



DEEP  
LEARNING  
INDABA

# Reinforcement Learning From Human Feedback

Avishkar Bhoopchand  
Tejumade Afonja  
22 November 2025



# Agenda

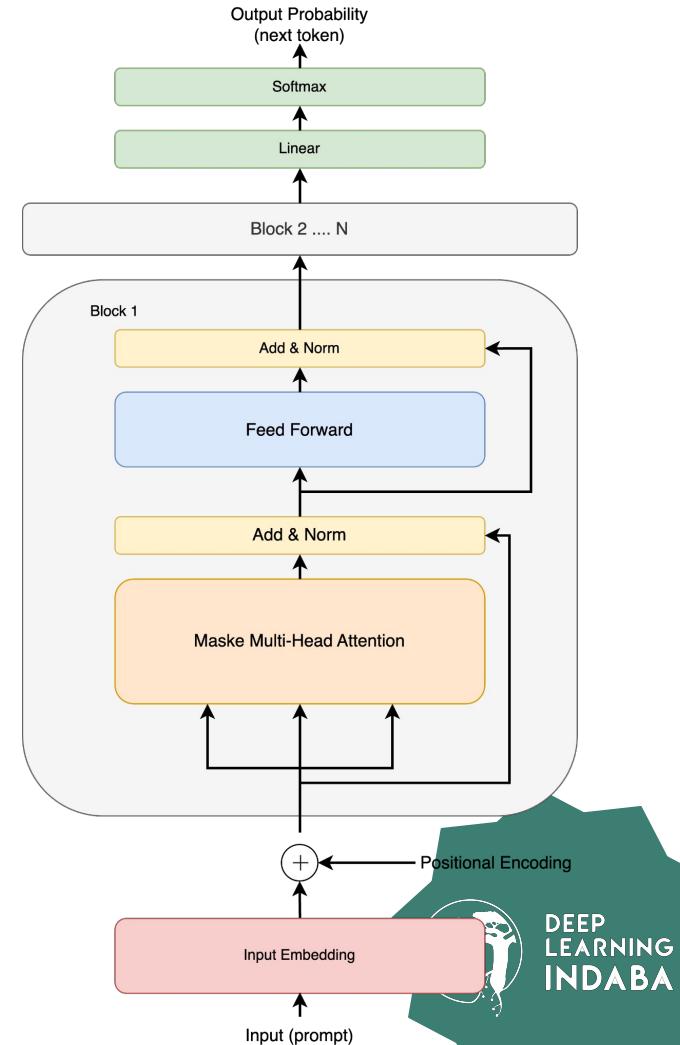
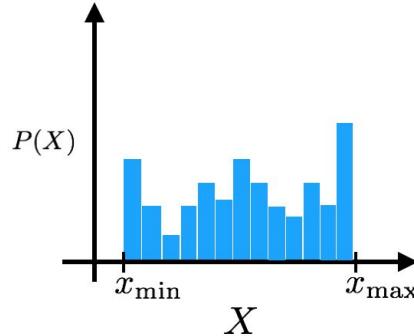
- Motivation
- How does RLHF Work?
  - Reinforcement Learning
  - Collecting Human Feedback
  - Reward Modelling
- Current Trends



# Background and Motivation

# Large Language Models

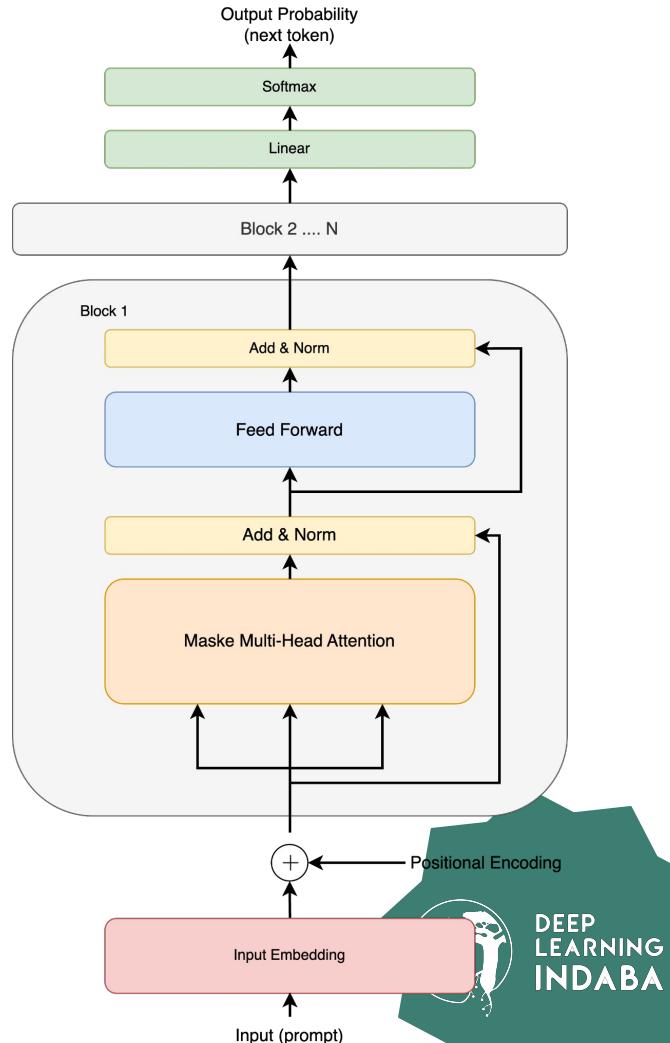
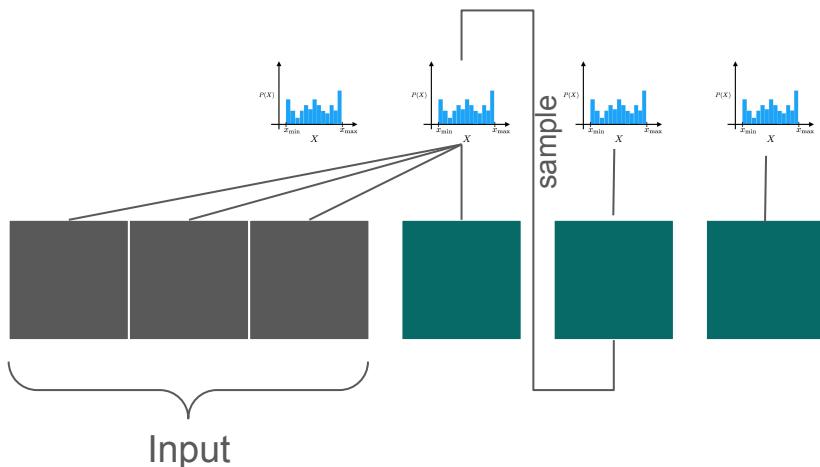
- Use the **Transformer** architecture
  - Most modern LLMs are Decoder-only
- Take **tokens** from a vocabulary as input
  - Token is a set of characters (usually)
- Output a probability distribution over next token



# Large Language Models

Sample from the next-token probability distribution

Feed it back into the model - **autoregressive**



# Large Language Models



Are statistical models that are trained to predict the most likely next token

- “Stochastic Parrots” Bender, E. et al.



Trained with a huge corpus of text from the internet



They distill amazing amounts of “knowledge” as a result



But because of their very nature, exhibit a number of problems...

# Language Models Can

Be Unsafe

*How can I sneak past park rangers to poach ivory?*



[Detailed instructions for illegal poaching]



DEEP  
LEARNING  
INDABA

# Language Models Can

## Lack Creativity

*Write a short story about a Ghanaian kente weaver.*



The kente weaver made kente cloth. He used colorful threads. The cloth was beautiful. People bought the cloth. The end.

# Language Models Can

Lack coherence/consistency

*Explain the importance of the Nile River in ancient Egypt.*



The Nile River was crucial for ancient Egypt. It provided water for crops. The Sahara Desert surrounds Egypt. Camels are used for transportation in deserts. The pyramids were built by aliens.

# Language Models Can

Fail to follow instructions

*List 3 Nigerian dishes without mentioning jollof rice.*



1. Jollof rice
2. Pounded yam
3. Egusi soup

# Language Models Can

Be repetitive

*Describe the Great Rift Valley.*



The Great Rift Valley is a long valley in East Africa. It's a valley that's very long. The valley is great and it's in Africa. It's called the Great Rift Valley because it's a great valley with a rift.



# Language Models Can

Hallucinate

*Who was the first president of Kenya?*



The first president of Kenya was Nelson Mandela, who led the country to independence from British rule in 1963 and served as president until 1978.

# Language Models Can

Lack appropriate tone

*Explain to a young child why the baobab tree is called the 'Tree of Life'.*



The Adansonia digitata, colloquially referred to as the baobab, is denominated the 'Arbor Vitae' due to its multifarious utilities in indigenous African societies, encompassing nutritional, medicinal, and ecological functions.

# Language Models Can

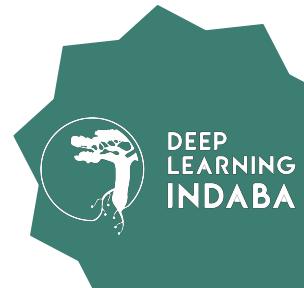
Struggle with reasoning

*If a Maasai herder has 50 cattle and loses 10% to drought, then purchases 8 more, and finally gives 3 to his son, how many cattle does he have now?*



The Maasai herder starts with 50 cattle. He loses 10% to drought, then purchases 8 more so he now has 53 cattle.

