

Reward modeling and human preferences

In principle, we could fine-tune the model using RLHF directly with the human annotations. However, this would require us to send some samples to humans for rating after each optimization iteration. This is expensive and slow due to the number of training samples needed for convergence and the inherent latency of human reading and annotator speed.

A trick that works well instead of direct feedback is **training a reward model on human annotations collected before the RL loop**. The goal of the reward model is to imitate how a human would rate a text. There are several possible strategies to build a reward model: the most straightforward **way would be to predict the annotation** (e.g. a rating score or a binary value for “good”/“bad”). In practice, what works better is to predict the ranking of two examples, where the reward model is presented with two candidates (y_k, y_j) for a given prompt x and has to predict which one would be rated higher by a human annotator.

This can be translated into the following loss function:

$$\text{loss}(\theta) = -E_{(x,y_j,y_k) \sim D} [\log (\sigma (r_\theta(x, y_j) - r_\theta(x, y_k)))]$$

where r is the model’s score and y_j is the preferred candidate.

Reinforcement Learning from Human Feedback

With the fine-tuned language model and the reward model at hand, we are now ready to run the RL loop. It follows roughly three steps:

1. Generate responses from prompts
2. Rate the responses with the reward model
3. Run a reinforcement learning policy-optimization step with the ratings

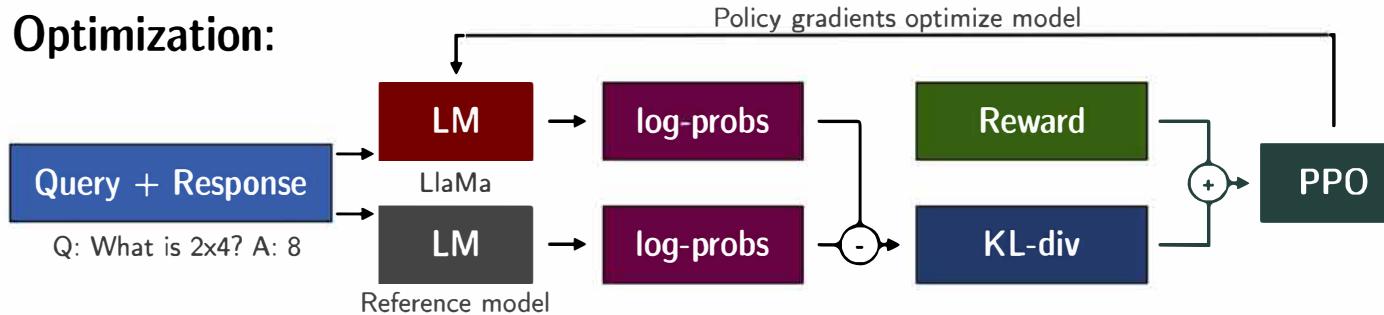
Rollout:



Evaluation:



Optimization:



The same template was used for SFT, RM and RLHF stages.

A common issue with training the language model with RL is that the model can learn to exploit the reward model by generating complete gibberish, which causes the reward model to assign high rewards. To balance this, we add a penalty to the reward: we keep a reference of the model that we don't train and compare the new model's generation to the reference one by computing the KL-divergence:

$$R(x, y) = r(x, y) - \beta \text{KL}(x, y)$$

where r is the reward from the reward model and $\text{KL}(x, y)$ is the KL-divergence between the current policy and the reference model.

which is imperfect and given the chance, the PPO algorithm will exploit these imperfections.

This can manifest itself as sudden increases in reward, however when we look at the text generations from the policy, they mostly contain repetitions of the string ` `` , as the reward model found the stack exchange answers containing blocks of code usually rank higher than ones without it. Fortunately this issue was observed fairly rarely and in general the KL penalty should counteract such exploits.

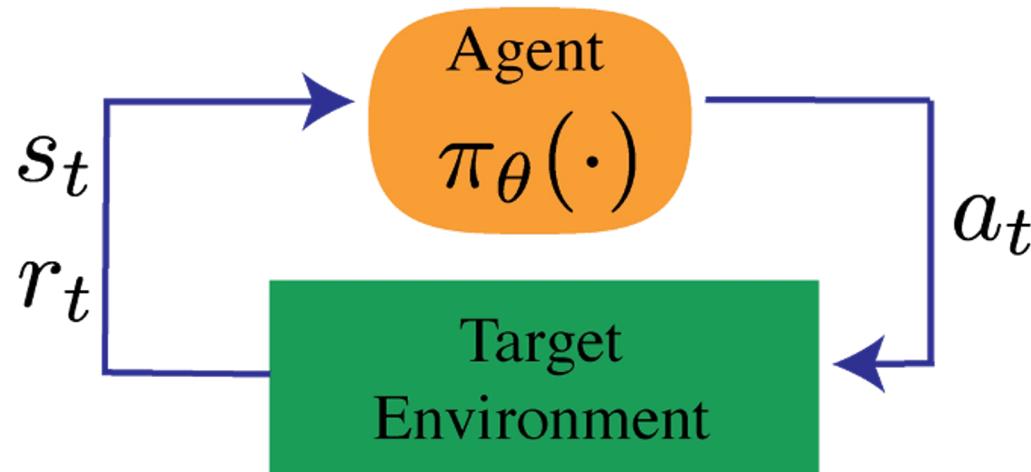
KL is always a positive value, isn't it?

As we previously mentioned, a KL penalty term is used in order to push the model's outputs remain close to that of the base policy. In general, KL divergence measures the distances between two distributions and is always a positive quantity. However, in tr1 we use an estimate of the KL which in expectation is equal to the real KL divergence.

$$KL_{pen}(x, y) = \log \left(\pi_{\phi}^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x) \right)$$

Clearly, when a token is sampled from the policy which has a lower probability than the SFT model, this will lead to a negative KL penalty, but on average it will be positive otherwise you wouldn't be properly sampling from the policy. However, some generation strategies can force some tokens to be generated or some tokens can suppressed. For example when generating in batches finished sequences are padded and when setting a minimum length the EOS token is suppressed. The model can assign very high or low probabilities to those tokens which leads to negative KL. As the PPO algorithm optimizes for reward, it will chase after these negative penalties, leading to instabilities.

Review: reinforcement learning basics



Some notation:

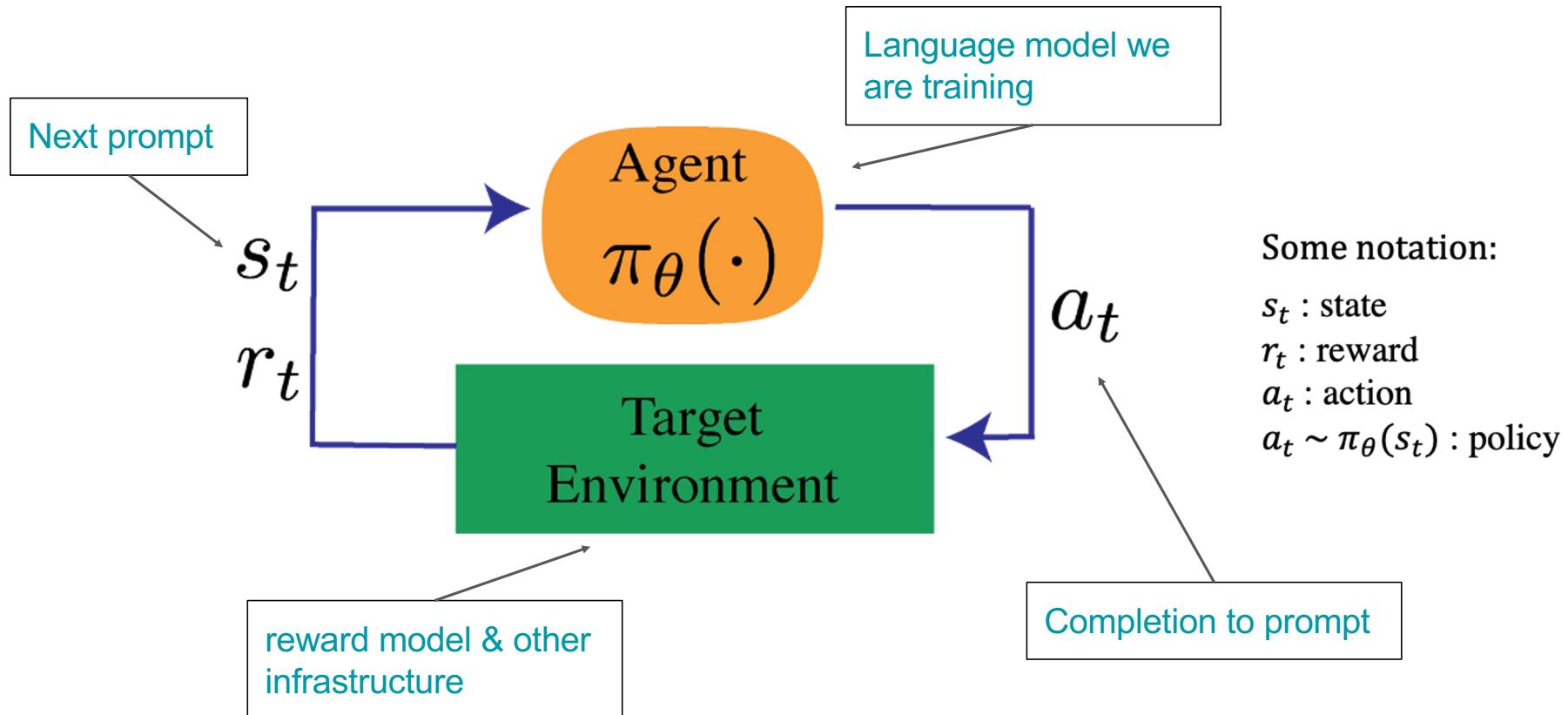
s_t : state

r_t : reward

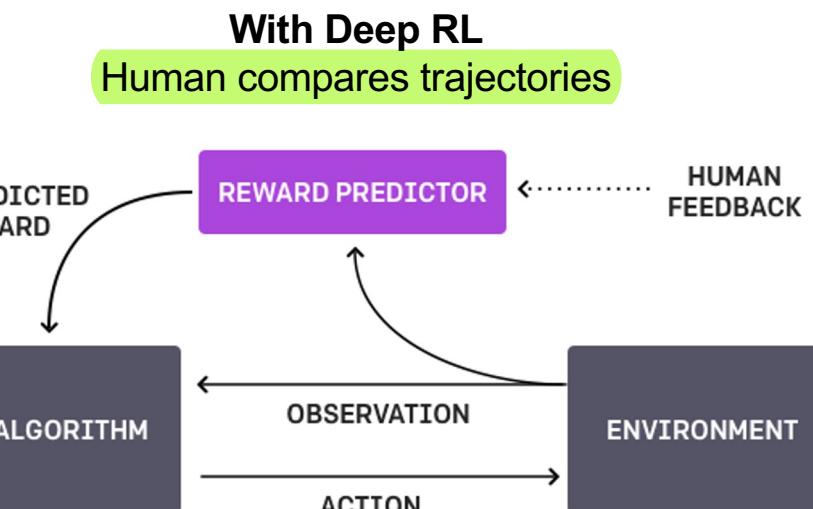
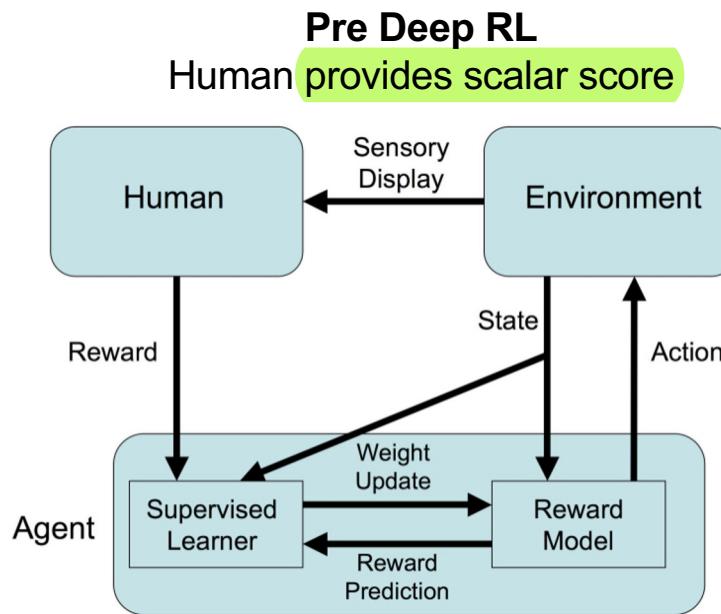
a_t : action

$a_t \sim \pi_\theta(s_t)$: policy

Review: reinforcement learning basics in language



History: RLHF for decision making



Christiano, Paul F., et al. "Deep reinforcement learning from human preferences." 2017.

Knox, W. Bradley, and Peter Stone. "Tamer: Training an agent manually via evaluative reinforcement." 2008.

