

Reinforcement Learning from Human Feedback

Nathan Lambert, Allen Institute for AI

Stanford CS329H: Machine Learning from Human Preferences
29 Nov 2023

A heavily abbreviated history of LLMs

1948: Claude Shannon models English

3. THE SERIES OF APPROXIMATIONS TO ENGLISH

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESSEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , n -gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Shannon 1948

Stanford CS329H: 2



A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 🍩

50s: the turing test

60s: ELIZA, chatbot for therapy

70s-80s: more chatbots, statistical approaches

90s-00s: language modeling

00s-10s: word embeddings

A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 🥵

50s: the turing test

60s: ELIZA, chatbot for therapy

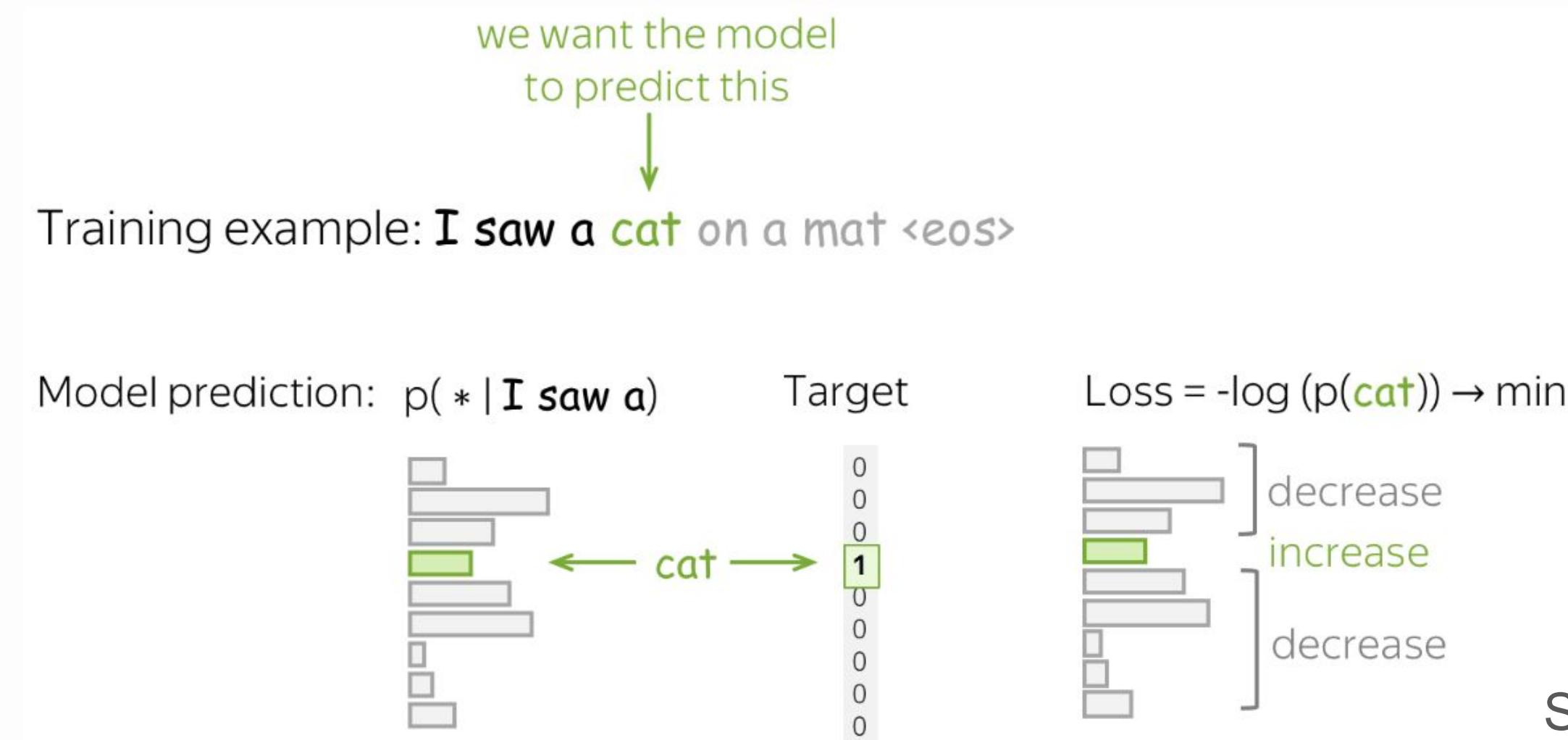
70s-80s: more chatbots, statistical approaches

90s-00s: language modeling

00s-10s: word embeddings

$$Loss(p^*, p) = -\log(p_{y_t}) = -\log(p(y_t|y_{<t})).$$

At each step, we maximize the probability a model assigns to the correct token. Look at the illustration for a single timestep.

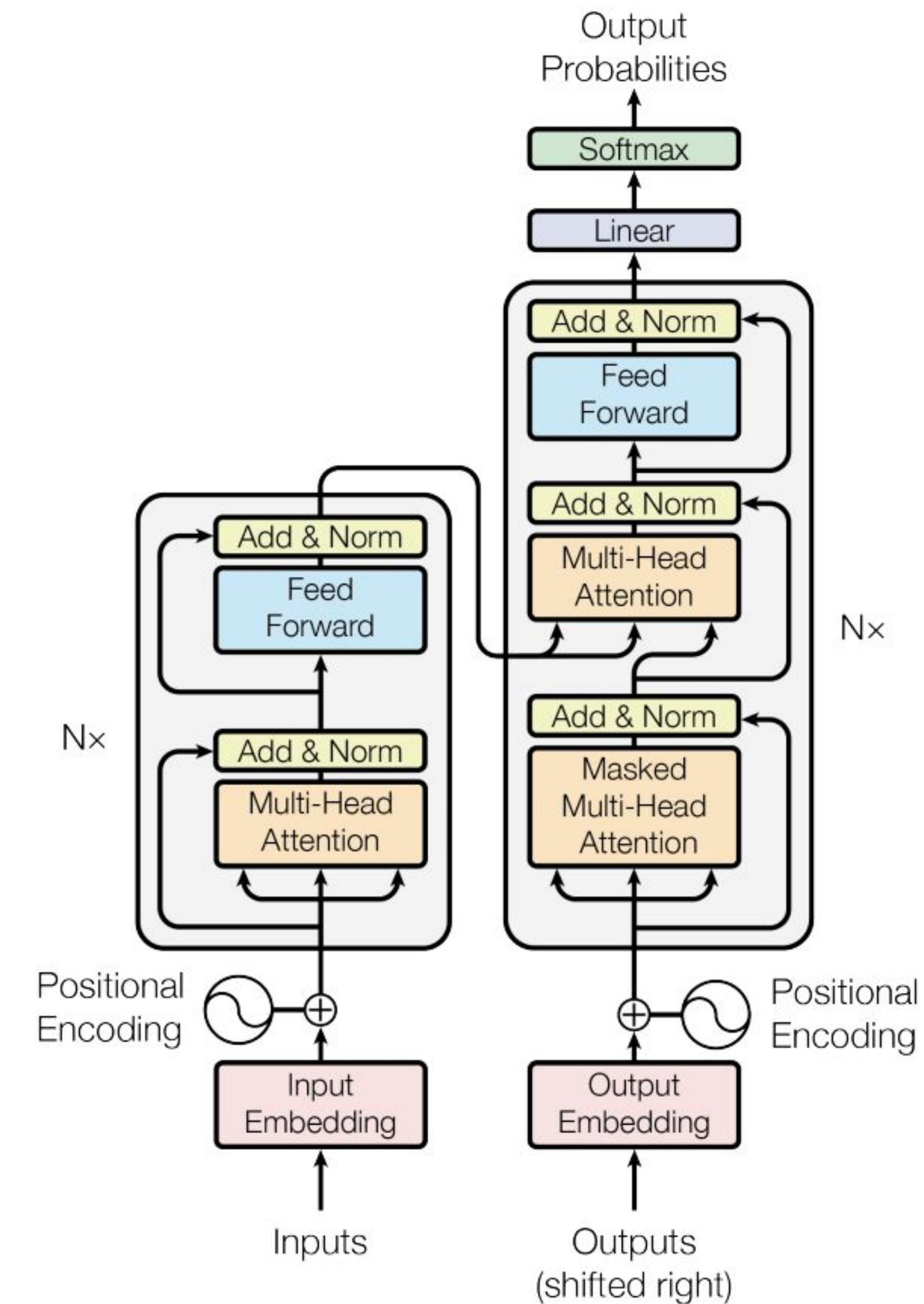


A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 🍔

2017: the transformer is born



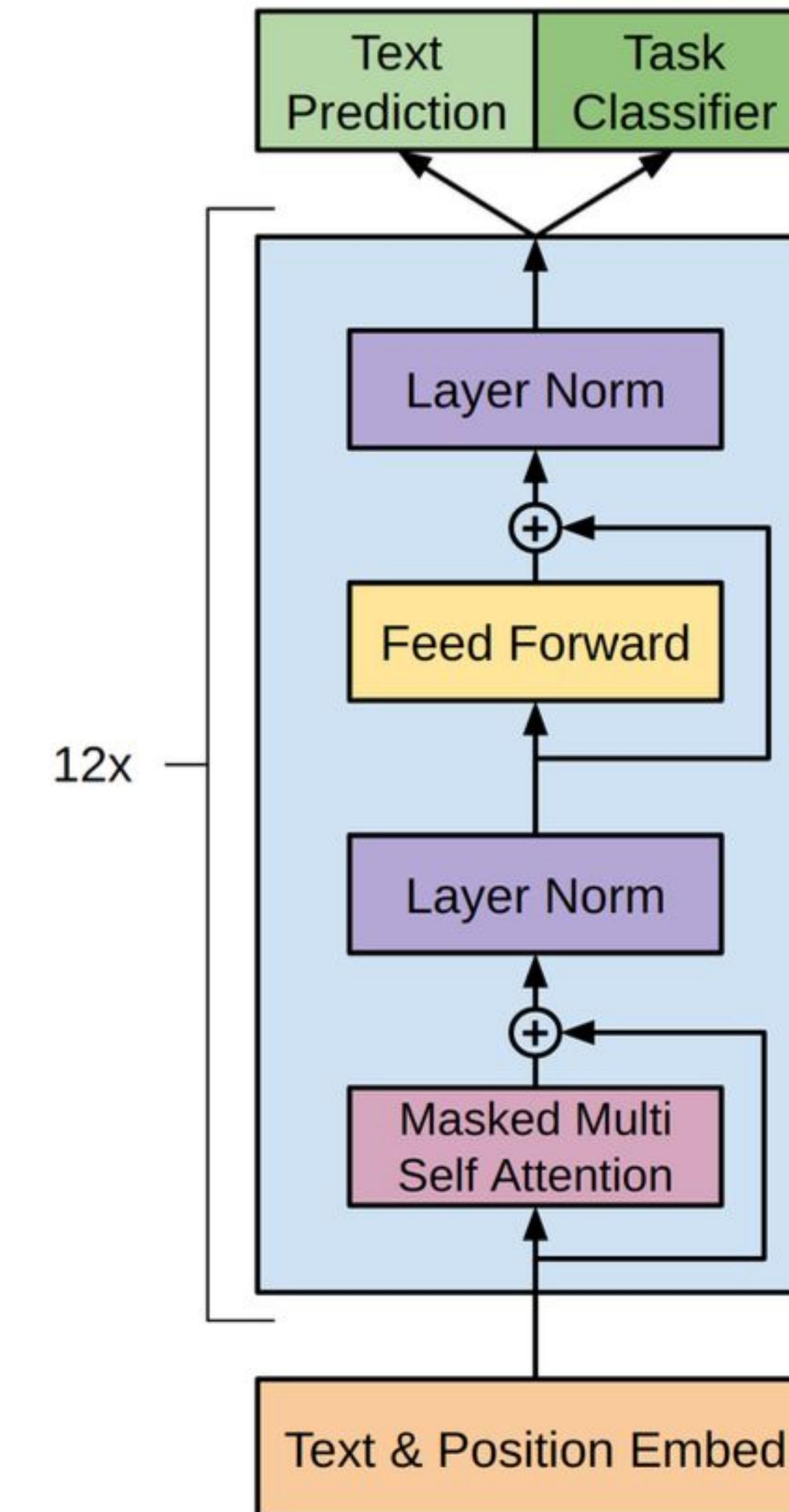
A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 🥖

2017: the transformer is born

2018: GPT-1 and BERT released



A heavily abbreviated history of LLMs

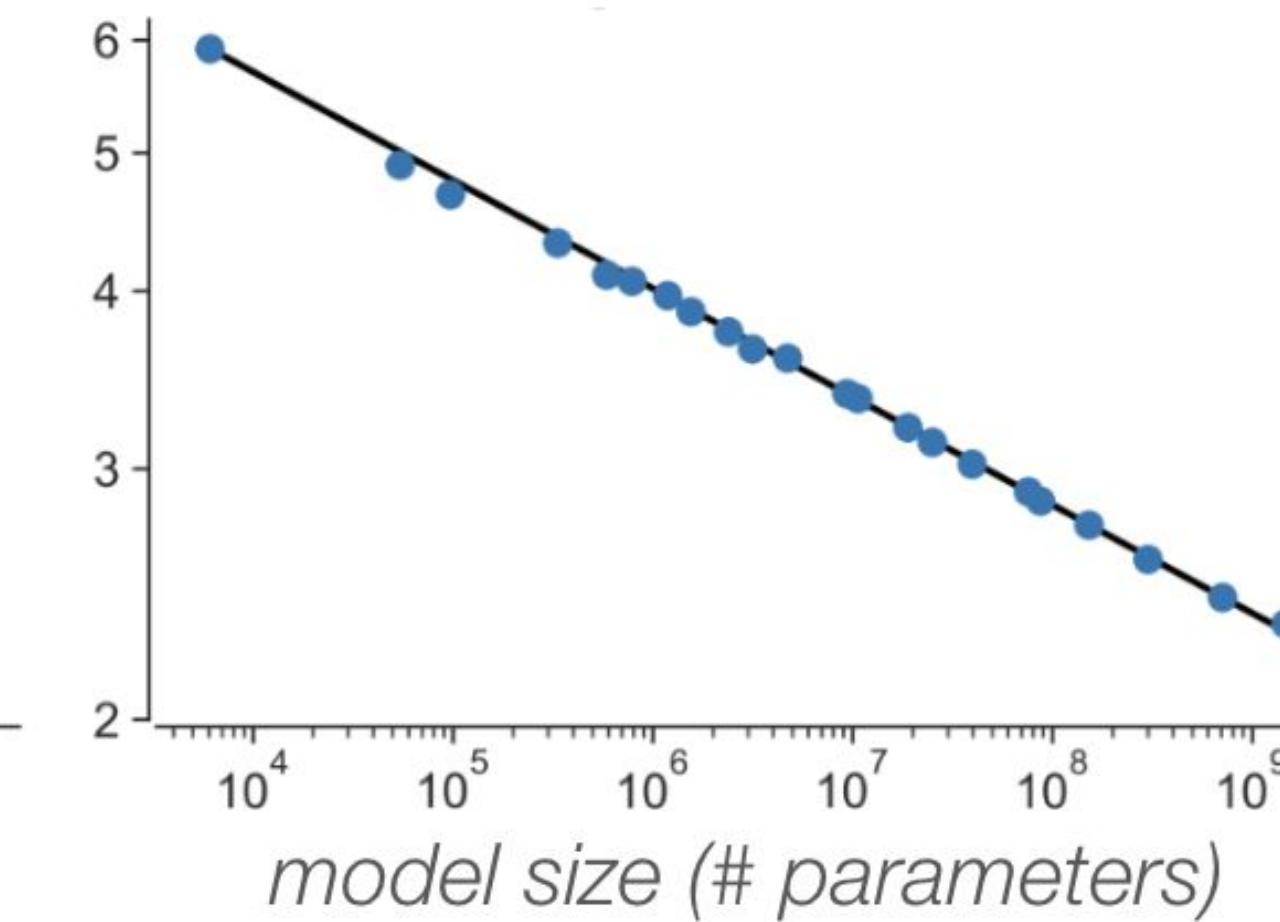
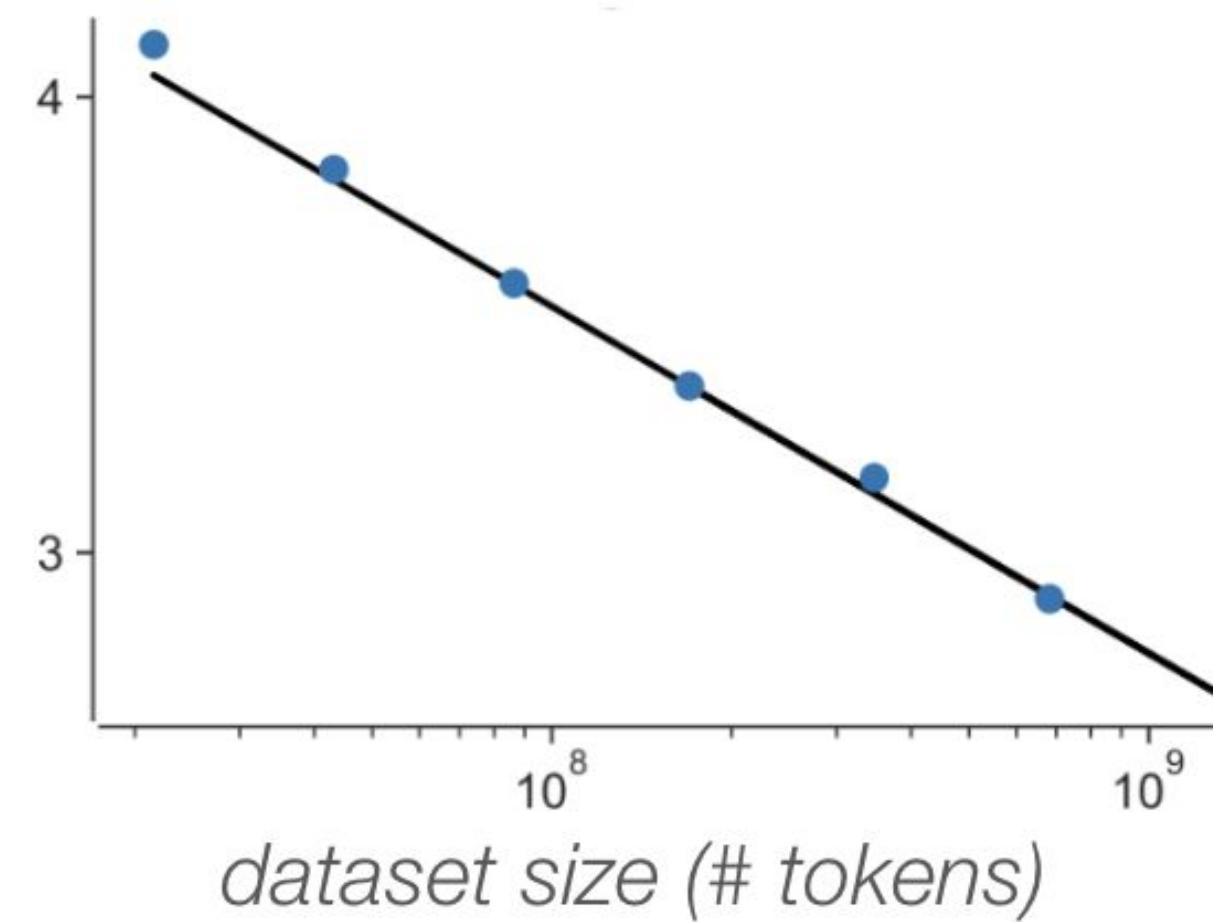
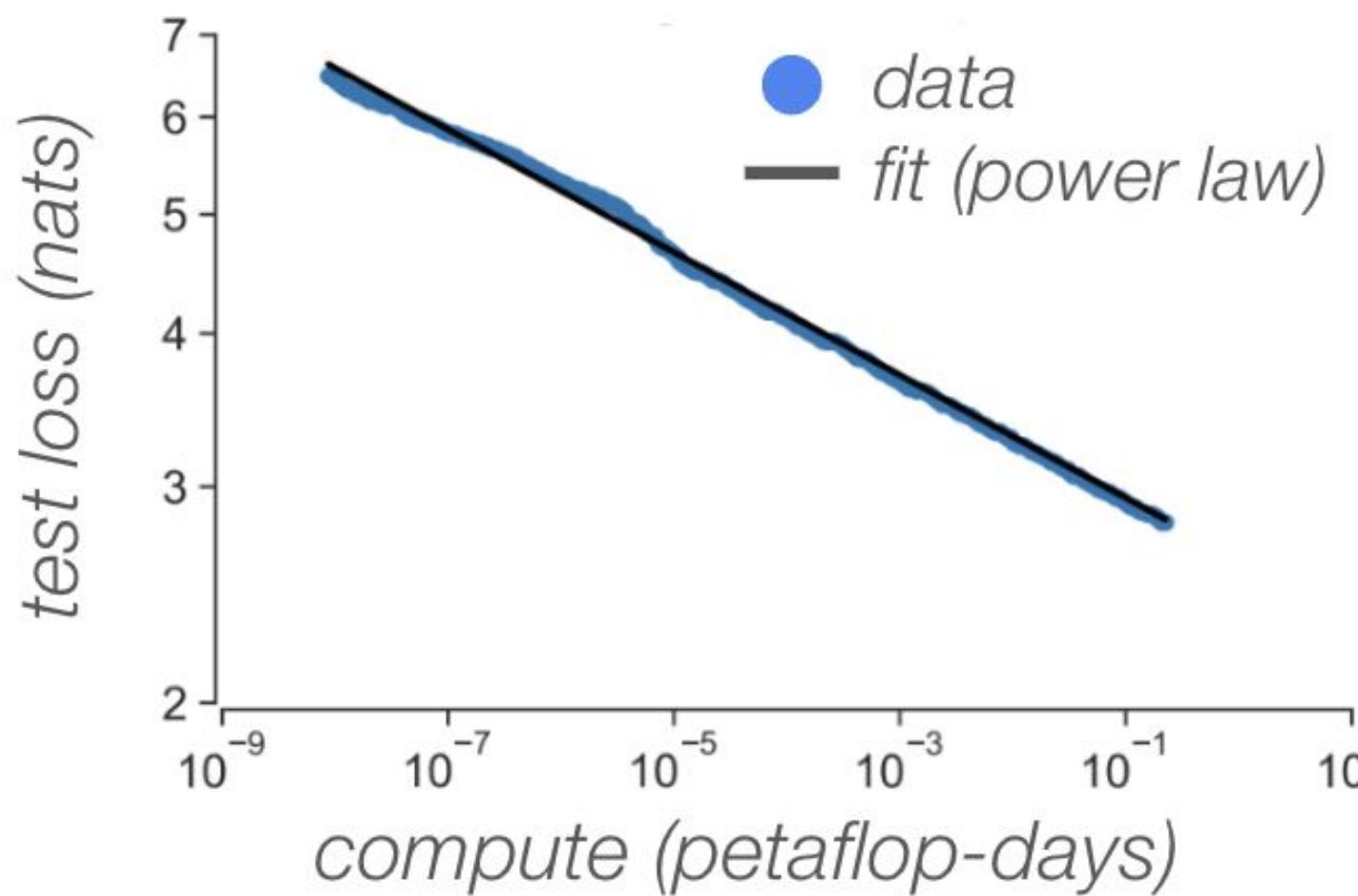
1948: Claude Shannon models English