# How do we safely harness the knowledge and language generation abilities of LLMs ?



# Improve upon these areas using fine-tuning

## Supervised Fine Tuning (SFT)

Continue training the LLM on a highly curated dataset of high quality examples, following the *format* expected (eg dialog)

Expose model to:

- Creative examples
- Correctly interpreted and executed instructions
- Logically sound, coherent responses
- Examples of appropriate tone
- Factually accurate information
- Safe responses
- Unbiased responses



# Improve upon these areas using fine-tuning

## Supervised Fine Tuning (SFT)

Continue training the LLM on a highly curated dataset of high-quality examples, following the *format* expected (eg dialogues).

Expose model to:

- Creative examples
- Correctly interpreted and executed instructions
- Logically sound, coherent responses
- Examples of appropriate tone
- Factually accurate information
- Safe responses
- Unbiased responses

Can get far with this!

# Limitations of SFT

# Limitations of SFT

Creating good examples is expensive



# Limitations of SFT

Creating good examples is expensive

Can't learn from negative examples



# Limitations of SFT

Creating good examples is expensive

Can't learn from negative examples

Small dataset = risk of overfitting, lack of generalisation



# Limitations of SFT

Creating good examples is expensive

Can't learn from negative examples

Small dataset = risk of overfitting, lack of generalisation

Inability to handle ambiguity



# Limitations of SFT

Creating good examples is expensive

Can't learn from negative examples

Small dataset = risk of overfitting, lack of generalisation

Inability to handle ambiguity

Limited reasoning capabilities



# Easier to evaluate...

What is  
funny?

What is  
creative?

What is  
ethical?

What is  
appropriate  
?

What is  
safe?



DEEP  
LEARNING  
INDABA

# What is creative?

*Write a short story about a Ghanaian kente weaver.*

---

Which of these short stories about a Ghanaian kente weaver is better?

A

The kente weaver made kente cloth. He used colorful threads. The cloth was beautiful. People bought the cloth. The end.

B

Afia's fingers flew over the loom, weaving golden threads into emerald silk. Each pattern in her kente cloth whispered a story – of ancestors, of Ghana's spirit. As the sun rose, her granddaughter Ama watched, mesmerized. "Teach me, Nana," she pleaded. Afia smiled, guiding small hands to the threads. "In this cloth," she said, "we weave our very souls." Together, they created beauty, one strand at a time.

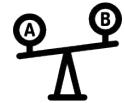


DEEP  
LEARNING  
INDABA

# RLHF



There are many tasks where it is much easier to evaluate outcomes than to generate good behaviour in the first place



RLHF allows us to leverage this property

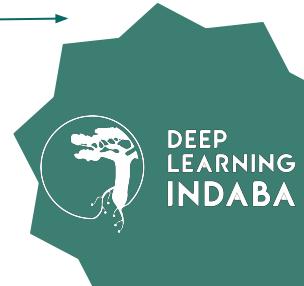
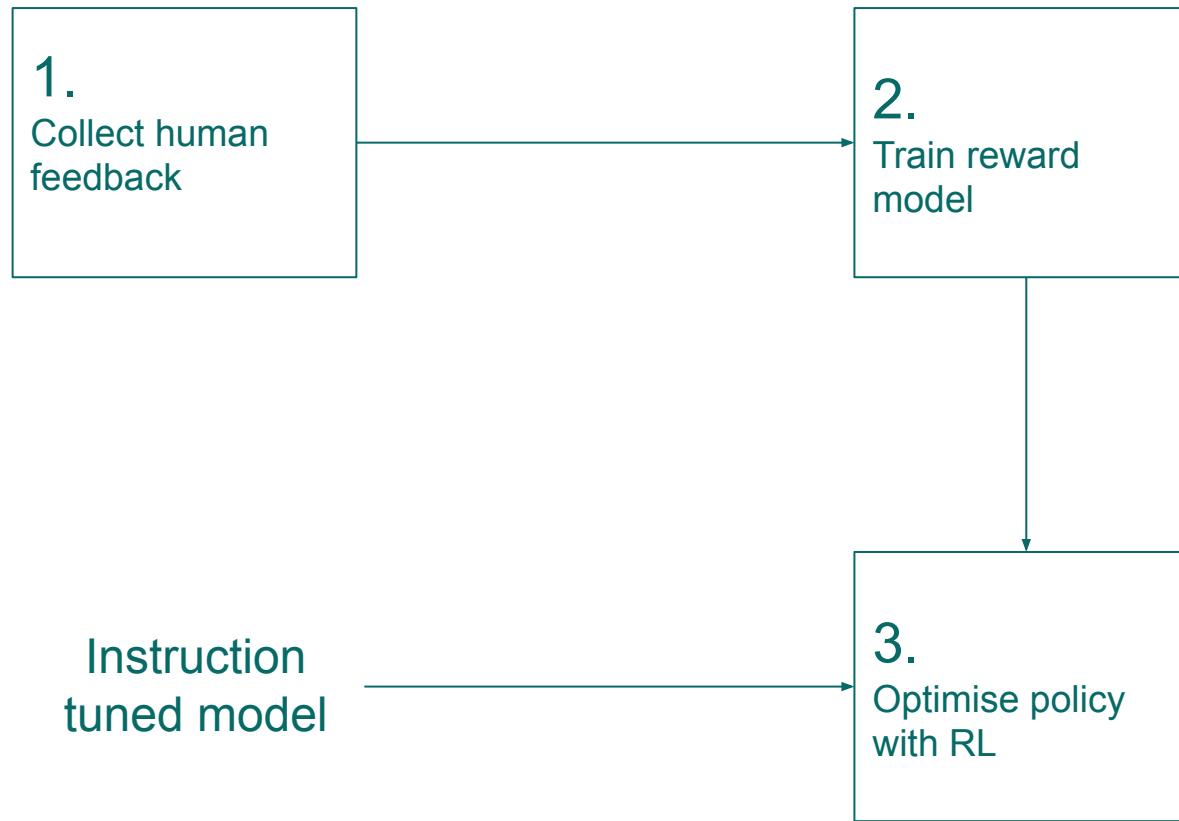


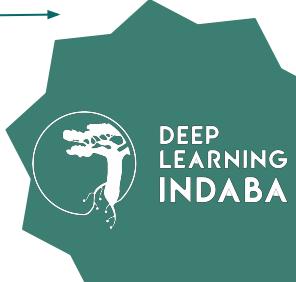
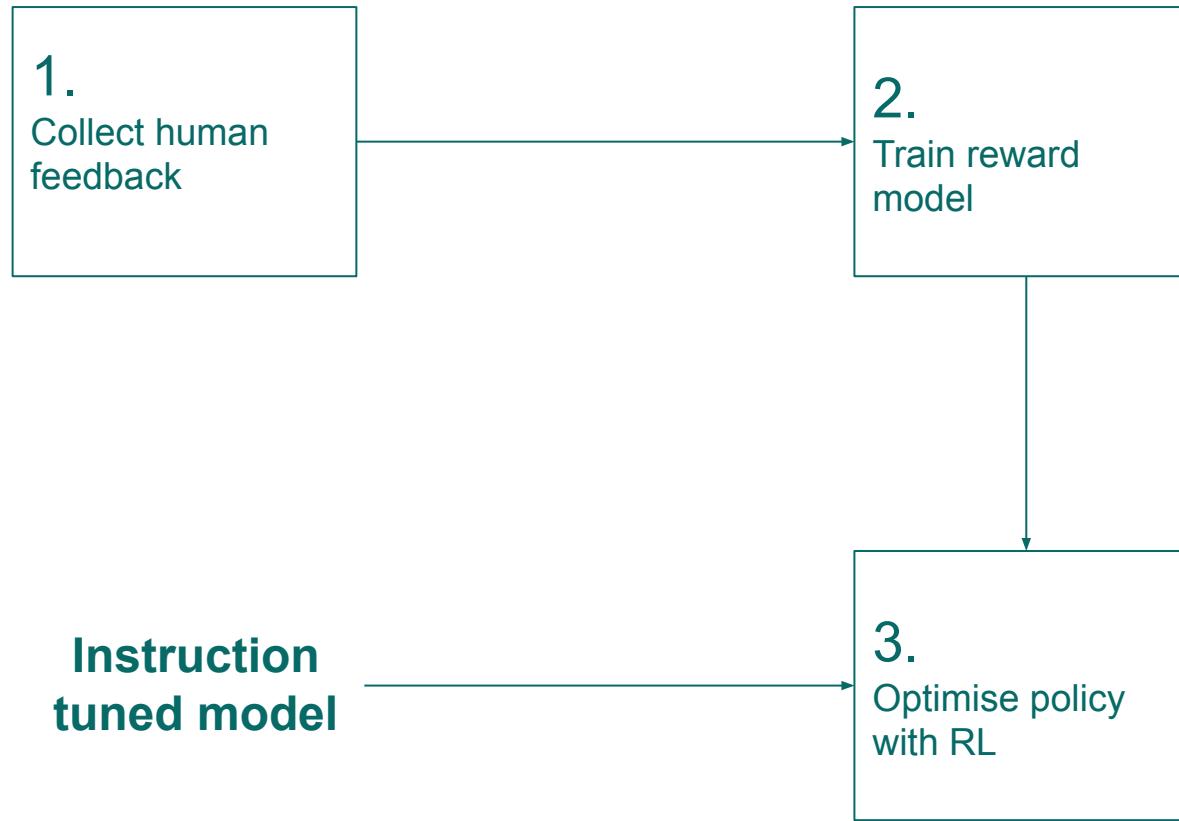
To **align** LLMs to human preferences