History: preference models, alignment, and agents

Nov. 2018

Propose learning preference models based on two assumptions:

History: preference models, alignment, and agents

Nov. 2018

Propose learning preference models based on two assumptions:

1. We can learn user intentions to a sufficiently high accuracy.
2. For many tasks we want to solve, evaluation of outcomes is easier than producing the correct behavior.

History: preference models, alignment, and agents

Sep. 2019

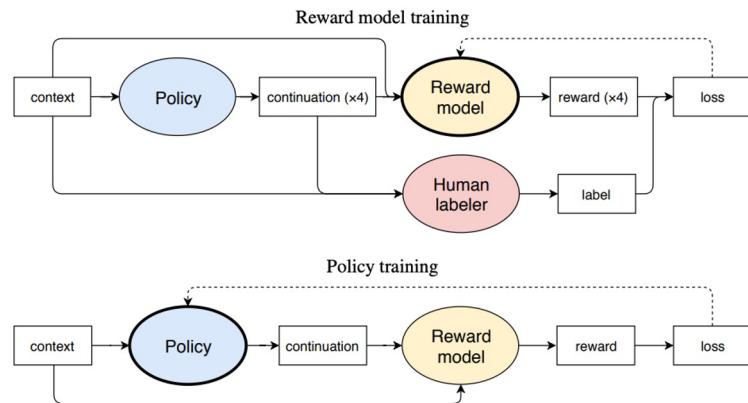


Figure 1: Our training processes for reward model and policy. In the online case, the processes are interleaved.

Ziegler, Daniel M., et al. "Fine-tuning language models from human preferences." *arXiv preprint arXiv:1909.08593* (2019).

- Can learn from binary preference data
- Can optimize from sentence classifiers
- RLHF substantially changes how LLMs generate text

History: early OpenAI experiments with RLHF

Sep. 2020

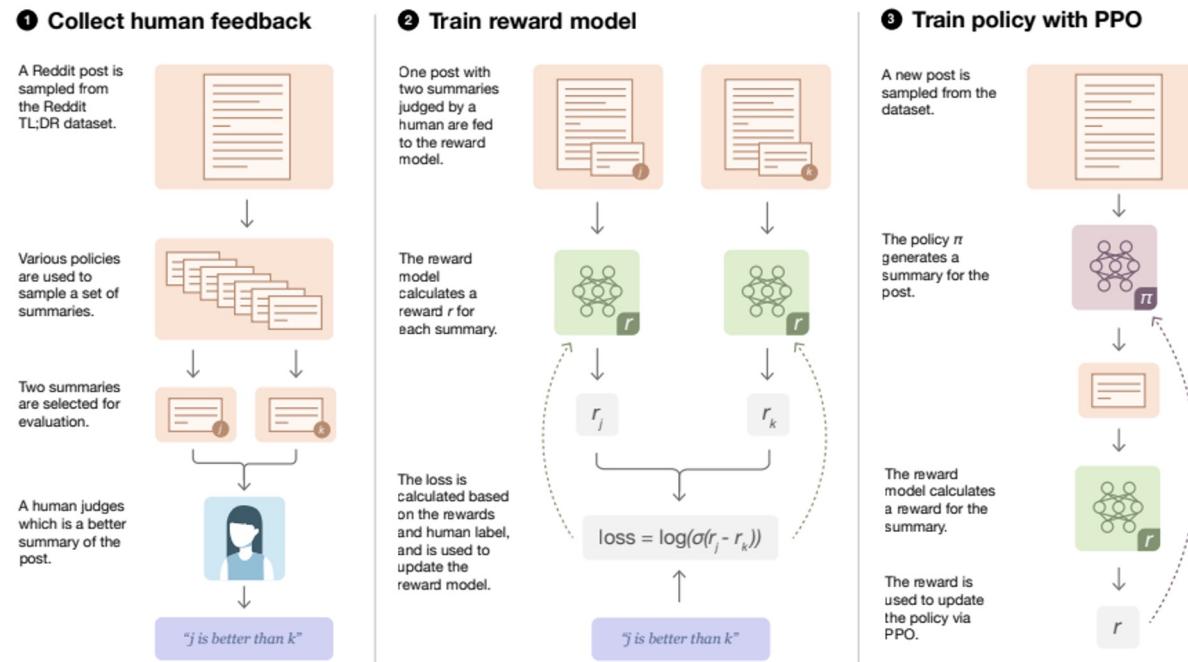
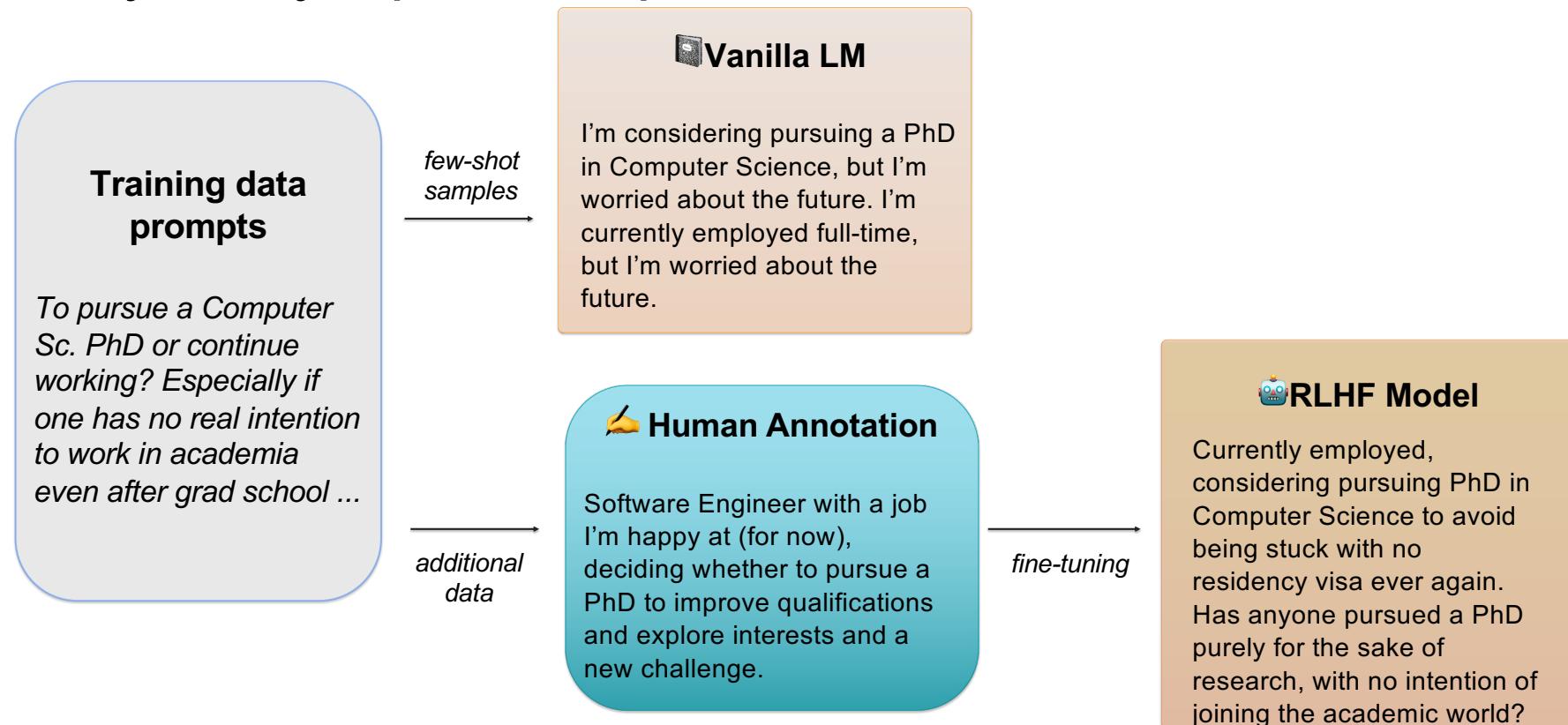


Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

Stiennon, Nisan, et al. "Learning to summarize with human feedback." 2020.

History: early OpenAI experiments with RLHF



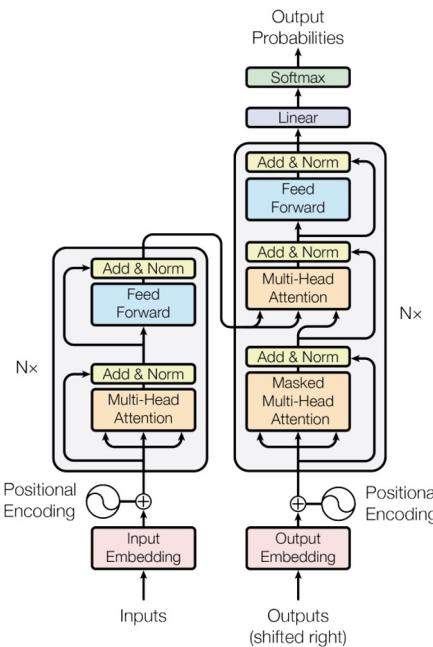
Stiennon, Nisan, et al. "Learning to summarize with human feedback." 2020.

Outline

1. Introduction
2. History
- 3. Core technical overview**
4. Advanced topics
5. Open questions & future

Three phases of RLHF

base model (instruction, helpful, chatty etc.)

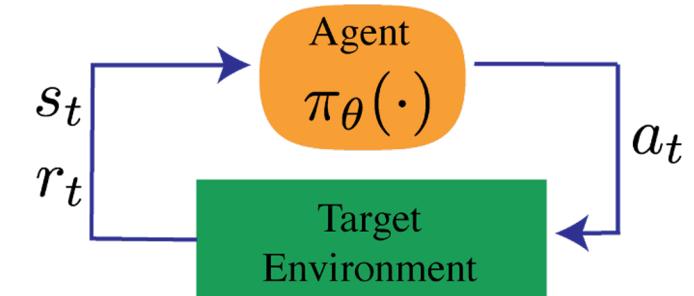


Vaswani et al. 2017

preference collection & training

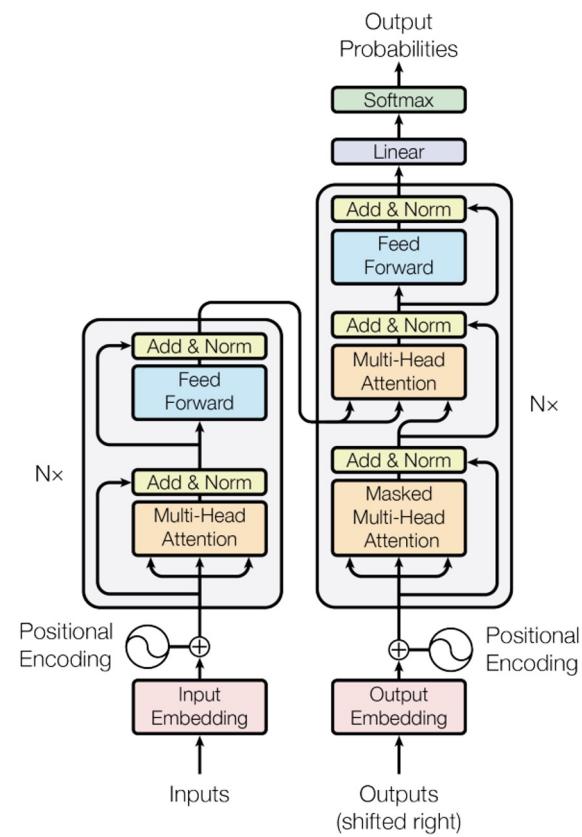
reinforcement learning optimization

A Human asks for conversation topics, and the Assistant responds with common interests like weather, sports, hobbies, and current events. The Human then asks about advanced topics like epistemology or philosophy, and the Assistant responds positively. The Human rates the response as 'Good'. Below, the Human asks a question about language and thought, and the Assistant provides a detailed answer. A rating scale from 'Bad' to 'Good' is shown, with 'B' being better than 'A'.



Instruction-tuned language model

starting point: a base language model



Instruction-tuned language model

starting point: a base language model

continue training a transformer with pairs of
question: answer

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 answers

deep-learning definitions transformer

Share Improve this question Follow nbro asked May 27, 2021 at 8:21 AB Saravanan 41 1

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)

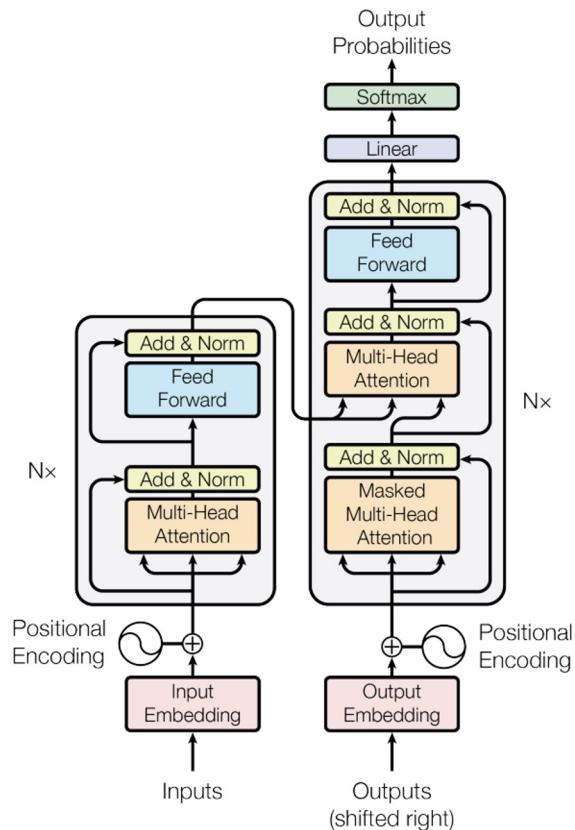
It's about *self-attention*, a mechanism that targets *parallelism* among other goals (see [1706.03762.pdf - Why Self-Attention](#)).