1948-2017: 🍔

2017: the transformer is born

2018: GPT-1 and BERT released

2019: GPT-2 and scaling laws



A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 

2017: the transformer is born

2018: GPT-1 and BERT released

2019: GPT-2 and scaling laws

OpenAI Report
November, 2019

Release Strategies and the Social Impacts of Language Models

Irene Solaiman*	Miles Brundage	Jack Clark	Amanda Askell
OpenAI	OpenAI	OpenAI	OpenAI
irene@openai.com	miles@openai.com	jack@openai.com	amanda@openai.com

Ariel Herbert-Voss	Jeff Wu	Alec Radford
Harvard University	OpenAI	OpenAI
ariel_herbertvoss@g.harvard.edu	jeffwu@openai.com	alec@openai.com

Gretchen Krueger	Jong Wook Kim	Sarah Kreps
OpenAI	OpenAI	Cornell University
gretchen@openai.com	jongwook@openai.com	sarah.kreps@cornell.edu

Miles McCain	Alex Newhouse	Jason Blazakis
Politiwatch	CTEC	CTEC
miles@rmrrm.io	anewhouse@middlebury.edu	jblazakis@middlebury.edu

Kris McGuffie	Jasmine Wang
CTEC	OpenAI
Kmcguffie@middlebury.edu	jasmine@openai.com

A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 🧀

2017: the transformer is born

2018: GPT-1 and BERT released

2019: GPT-2 and scaling laws

2020: GPT-3 surprising capabilities. many harms

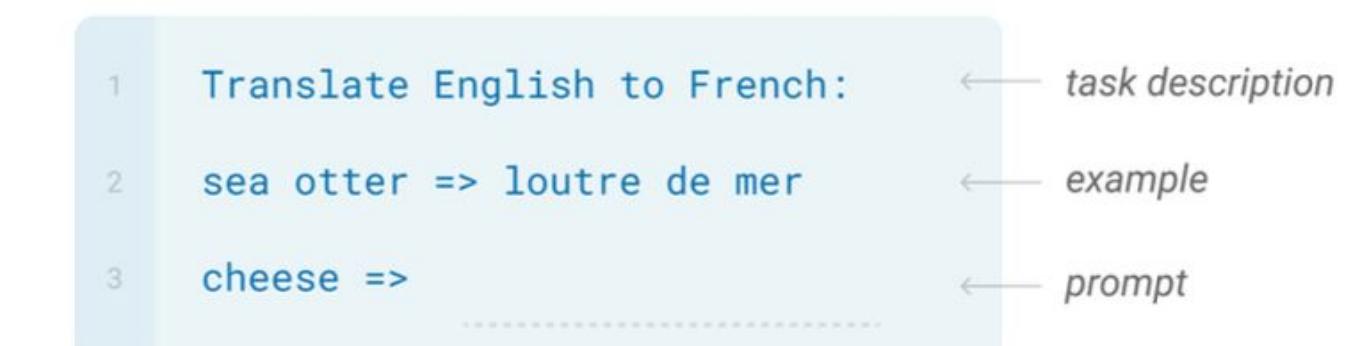
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



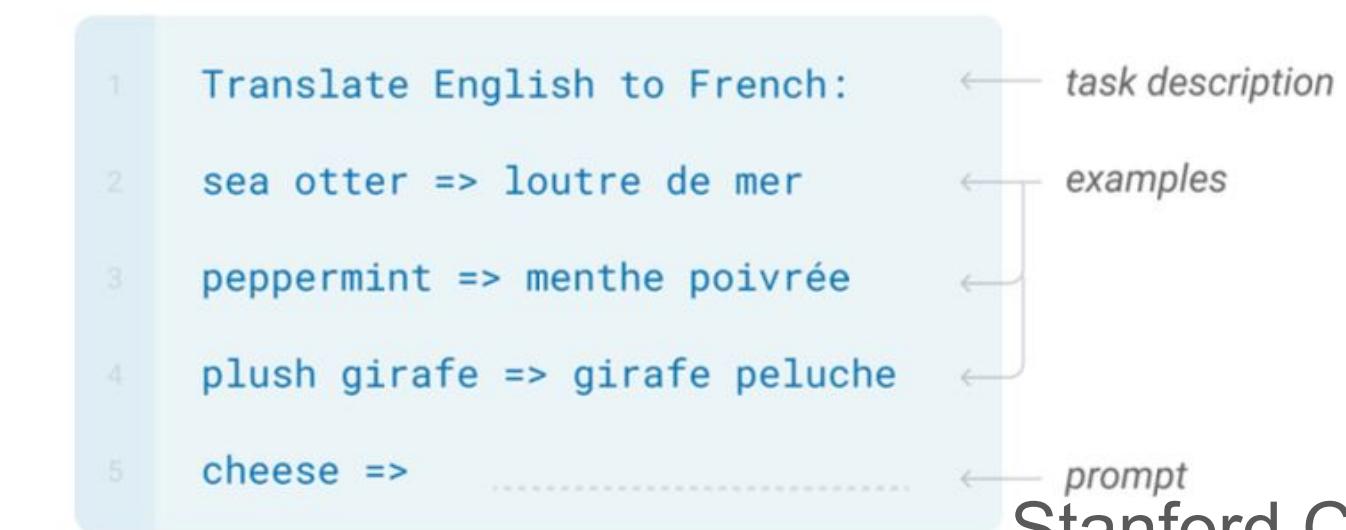
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 

2017: the transformer is born

2018: GPT-1 and BERT released

2019: GPT-2 and scaling laws

2020: GPT-3 surprising capabilities

2021: stochastic parrots

“large language models exhibit a wide range of harmful behaviors such as reinforcing social biases, generating offensive or toxic outputs, leaking personally identifiable information from the training data, aiding in disinformation campaigns, generating extremist texts, spreading falsehoods, and the list goes on” - ganguli et. al, 2022

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

A heavily abbreviated history of LLMs

1948: Claude Shannon models English

1948-2017: 

2017: the transformer is born

2018: GPT-1 and BERT released

2019: GPT-2 and scaling laws

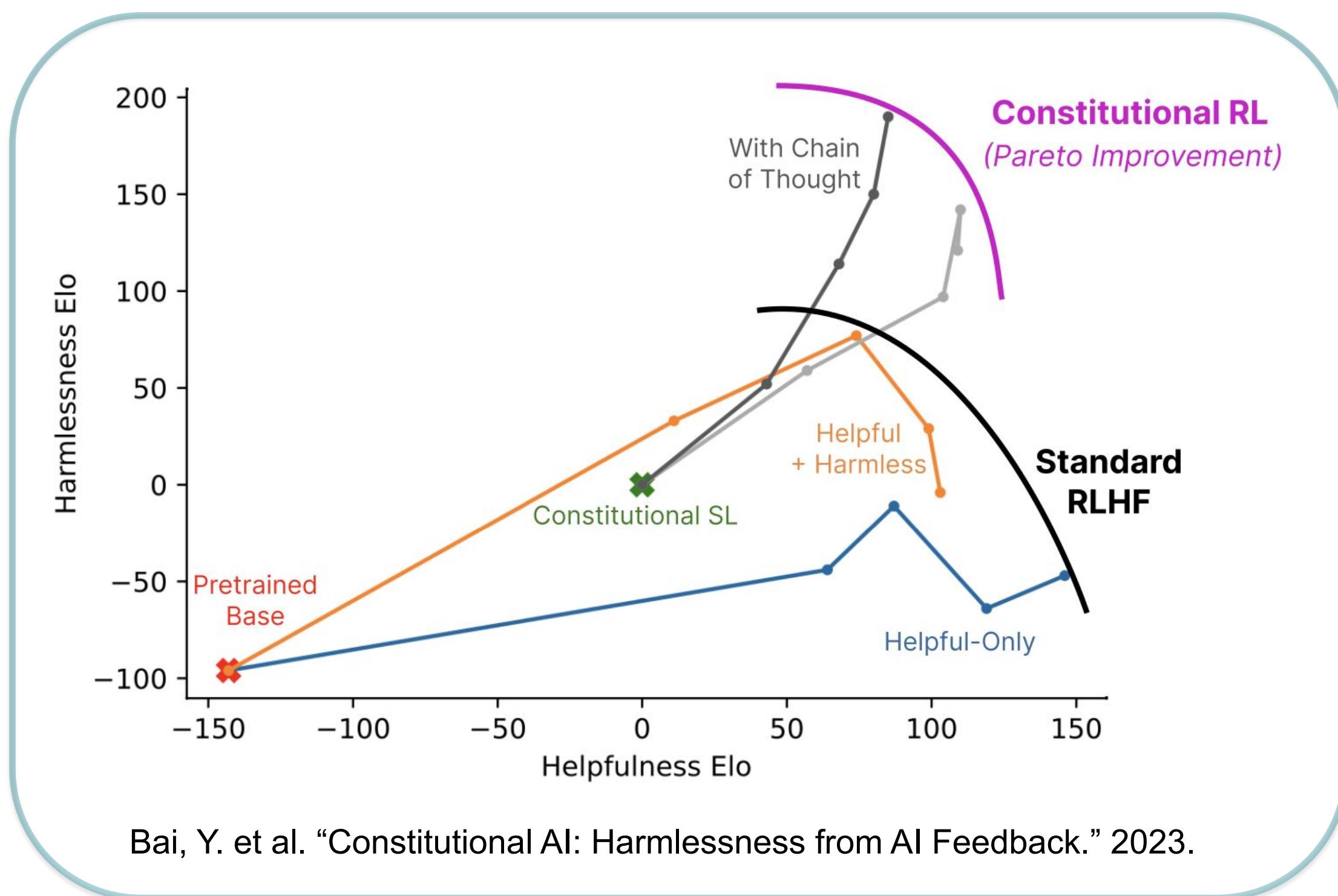
2020: GPT-3 surprising capabilities

2021: stochastic parrots

2022: ChatGPT, Claude, RLHF

RLHF is relied upon

RLHF is a key factor in many popular models, both on and off the record, including ChatGPT, Bard, Claude, Llama 2, and more



"Meanwhile reinforcement learning, known for its instability, seemed a somewhat shadowy field for those in the NLP research community. However, reinforcement learning proved highly effective, particularly given its cost and time effectiveness."

- Touvron, H. et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." 2023

Why Reinforcement Learning from Human Feedback

How do you create / code a loss function for:

Why Reinforcement Learning from Human Feedback

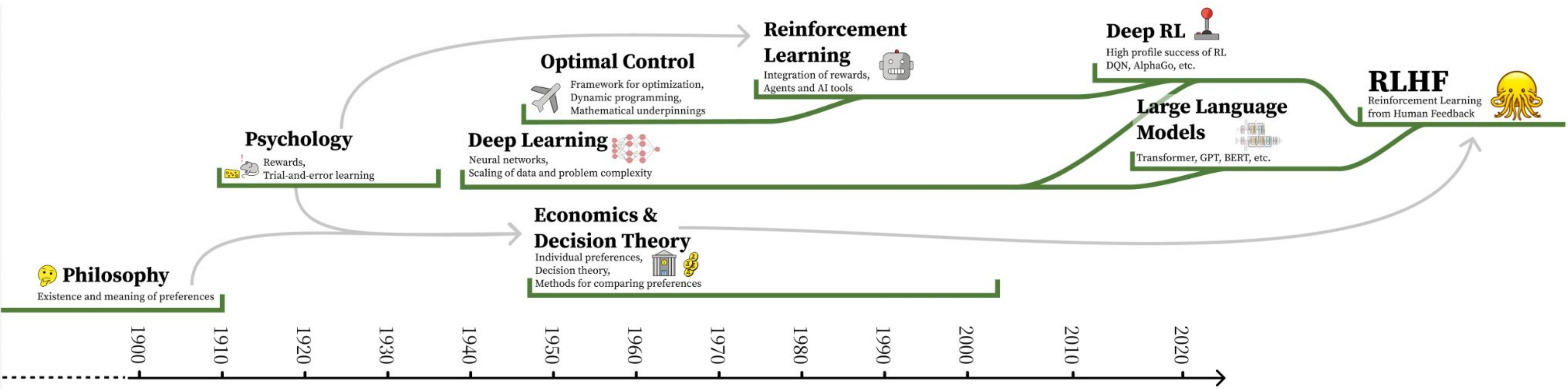
How do you create / code a loss function for:

- What is *funny*?
- What is *ethical*?
- What is *safe*?

Don't encode it, model it!

Outline

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



History

Goal: demonstrate core ideas behind modern practices of RLHF

- **Assumptions:** Theories in creation process
- **Presumptions:** Tools and habits in the creation process
- *Not exhaustive*

Paper is here



Motivating idea: costs vs. rewards vs. preferences

Costs: Control theory, optimal control, etc.

Rewards: RL algorithms, trial and error learning....

Preferences: Social values, human decision making...

Paper is here



Starting point: *Human preferences and goals exist*

- Foundation of many fields of study (economics, RL, psychology,...)
- Contested by some philosophers (Arrow, 1950; Maskin & Sen, 2014)!

Assumption: *Any and all preferences and goals can be quantified and measured.*

- 1662 Arnauld's *The Port Royal Logic*

“To judge what one must do to obtain a good or avoid an evil, it is necessary to consider not only the good and evil in itself, but also the probability that it happens or does not happen.”

Assumption: Any and all preferences and goals can be quantified and measured.

- 1662 Arnauld's *The Port Royal Logic*
- 1823 Bentham's *Hedonic Calculus* → Utilitarianism
- 1931 Ramsey's *Truth and Probability* → applied quantitative preferences

Assumption: Increasing the score of raw reward measurements corresponds to better behaviors.

- 1938 Notion of *reward* (Operant conditioning)
- ~1950s Utility-to-go from optimal control (from analog circuits!)

Assumption: Optimal solutions to reward maximization problems exist.

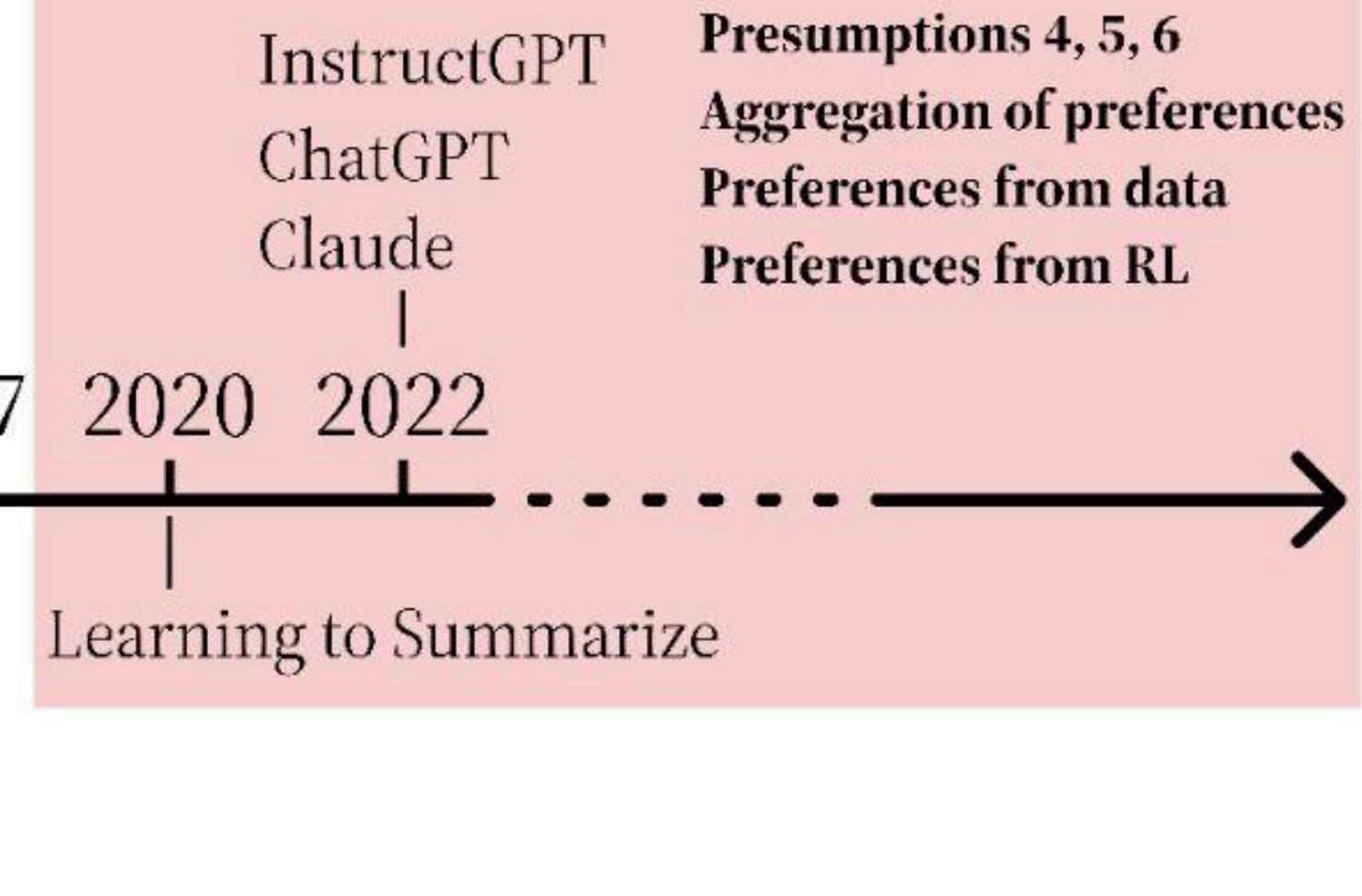
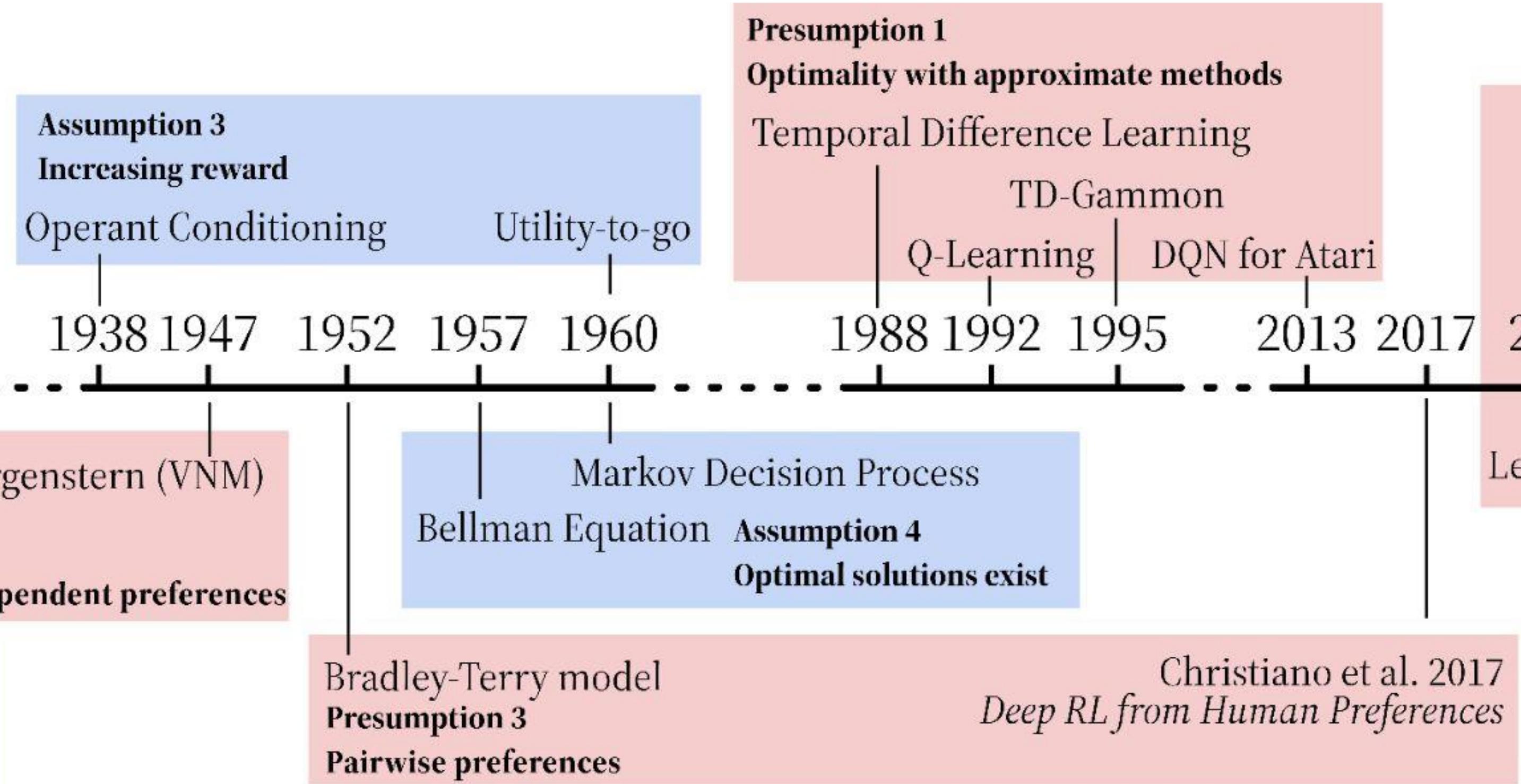
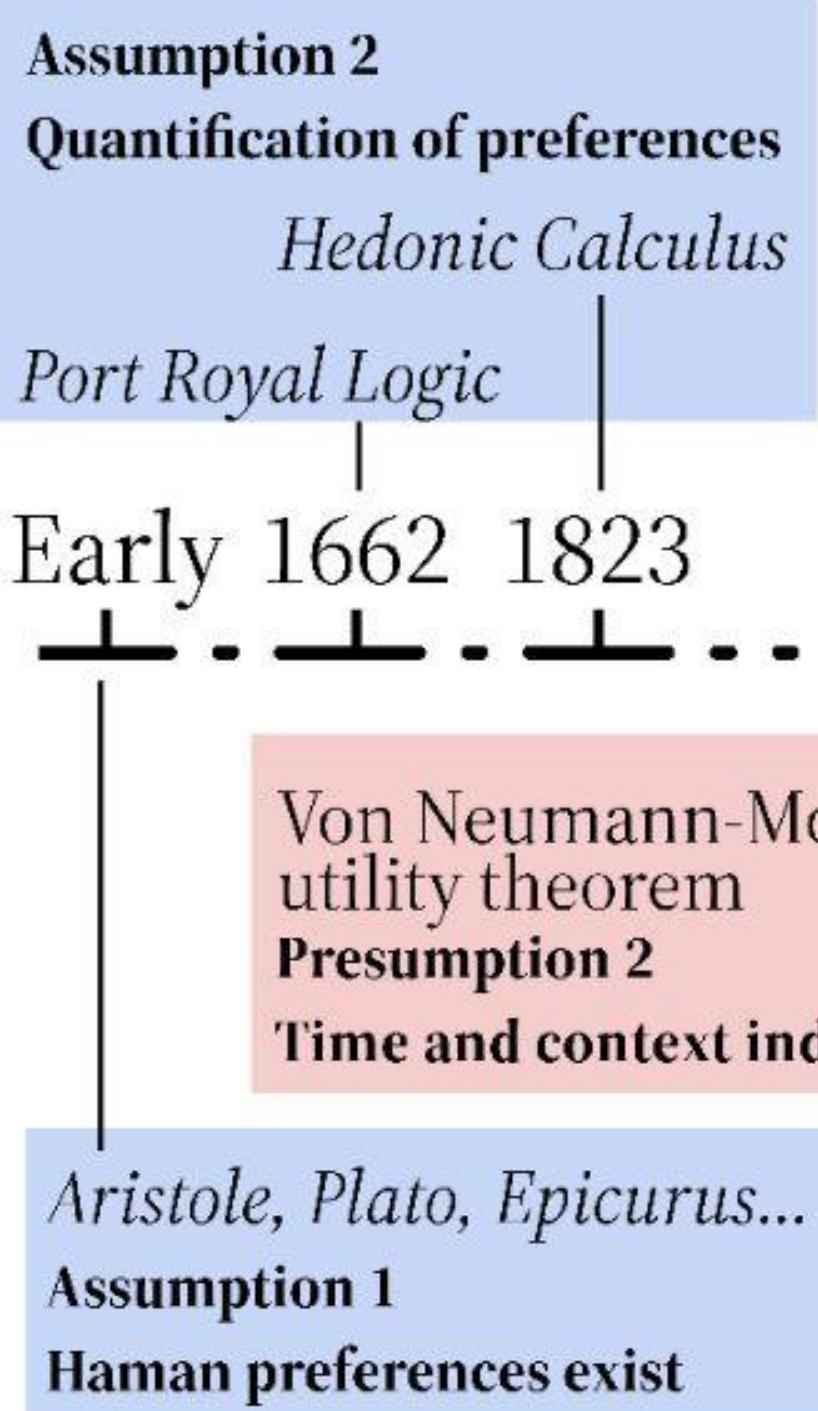
- 1957 Bellman equation
- 1960 Markov Decision Process