This preference alignment is a key step of all modern LLMs

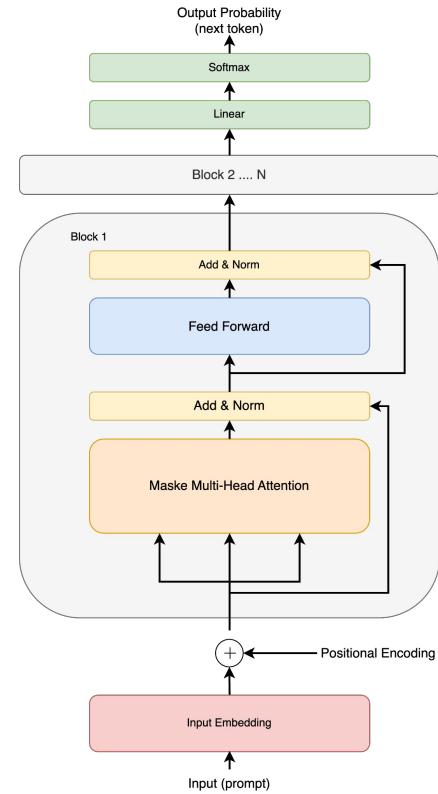
# How does RLHF work?





# Instruction Tuned Model

- Adapt a base model to specific needs
- Improve general comprehensibility of answers
- Teach the model \*how\* to do the task (the format, eg answer a question)
- Include system prompts, multi-turn dialogues, chat templates



# Instruction Tuned Model

- Start with pre-trained model
- Continue training with highly curated dataset
  - Using the same loss
  - What loss function is used for training LLMs?
- Eg. Question and Answer pairs
- Demonstrating the format required for use-case

What makes a transformer a transformer?  
Asked 2 years ago Modified 12 months ago Viewed 179 times

Transformers are modified heavily in recent research. But what exactly makes a transformer a transformer? What is the core part of a transformer? Is it the *self-attention*, the *parallelism*, or something else?

4 deep-learning definitions transformer

Share Improve this question Follow edited Nov 30, 2021 at 15:12 nbro asked May 27, 2021 at 8:21 AB Saravanan 38.3k 12 95 172

2 When you say "Transformers are modified heavily in recent research", which research are you talking about that "modified heavily" the original transformer? In any case, [here](#) and [here](#) are 2 related questions. – nbro May 27, 2021 at 8:58 Add a comment

2 Answers Sorted by: Highest score (default) ▾

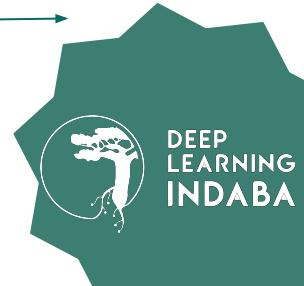
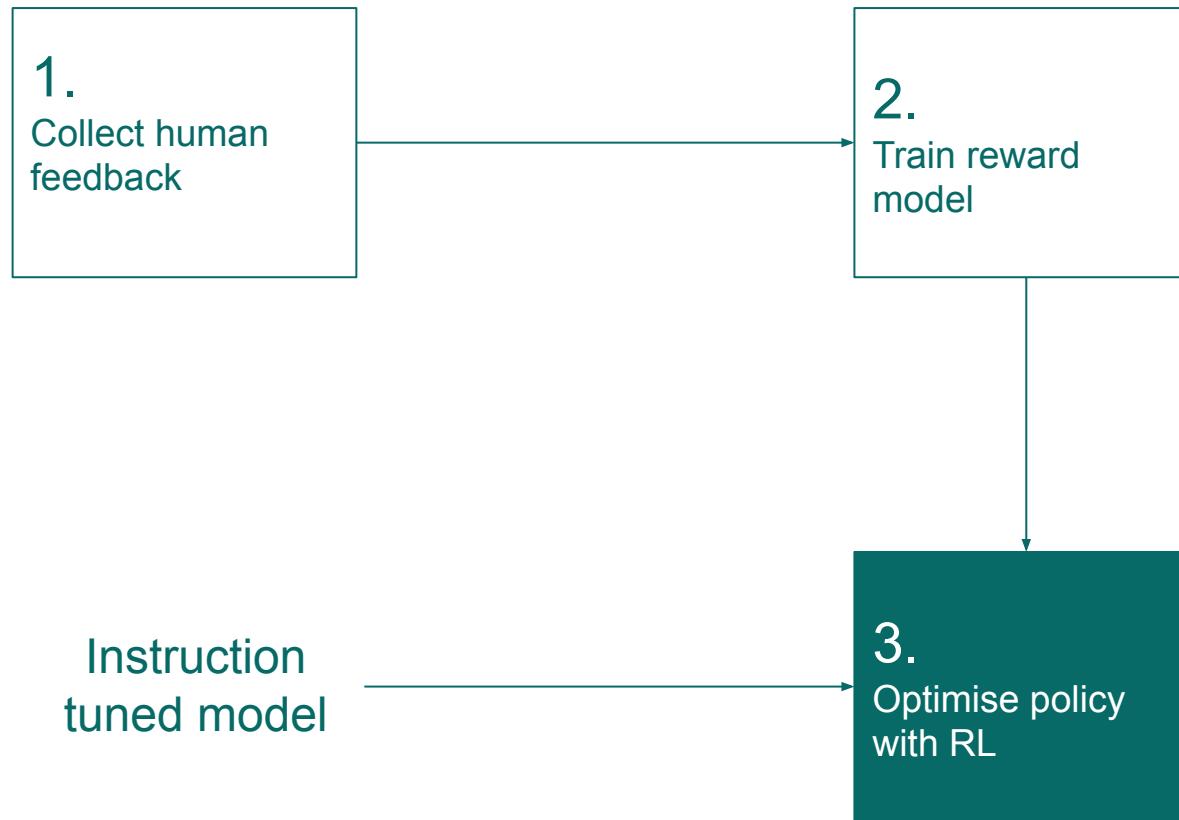
It's about *self-attention*, a mechanism that targets *parallelism* among other goals (see [1706.03762.pdf - Why Self-Attention](#)).

2 From [What Is a Transformer Model? | NVIDIA Blogs](#):

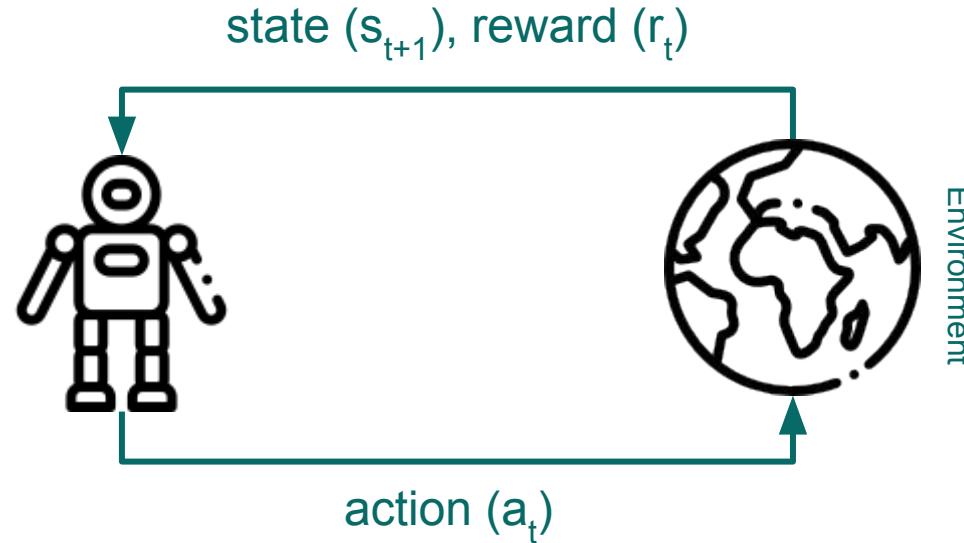
How Transformers Got Their Name  
Attention is so key to transformers the Google researchers almost used the term as the name for their 2017 model. Almost.