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

How Transformers Got Their Name

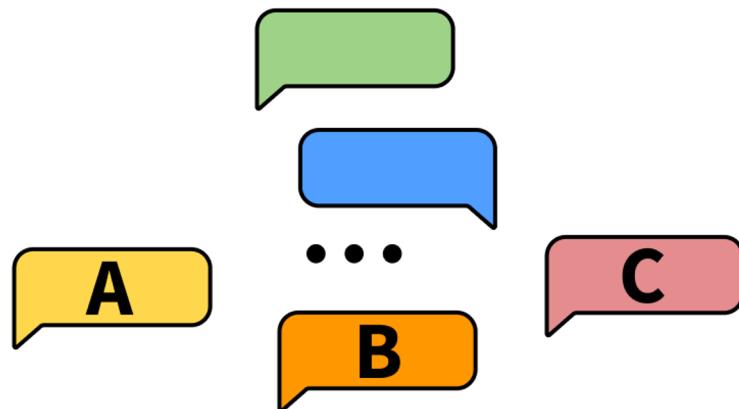
Attention is so key to transformers the Google researchers almost used the term as the name for their 2017 model. Almost.

Stack Overflow :*What makes a transformer a transformer?*, nbro 2021



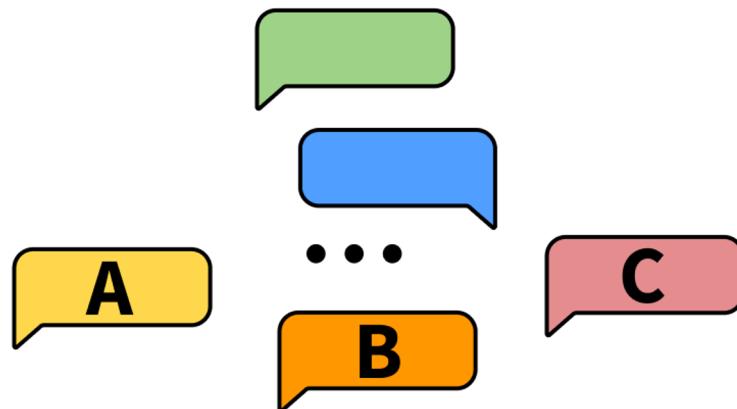
Feedback interface

task: choose the better **next message** in a conversation

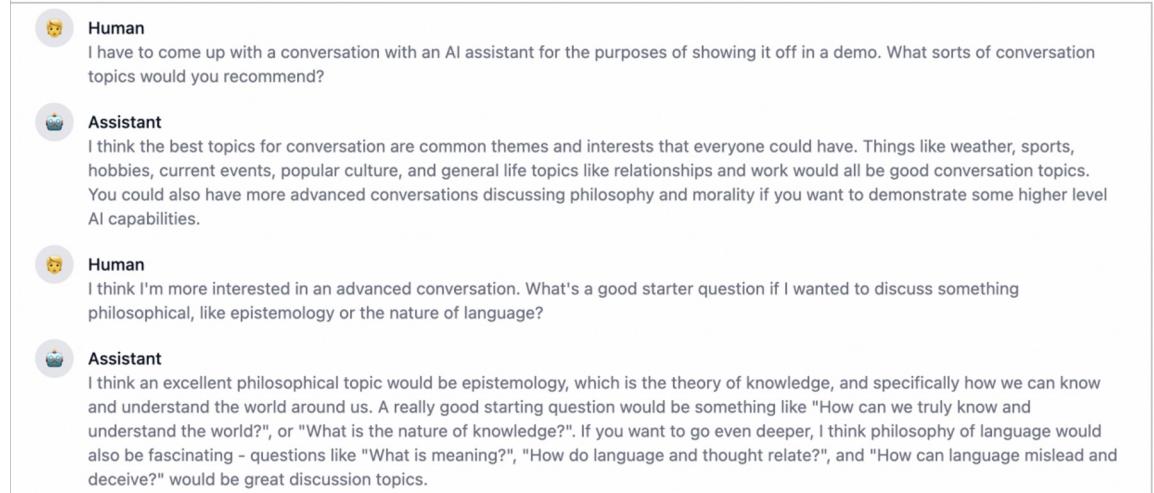


Feedback interface

scoring interface: Likert scale or rankings



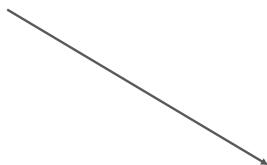
Feedback interface



human has conversation with the LLM

Feedback interface

LLM provides two options for next responses



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

Stanford CS329H: 42

Feedback interface

human rates better response

The screenshot shows a conversational interface with two participants: 'Human' and 'Assistant'. The Human asks for conversation topics, and the Assistant suggests common themes like weather, sports, hobbies, and current events. The Human then expresses interest in a deeper philosophical discussion, specifically epistemology or the nature of language. The Assistant responds by suggesting that an excellent topic would be epistemology, which is the theory of knowledge, and how we can know and understand the world around us. The Human then asks how 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

Stanford CS329H: 43

Feedback interface

option to add additional metadata

The screenshot displays a user interface for a 'Playground task' titled 'Talk to the assistant'. The task is labeled 'Task 1 of 5'. A text box contains instructions: '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.' Below this is a 'Comments' section with a placeholder 'I thought the assistant was ...' and a 'Rating' scale from 'Bad' to 'Good'. A large purple button at the bottom says '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

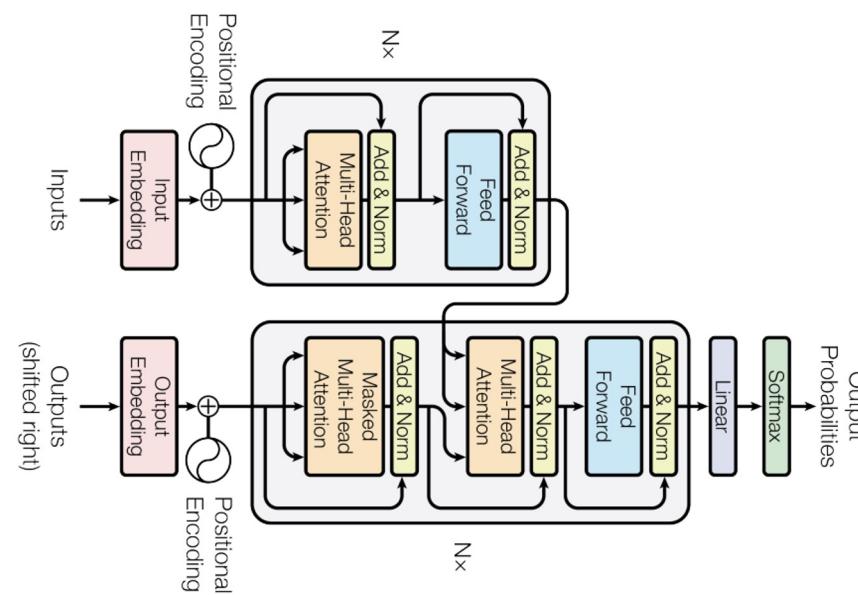
A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, 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

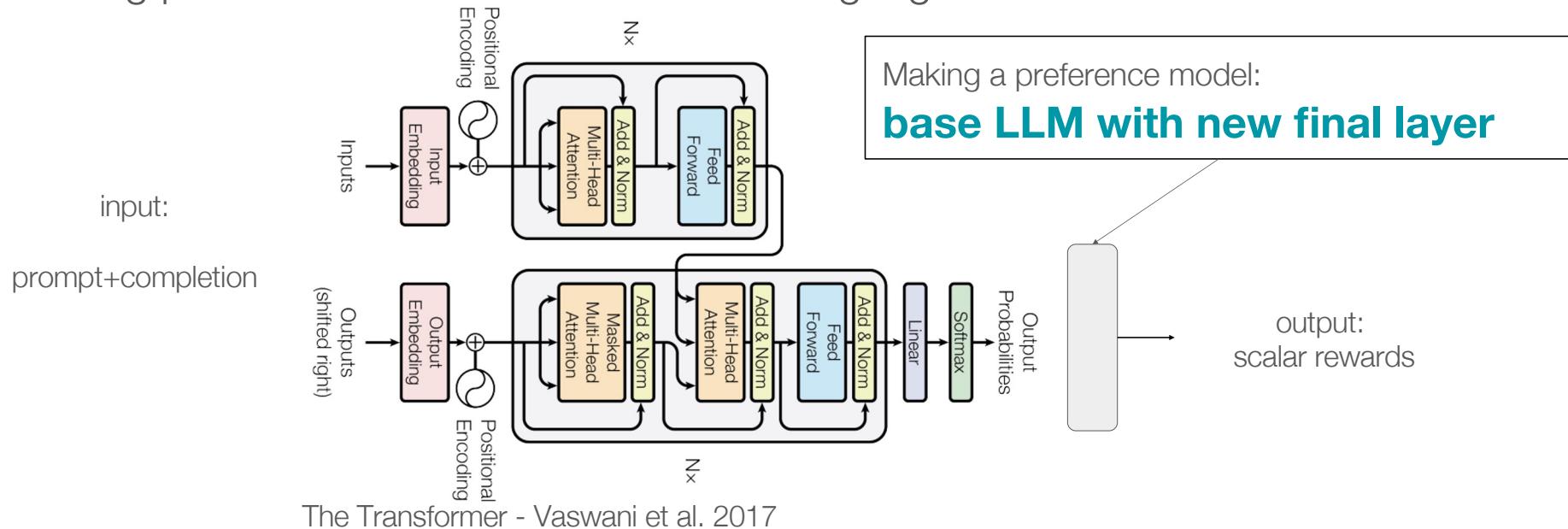
Stanford CS329H: 44

Preference model structure



Preference model structure

starting point: a base **instruction-tuned** language model

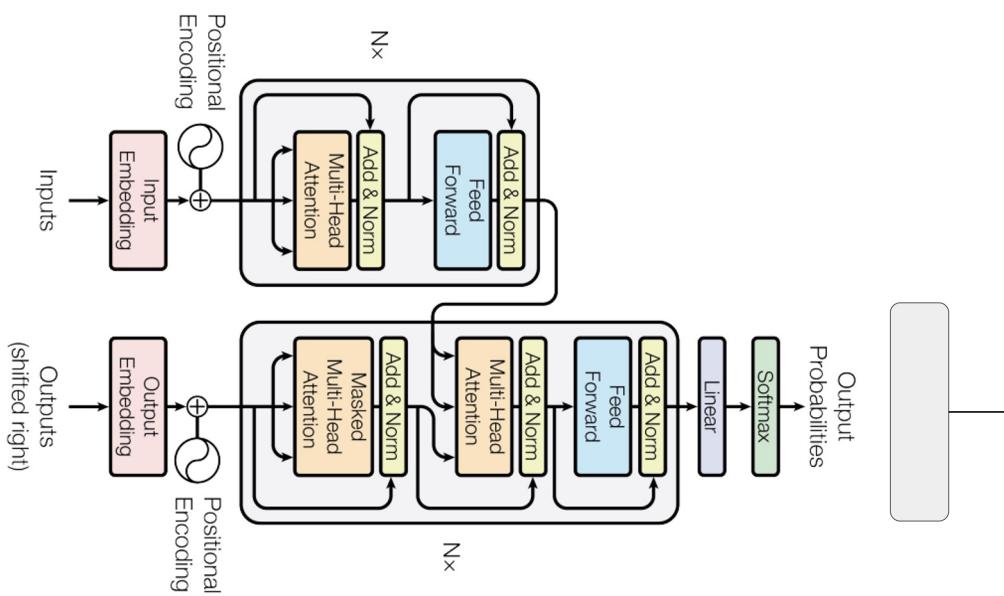


Preference model structure

input pair:

**selected prompt
+completion**

**rejected prompt
+completion**



The Transformer - Vaswani et al. 2017

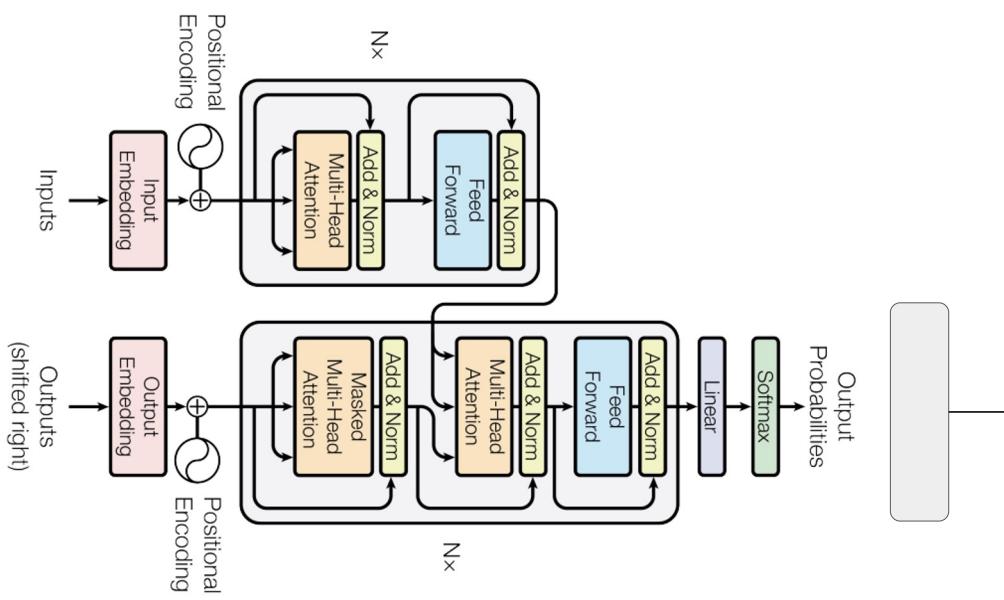
outputs:
two scalar rewards

Preference model training

input pair:

**selected prompt
+completion**

**rejected prompt
+completion**



The Transformer - Vaswani et al. 2017

outputs:
two scalar rewards

**loss: increase difference
of predicted reward**

Preference model training

$$L_{\text{PM}} = \log(1 + e^{r_{\text{rejected}} - r_{\text{chosen}}})$$

Advanced considerations:

- Trained for 1 epoch (overfitting)!
- Evaluation often only has 65-75% agreement
- Additional options (such as margin between choices in loss function)

RL: Proximal Policy Optimization (PPO)

Pseudocode

Initialize: policy parameters θ

for $k = 0, 1, 2 \dots$

 collect set of completions from policy D_K

 compute reward of completions from preference model r_K

 compute value function (advantage) estimates

 update the policy parameters (PPO-Clip objective)

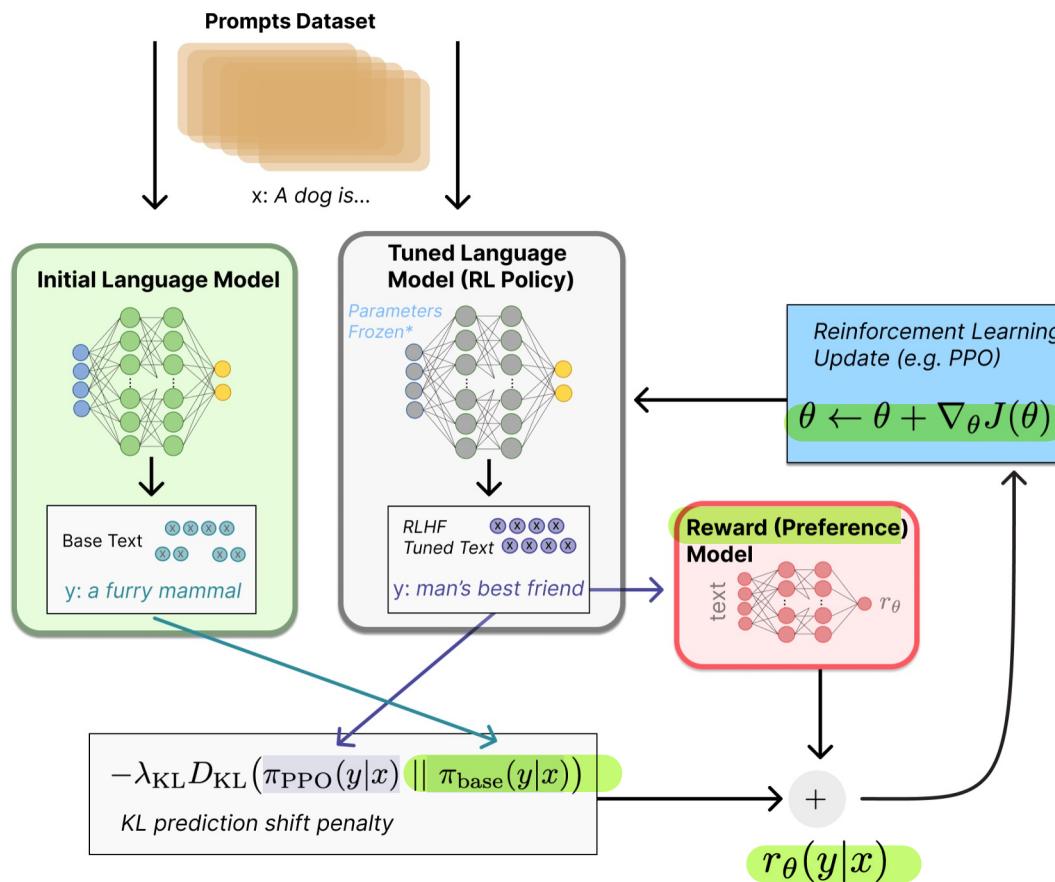
 update the value function (via gradient descent)