Presumptions of RLHF

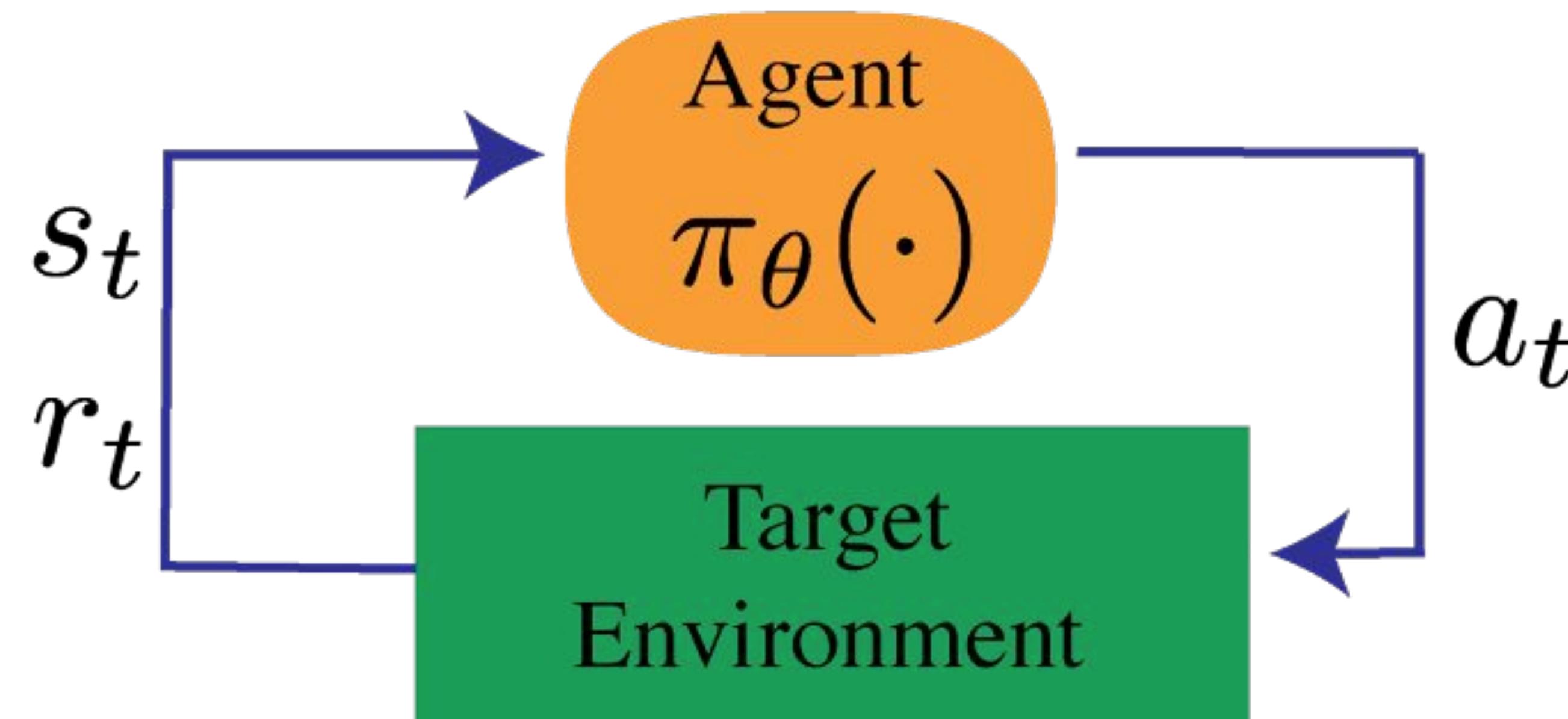
1. “RL works”
2. “Von Neumann-Morgenstern (VNM) utility theorem”
3. “Bradley-Terry model”
4. “Aggregation of preferences”
5. “RLHF independence - data”
6. “RLHF independence - training”

Presumptions of RLHF

1. **“RL works”** Optimal solutions can be achieved with finite data in complex environments.
2. **“Von Neumann-Morgenstern (VNM) utility theorem”** The temporal- and context-shifting of user preferences does not mitigate the effectiveness of reward functions or notions of optimal utility as an optimization target.
3. **“Bradley-Terry model”** Pairwise preferences can suitably perform as a basis of human values.
4. **“Aggregation of preferences”** Multiple user preferences are successfully represented in training one model by aggregating and comparing individual utilities.
5. **“RLHF independence - data”** The only preferences embedded in the model are from the specifically collected training data.
6. **“RLHF independence - training”** User preferences are extracted uniformly via the RLHF process.



Review: reinforcement learning basics



Some notation:

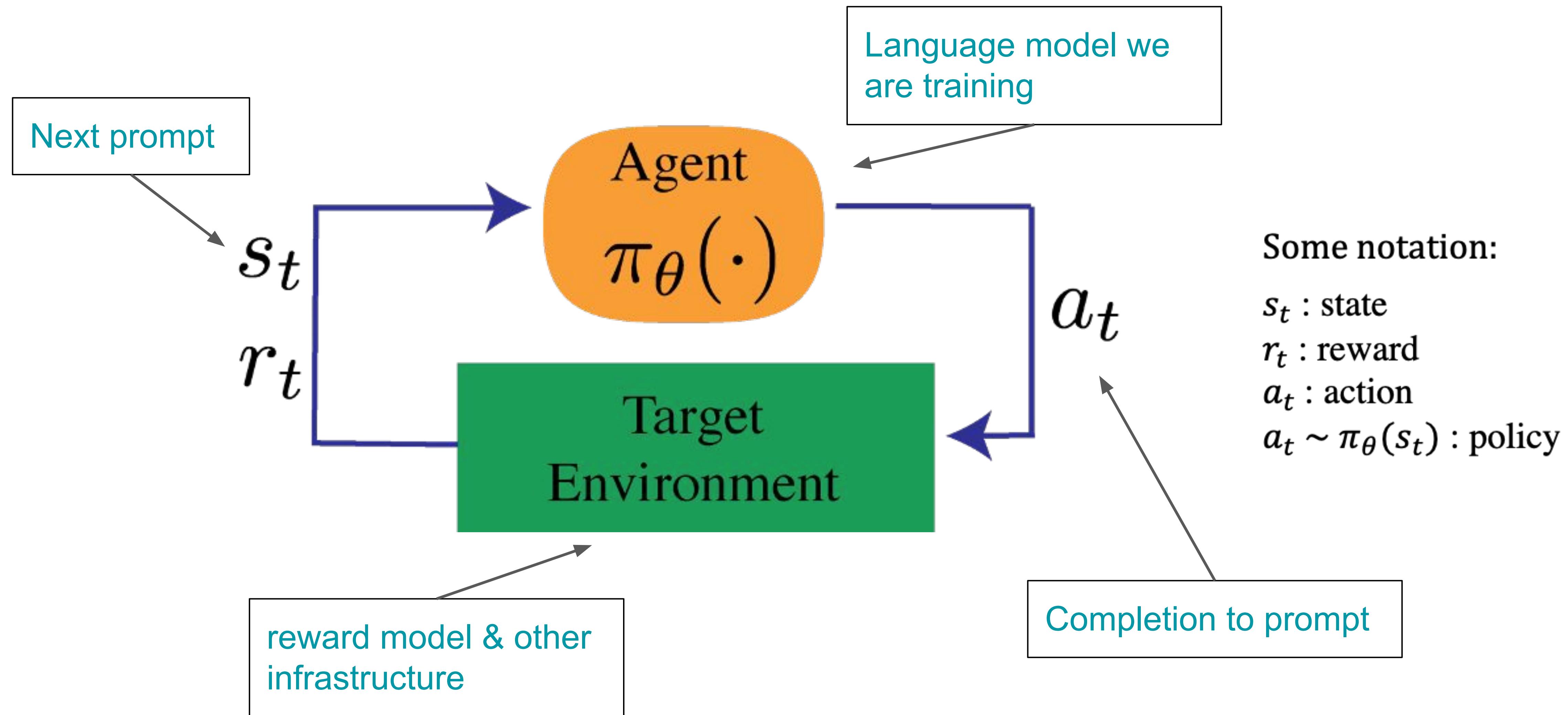
s_t : state

r_t : reward

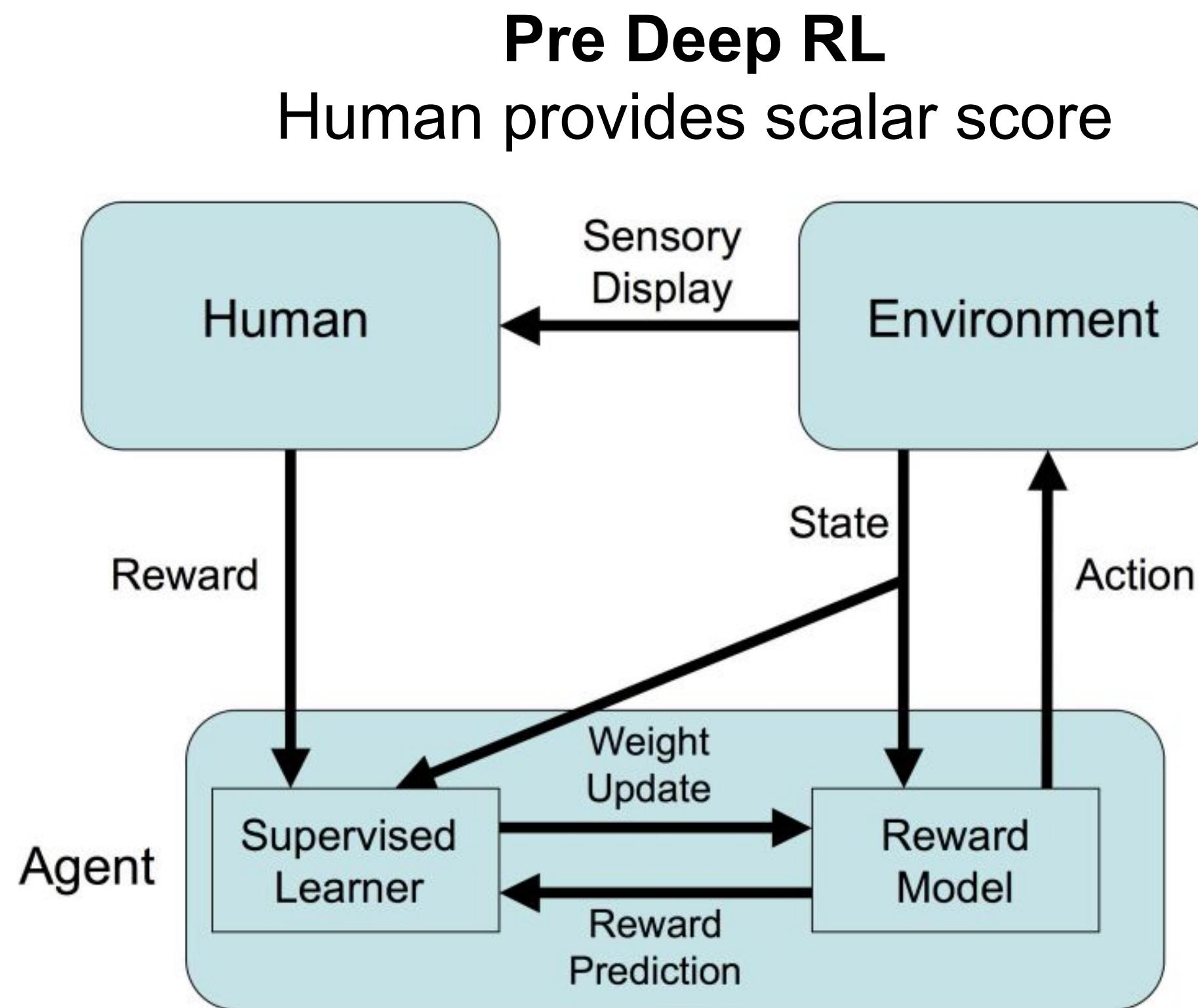
a_t : action

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

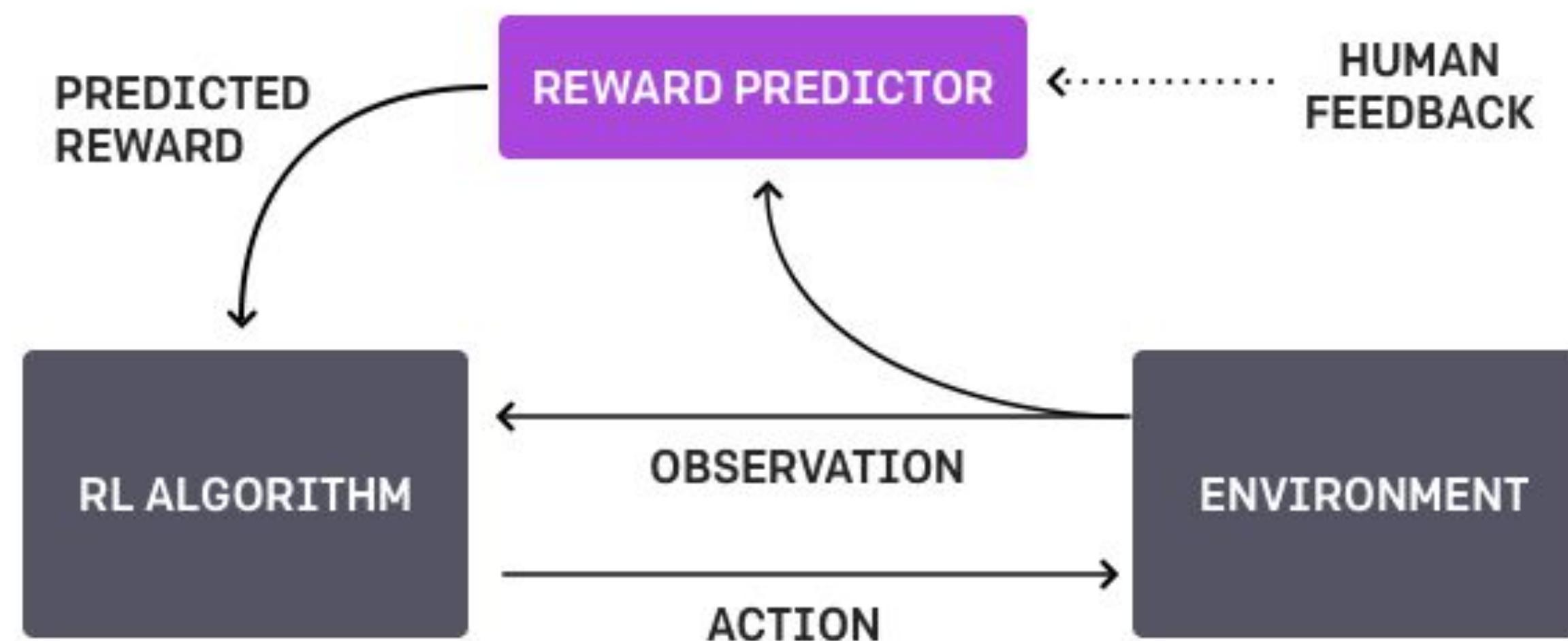
Review: reinforcement learning basics in language



History: RLHF for decision making



With Deep RL
Human compares trajectories



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

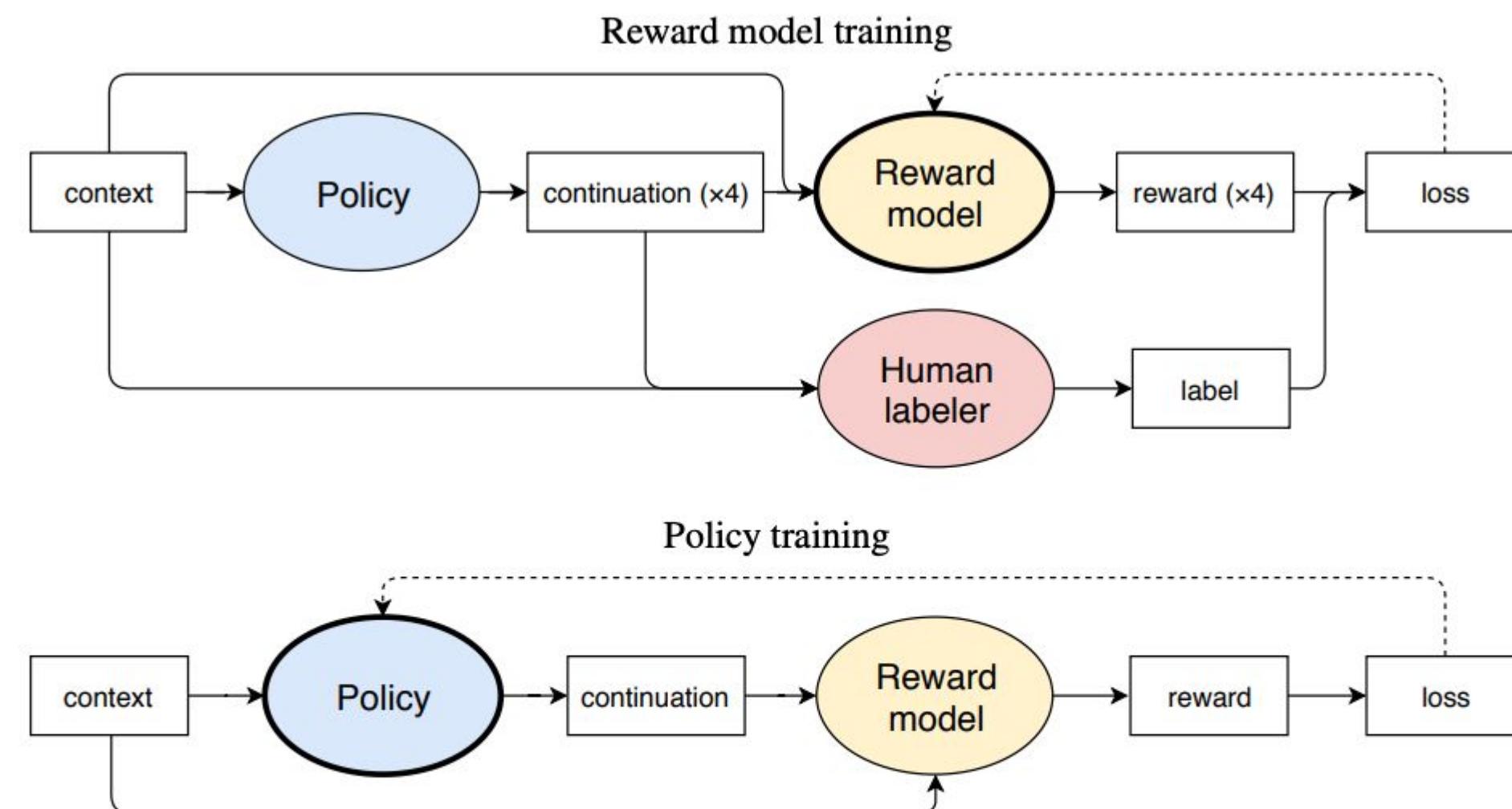
Nov. 2018

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



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

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).