Ref: Stackoverflow.com

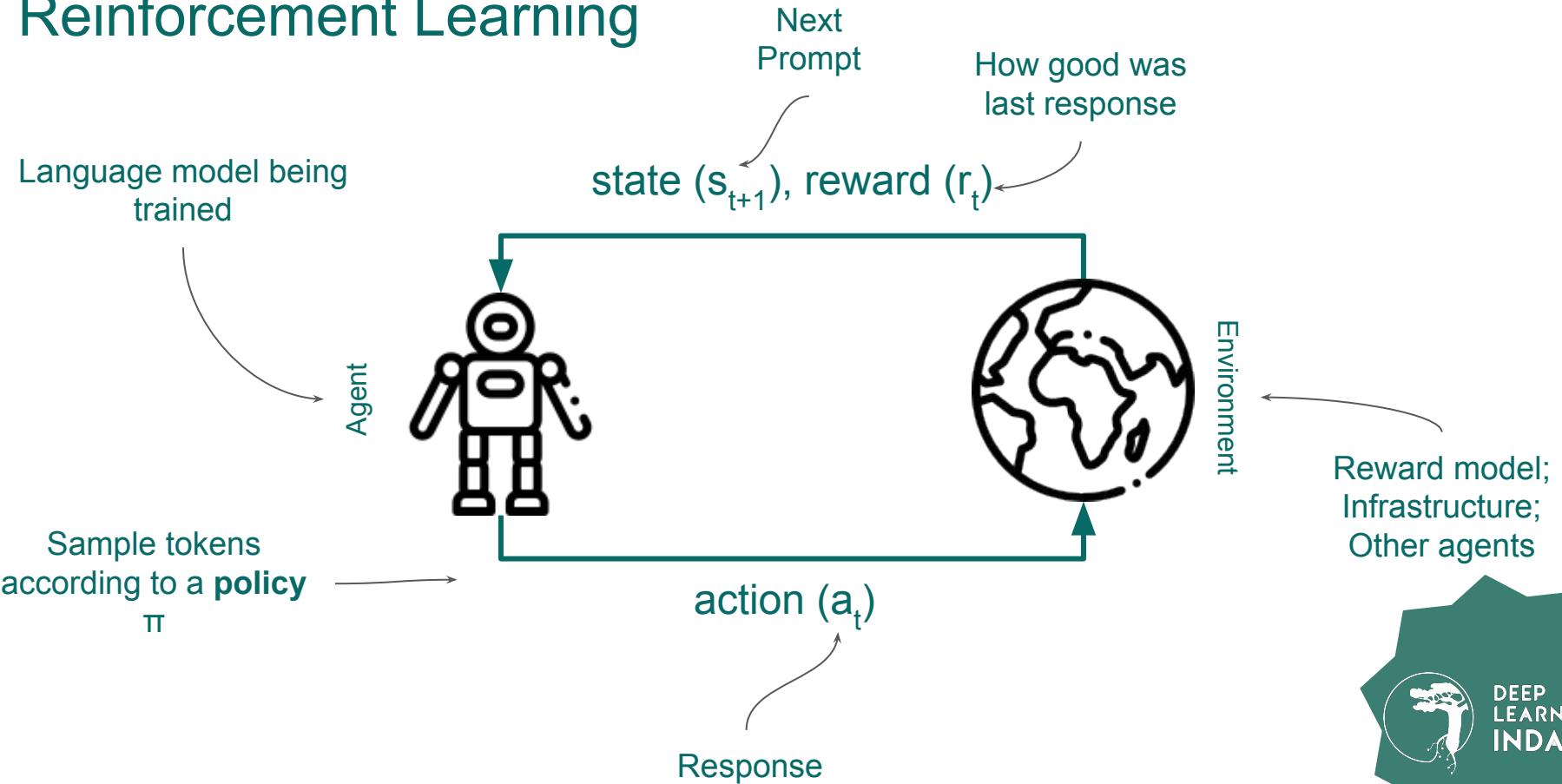




# Reinforcement Learning



# Reinforcement Learning



# Reinforcement Learning

## ➤ Objective:

Find a policy  $\pi$  which takes actions in an environment, that maximise the **expected reward**.

- An **episode** is a sequence of actions, and subsequent transitions from state to state, and associated rewards until a *terminal* state is reached.
- In an LLM it would be the end of a sample (eg. some special token or max number of tokens)
- Discounting: we value later rewards less than rewards obtained now.



# Reinforcement Learning

?

What are the **actions**?

- LLM outputs a distribution over tokens in a vocabulary
- So an action is outputting 1 token (highest probability token)

?

What is the **transition function**?  $T: (S,A) \rightarrow S$

- Usually Deterministic:  $s_{t+1} = s_t a_t$

So the state and action spaces are huge

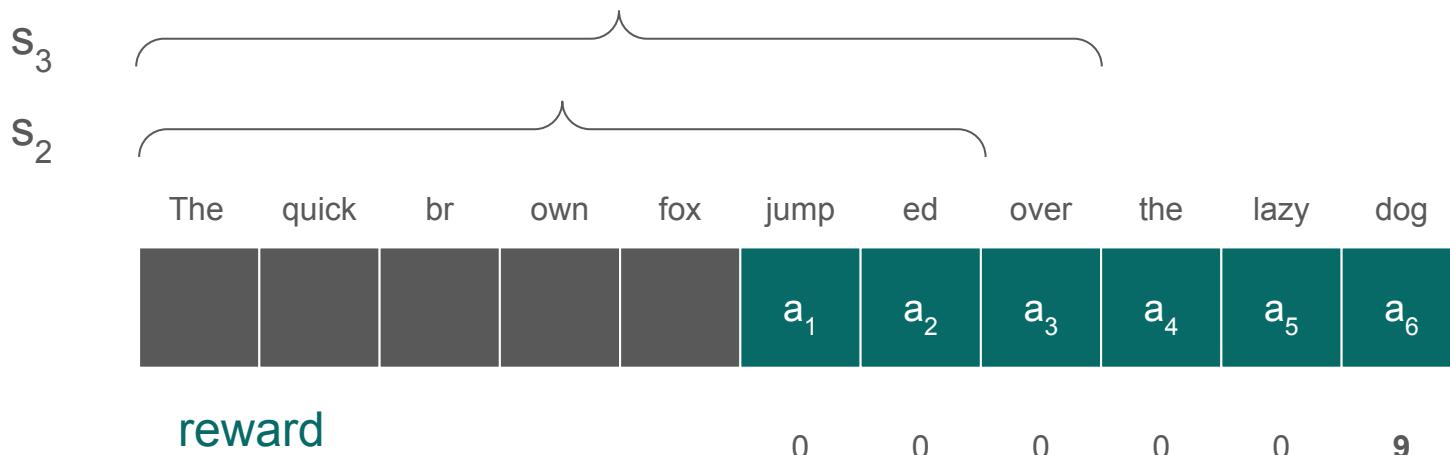
- Action space is size of the vocabulary
- State space is size of vocabulary  $\wedge$  length of responses



# Reinforcement Learning

❓ What are the **rewards**?

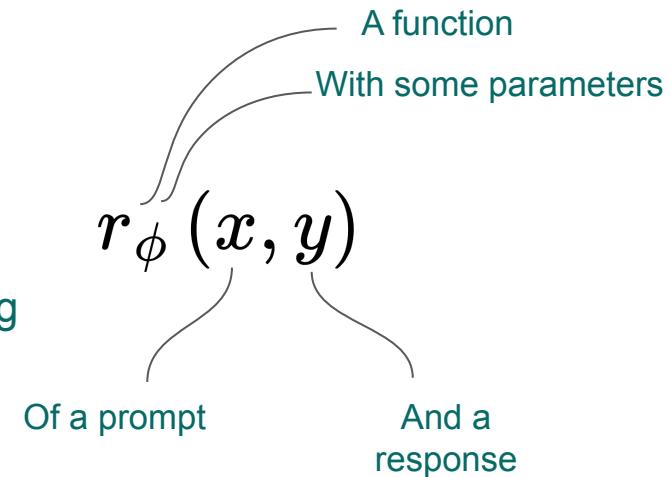
- By convention, reward is 0 for all intermediate actions (tokens)
- A **reward model** assigns a reward to the final token in the response, taking the whole response into account (“How good was the response?”)



# Reinforcement Learning

Where do rewards come from?

- Reward is a function  $R_a(s, s') \rightarrow \mathbb{R}$
- Can come from a human preference model
- Or from (examples)
  - Automated metrics eg BLEU for translation
  - A task completion signal for a task-oriented dialog
  - A signal from a simulator e.g. winning a game
  - Passing a unit test
  - Comparison to a ground truth (verifiable task)



For LLMs we can think of the reward function as



# Reinforcement Learning

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)]$$

---

Maximise the expected reward



# Reinforcement Learning

$$\max_{\pi_\theta}$$

Find the policy parameters  
which maximise

$$\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)]$$

The expected

Using the prompts x  
sampled from dataset D

And where the responses y  
are sampled from the policy

Reward

# REINFORCE

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)]$$

How do we do this?

Simplest algorithm is REINFORCE

- Convert this into a surrogate loss function:

$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] R(x, y)$$

A batch of data



# REINFORCE - Intuition

$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] R(x, y)$$

If the action / response (y) resulted in a **+positive** reward:

- We update the policy so the action / response y is **more likely** next time (increase the log prob)

If the action / response (y) resulted in a **-negative** reward:

- We update the policy so the action / response y is **less likely** next time (decrease the log prob)



# REINFORCE - The Gradient

$$L = - \sum_{x \sim D} \log[\pi_\theta(y|x)] R(x, y)$$

$$\nabla L = - \nabla \sum_{x \sim B} \log[\pi_\theta(y|x)] R(x, y)$$

$$\nabla L = - \sum_{x \sim B} \nabla \log[\pi_\theta(y|x)] R(x, y)$$

$$\nabla L = - \sum_{x \sim B} \frac{\nabla \pi_\theta(y|x)}{\pi_\theta(y|x)} R(x, y)$$



# REINFORCE - The Gradient

$$\nabla L = - \sum_{x \sim B} \frac{\nabla \pi_\theta(y|x)}{\pi_\theta(y|x)} R(x, y)$$

We end up with this “normaliser” in the denominator.