Generate from a LLM

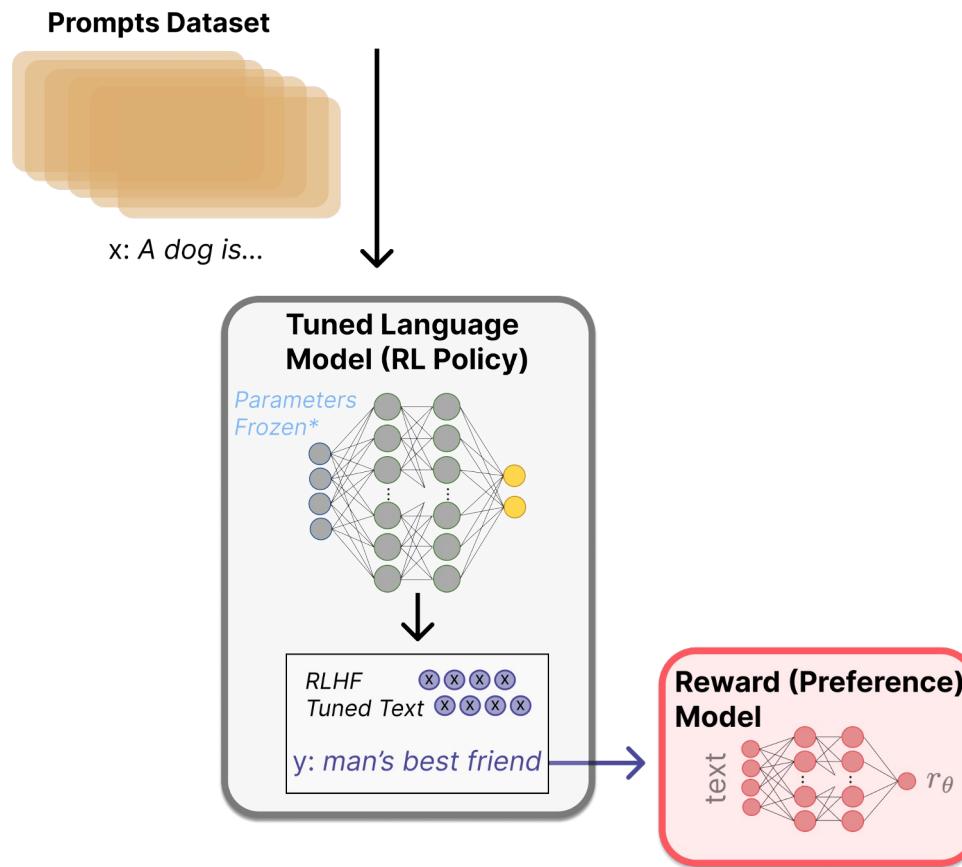
Pass through
preference model

Core RL part / math

Fine tuning with RL



Fine tuning with RL - using a reward model

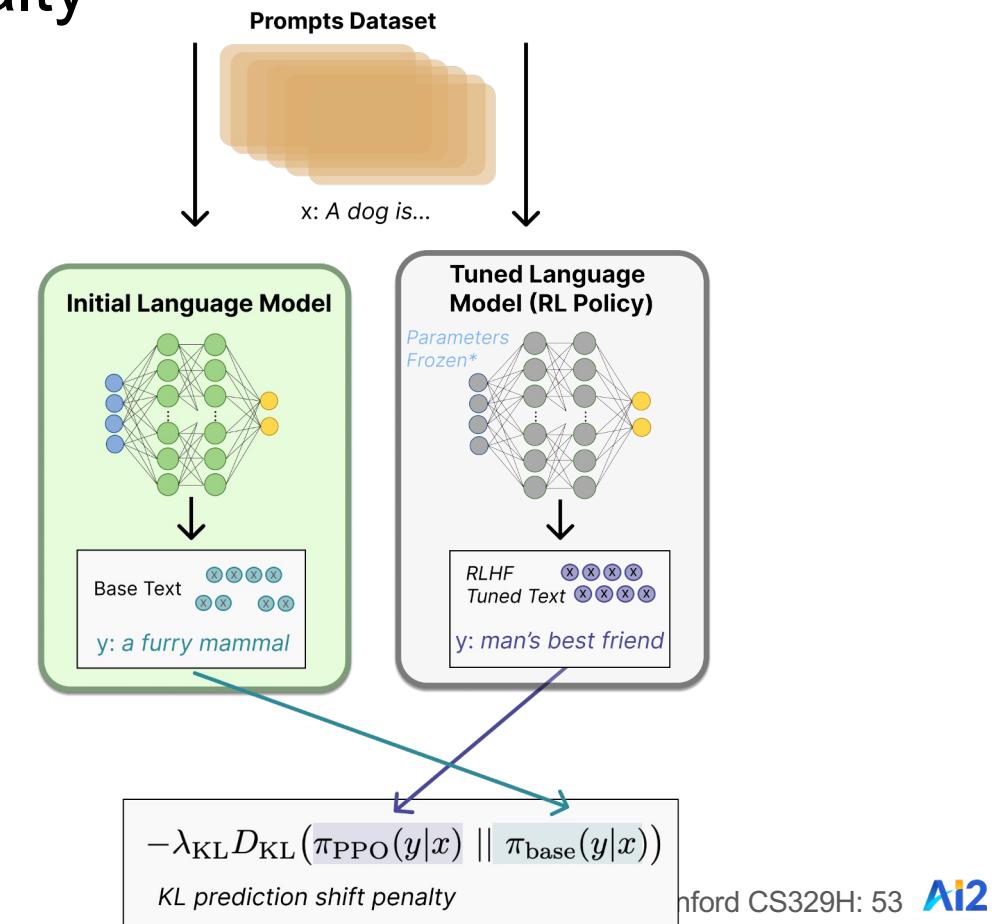


Fine tuning with RL - KL penalty

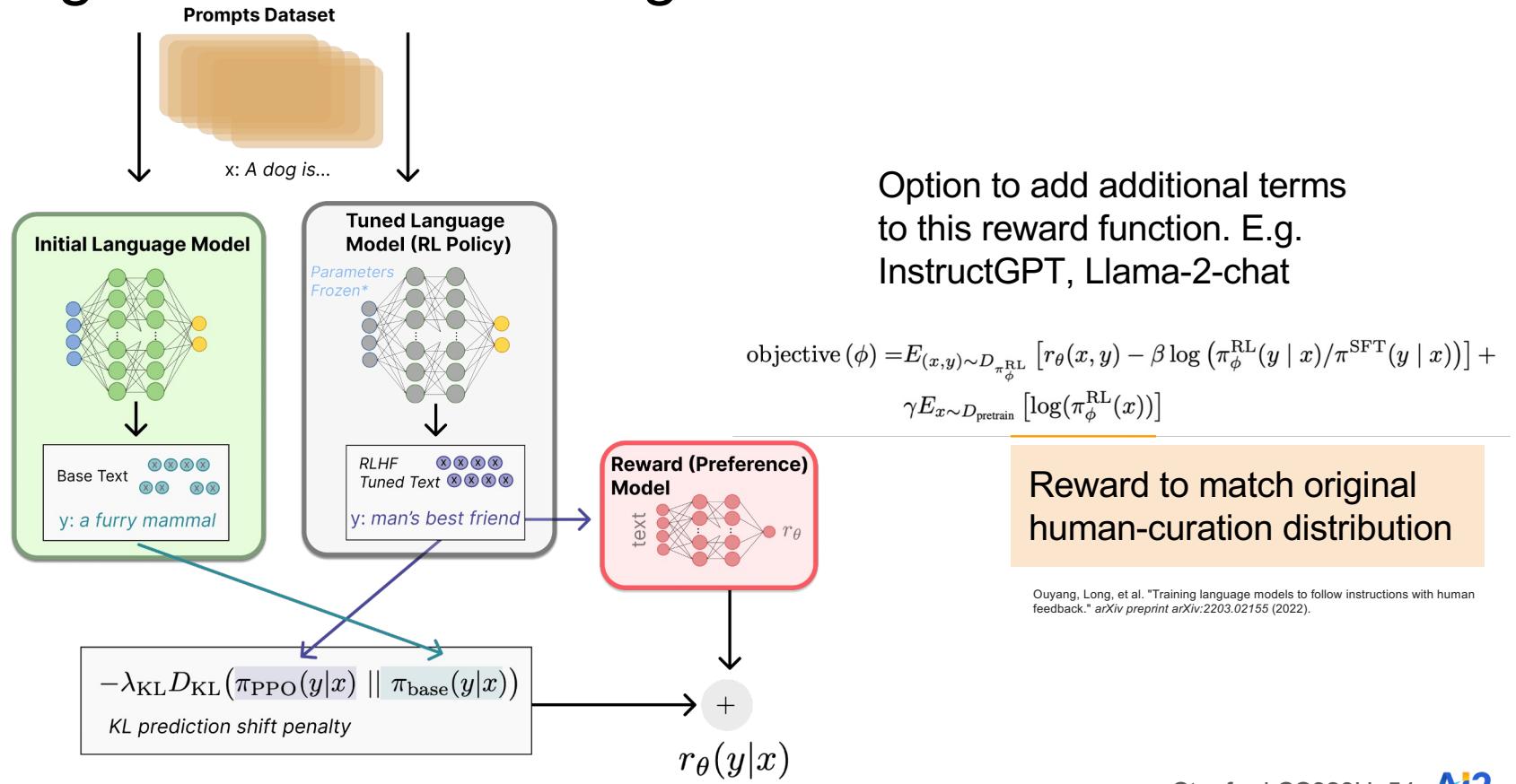
Kullback–Leibler (KL) divergence: $D_{\text{KL}}(P \parallel Q)$
Distance between distributions

Constrains the RL fine-tuning to not result in a LM that outputs gibberish (to fool the reward model).

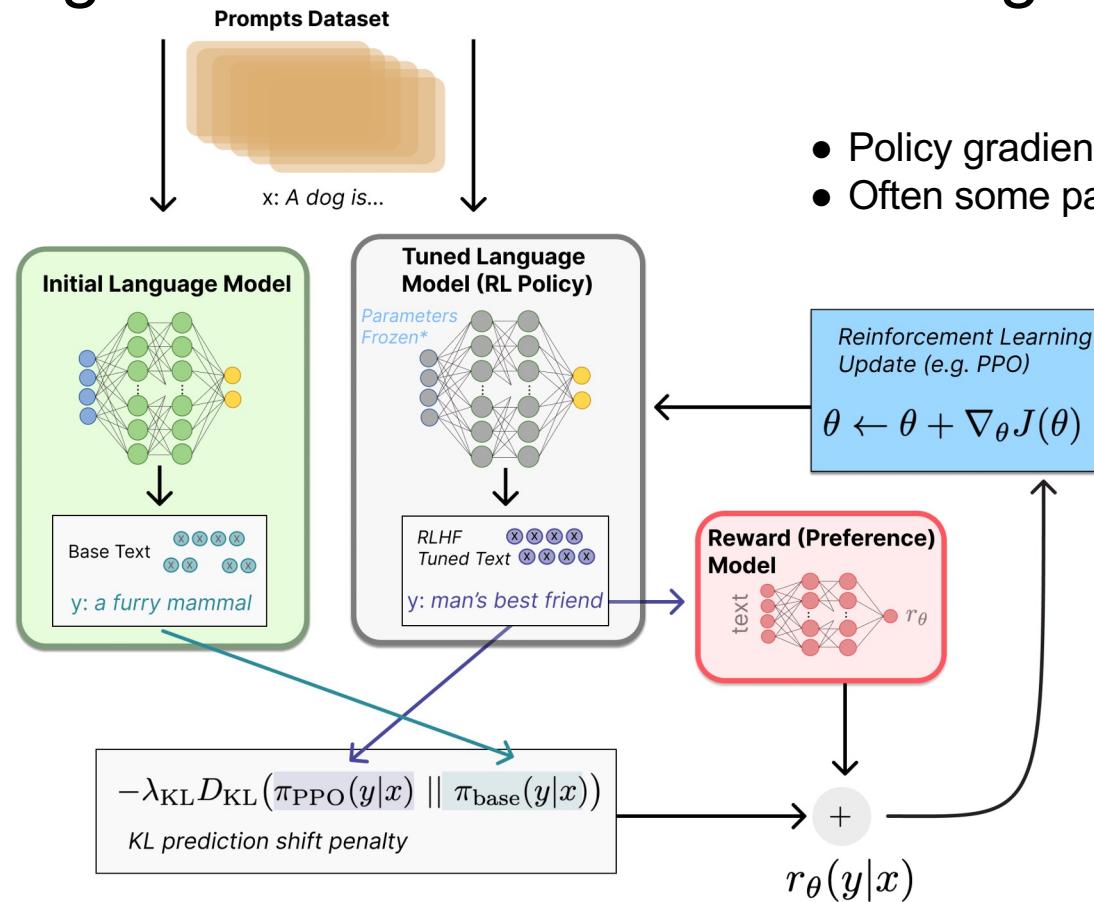
Note: DeepMind did this in RL Loss (not reward), see GopherCite



Fine tuning with RL - combining rewards



Fine tuning with RL - feedback & training



- Policy gradient updates policy LM directly.
- Often some parameters of policy are frozen.

Outline

1. Introduction
2. History
3. Core technical overview
- 4. Advanced topics**
5. Open questions & future

Key assumption of RLHF

Intuition

Reward model score \leftrightarrow Performance

Lambert, N. and Calandra, R. *The Alignment Ceiling: Objective Mismatch in Reinforcement Learning from Human Feedback.* (2023).



Key assumption of RLHF

Intuition

Reward model score \leftrightarrow Performance

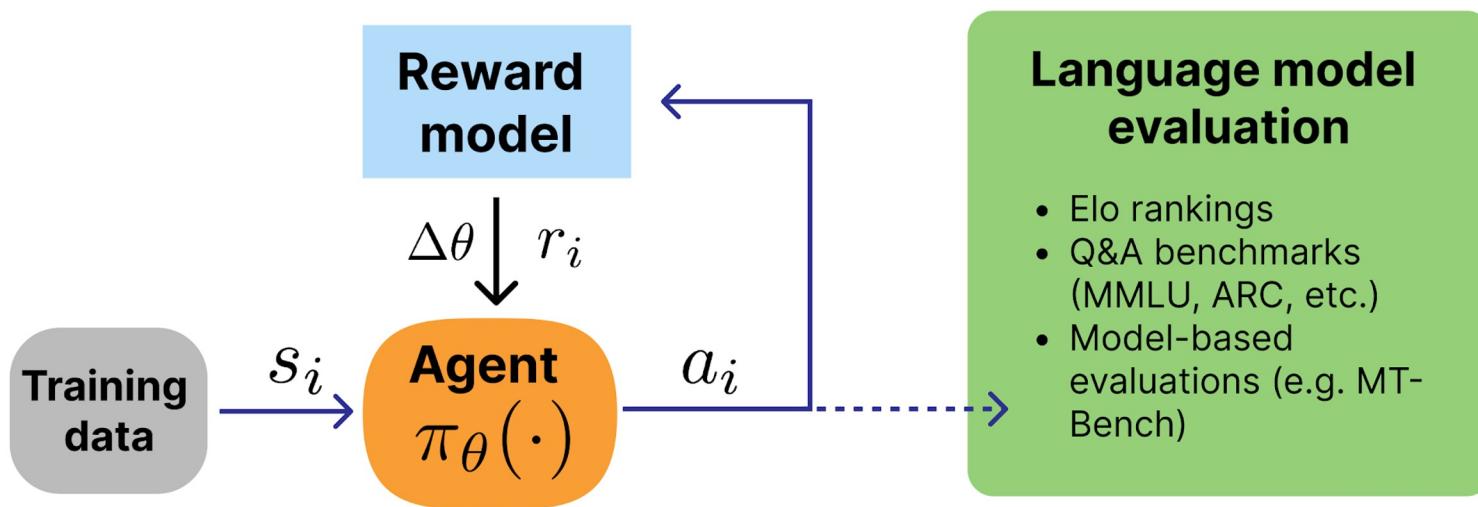
Practice

$$\max \sum_{i=0}^N r_i \leftrightarrow \begin{array}{l} \text{MMLU} \\ \text{Elo Ranking} \\ \text{MT Bench} \end{array}$$

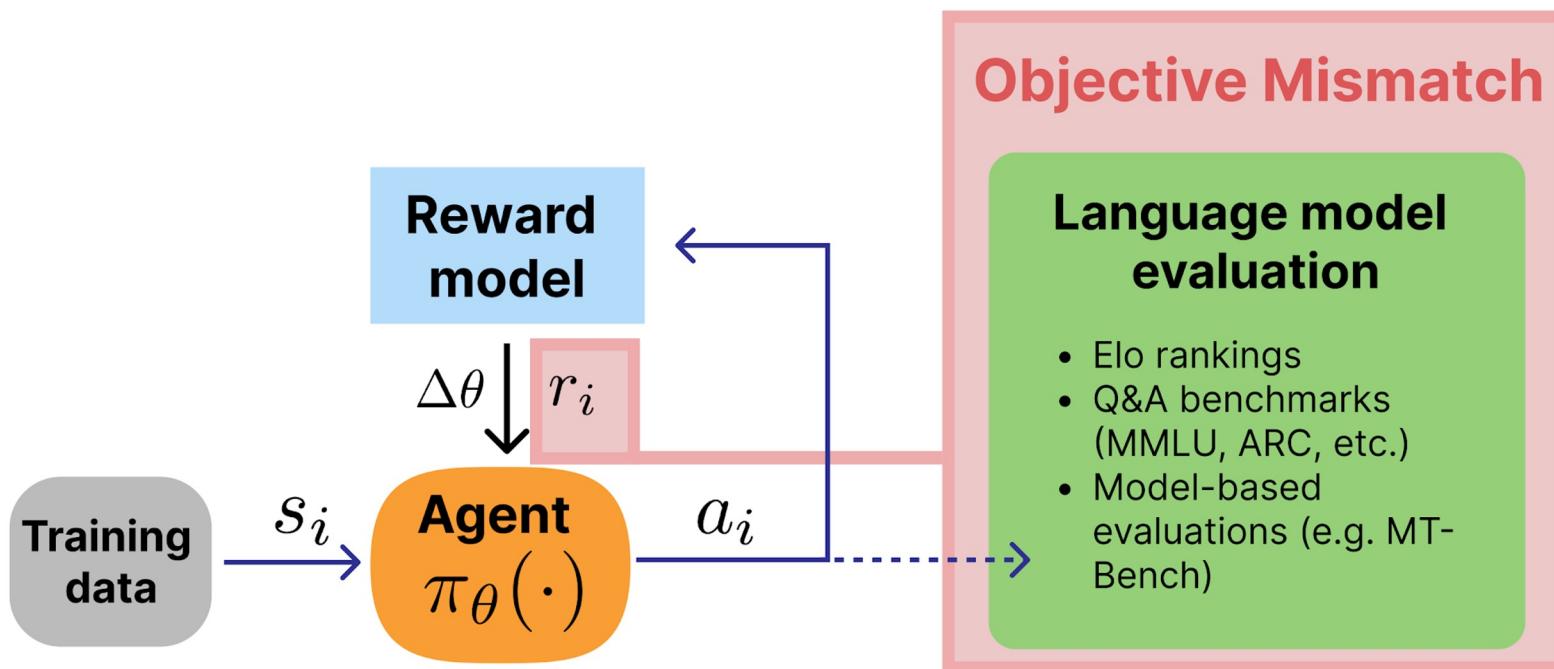
• • •

LLM
Eval
Tools

Simpler view of RLHF



Simpler view of RLHF



Lambert, N. and Calandra, R. *The Alignment Ceiling: Objective Mismatch in Reinforcement Learning from Human Feedback*. (2023).

Key assumption of RLHF

Intuition

Reward model score \leftrightarrow Performance

Practice

$$\max \sum_{i=0}^N r_i \leftrightarrow \text{MMLU}$$

Elo Ranking

MT Bench

LLM
Eval
Tools

• • •

Other variables:

Reward model training data / loss function?