History: early OpenAI experiments with RLHF

Sep. 2020

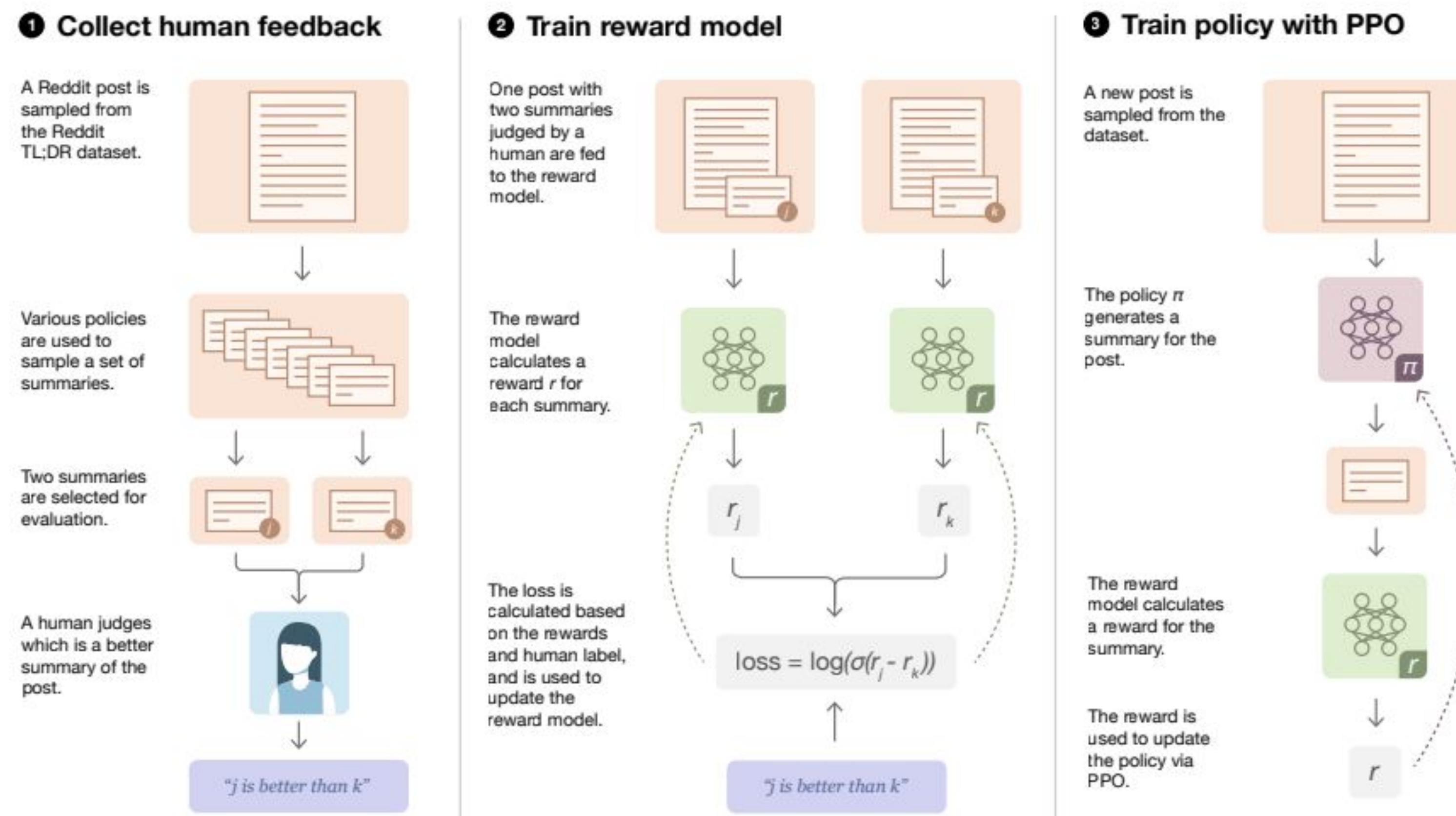
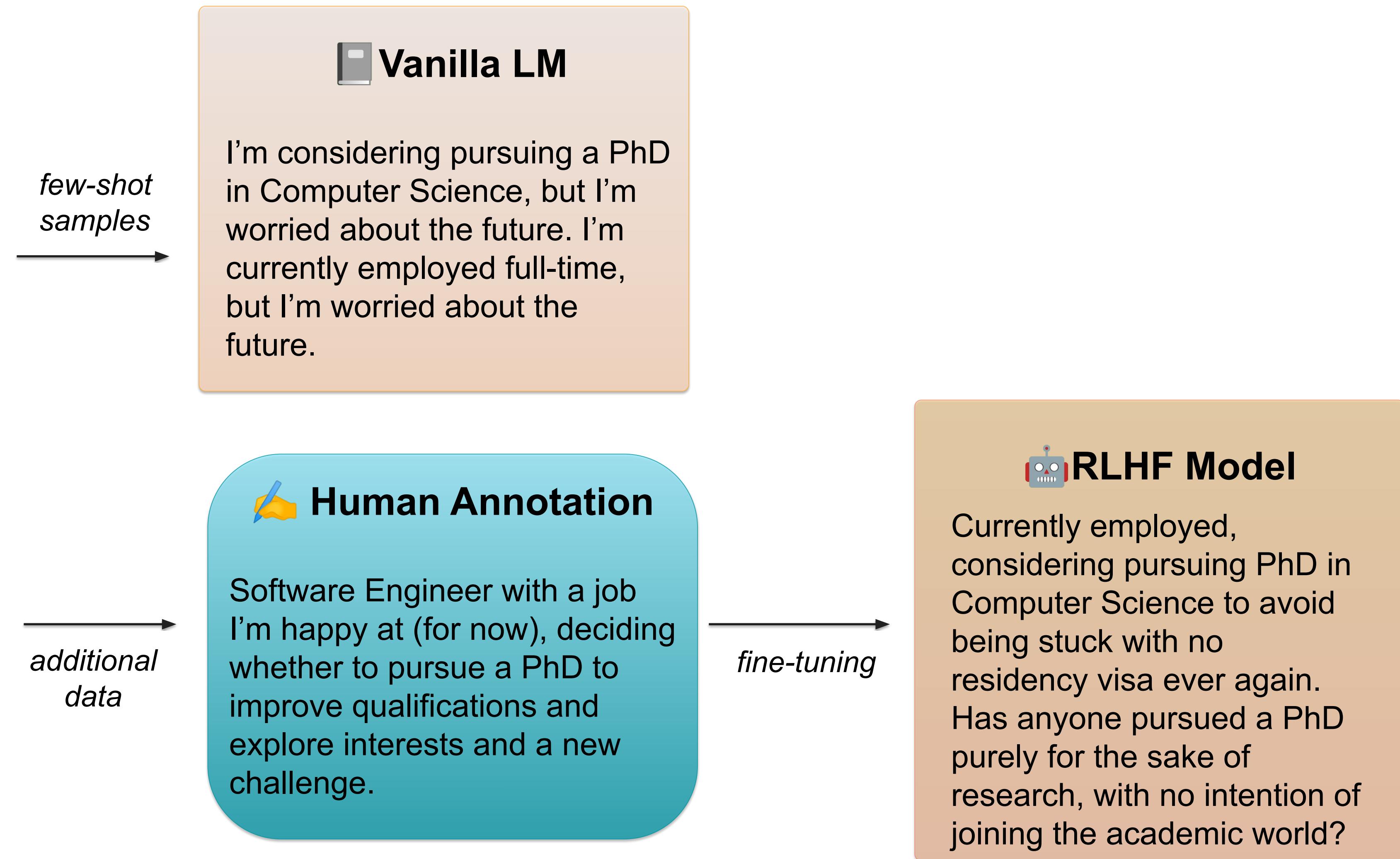
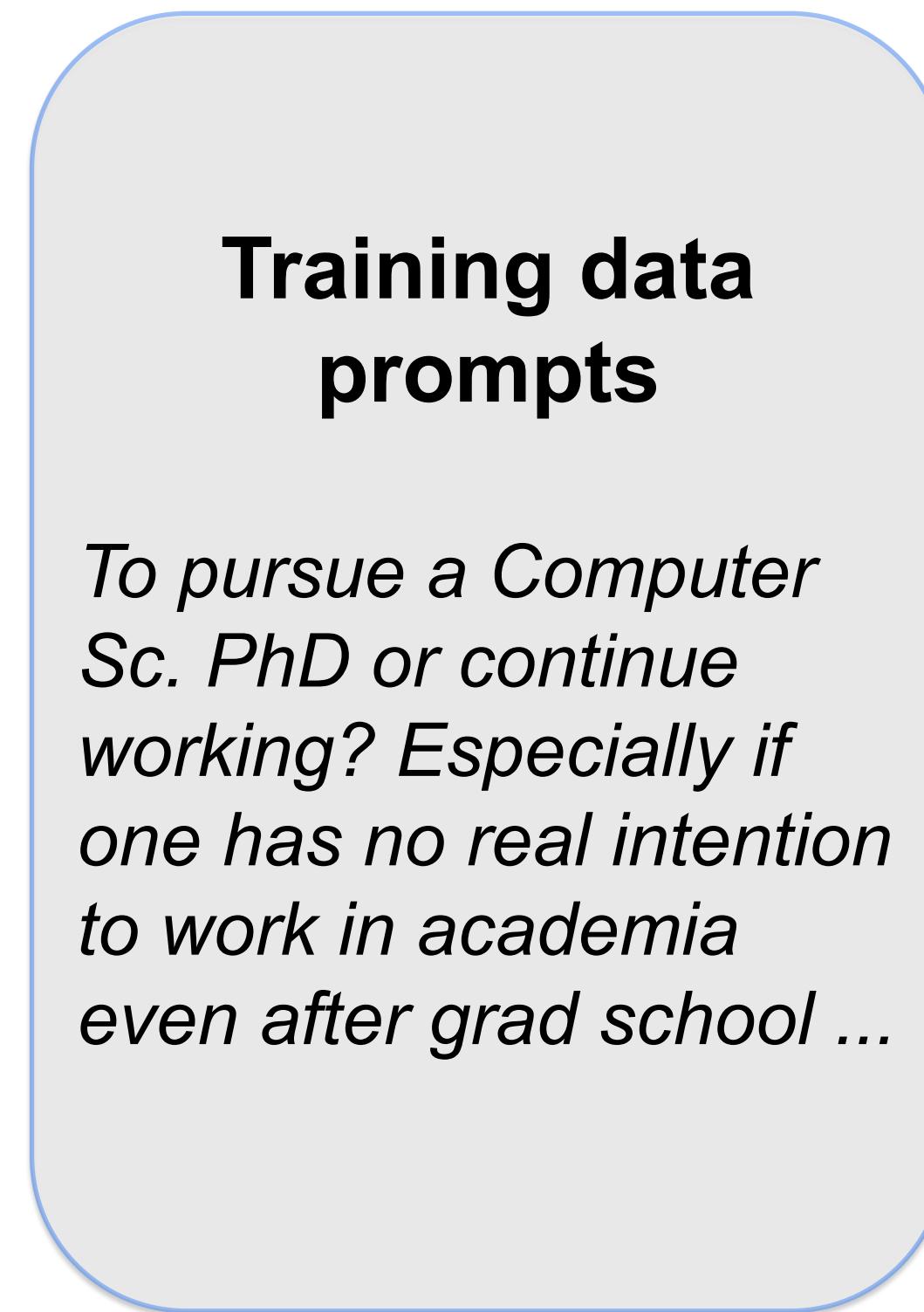


Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

History: early OpenAI experiments with RLHF

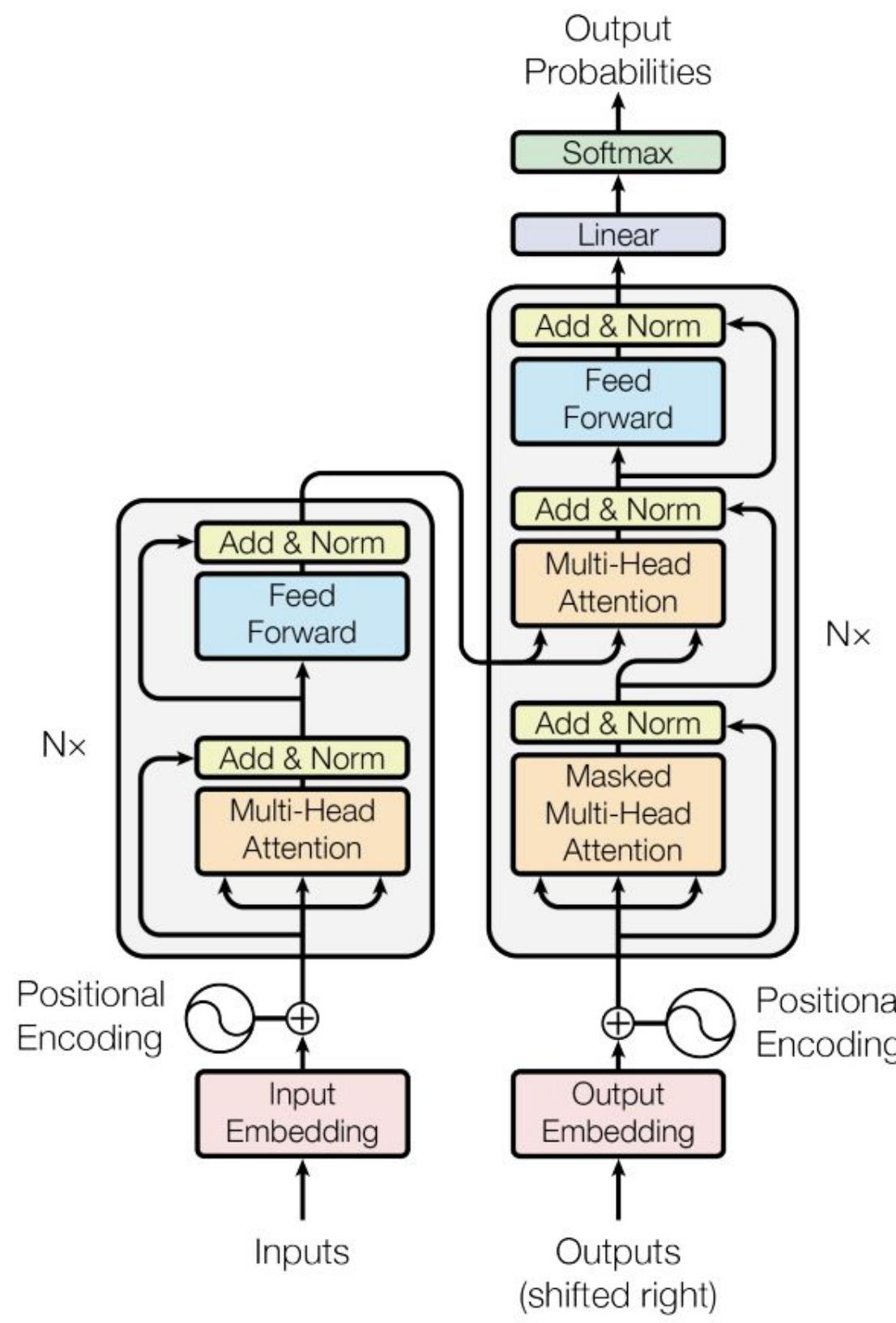


Outline

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

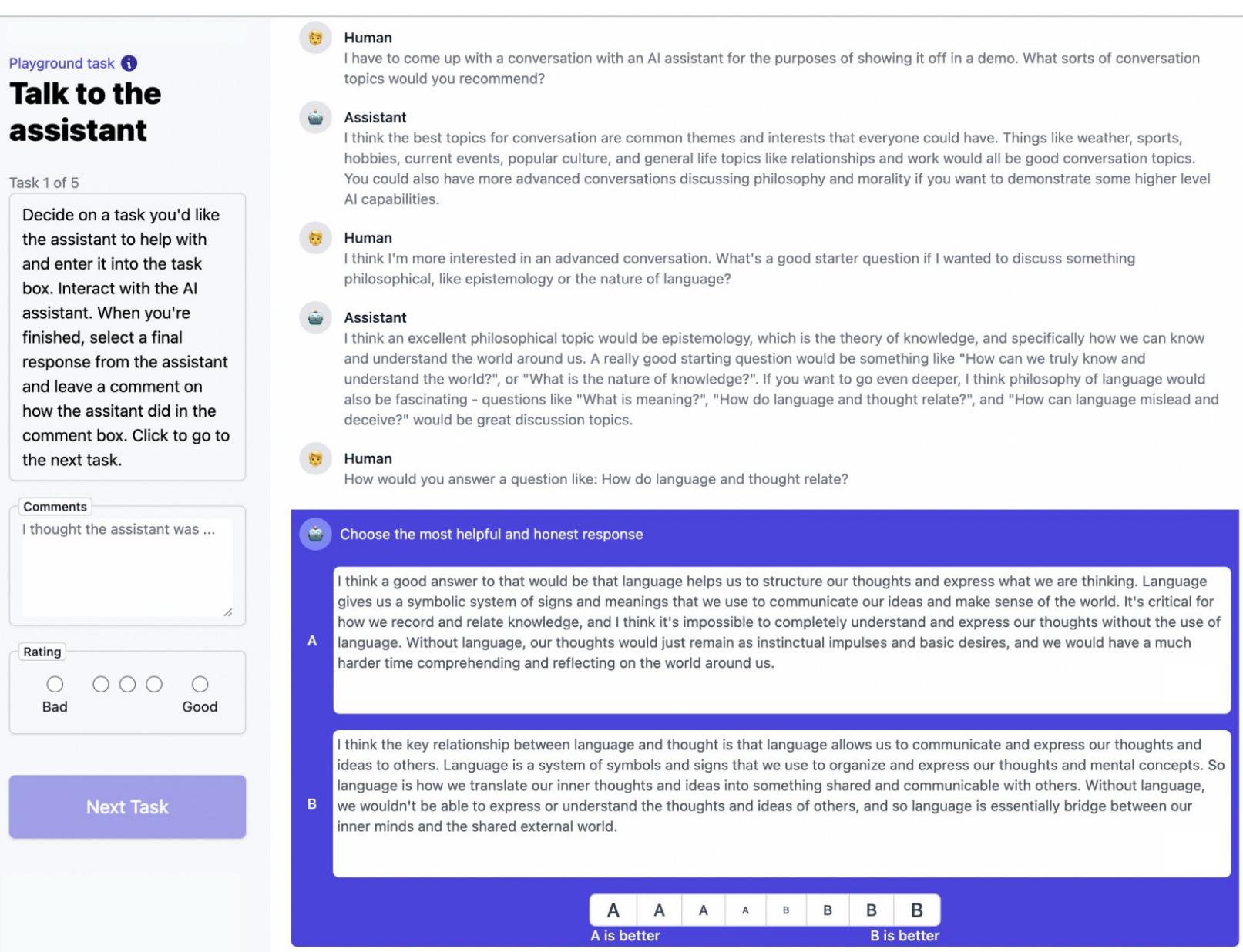
Three phases of RLHF

base model (instruction, helpful, chatty etc.)

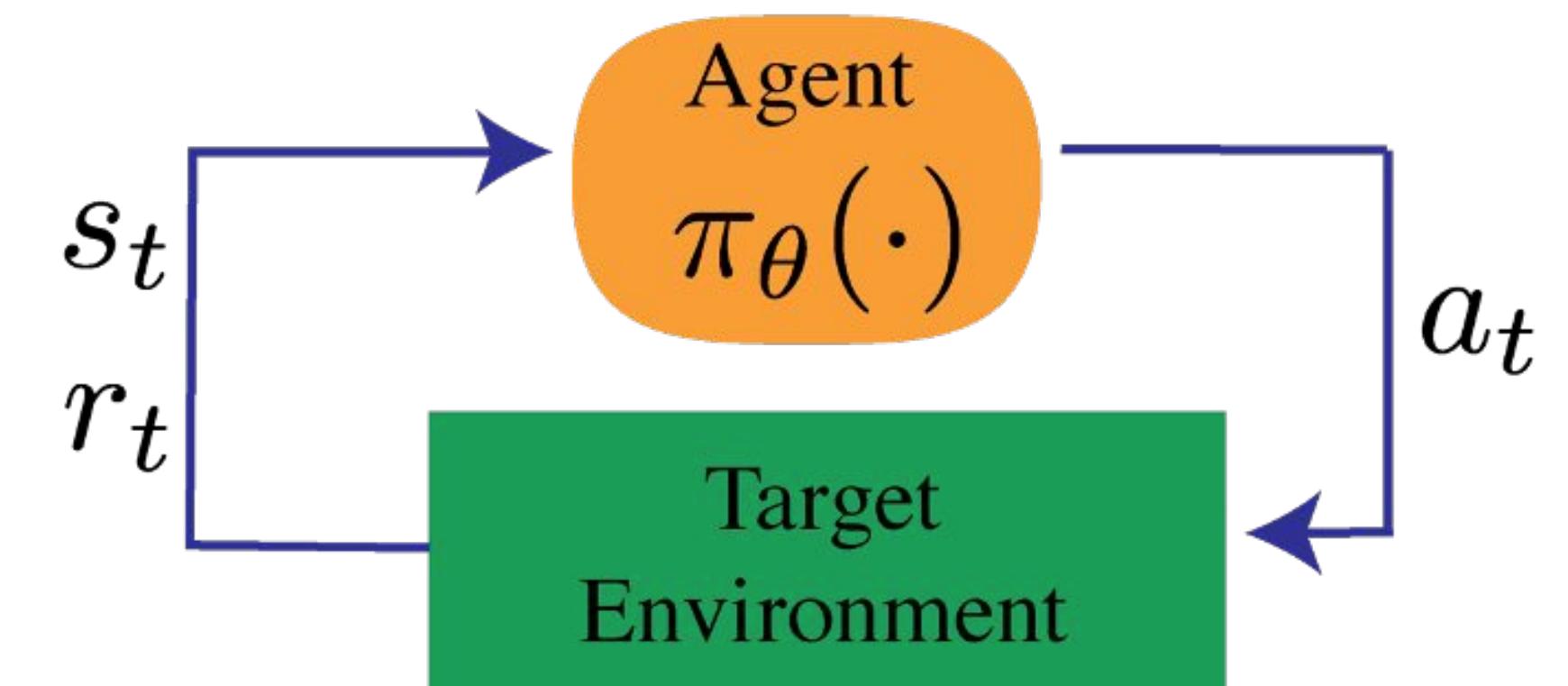


Vaswani et al. 2017

preference collection & training

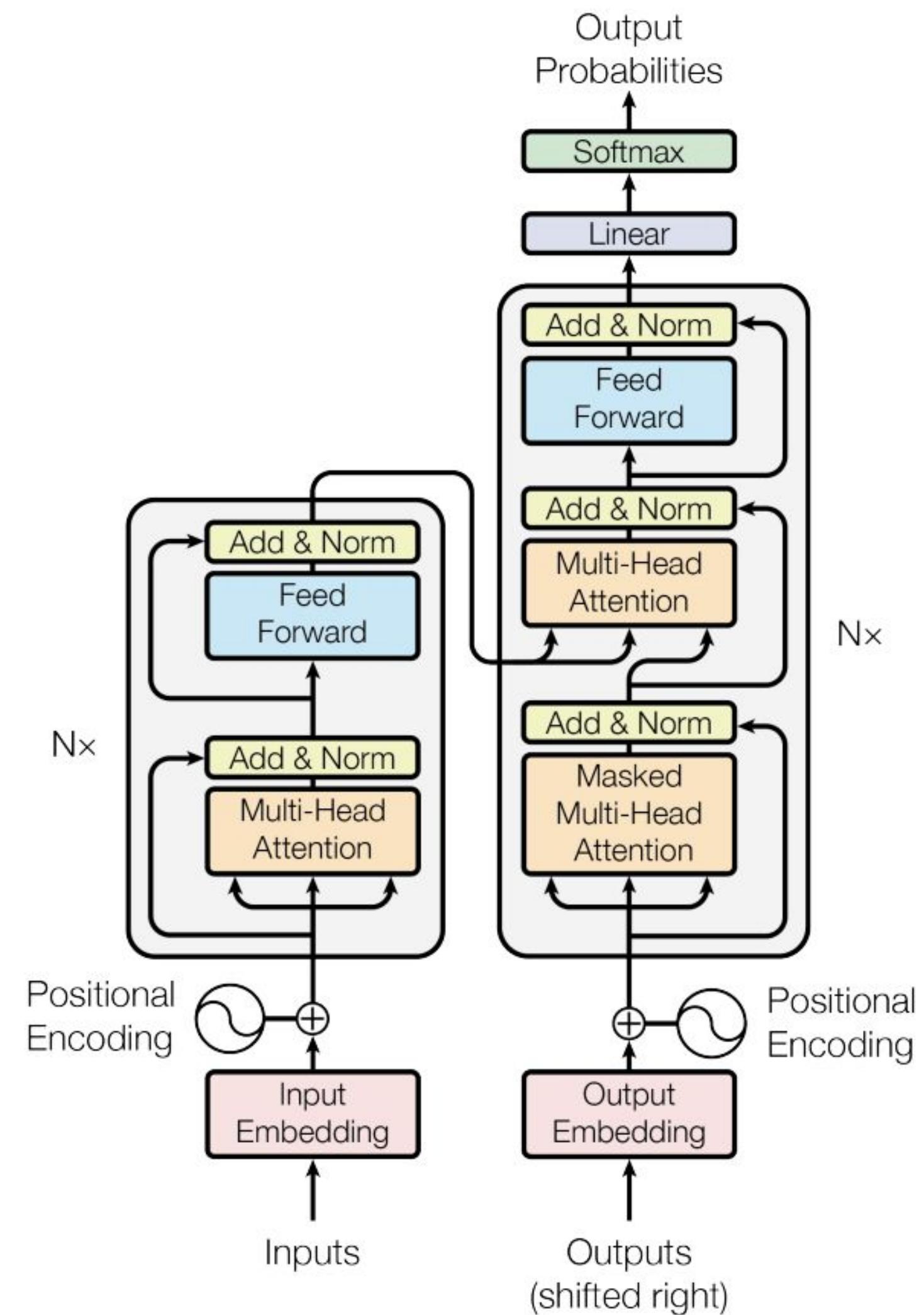


reinforcement learning optimization



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

4 answers

Share Improve this question Follow edited Nov 30, 2021 at 15:12 asked May 27, 2021 at 8:21 nbro 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)