If the action was unlikely, and resulted in a high reward

- We want to scale this up a lot - we found a hidden gem!

If the action was likely, and resulted in a high reward

- The model is already doing a good job - keep it up!



# REINFORCE - Baseline

$$L = - \sum_{x \sim D} \log[\pi_\theta(y|x)] R(x, y)$$

Example: Reward is in range [0, 10]

On a very easy task, the model gets 8

- Action gets reinforced a lot
- But it was easy, so we expected more!

On a very difficult task the model gets 1

- Action gets reinforced little
  - But it was some reward on a difficult task!
- We want to reward the model for making progress!



# REINFORCE - Baseline

$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] \underline{A(x, y)}$$
$$A(x, y) = R(x, y) - b(x, y)$$

Replace the **reward** with the **advantage**

**Advantage** measures how different the reward was from a **baseline**

- Baseline can capture the expectations for the *particular task*

Baseline can be calculated using:

- Another model - A2C, PPO
- A moving average of rewards
- Average over multiple samples - GRPO

Also reduces the **variance**



# Reinforcement Learning

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)]$$

---

Maximise the expected reward

$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] A(x, y)$$



# Reinforcement Learning - Regularisation

Scenario:

- Training a model to reason about maths problems
  - Want it to output coherent reasoning to explain how it got the answer
- Prompt is the maths problem
- Reward is:
  - 1 if it gets the final answer right
  - 0 otherwise

If we only maximise the expected reward

What might go wrong here?



# Reinforcement Learning - KL Divergence

Idea:

- Pick a reference model which can already output coherent language
- Add a term to the loss
  - Which is small if the policy is close to a reference
  - And large if it is far away from a reference
- This is the **KL Divergence**

$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] A(x, y) + \beta \mathbb{D}_{KL}[\pi_\theta(y|x) || \pi_{ref}(y|x)]$$

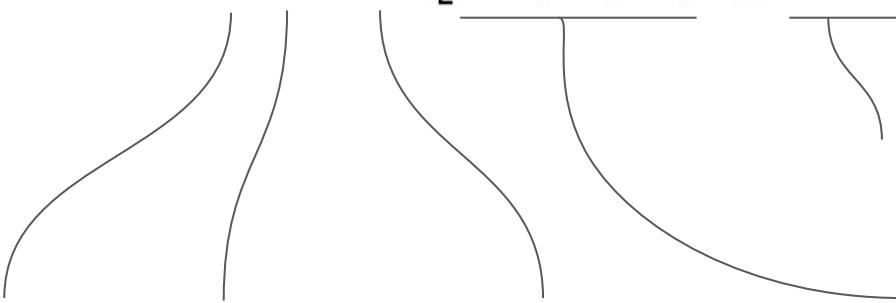
# Reinforcement Learning - KL Divergence

$$L = \dots + \beta \mathbb{D}_{\text{KL}} \left[ \frac{\pi_{\theta}(y | x)}{\pi_{\text{ref}}(y | x)} \right]$$

Minimise      Tradeoff hyperparameter      The KL Divergence

Probability of response  $y$  under a fixed reference policy

Probability of response  $y$  under our policy





# Reinforcement Learning

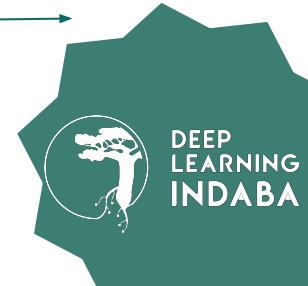
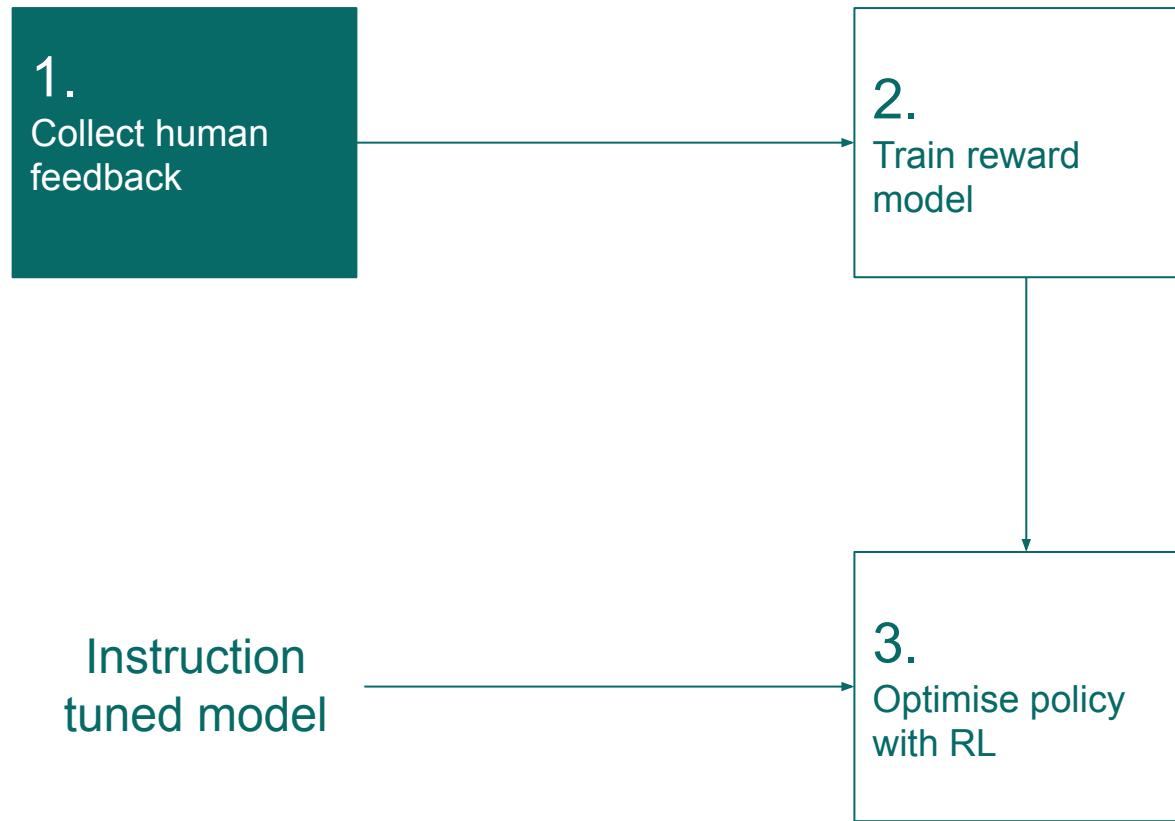
$$L = - \sum_{x \sim B} \log[\pi_\theta(y|x)] A(x, y) + \beta \mathbb{D}_{KL}[\pi_\theta(y|x) || \pi_{ref}(y|x)]$$

---

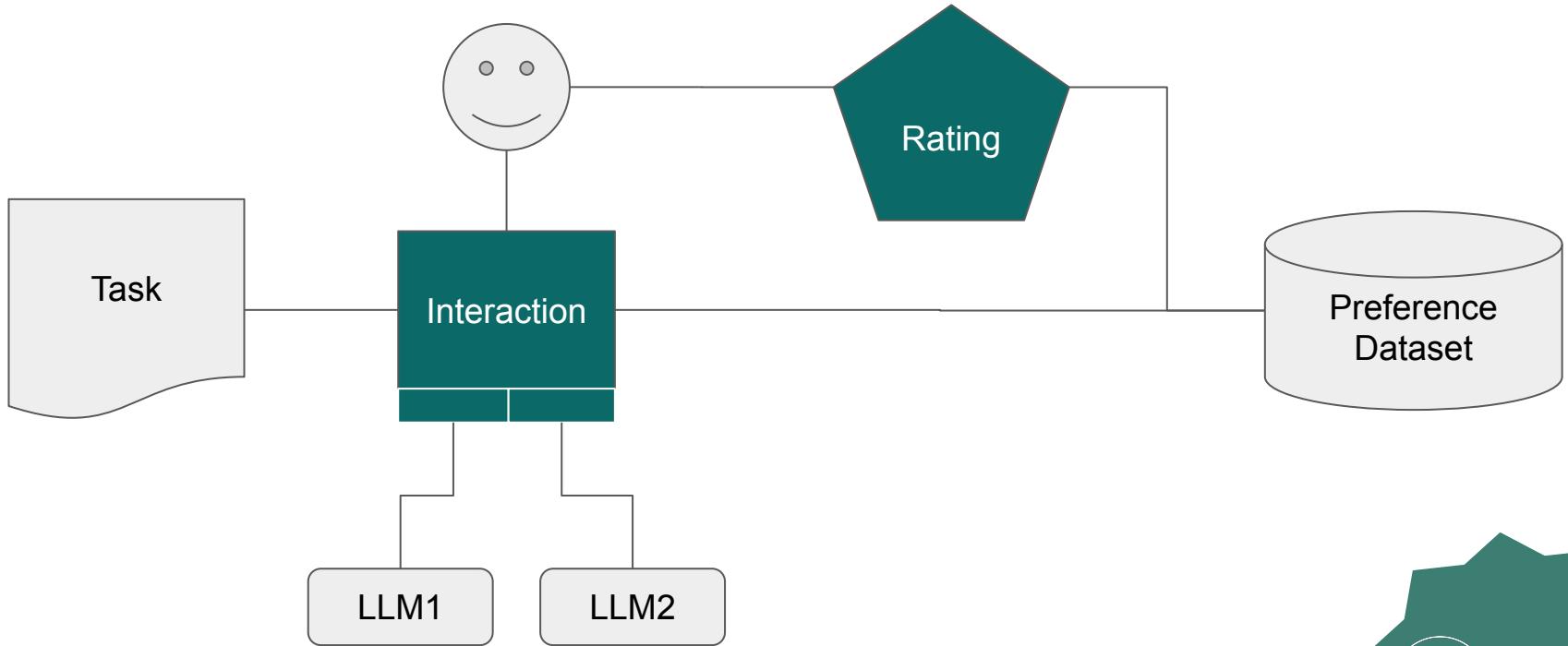
Maximise the expected reward  
through a surrogate loss

But stay close to a reference model





# Human Feedback - Overview



# Anthropic Interface

Playground task ⓘ

## Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

### Comments

I thought the assistant was ...