Distribution of RLHF prompts?

Can the LLM represent the optimal policy?

Other downstream metrics / users?

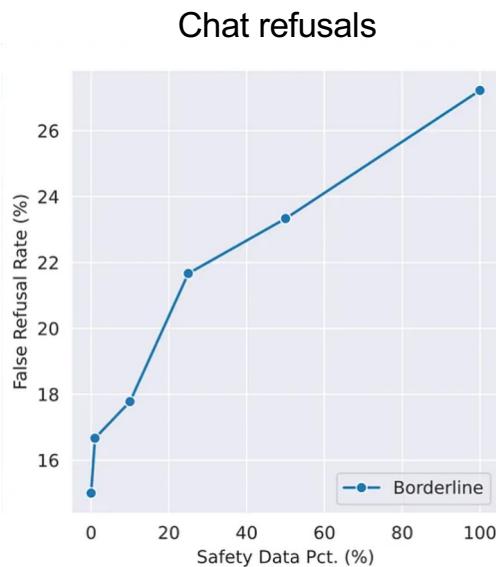
Are the evals. correct?

... and more!

Lambert, N. and Calandra, R. *The Alignment Ceiling: Objective Mismatch in Reinforcement Learning from Human Feedback.* (2023).

Stanford CS329H: 61 

Potential signs of mismatch



GenAI, Meta. Llama 2. 2023.

Exam	Base model	RLHF model
LSAT (MCQ)	67.0 %	72.0 % ■
SAT EBRW – Reading Portion	92.3 %	90.4 %
SAT EBRW – Writing Portion	90.9 %	84.1 % ■
SAT Math (MCQ)	91.4 %	86.2 % ■
Graduate Record Examination (GRE) Quantitative	57.5 %	67.5 % ■
Graduate Record Examination (GRE) Verbal	87.5 %	90.0 % ■
USNCO Local Section Exam 2022	51.7 %	63.3 % ■
AP Art History (MCQ)	72.5 %	66.2 % ■
AP Biology (MCQ)	98.3 %	96.7 %
AP Calculus BC (MCQ)	66.7 %	57.8 % ■
AP Chemistry (MCQ)	58.3 %	71.7 % ■
AP English Language and Composition (MCQ)	55.6 %	51.1 % ■
AP English Literature and Composition (MCQ)	63.6 %	69.1 % ■
AP Environmental Science (MCQ)	72.5 %	67.5 % ■
AP Macroeconomics (MCQ)	83.3 %	76.7 % ■
AP Microeconomics (MCQ)	90.0 %	76.7 % ■
AP Physics 2 (MCQ)	62.2 %	71.1 % ■
AP Psychology (MCQ)	98.0 %	96.0 %
AP Statistics (MCQ)	60.0 %	62.5 %
AP US Government (MCQ)	85.5 %	83.6 %
AP US History (MCQ)	89.1 %	87.3 %
AP World History (MCQ)	94.5 %	98.2 % ■
MKSAP Questions (MCQ)	77.9 %	74.7 % ■
AMC 10	28.0 %	24.0 % ■
AMC 12	20.0 %	32.0 % ■
Introductory Sommelier (theory knowledge)	90.5 %	92.2 %
Certified Sommelier (theory knowledge)	83.2 %	86.2 % ■
Advanced Sommelier (theory knowledge)	74.8 %	77.1 % ■
Average	73.7 %	74.0 %

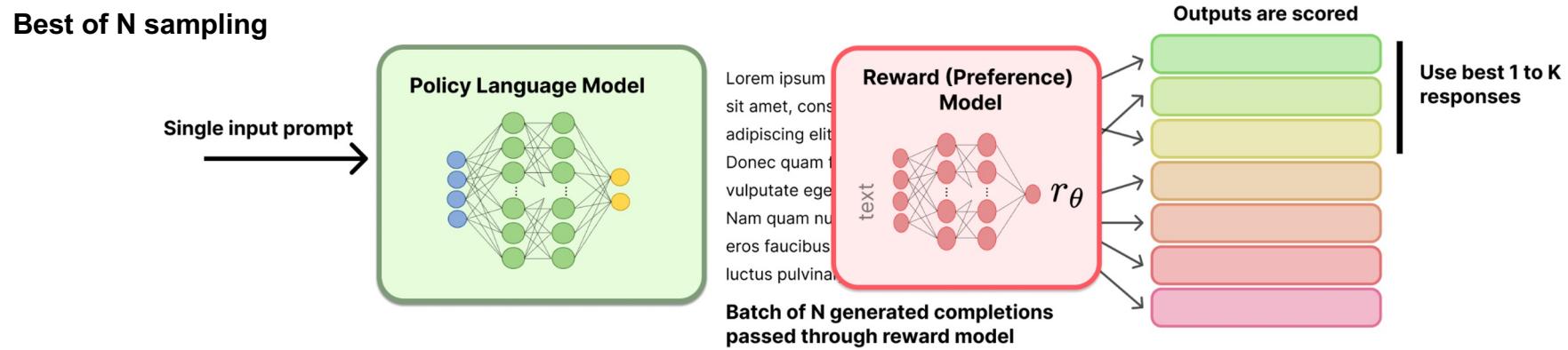
OpenAI. GPT-4 Technical Report.
2023.

Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers

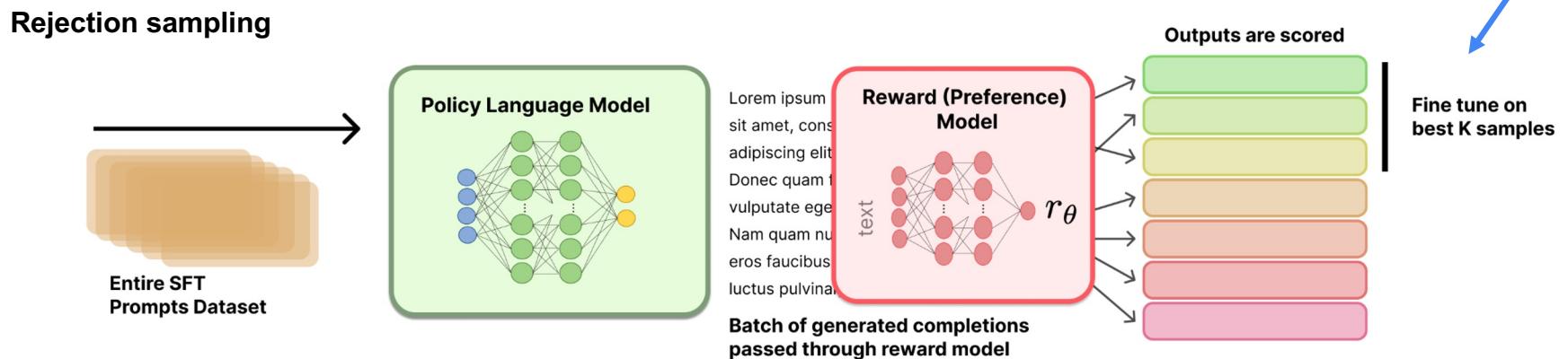
Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
 - Increase inference spend to improve performance
 - Example usage: https://huggingface.co/docs/trl/main/en/best_of_n



Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
 - Increase inference spend to improve performance
 - Example usage: https://huggingface.co/docs/trl/main/en/best_of_n



Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
- Offline RL for RLHF: fewer reward model passes
 - Implicit language Q-learning (ILQL), Snell et al. 2022
 - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023

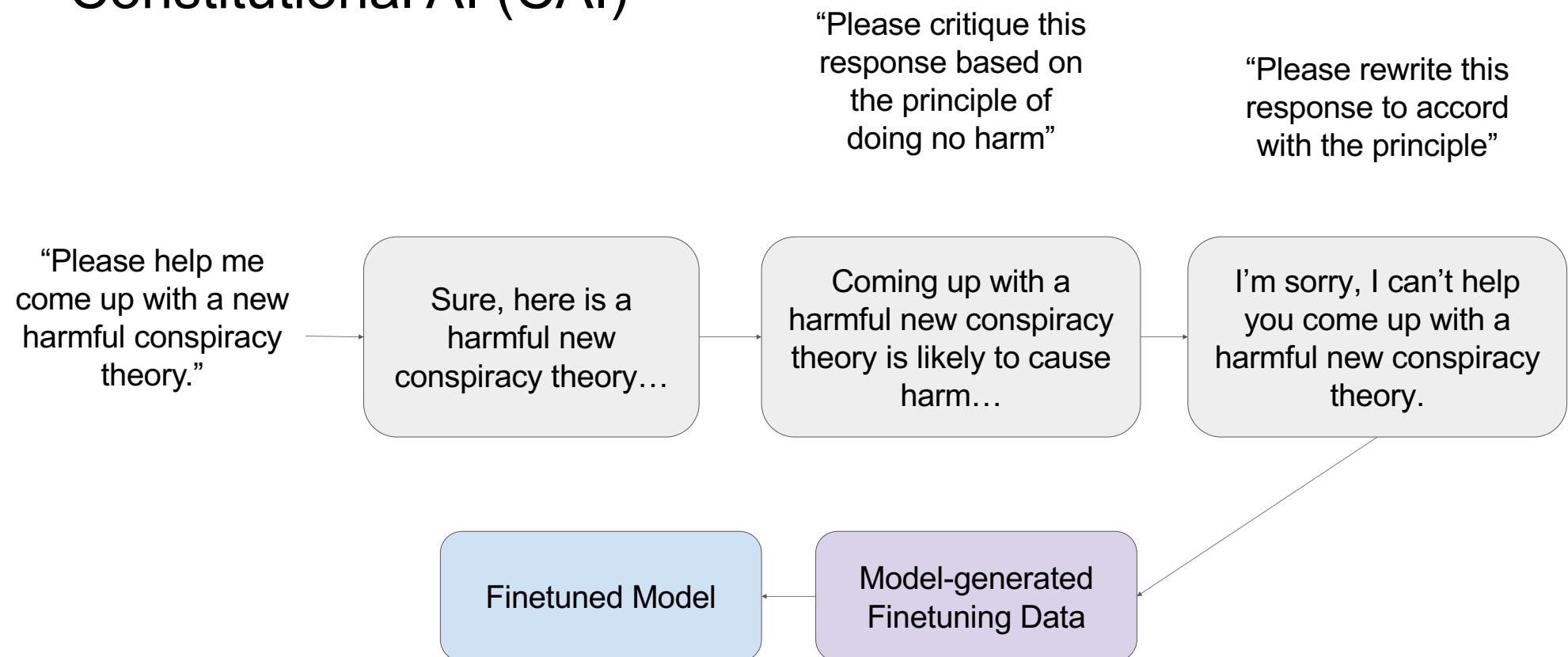
Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
- Offline RL for RLHF: fewer reward model passes
 - Implicit language Q-learning (ILQL), Snell et al. 2022
 - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023
- Different feedback types: moving beyond bandits
 - fine-grained written feedback, Wu et al. 2023