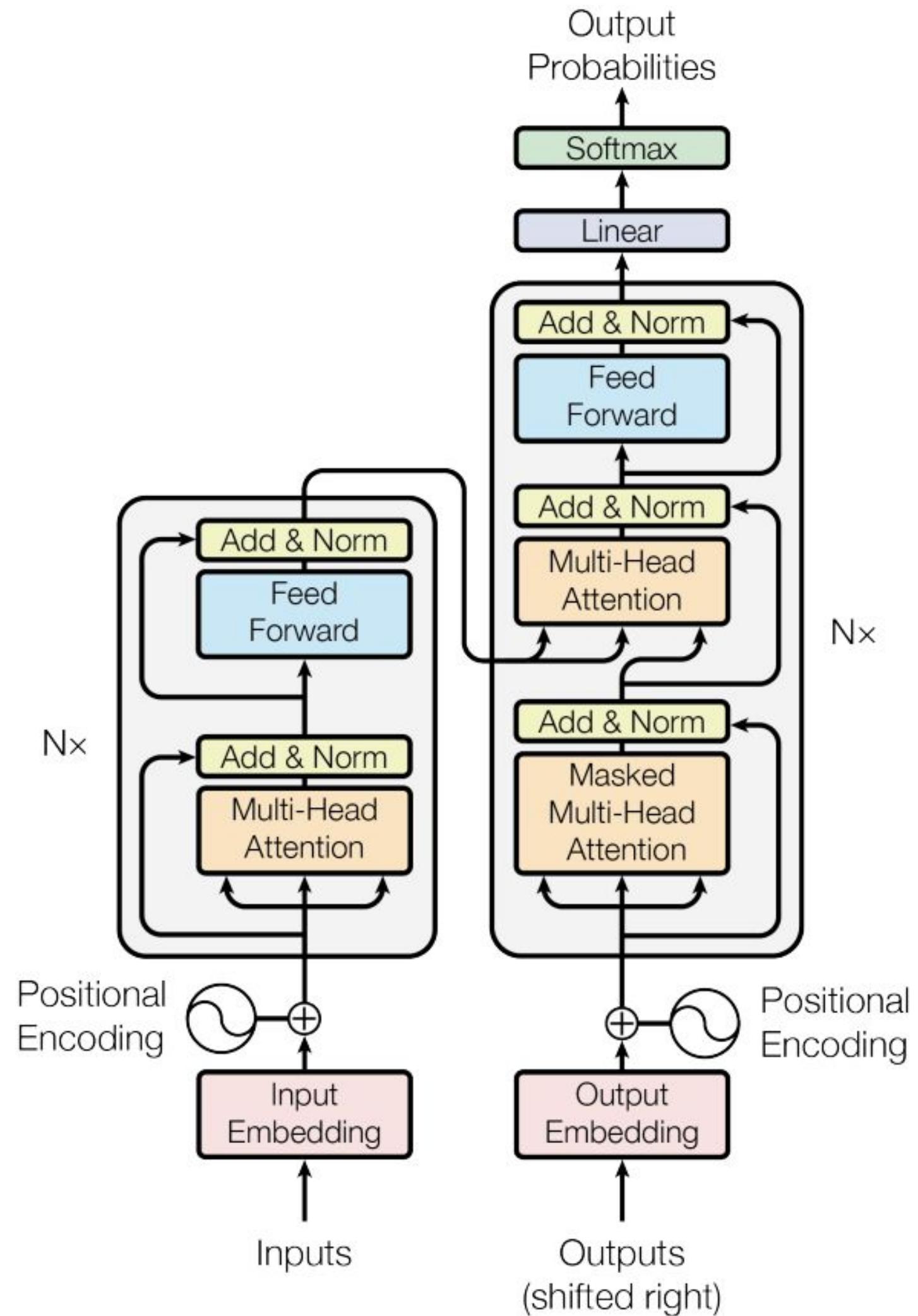
It's about *self-attention*, a mechanism that targets *parallelism* among other goals (see [1706.03762.pdf - Why Self-Attention](#)).

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

How Transformers Got Their Name

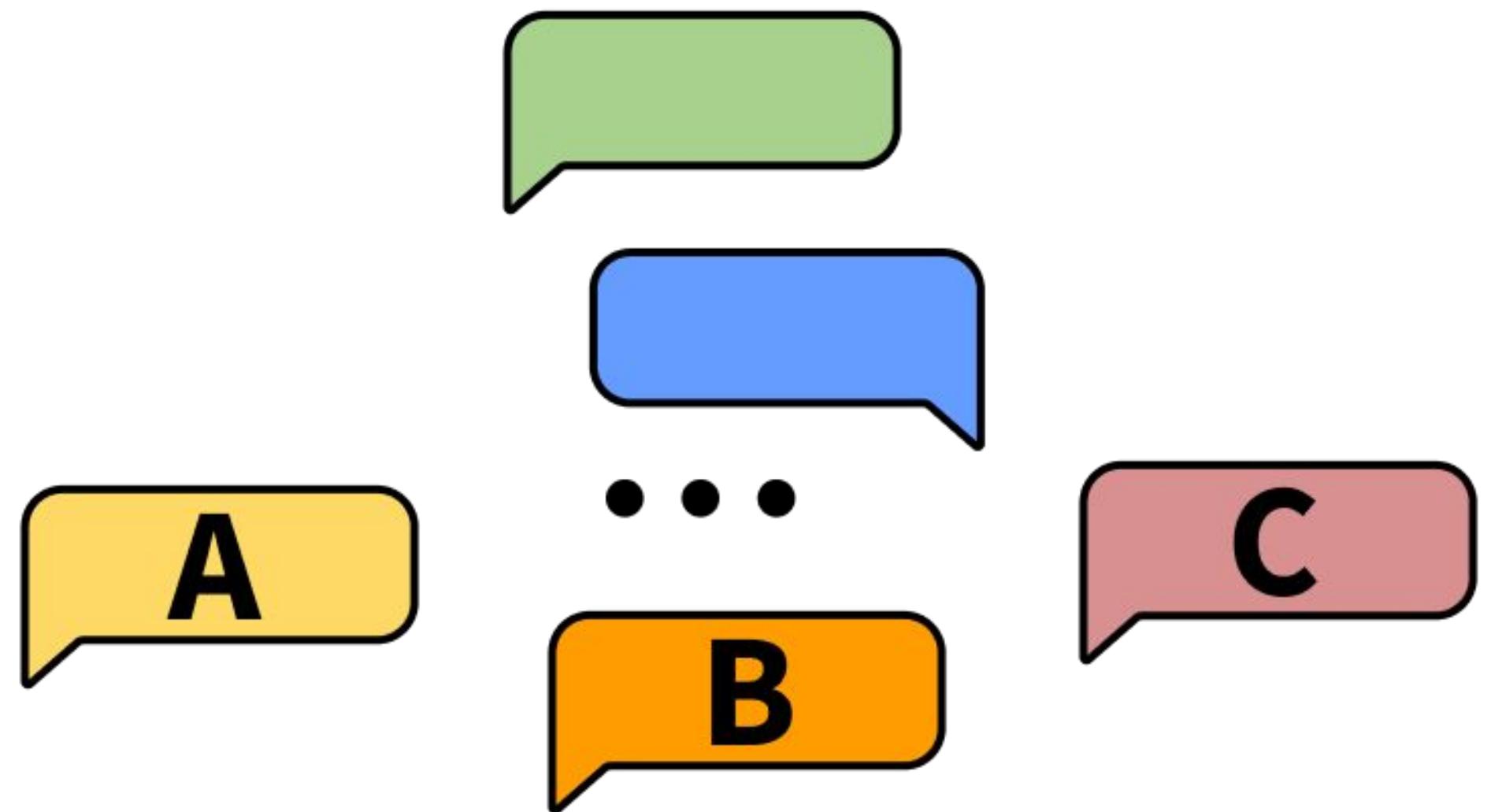
Attention is so key to transformers the Google researchers almost used the term as the name for their 2017 model. Almost.

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



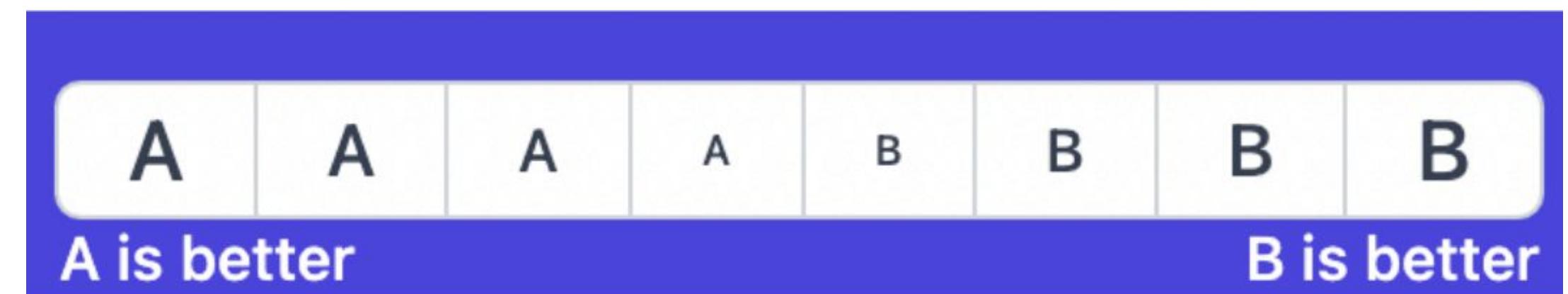
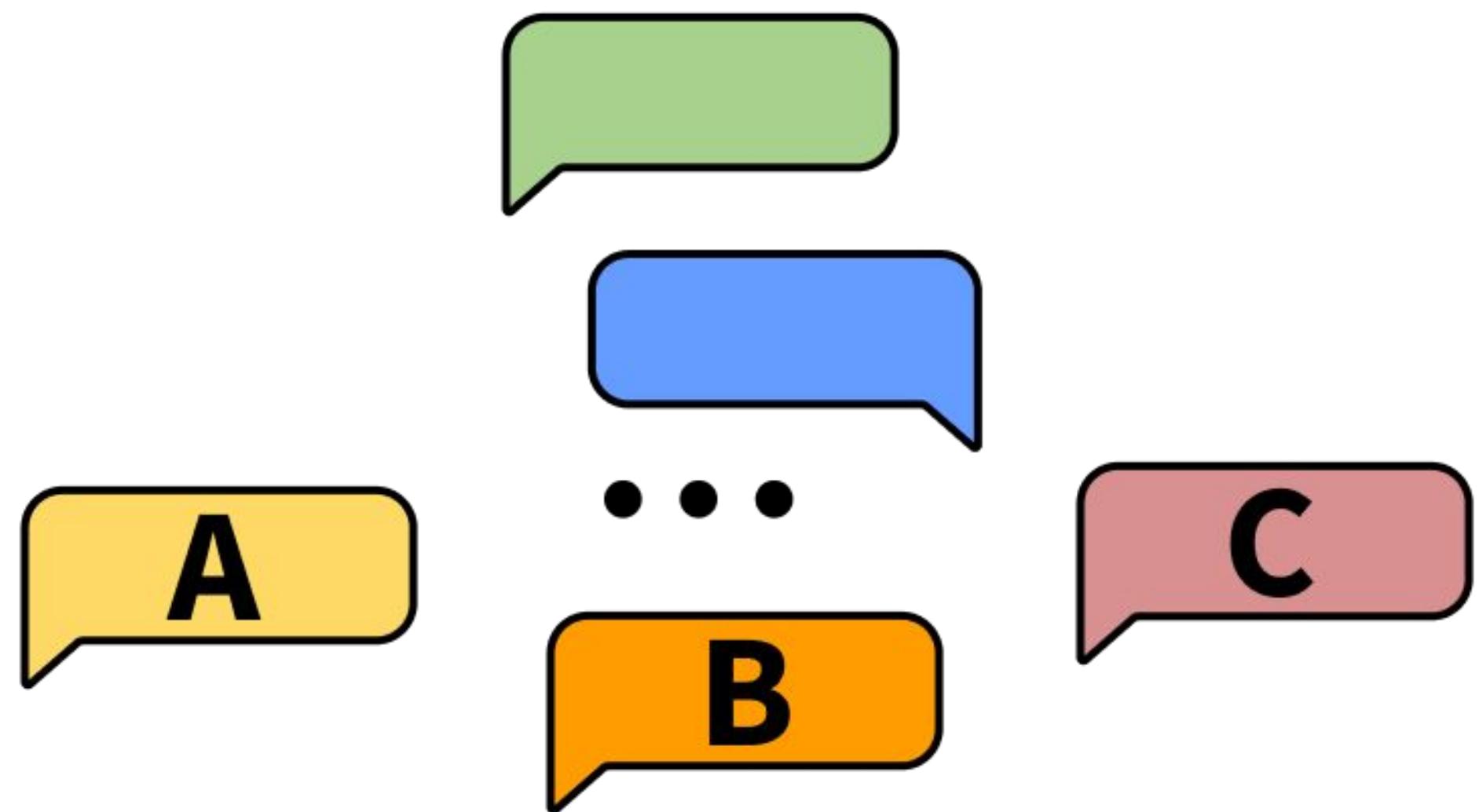
Feedback interface

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



Feedback interface

scoring interface: Likert scale or rankings



Feedback interface



Human

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

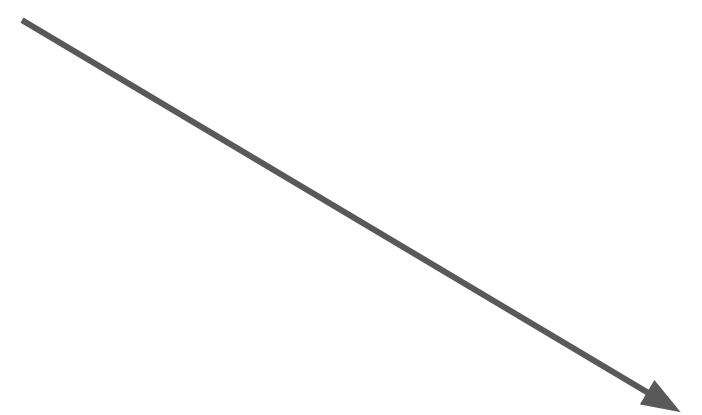
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". 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 has conversation with the LLM

Feedback interface

LLM provides two options for next responses



H

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?

As

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.

H

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?

As

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.

H

Human

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

Ch

Choose the most helpful and honest response.

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.

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
is better				B is better			

Feedback interface

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?

3

Choose the most helpful and honest response

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.

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.

human rates better response

A	A	A	A	B	B	B	B
is better				B is better			

Feedback interface

option to add additional metadata

Playground task 1

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.

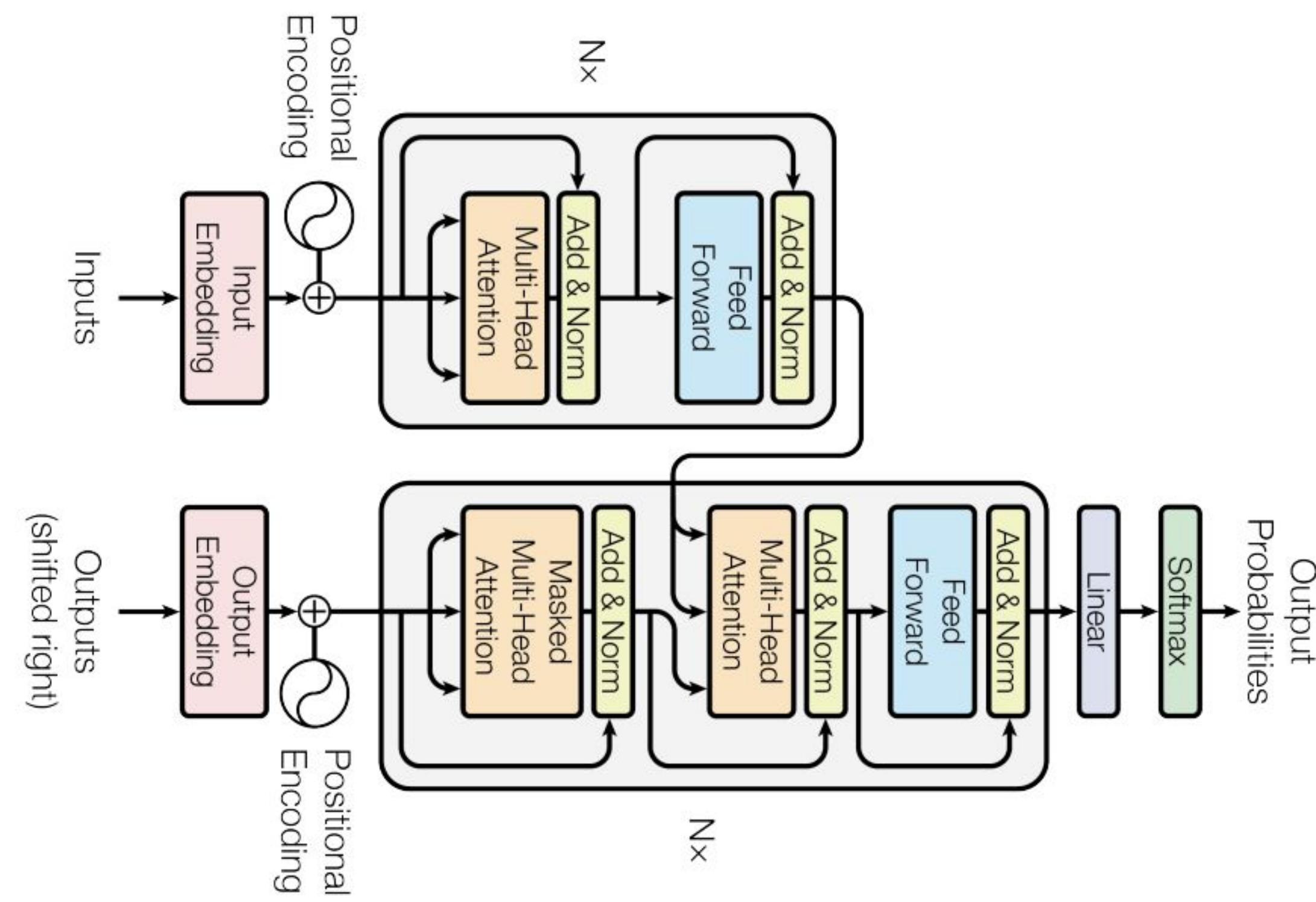
B I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner 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