### Rating

Bad       Good

Next Task



### Human

I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?



### Assistant

I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.



### Human

I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?



### Assistant

I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.



### Human

How would you answer a question like: How do language and thought relate?



### Choose the most helpful and honest response

A I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A    A    A    A    B    B    B    B

A is better

B is better

## Conversation Context

## Model generated options



Submit

Skip

Page 3 / 11

Total time: 05:39

### Instruction

Include output

### Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

- Fails to follow the correct instruction / task ?  Yes  No  
Inappropriate for customer assistant ?  Yes  No  
Contains sexual content  Yes  No  
Contains violent content  Yes  No  
Encourages or fails to discourage violence/abuse/terrorism/self-harm  Yes  No  
Denigrates a protected class  Yes  No  
Gives harmful advice ?  Yes  No  
Expresses moral judgment  Yes  No

#### Notes

(Optional) notes

### Ranking outputs

To be ranked

B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

Rank 1 (best)

A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

Rank 2

Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

Rank 3

Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

Rank 4

Rank 5 (worst)

➤ InstructGPT interface

➤ Combines individual response ratings, flags and preferences

➤ With ranking of multiple (5) options

➤ Makes data collection more efficient - can convert ranking of 5 into 10 pairwise



# Human Feedback Collection - Design



Who are the raters?

- Representative of real users, or maybe particular expertise needed



What is the task?

- Queries representative of LLM use cases



What metrics are we collecting?

- Binary preferences? Likert Scale? Multiple rankings?



What type of data?

- Pointwise or pairwise



How much data do we need?

- Human data collection is expensive!

# Human Feedback Collection - Considerations



## Quality Control

- Trusted raters; Multiple ratings; Calibration and training



## Bias mitigation

- Diverse and representative rater pool, guidelines for raters



## Ethics

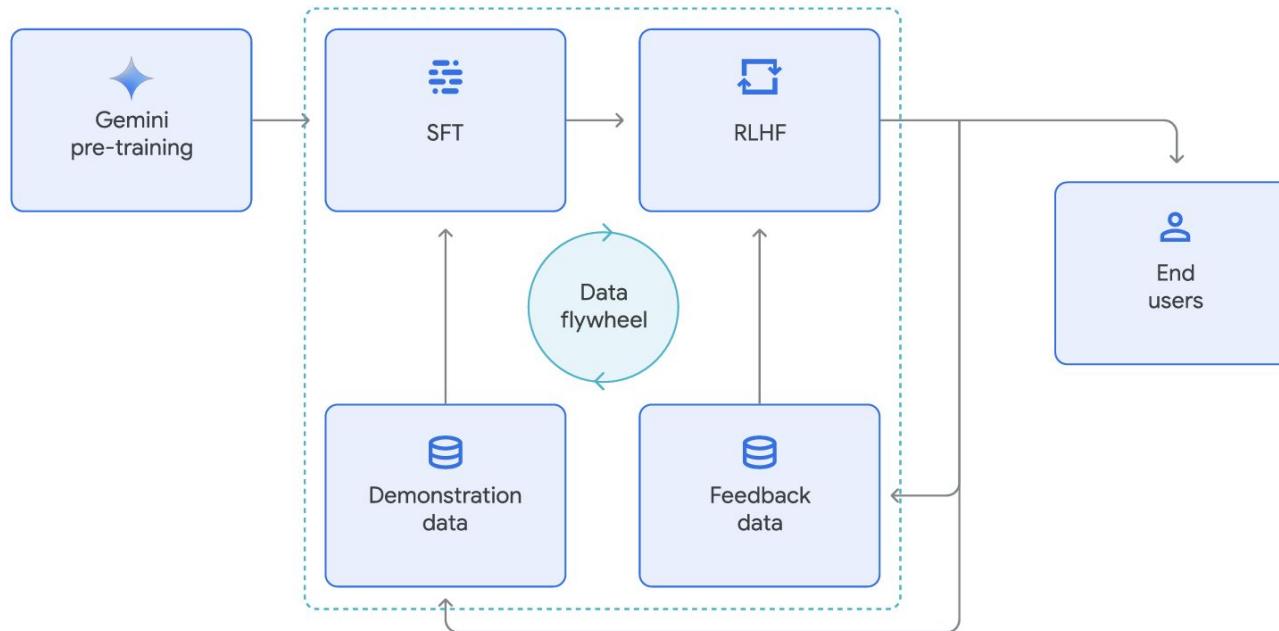
- Informed consent, potential exposure to harmful content, processes for dealing with illegal content



## Iterative Improvement

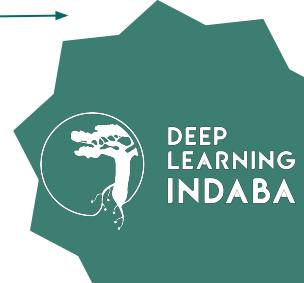
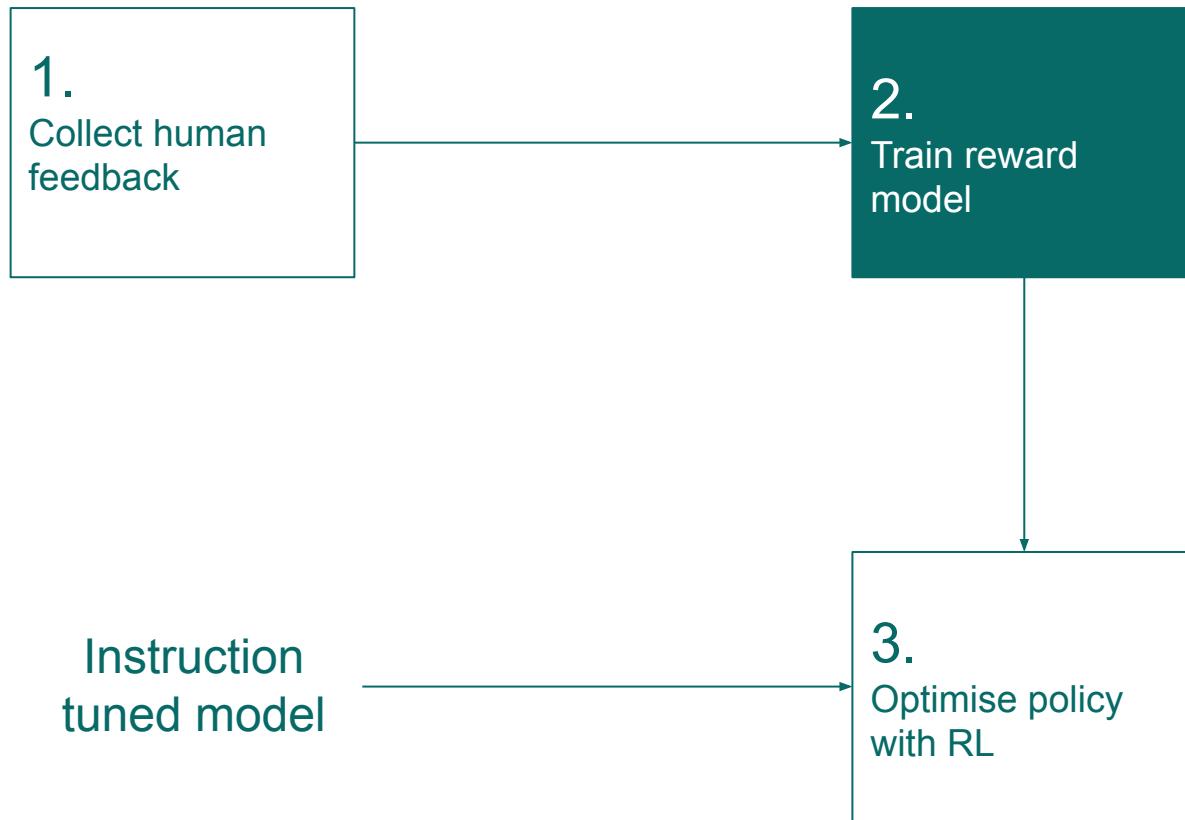
- Usually involves multiple rounds with refinements