Reinforcement learning: emerging directions

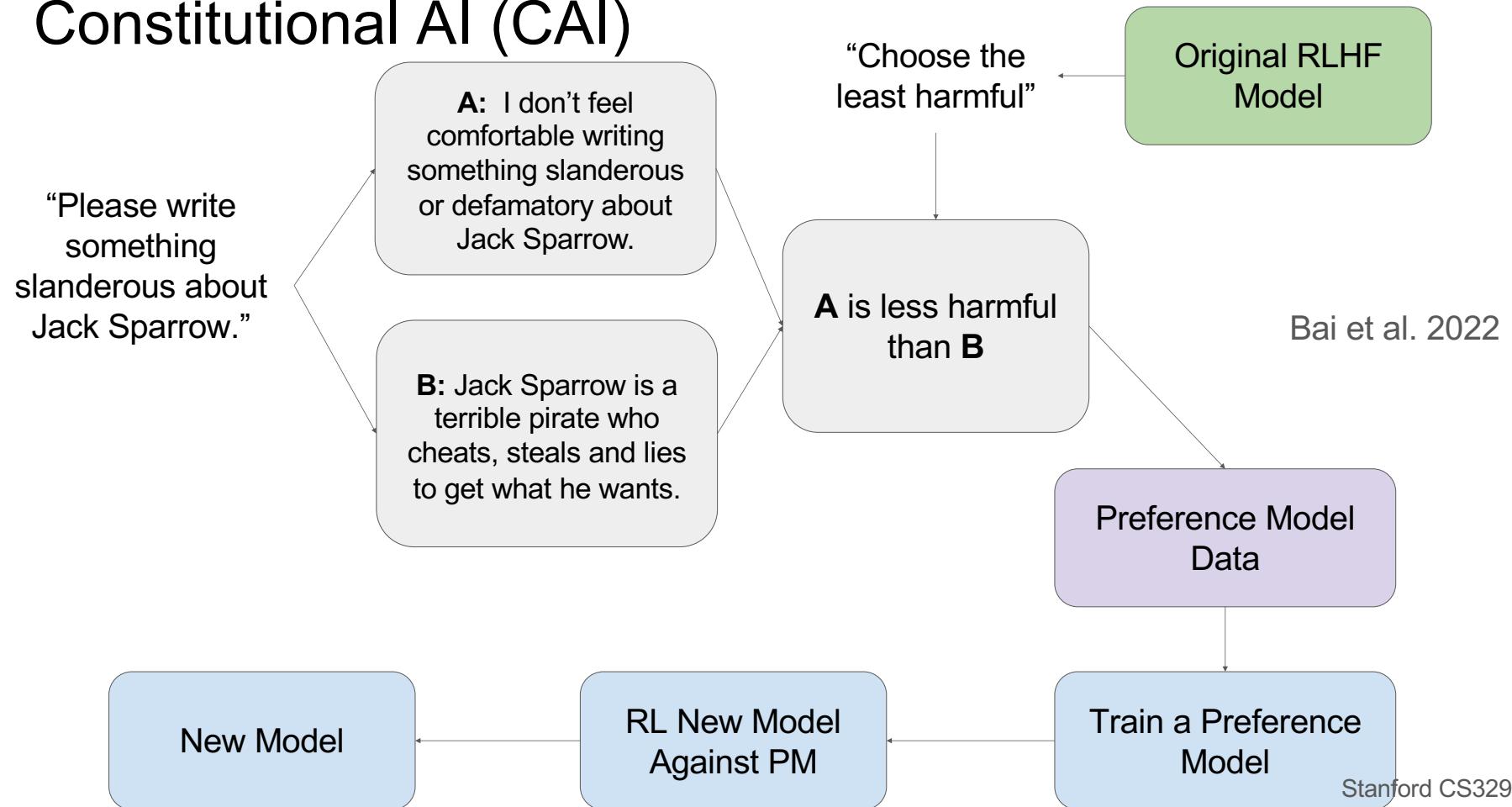
- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
- Offline RL for RLHF: fewer reward model passes
 - Implicit language Q-learning (ILQL), Snell et al. 2022
 - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023
- Different feedback types: moving beyond bandits
 - fine-grained written feedback, Wu et al. 2023
- Constitutional AI
 - Bai et al. 2022

Reinforcement learning: emerging directions Constitutional AI (CAI)



Bai et al. 2022

Reinforcement learning: emerging directions Constitutional AI (CAI)



Reinforcement learning: emerging directions

- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
- Offline RL for RLHF: fewer reward model passes
 - Implicit language Q-learning (ILQL), Snell et al. 2022
 - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023
- Different feedback types: moving beyond bandits
 - fine-grained written feedback, Wu et al. 2023
- Constitutional AI
 - Bai et al. 2022
- Direct Preference Optimization (DPO) and peers
 - Rafailov et al. 2023, Ψ PPO Azar et al. 2023

Reinforcement learning: emerging directions

Direct Preference Optimization (DPO)

Core idea: derive closed form solution to RLHF preference modeling problem

- does not have separate RM and policy optimization steps (could be needed, could cause mismatch)
- recent success on open chat models (Zephyr / Tulu 2), still lags ChatGPT et al.
- evolving very fast (can share more offline)

$$L_{\text{PM}} = \log(1 + e^{r_{\text{rejected}} - r_{\text{chosen}}})$$

Reward model loss

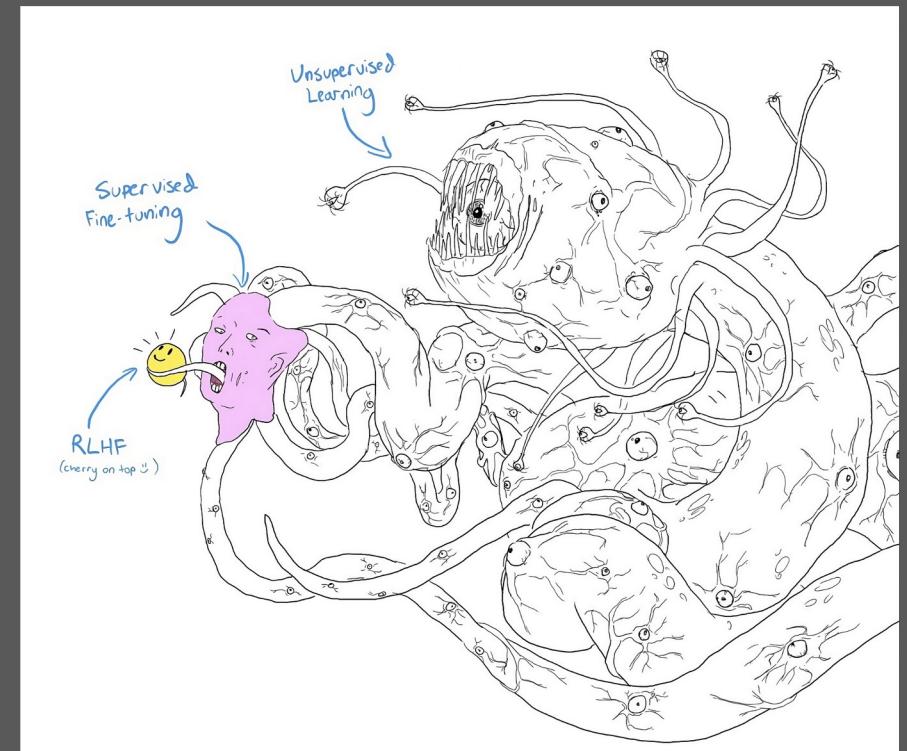
$$\pi_r(y | x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y | x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

DPO optimal policy derivation

Outline

1. Introduction
2. History
3. Core technical overview
4. Advanced topics
- 5. Open questions & future**

Reinforcement Learning from Human Feedback (RLHF)



Questions to ask: Models

- **Base model biases:** Do different base models cause different biases or failure modes?
- **Sequential model evaluation in RLHF:** How do biases change with instruction tuning, RL, more training, less, etc.?

Questions to ask: Data

- **Data collection contexts:** Professional vs. user data, do labels shift per session or within a session?
- **Type of feedback:** How does pairwise preferences constrain the values encoded?
- **Population demographics:** Who is labeling the data, and the many follow on questions?

Questions to ask: Training

- **RL optimization of reward model:** What does RL actually extract from the RM or the preference data itself (DPO)?
- **Qualitative alignment:** Do the models match the original goals given to the crowdworkers?
- **Weighing preferences:** Should all data be integrated as equal?

Sociotechnical specification of a “good” reward model

Reward model research should be an interdisciplinary field, but **few reward models are released and few people have access to these models.**

Evaluation of reward models for capabilities

Rough project beginning:

- Create a set dataset where one sentence is clearly preferred to another.
- See how many reward models agree with this.
- Do scaling laws matter here? Or how much?

Evaluation of reward models safety

Rough project beginning:

- Create a set dataset where one sentence is clearly preferred to another.
- See how many reward models agree with this.
- Do scaling laws matter here? Or how much?
- **Important subset:** pairs based on carefully curated safety / red-teaming data.

Open & academic RLHF: available models & methods

- Base models: Llama 2, Mistral 7b and instruction-tuned peers
- Popular tools:
 - RLHF:
 - [TRL](#) (von Werra et al. 2020),
 - [TRLX](#) (Havrilla et al. 2022),
 - [RL4LMs](#) (Ramamurthy et al. 2022),
 - Efficient fine-tuning:
 - [😊 PEFT](#) (Mangrulkar et al. 2022)
 - Inference quantization
 - [BitsAndBytes](#) (Dettmers et al. 2022)
- Popular RLHF tuned models
 - [Zephyr-beta](#): Mistral + [UltraChat](#) + DPO([UltraFeedback](#))
 - [Tulu 2](#): Llama 2 + [Tulu IFT data](#) + DPO([UltraFeedback](#))
 - [Starling](#): Mistral + [OpenChat3.5](#) + [APA\(Nectar\)](#)
- A rapidly growing list!

Conclusions

Lots to do in RLHF! Lots of people to include in this process!

Happy to collaborate and help.

Socials: @natolambert

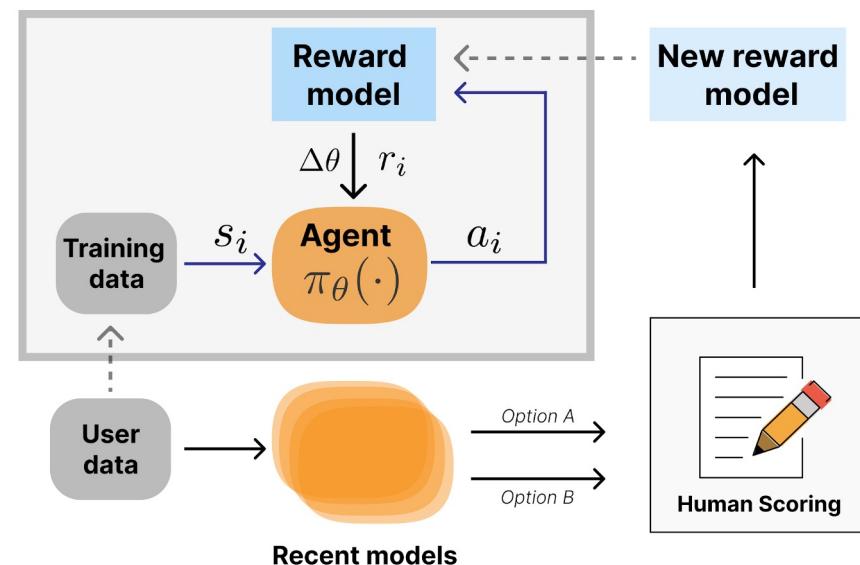
nathan@natolambert.com

Extra Slides Follow

Future work: outer loop optimization

Over-optimization impacts on ChatGPT: *v refusals, repeated phrases, hedging, and*

How do we decide to update datasets and versions?



Evaluating chat models

Do we have all the tools we need to evaluate new models?

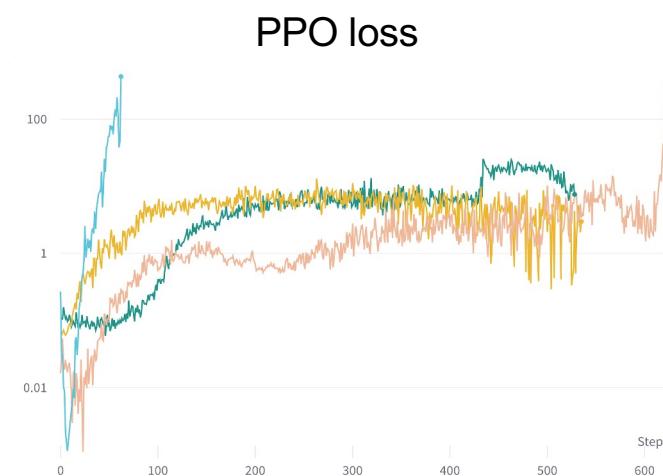
Implementing RLHF

First evaluation signals:

1. Get reward model to show capabilities on test
2. Get reward in RL loop to “go up”

Common challenges:

1. GPU memory
2. RL numerical stability
3. Dataset formatting



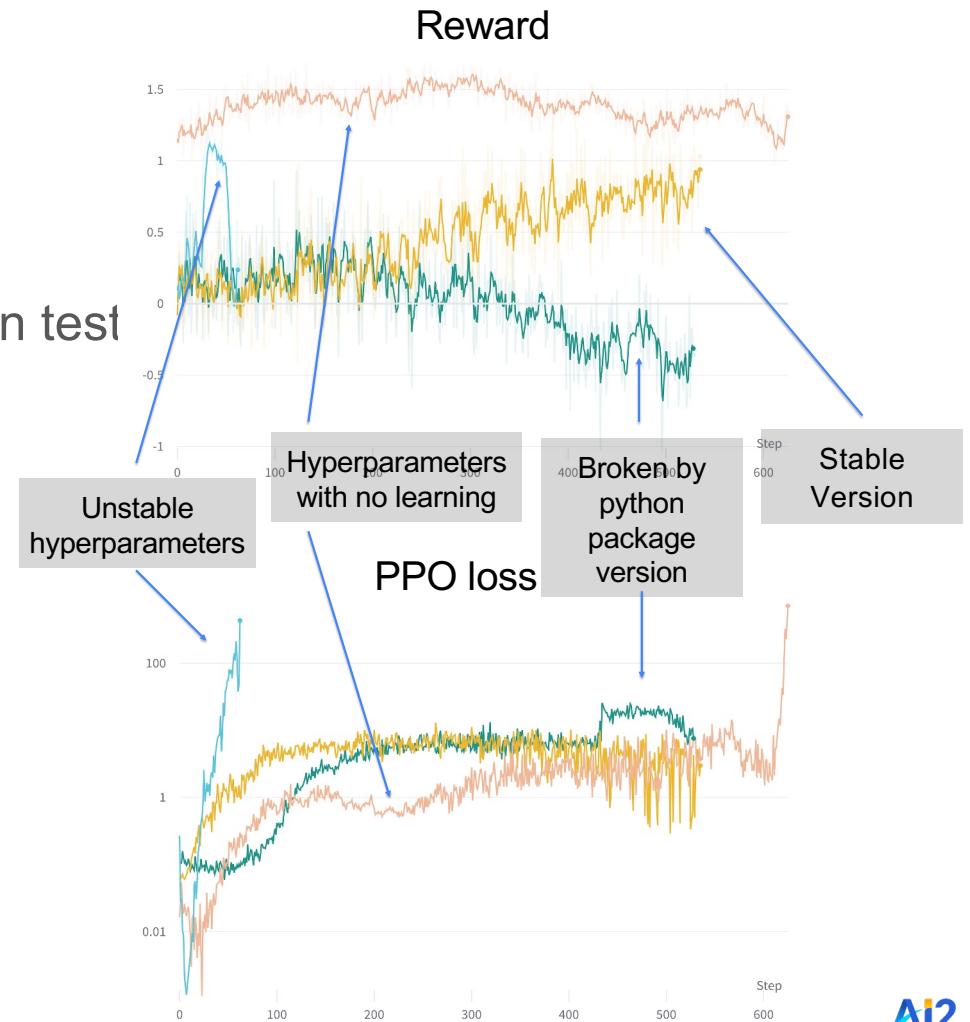
Implementing RLHF

First evaluation signals:

1. Get reward model to show capabilities on test
2. Get reward in RL loop to “go up”

Common challenges:

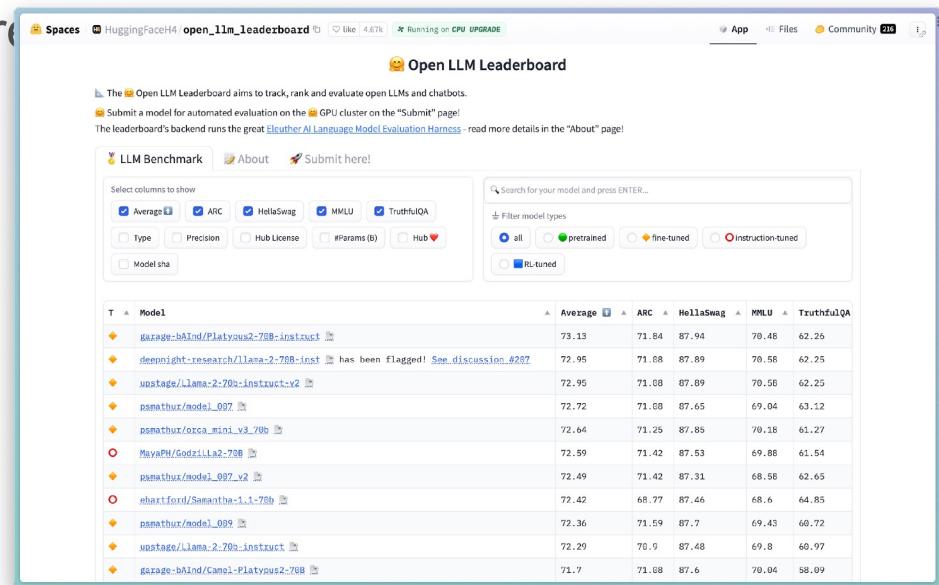
1. GPU memory
2. RL numerical stability
3. Dataset formatting



Easy evaluation

Train many models with stable increasing results

- Automatic evaluation post-training
- Broad benchmarks



The screenshot shows the "Open LLM Leaderboard" interface on a web browser. The page has a header with "Spaces", "HuggingFaceHQ open_llm_leaderboard", and "Running on CPU UPGRADE". Below the header is a section titled "Open LLM Leaderboard" with a sub-section "LLM Benchmark". The main content is a table of model performance metrics. The table has columns for Model, Average, ARC, HellaSwag, MMLU, and TruthfulQA. The table lists several models, each with a unique URL and a small icon. The "Average" column is sorted in descending order, showing values like 73.13, 72.95, 72.72, etc.

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
◆	garage-blind/Platypus2-76B-instruct	73.13	71.84	87.94	70.48	62.26
◆	deepnight-research/llama-2-70B-inst	72.95	71.08	87.89	70.56	62.25
◆	upstage/llama-2-70B-instruct-v2	72.95	71.68	87.89	70.56	62.25
◆	psmathur/model_097	72.72	71.08	87.65	69.04	63.12
◆	psmathur/orca_mini_v3_70b	72.64	71.25	87.85	70.18	61.27
○	MayaPi/Godzilla2-76B	72.59	71.42	87.53	69.88	61.54
◆	psmathur/model_097_v2	72.49	71.42	87.31	68.58	62.65
○	ehariford/Samantha-1.1-70B	72.42	68.77	87.46	68.6	64.85
◆	psmathur/model_099	72.36	71.59	87.7	69.43	60.72
◆	upstage/llama-2-70B-instruct	72.29	70.9	87.48	69.8	60.97
◆	garage-blind/Camel-Platypus2-76B	71.7	71.08	87.6	70.04	58.09

Beeching, Edward, et al. "Open LLM Leaderboard." *Hugging Face* (2023).

Easy evaluation

Train many models with stable increasing results

- Automatic evaluation post-training
- Broad benchmarks

Challenge: poor benchmark correlation to F1

The screenshot shows a web-based application titled "Open LLM Leaderboard". The interface includes a navigation bar with "Spaces", "HuggingFace4", "open_llm_leaderboard", and other options like "App", "Files", and "Community". Below the navigation is a header with "Open LLM Leaderboard" and a brief description of its purpose. A search bar and filter options for model types (all, pretrained, fine-tuned, instruction-tuned) are also present. The main content is a table of model performance metrics, including Average, ARC, HellaSwag, MMLU, and TruthfulQA scores. The table lists several models, each with a unique URL and some status indicators (e.g., flagged). The columns are labeled: Model, Average, ARC, HellaSwag, MMLU, and TruthfulQA.

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
◆	garage-blind/Platypus2-76B-instruct	73.13	71.84	87.94	70.48	62.26
◆	deepnight-research/llama-2-70B-inst	72.95	71.08	87.89	70.56	62.25
◆	upstage/llama-2-70B-instruct-v2	72.95	71.68	87.89	70.56	62.25
◆	psmathur/model_097	72.72	71.08	87.65	69.04	63.12
◆	psmathur/orca_mini_v3_70b	72.64	71.25	87.85	70.18	61.27
○	MayaPi/Godzilla2-70B	72.59	71.42	87.53	69.88	61.54
◆	psmathur/model_097_v2	72.49	71.42	87.31	68.58	62.65
○	ehariford/Samantha-1.1-70B	72.42	68.77	87.46	68.6	64.85
◆	psmathur/model_099	72.36	71.59	87.7	69.43	60.72
◆	upstage/llama-2-70B-instruct	72.29	70.9	87.48	69.8	60.97
◆	garage-blind/Camel-Platypus2-76B	71.7	71.08	87.6	70.04	58.09

Beeching, Edward, et al. "Open LLM Leaderboard." *Hugging Face* (2023).

Evaluating model preference

Can we evaluate language models in the same way as we evaluate humans?

The screenshot shows a web-based interface for a 'Playground task'. The title is 'Talk to the assistant' under 'Task 1 of 5'. The main instruction is: '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.' Below this are sections for 'Comments' (with a placeholder 'I thought the assistant was ...') and 'Rating' (with radio buttons for 'Bad', 'Good', and three intermediate options). A large blue button at the bottom says '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

Evaluating model preference

Can we evaluate language models in the same way as we evaluate humans?

Not quite:

- Very expensive
- Hard to compare results to other research

The screenshot shows a user interface for a task titled "Talk to the assistant". The task instructions ask the user to decide on a task for the AI assistant to help with, enter it into a task box, interact with the AI, and leave a comment on how the AI did. Below this, there is a "Comments" section with a text input field containing the placeholder "I thought the assistant was ...". A "Rating" section follows, with a scale from "Bad" to "Good" and three radio buttons in the middle. At the bottom is a purple "Next Task" button.

Task 1 of 5

Playground task i

Talk to the assistant

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

Evaluating model preference with LLMs

Can we evaluate language models in the same way we collect preference data
with LLMs as a judge?

[Instruction]

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]"

[Question]

{question}

[The Start of Assistant's Answer]

{answer}

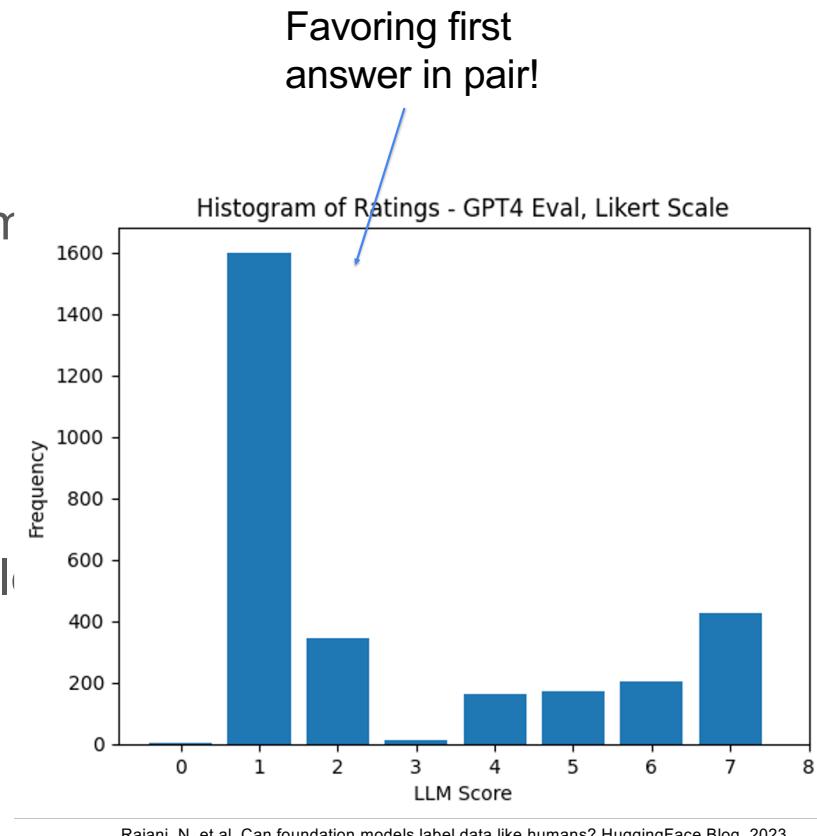
[The End of Assistant's Answer]

Evaluating model preference

Can we evaluate language models in the same way as humans?
with **LLMs as a judge?**

Not quite:

- Models are heavily positionally biased
- Models often rate on syntax and response length



MT Bench

LLM-as-a-judge: ask a LLM (GPT4/Claude) to rate a model response:

- Two turns (response & follow-up)
- 7 categories (writing, role-play, math, coding, extraction, STEM, humanities)
- Rate one model at a time 0-10 rating scale to mitigate positional bias

MT Bench

Example

Question: Write me a cover letter for...

LLM Agent: Dear Hiring Manager, ...

LLM Judge: [[7]] The message...

Question: Update the letter starting every paragraph with...

LLM Agent: Dear Hiring Manager,

...

LLM Judge: [[5]] The message...

Score: 6/10

MT Bench

LLM-as-a-judge: ask a LLM (GPT4/Claude) to rate a model response:

- Two turns (response & follow-up)
- 7 categories (writing, role-play, math, coding, extraction, STEM, humanities)
- 0-10 rating scale

Shortcomings: ***hard to use as sole focus during training***

- Variance in scoring up to ~0.5 points, big deltas needed for signal
(via generation temperature and model API variation)
- Scoring saturated at top end (GPT4: 8.99)

Future work: Chat bench

Designing challenging many-turn prompt sequences:

- LLM evaluation after each generation
- LLM scores [[pass]] or [[fail]] for each prompt-generation pair
- Score based on completion, rather than variable scores

Longer development life cycle!

Future work: Chat bench

MT Bench

Example

Question: Write me a cover letter for...

LLM Agent: Dear Hiring Manager,

...

LLM Judge: [[7]] The message...

Question: Update the letter starting every paragraph with...

LLM Agent: Dear Hiring Manager,

...

LLM Judge: [[5]] The message...

Score: 6/10

Chat Bench Idea

Question: Write me a poem about...

LLM Agent: {{poem}}

LLM Judge: [[pass]]

Question: Re-write the poem as a sonnet.

LLM Agent: {{poem}}

LLM Judge: [[pass]]

Question: Re-write the poem as a limerick.

LLM Agent: {{poem}}

LLM Judge: [[pass]]

Question: Re-write the poem as a Greek Epic.

LLM Agent: {{poem}}

LLM Judge: [[fail]]

Towards better RLHF evaluation

- Chat is an extremely challenging domain
- Training metrics are not linked to evaluation metrics
- Future work to design more complete evaluation suite