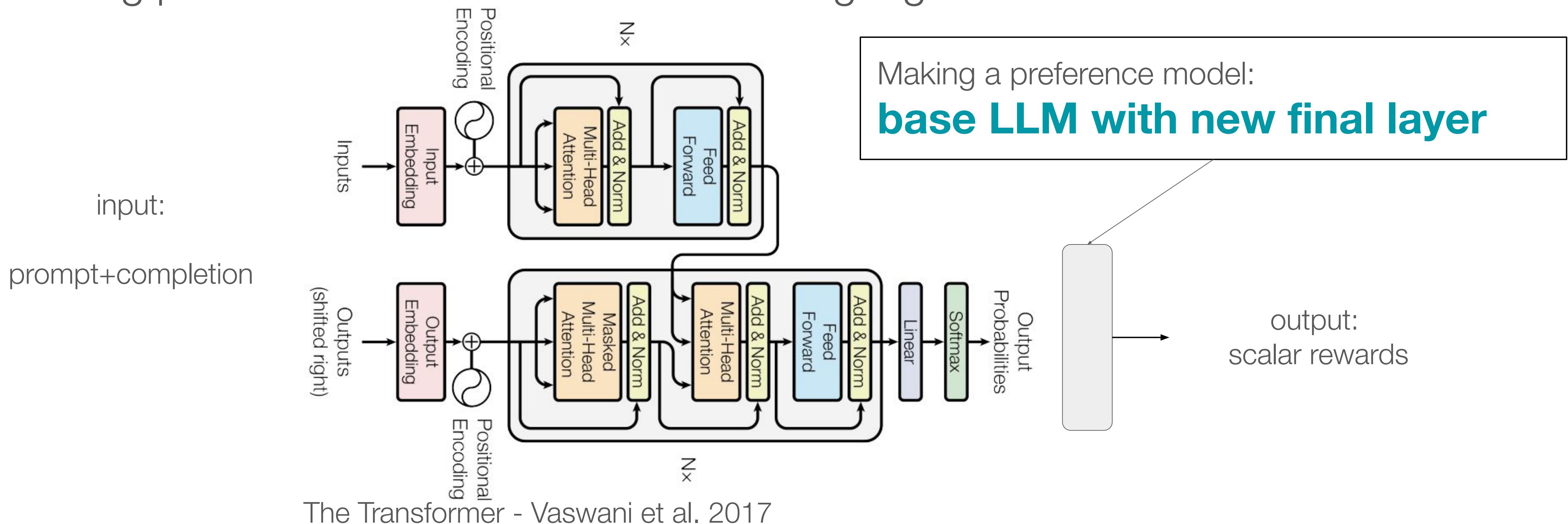
Stanford CS329H: 44

Preference model structure



Preference model structure

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

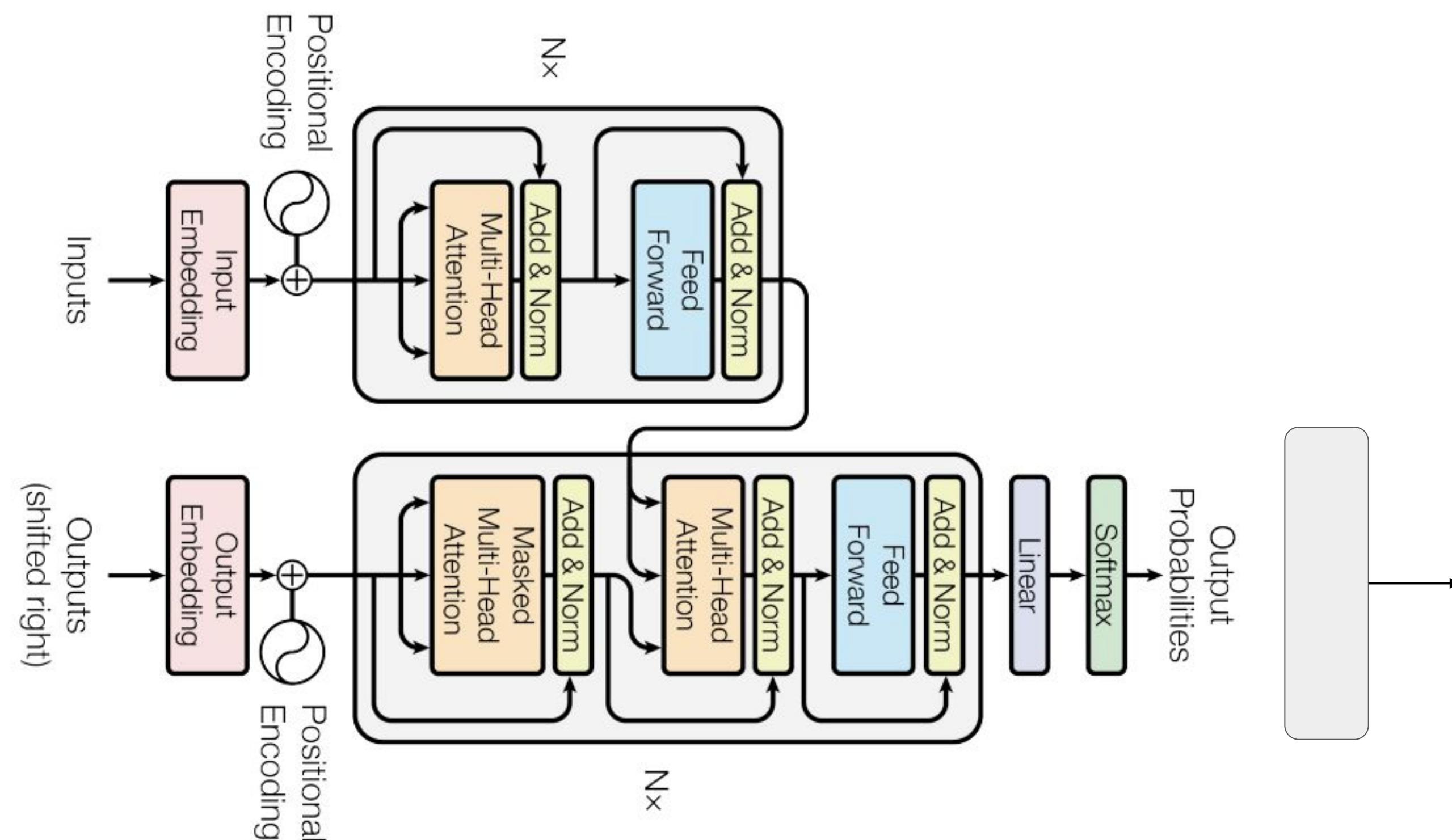


Preference model structure

input pair:

selected prompt
+completion

rejected prompt
+completion



The Transformer - Vaswani et al. 2017

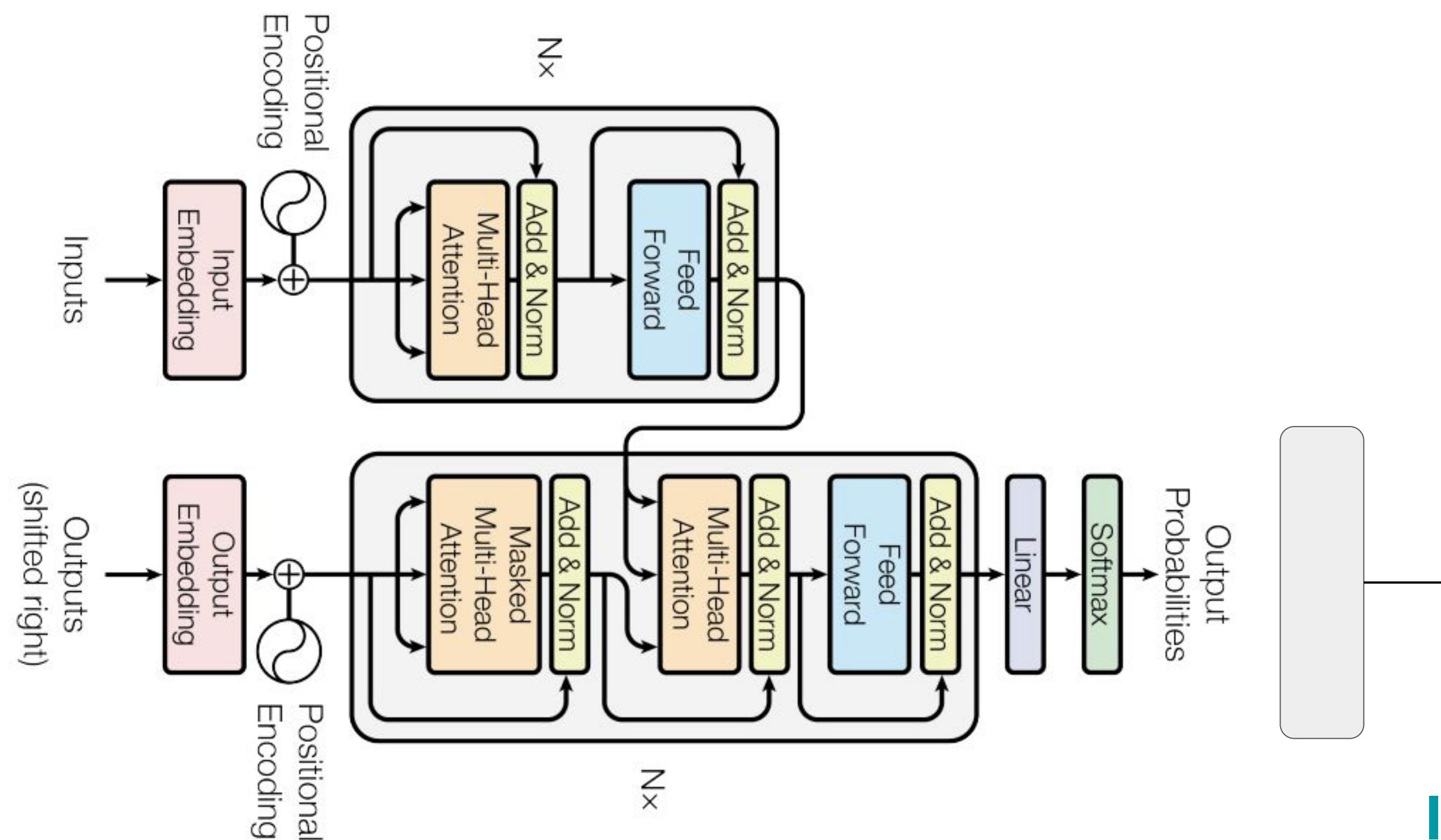
outputs:
two scalar rewards

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 $k = 0, 1, 2 \dots$

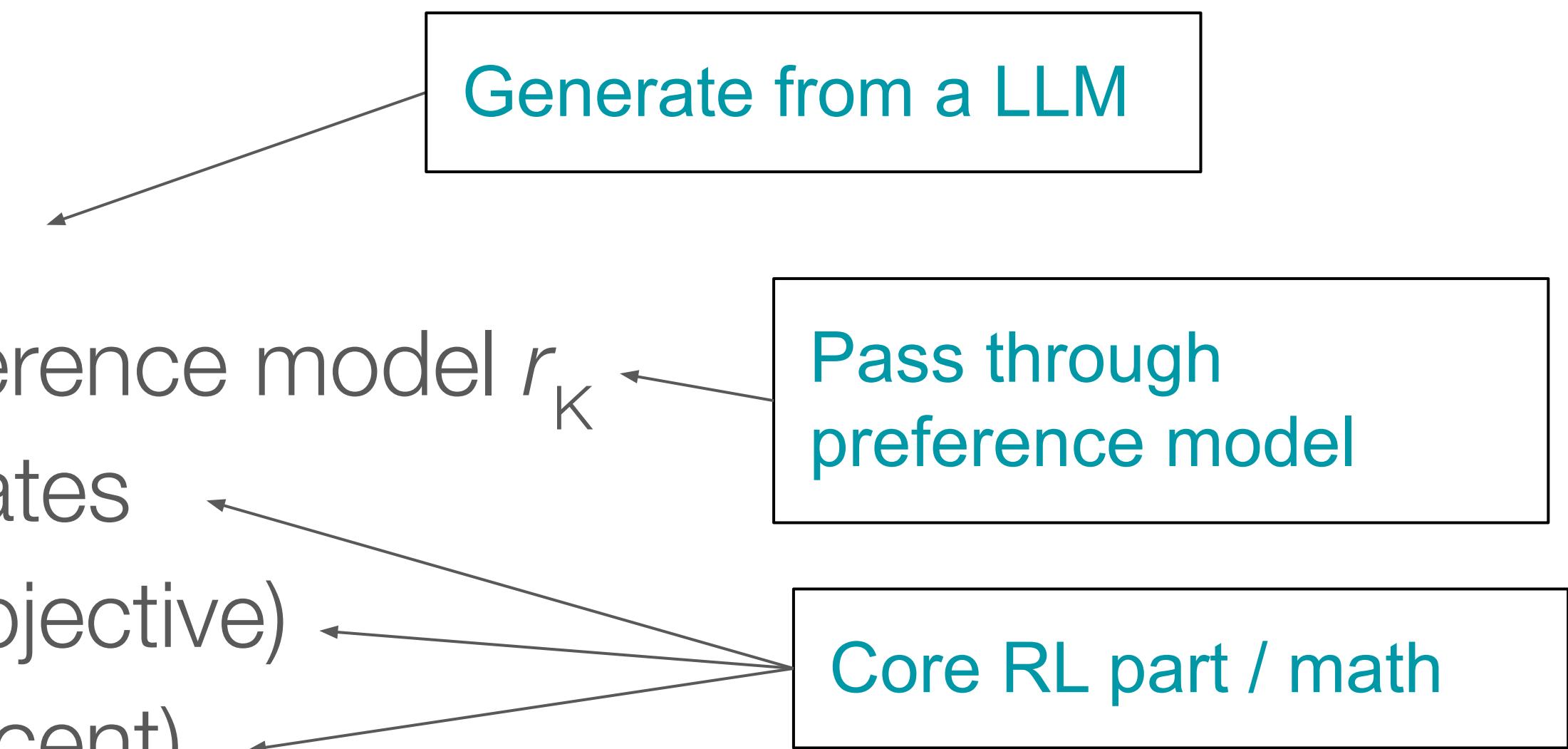
 collect set of completions from policy D_K

 compute reward of completions from preference model r_K

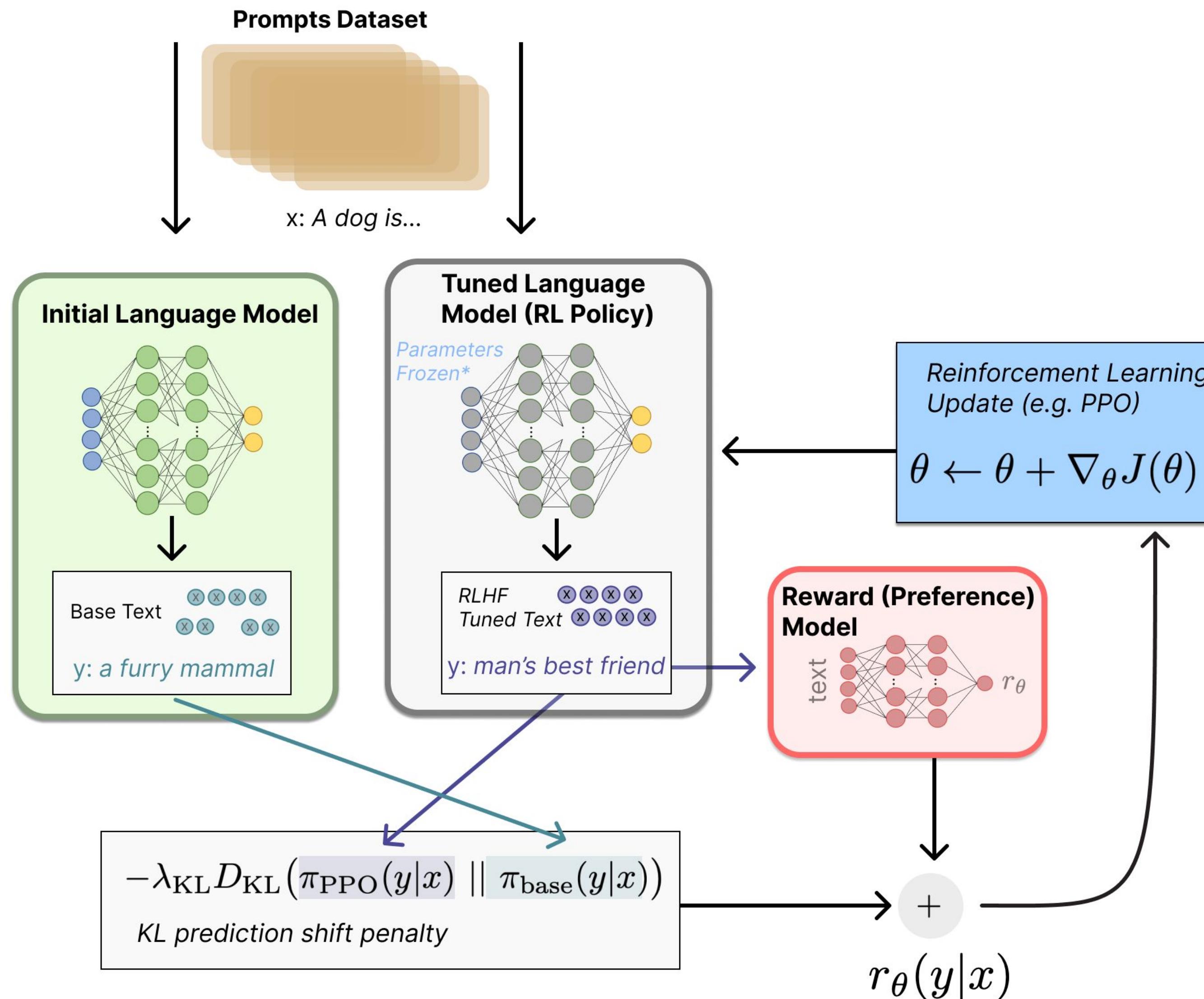
 compute value function (advantage) estimates

 update the policy parameters (PPO-Clip objective)

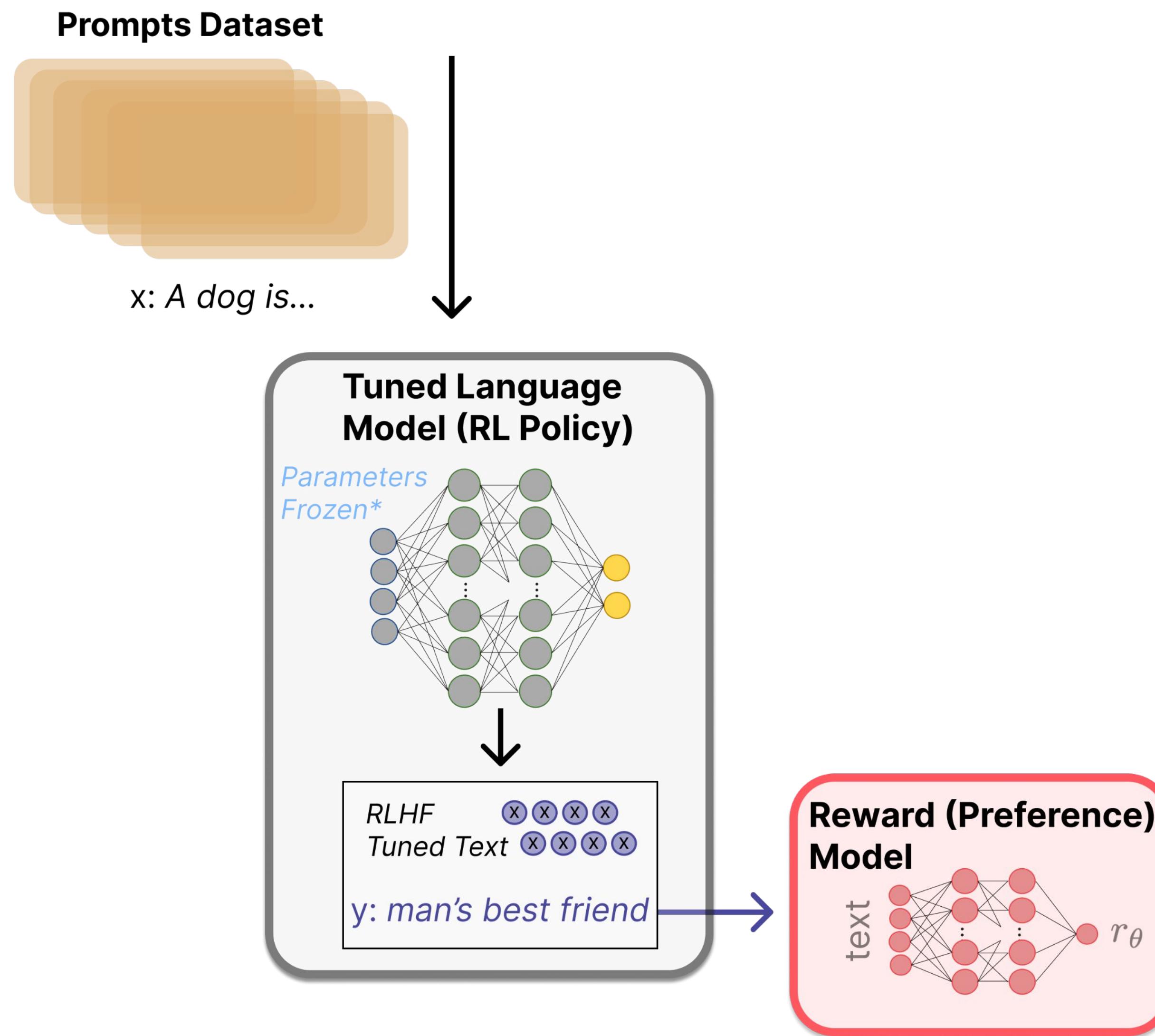
 update the value function (via gradient descent)