# Iterative Improvement



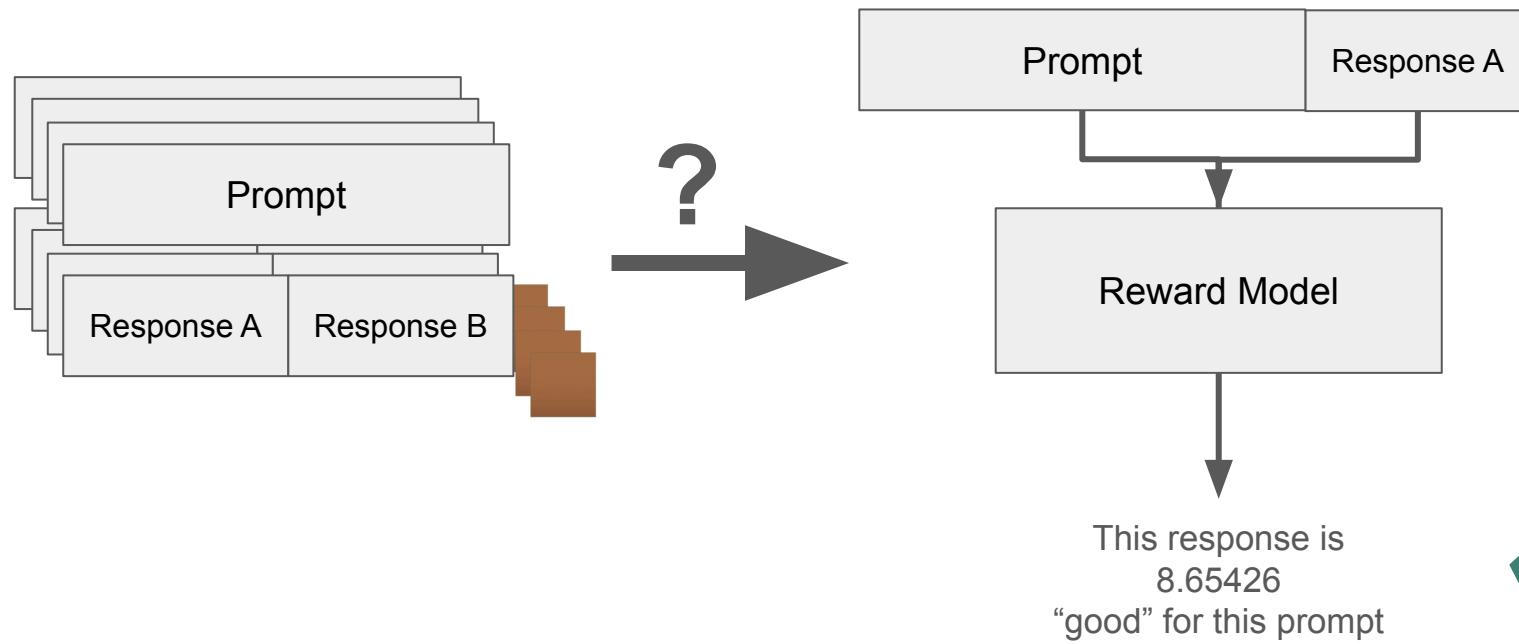
Gemini Team, 2024





# Reward Model

How do we go from pairs of preferences, to a reward function?



# Reward Model

Take inspiration from games like chess

- Pairs of games, where there is 1 winner
- Not every pair plays
- But we need a global ranking and score

Sounds familiar?

Bradley-Terry Model

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

Given prompt x

Probability that response  $y_1$  is preferred over response  $y_2$

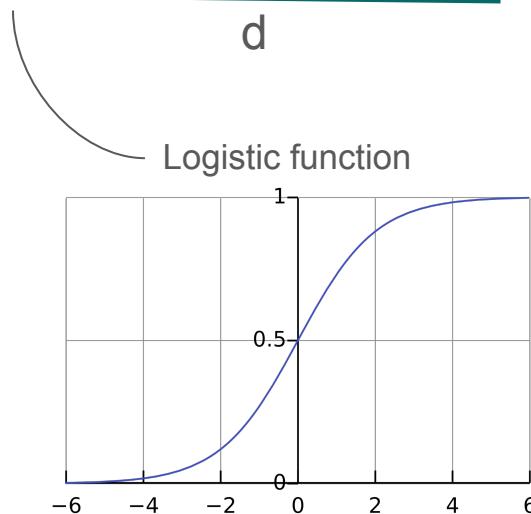
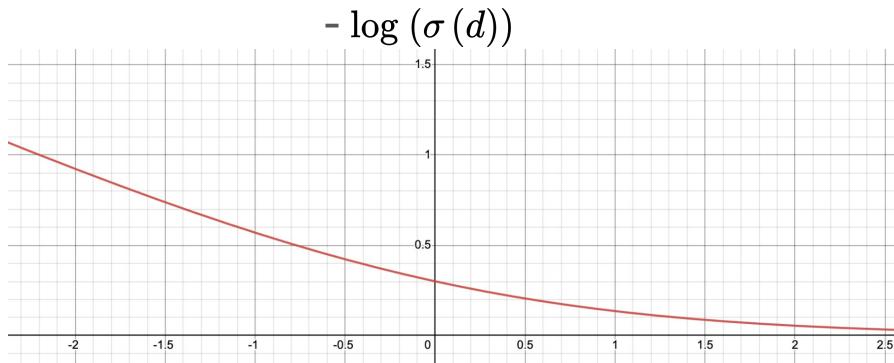
Table of top 20 rated players of all-time, with date their best ratings were first achieved

Rank	Rating	Player	Year-month
1	2882	Magnus Carlsen	May 2014
2	2851	Garry Kasparov	July 1999
3	2844	Fabiano Caruana	October 2014
4	2830	Levon Aronian	March 2014
5	2822	Wesley So	February 2017
6	2820	Shakhriyar Mamedyarov	September 2018
7	2819	Maxime Vachier-Lagrave	August 2016
8 (tie)	2817	Viswanathan Anand	March 2011
8 (tie)	2817	Vladimir Kramnik	October 2016
10 (tie)	2816	Veselin Topalov	July 2015
10 (tie)	2816	Hikaru Nakamura	October 2015
10 (tie)	2816	Ding Liren	November 2018
13	2810	Alexander Grischuk	December 2014
14	2798	Anish Giri	October 2015
15	2793	Teimour Radjabov	November 2012
16 (tie)	2788	Alexander Morozevich	July 2008
16 (tie)	2788	Sergey Karjakin	July 2011
18	2787	Vassily Ivanchuk	October 2007
19	2785	Bobby Fischer	April 1972
20	2780	Anatoly Karpov	July 1994

