF and SFT models**. For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Error bars are standard deviation of the per-input diversity score across different inputs. Note that some plots have broken y-axis for better visualisation.

Thank you!