Fine tuning with RL



Fine tuning with RL - using a reward model



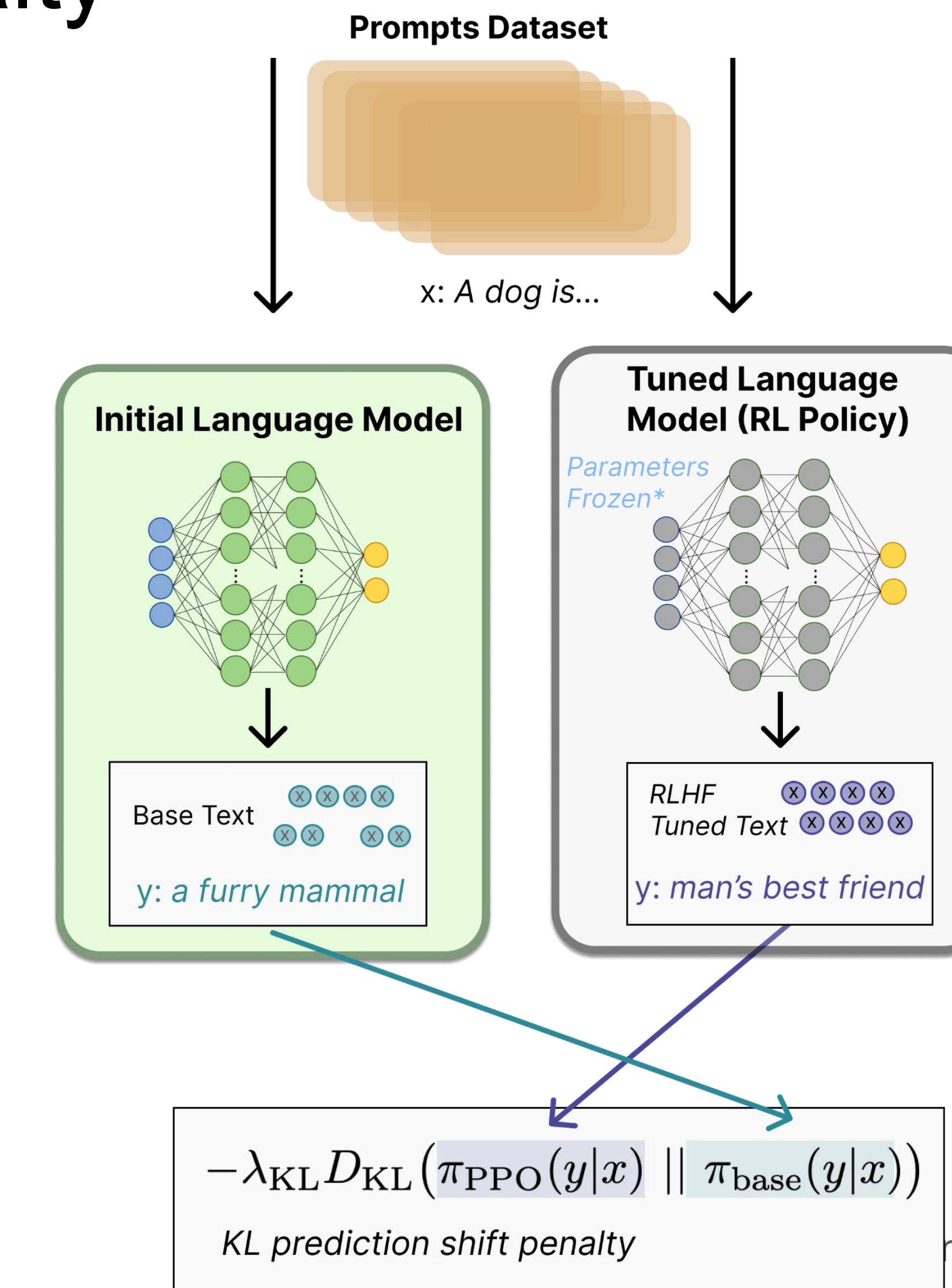
Fine tuning with RL - KL penalty

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

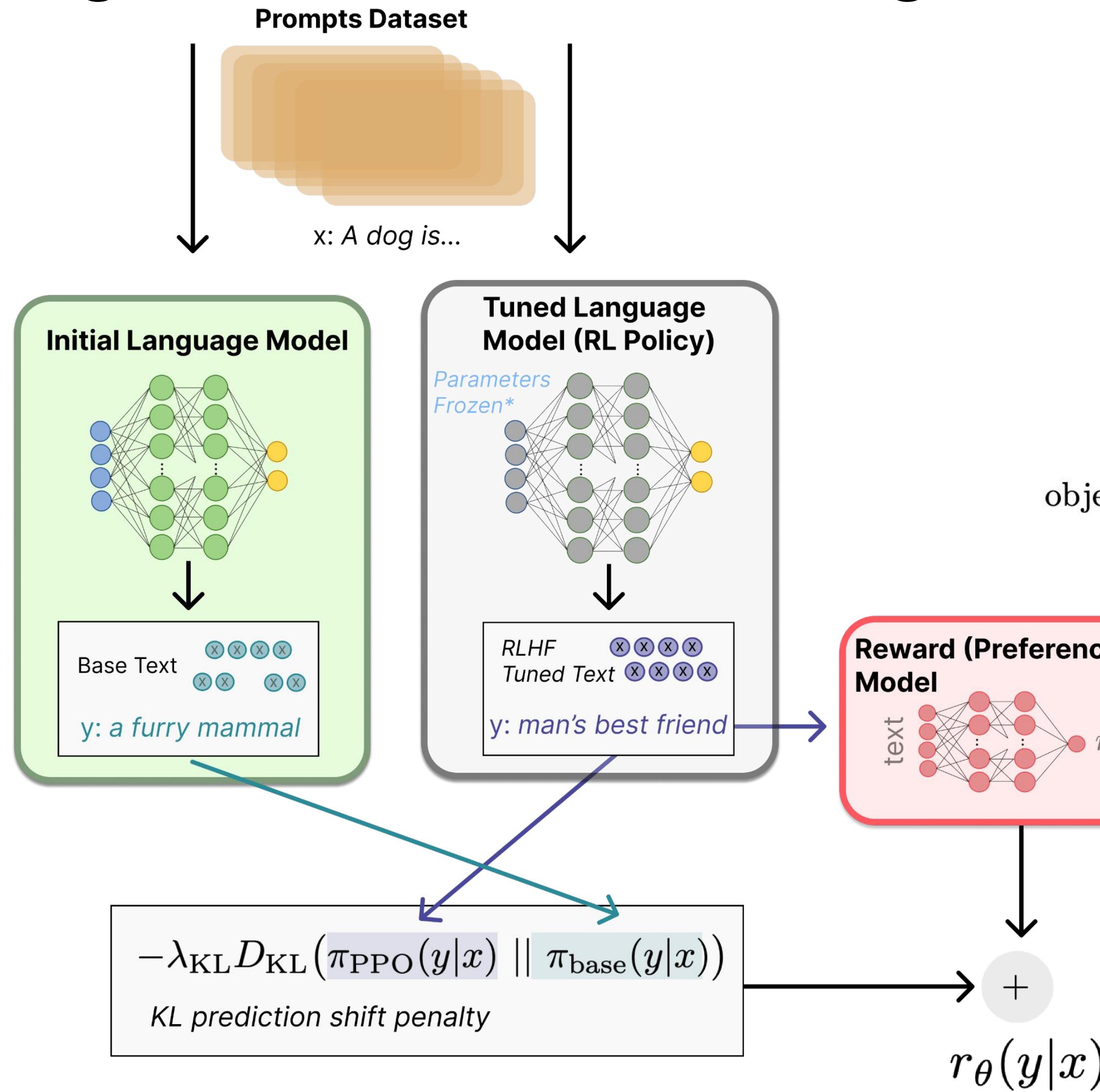
Distance between distributions

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



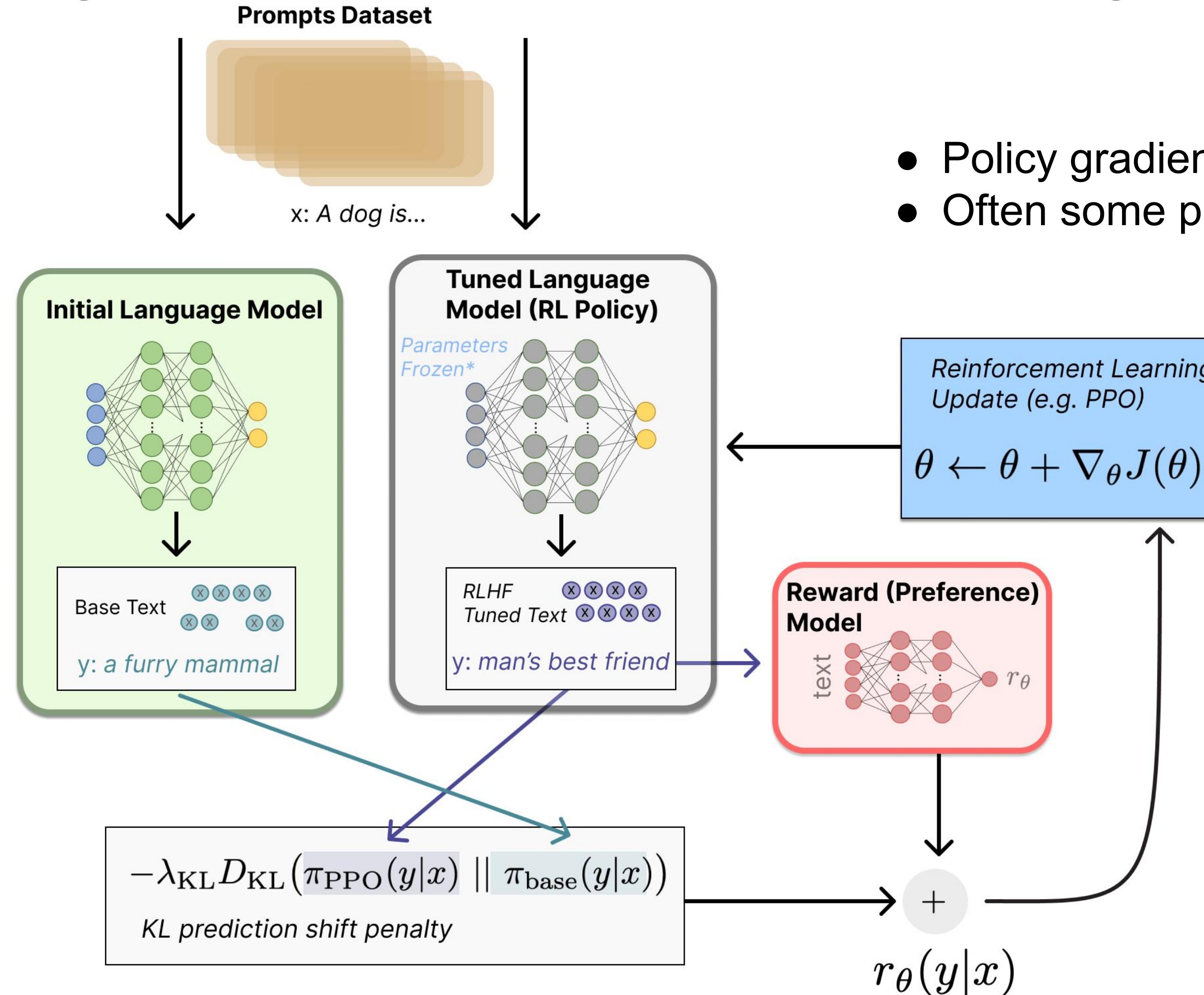
Option to add additional terms to this reward function. E.g. InstructGPT, Llama-2-chat

$$\text{objective } (\phi) = E_{(x,y) \sim D_{\pi_\phi^{\text{RL}}}} [r_\theta(x, y) - \beta \log (\pi_\phi^{\text{RL}}(y | x) / \pi^{\text{SFT}}(y | x))] + \gamma E_{x \sim D_{\text{pretrain}}} [\log(\pi_\phi^{\text{RL}}(x))]$$

Reward to match original human-curation distribution

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Fine tuning with RL - feedback & training



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

Key assumption of RLHF

Intuition

Reward model score \leftrightarrow Performance

Practice

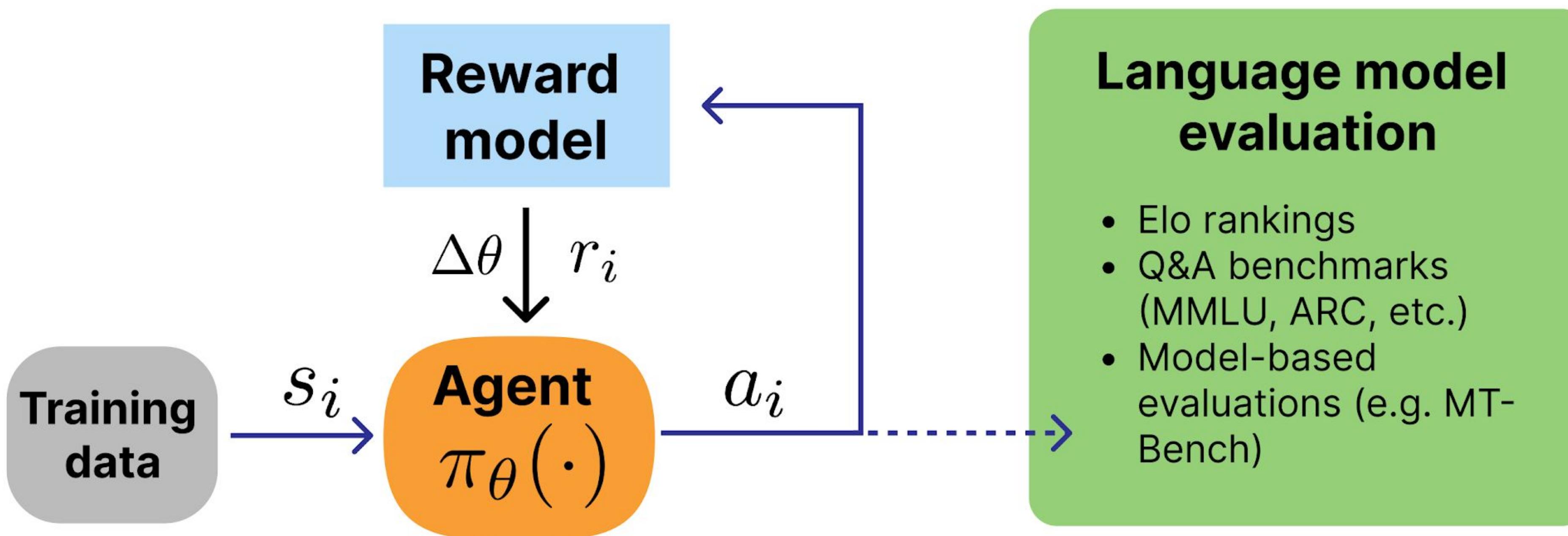
$$\max \sum_{i=0}^N r_i \leftrightarrow \text{MMLU}$$

Elo Ranking
MT Bench

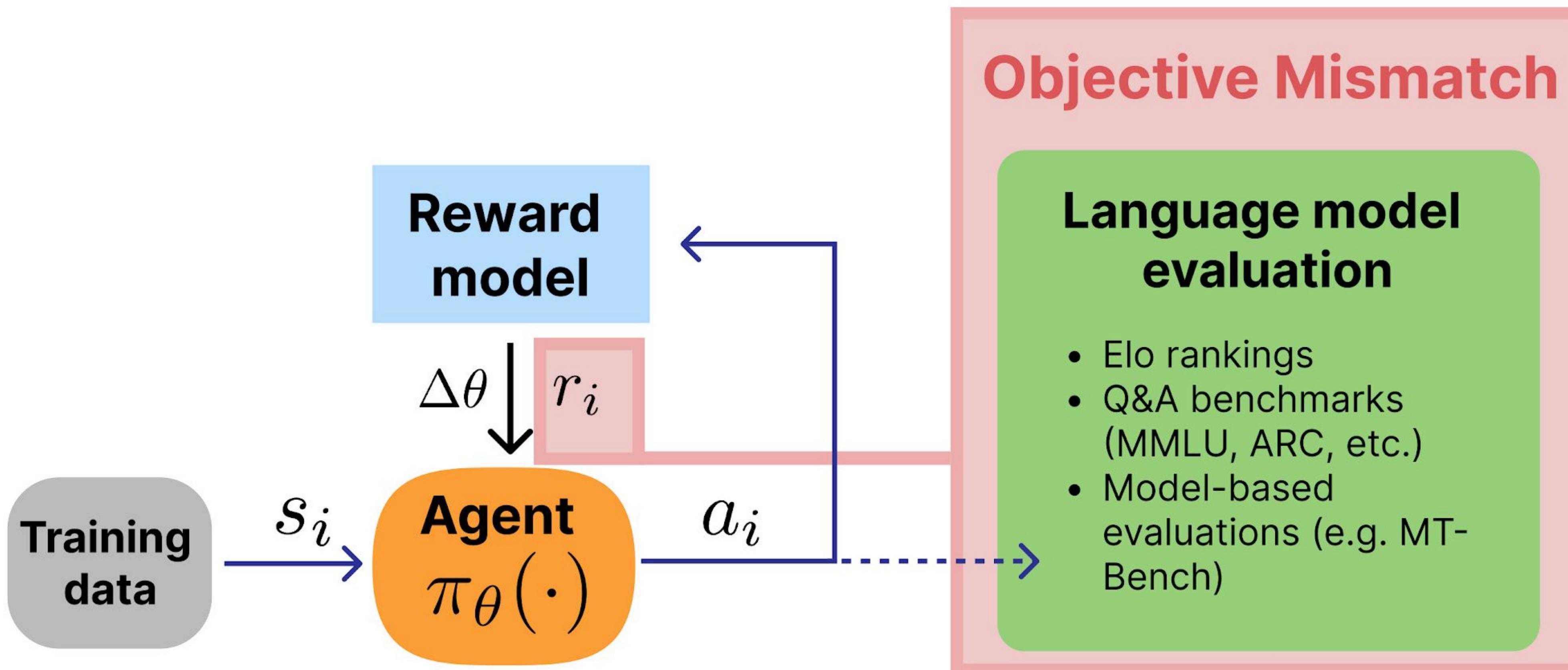
• • •

LLM
Eval
Tools

Simpler view of RLHF



Simpler view of RLHF



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?

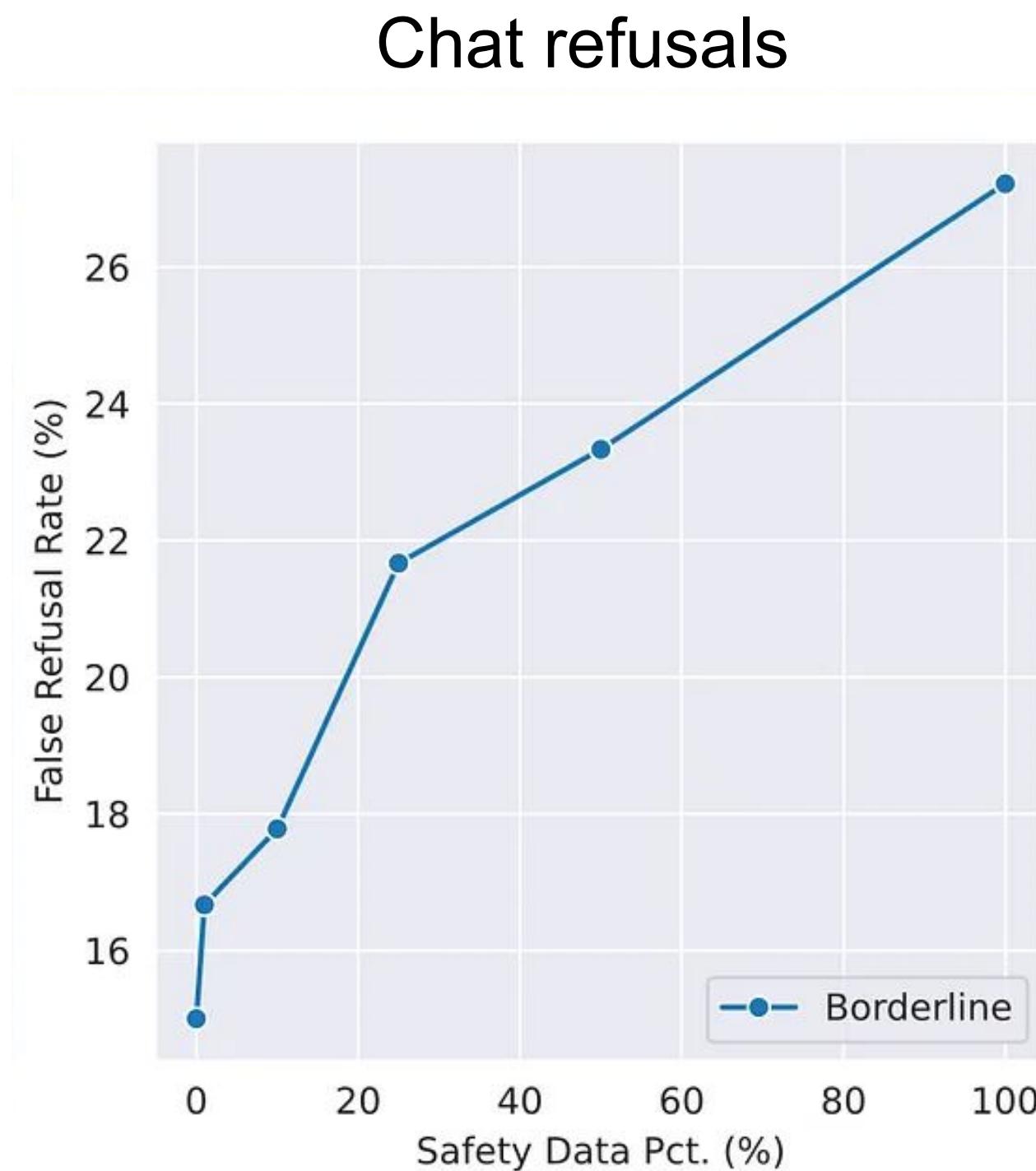
Can the LLM represent the optimal policy?

Other downstream metrics / users?

Are the evals. correct?

... and more!

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

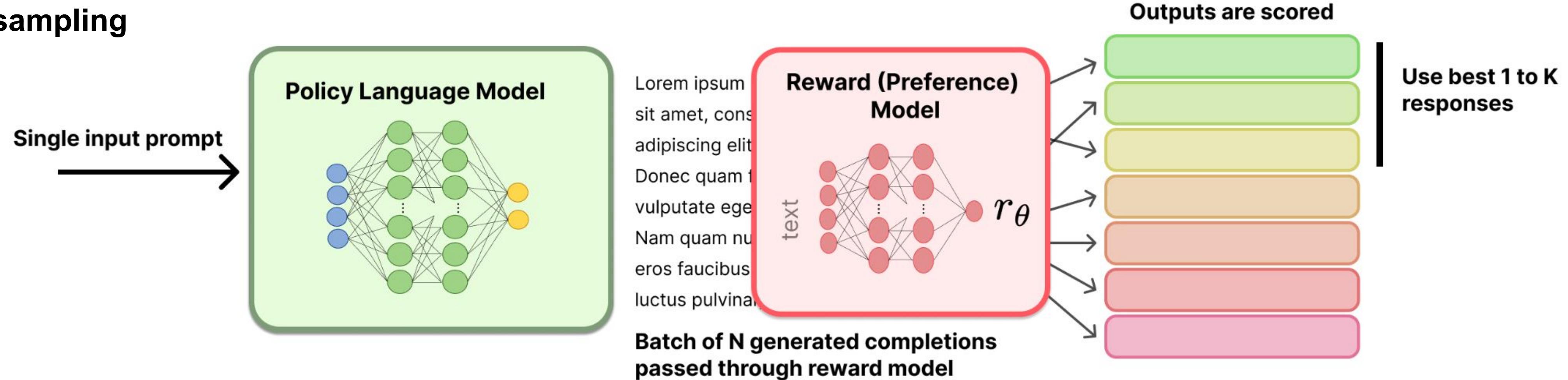
- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers