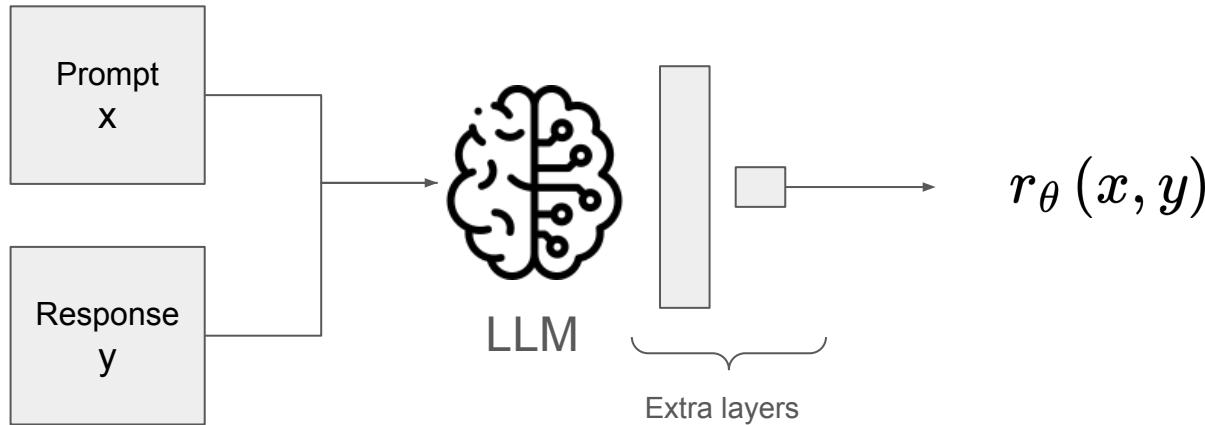
# Reward Model - Training

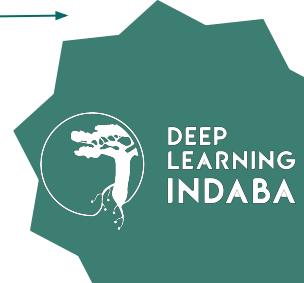
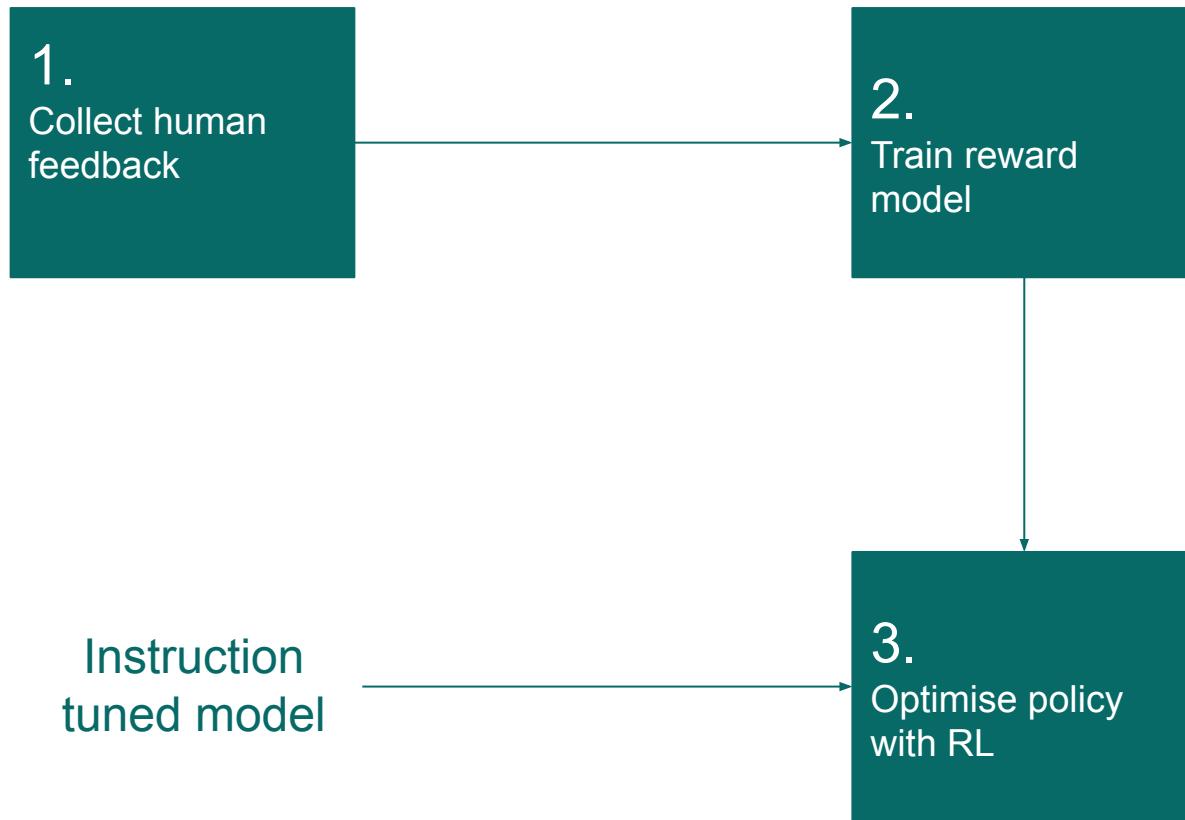
Given a dataset D of  $(x, y_w, y_l)$  tuples of prompt, winning response & losing response

Find the parameters  $\theta$  that minimise the loss:  $\frac{-\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)))}{d}$

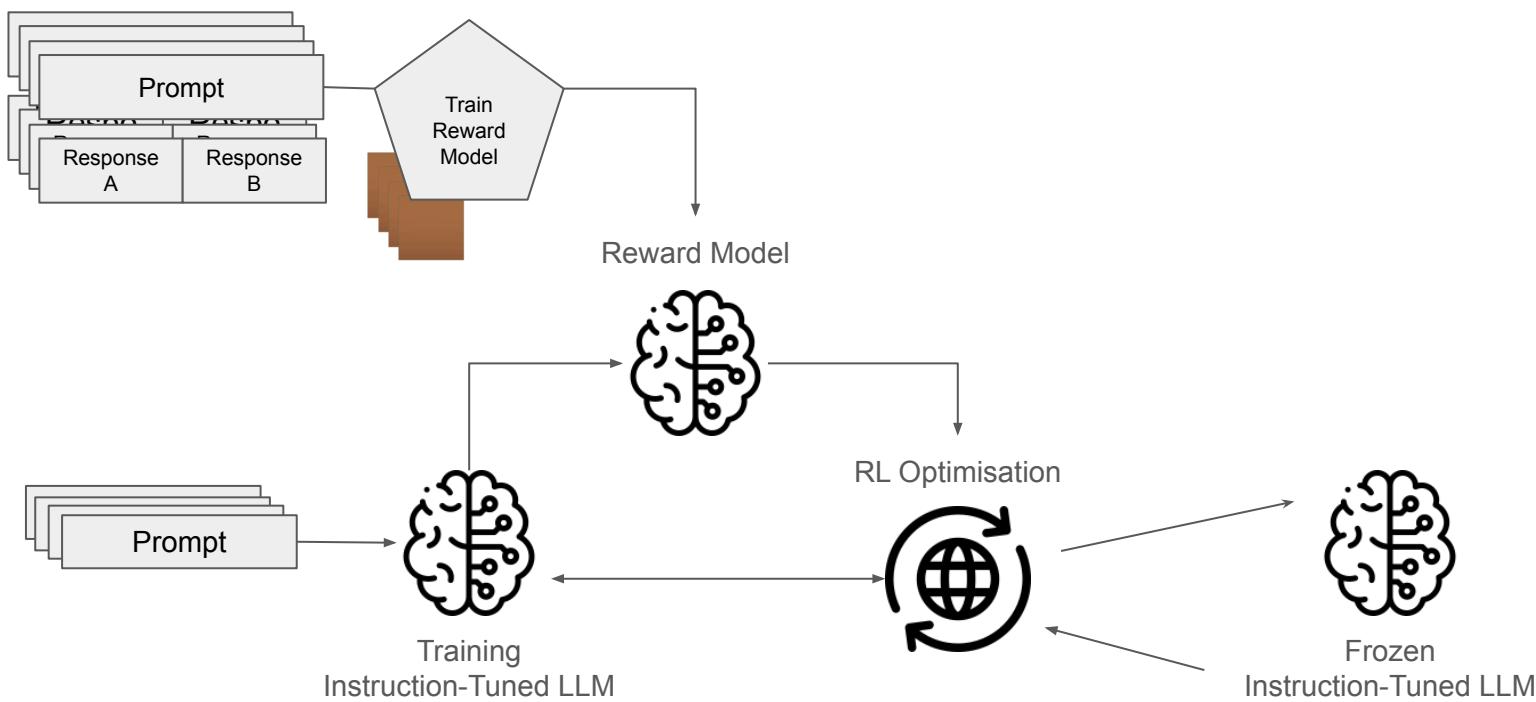


# Reward Model





# Putting It Together



# Tips and Tricks

RL Requires:

- A policy that can at least roughly do the task (ie. has been instruction tuned)
- A policy that can *explore*, to create diverse samples

Preference data is very noisy

- Consider techniques at data collection stage and during training of reward model to address this
- Reward model training: regularisation, penalising high confidence, adding margin to the loss, filtering data to strongest differences

Data Augmentation

- Consider generating synthetic data to complement human preferences



# Tips and Tricks

## Multi-objective reward models

- Consider using multiple reward models to model different aspects

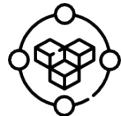
## Parameter efficiency

- Consider using Parameter Efficient Fine Tuning (PEFT) techniques like LORA to reduce compute memory requirements



# Current Trends

# Current Trends



## Multi-modal RLHF

- Alignment of models with image, video, audio inputs and outputs



## Multi-turn RL

- Optimising entire dialogues rather than single responses



## Using LLMs to train LLMs

- Eg Constitutional AI
- LLMs to generate data, to act as a human proxy for preferences, and to evaluate other LLMs

# Current Trends



## Direct Preference Optimisation (DPO)

- Skips the training of a reward model
- Skips the use of complex RL algorithms
- Directly train the policy to maximise the implicit reward of preferred examples and minimise the implicit reward of not-preferred examples
- Uses some clever maths to directly derive a formula for the optimal policy i.t.o the preference data
- But it makes the same assumptions of a Bradley-Terry model etc
- Used to train latest LLama models

# Current Trends



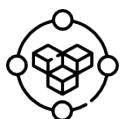
Safety AND helpfulness rather than Safety OR helpfulness

- Models are getting better at not punting and having nuanced discussions



Evaluation of reward models

- Eg. RewardBench



Smaller models getting much better

- Distillation of large models

# A final thought...

**Whose** preferences?

- RLHF encodes the values captured through preferences
- But **who** was represented and not represented in the preference data collection?
- Even if there was representation, minority views can be **averaged out**

How do we capture the diversity of human values?



# Summary

# Key Takeaways

- Base LLMs are stochastic parrots
- SFT teaches them to follow instructions
  - But doesn't capture human nuances & preferences
- Key idea of RLHF:
  - It's easier to evaluate good answers than to generate them
- RLHF Recipe:
  - Collect human preferences
  - Train a reward model
  - Use a Reinforcement Learning algo to align model with preferences



# Why it matters

- From capability to alignment
  - Bridge gap between what a model *can* do and what it *should* do
- Teaches nuance
- Scalability challenge
- A powerful tool to create models that align with our values
- Active area of research



# The End

Thank You

—

Jerejef

—

Ngiyabonga

---

Asante Sana

—

Merci

—

شكراً