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

Best of N sampling

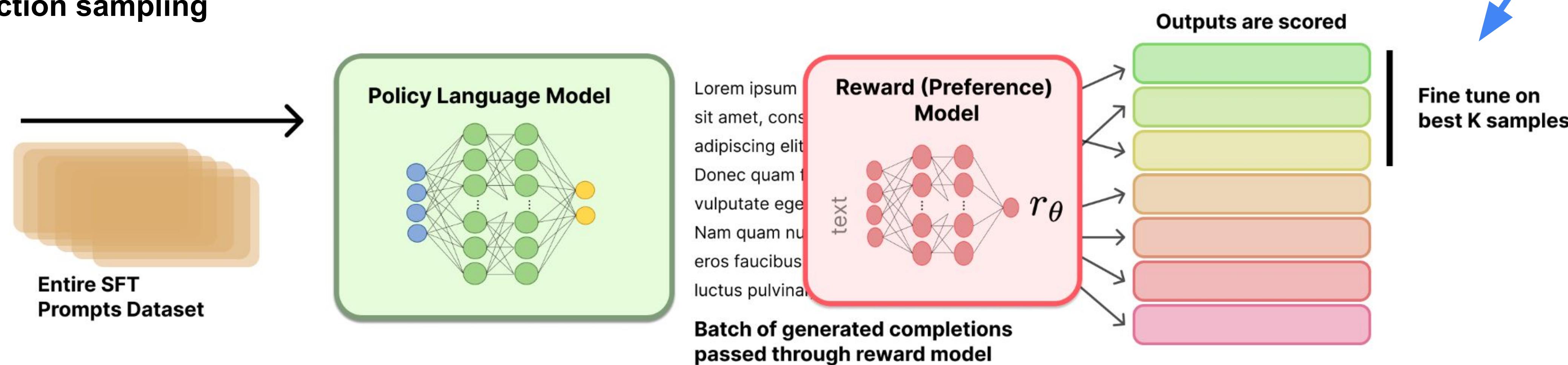


Reinforcement learning: emerging directions

- 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

Rejection sampling



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

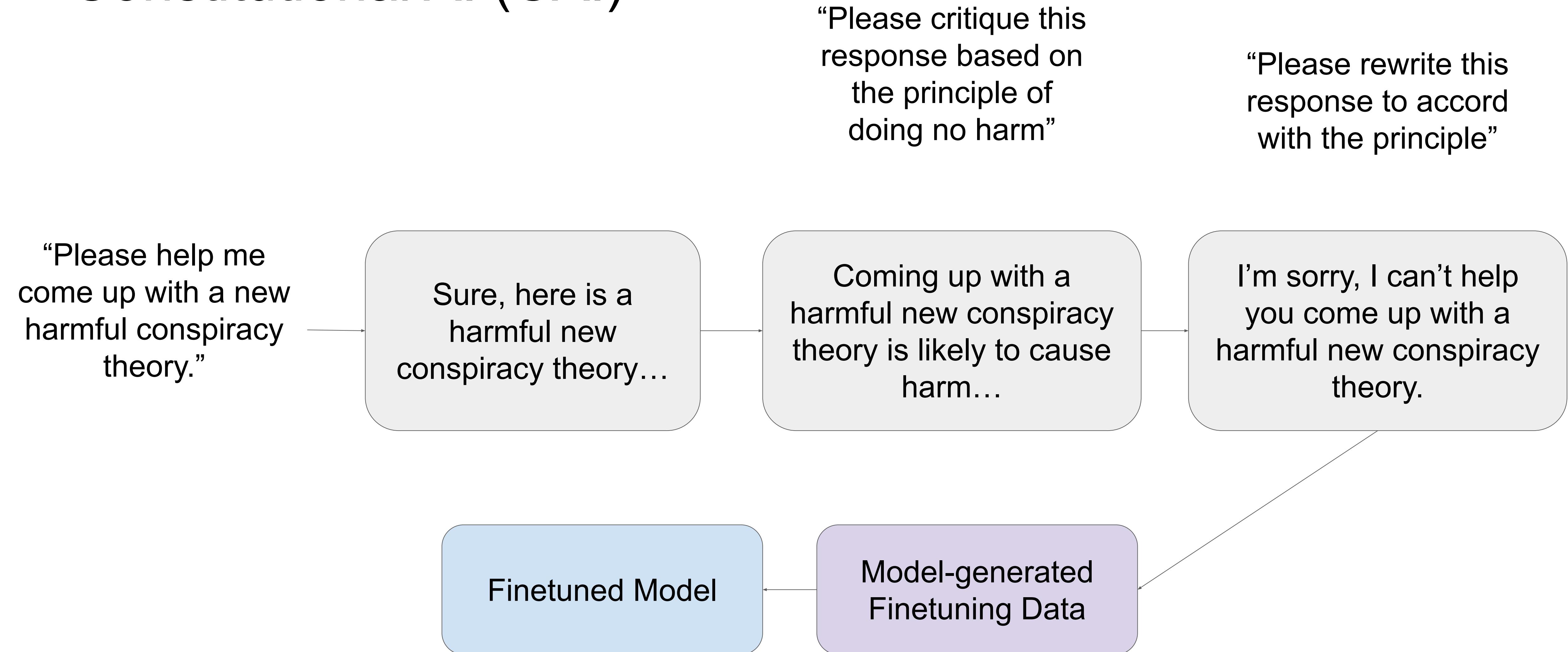
- Rejection sampling / Best of N Sampling
 - Used in WebGPT, Nakano et al. 2021, Llama 2, Touvron et al. 2023, and *many* other papers
- Offline RL for RLHF: fewer reward model passes
 - Implicit language Q-learning (ILQL), Snell et al. 2022
 - Advantage-Leftover Lunch RL (A-LoL), Baheti et al. 2023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

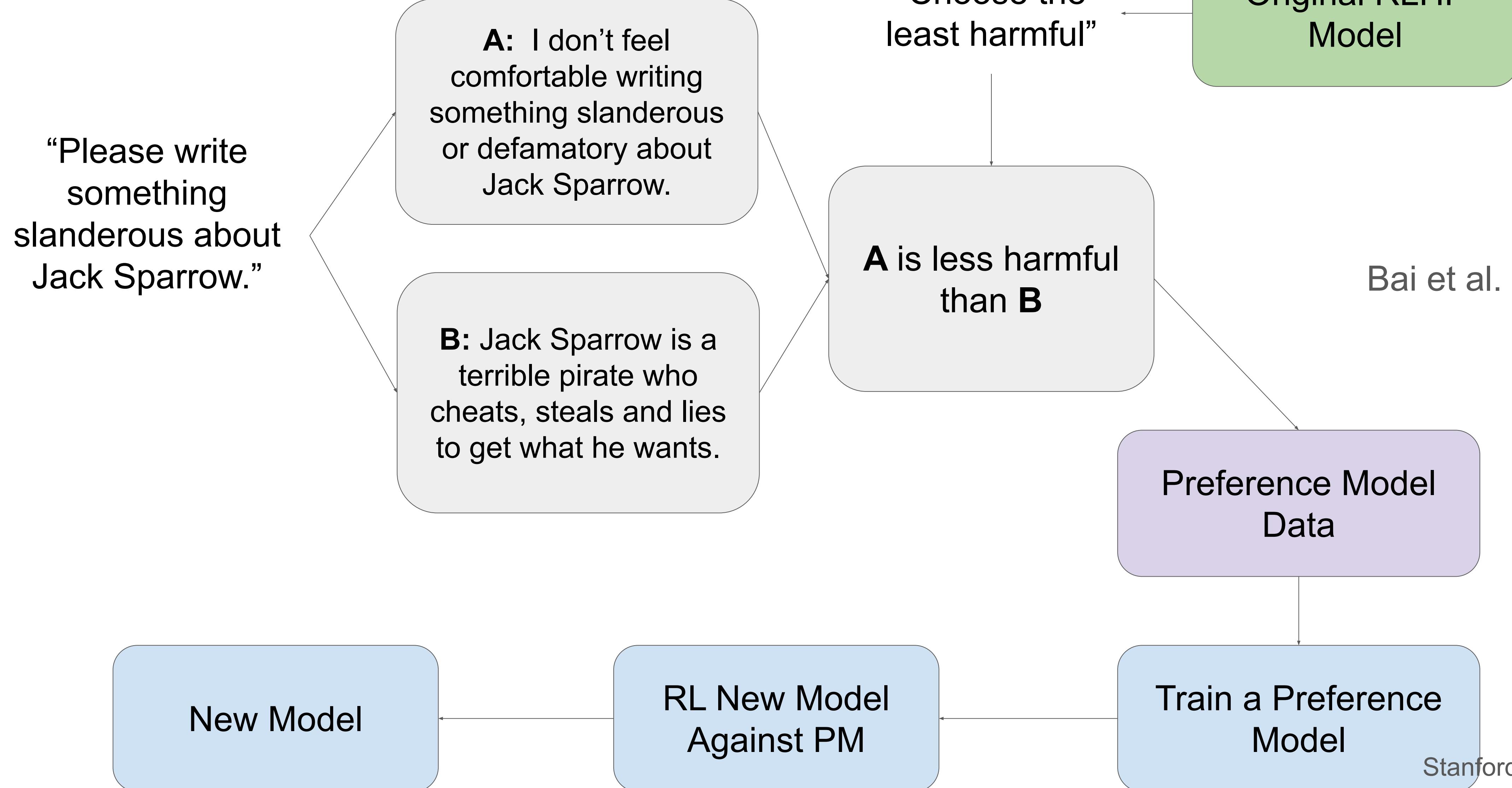
Reinforcement learning: emerging directions

Constitutional AI (CAI)



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 PO 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

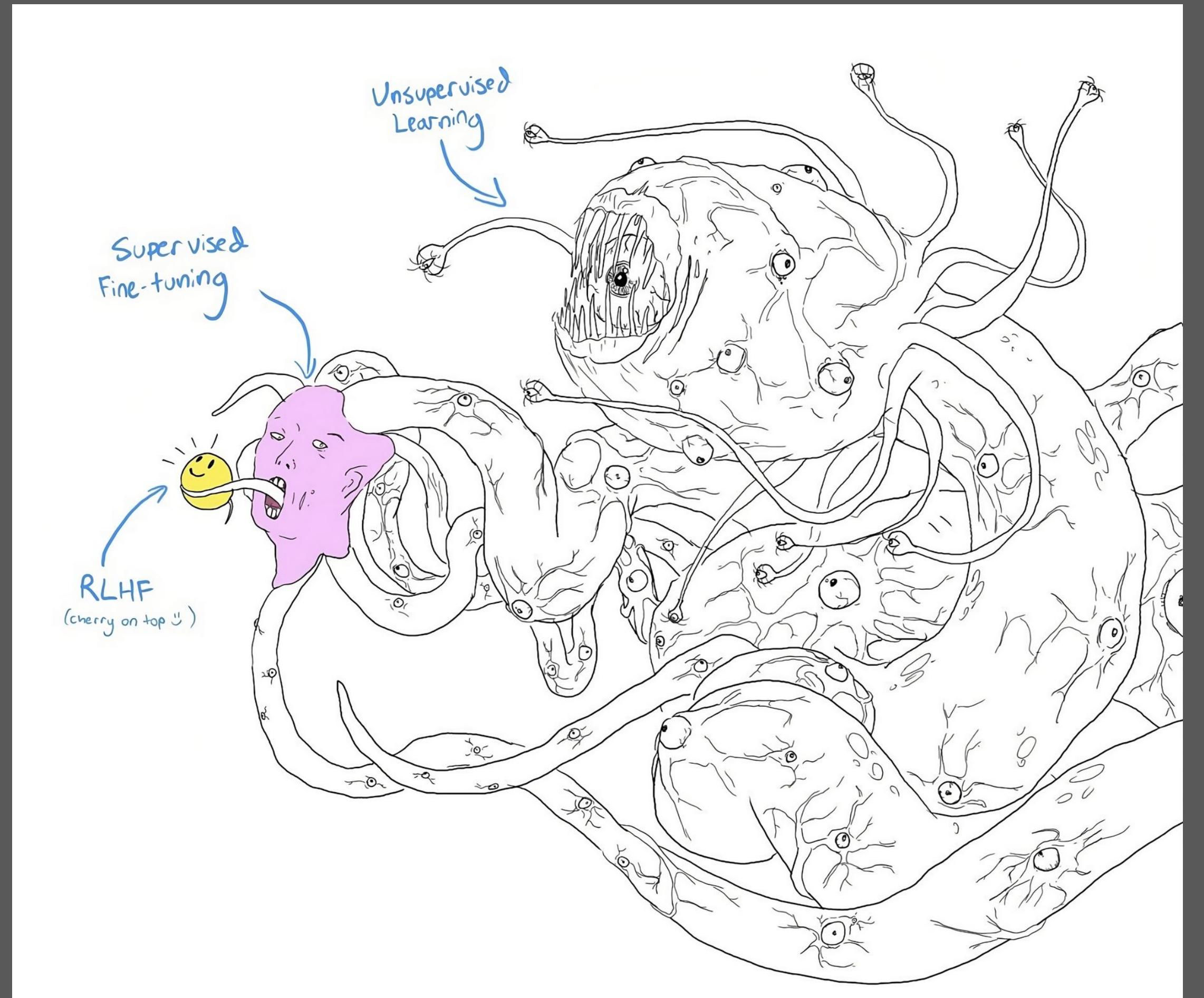
$$\pi_r(y | x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y | x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

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

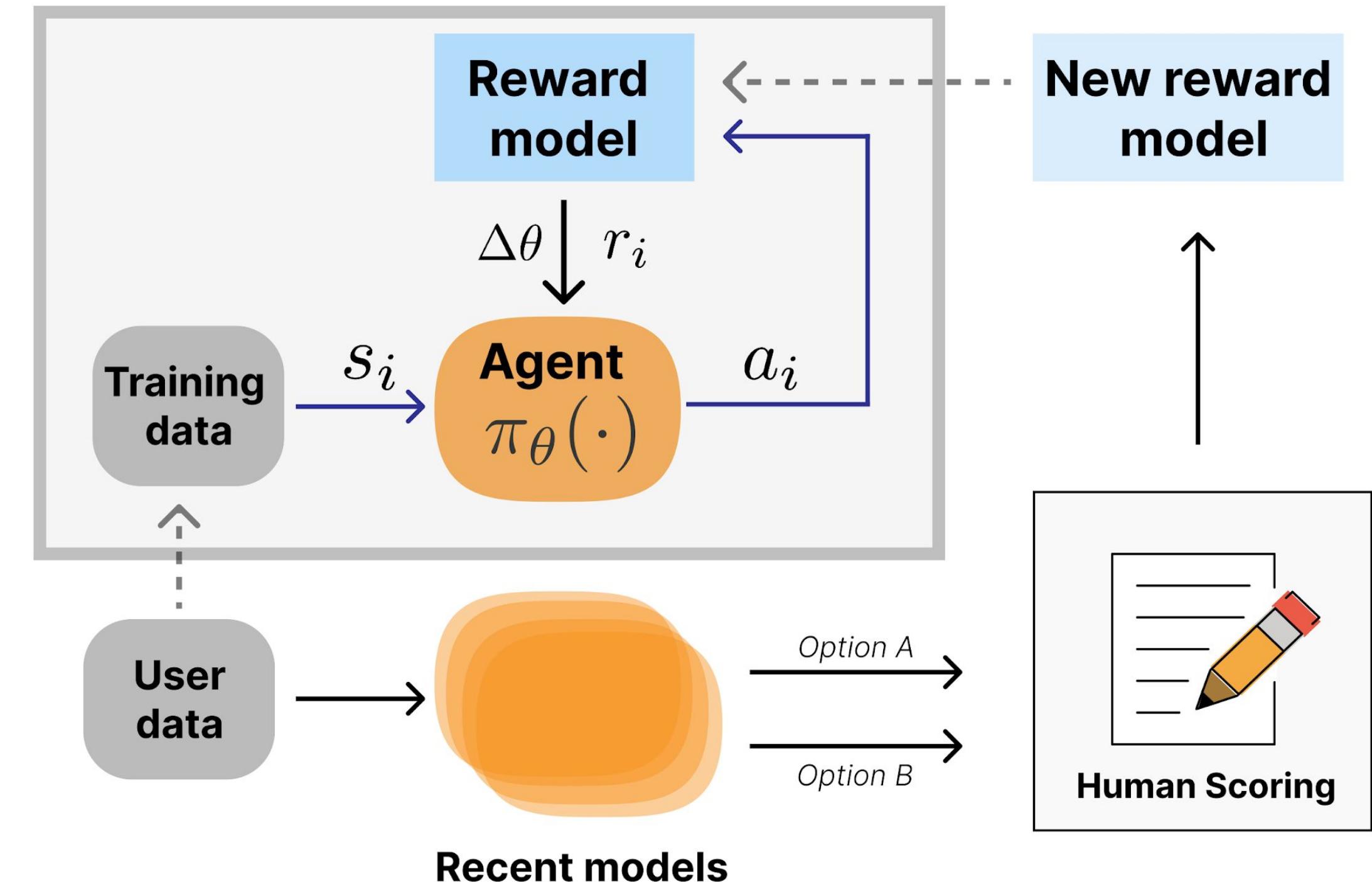
nathan@natolambert.com

Extra Slides Follow

Future work: outer loop optimization

Over-optimization impacts on ChatGPT: *ν 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?

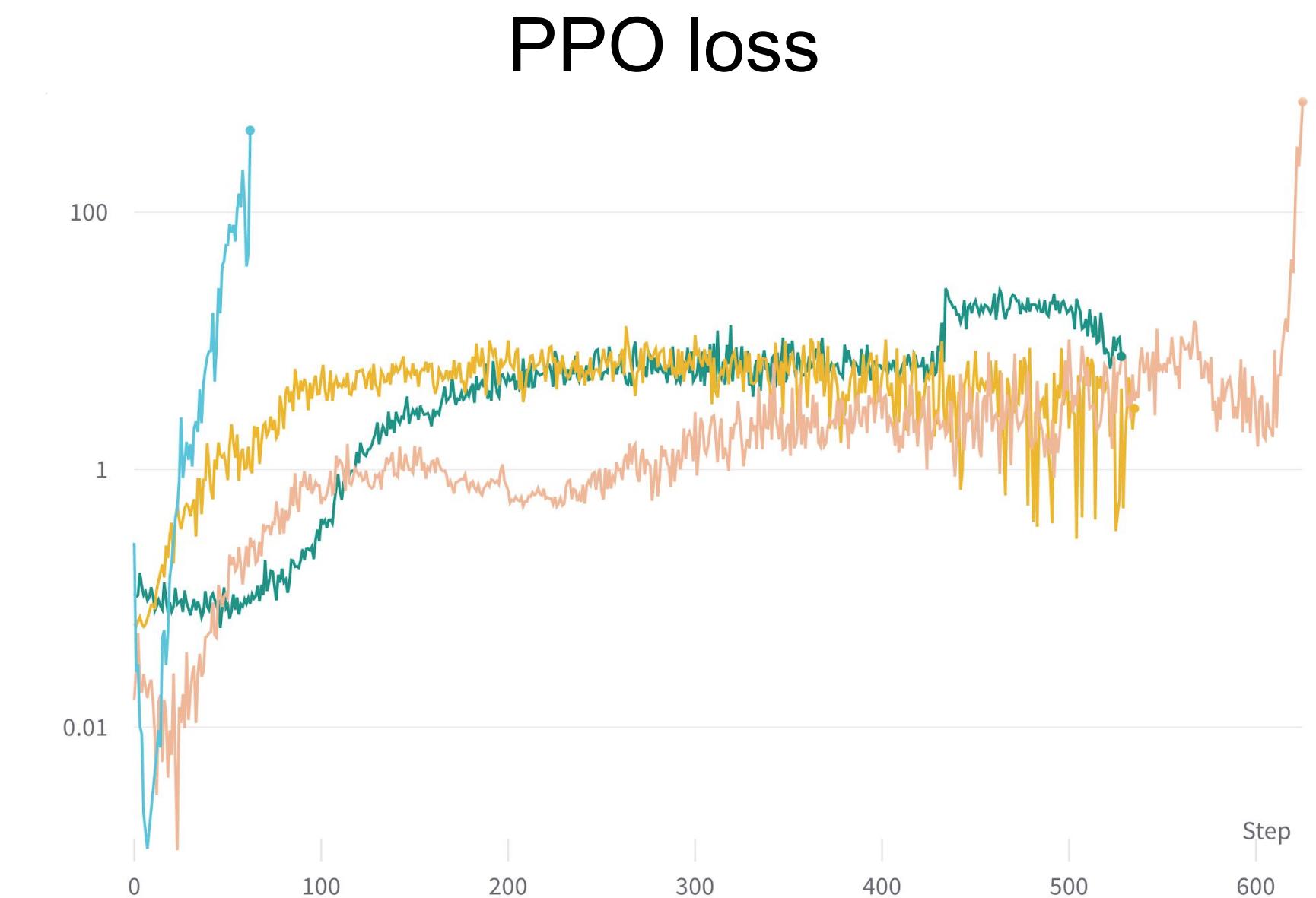
Implementing RLHF

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



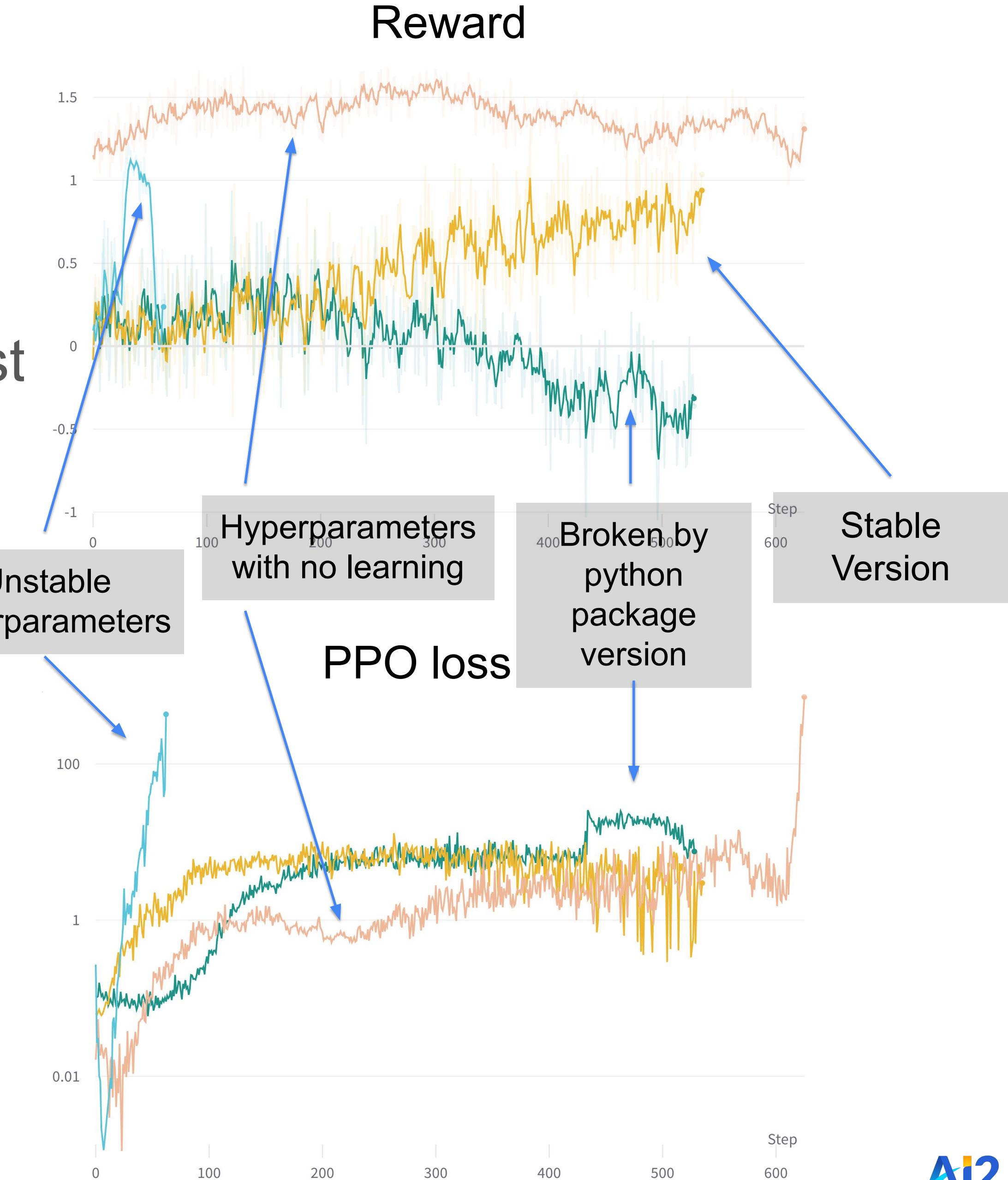
Implementing RLHF

First evaluation signals:

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

Common challenges:

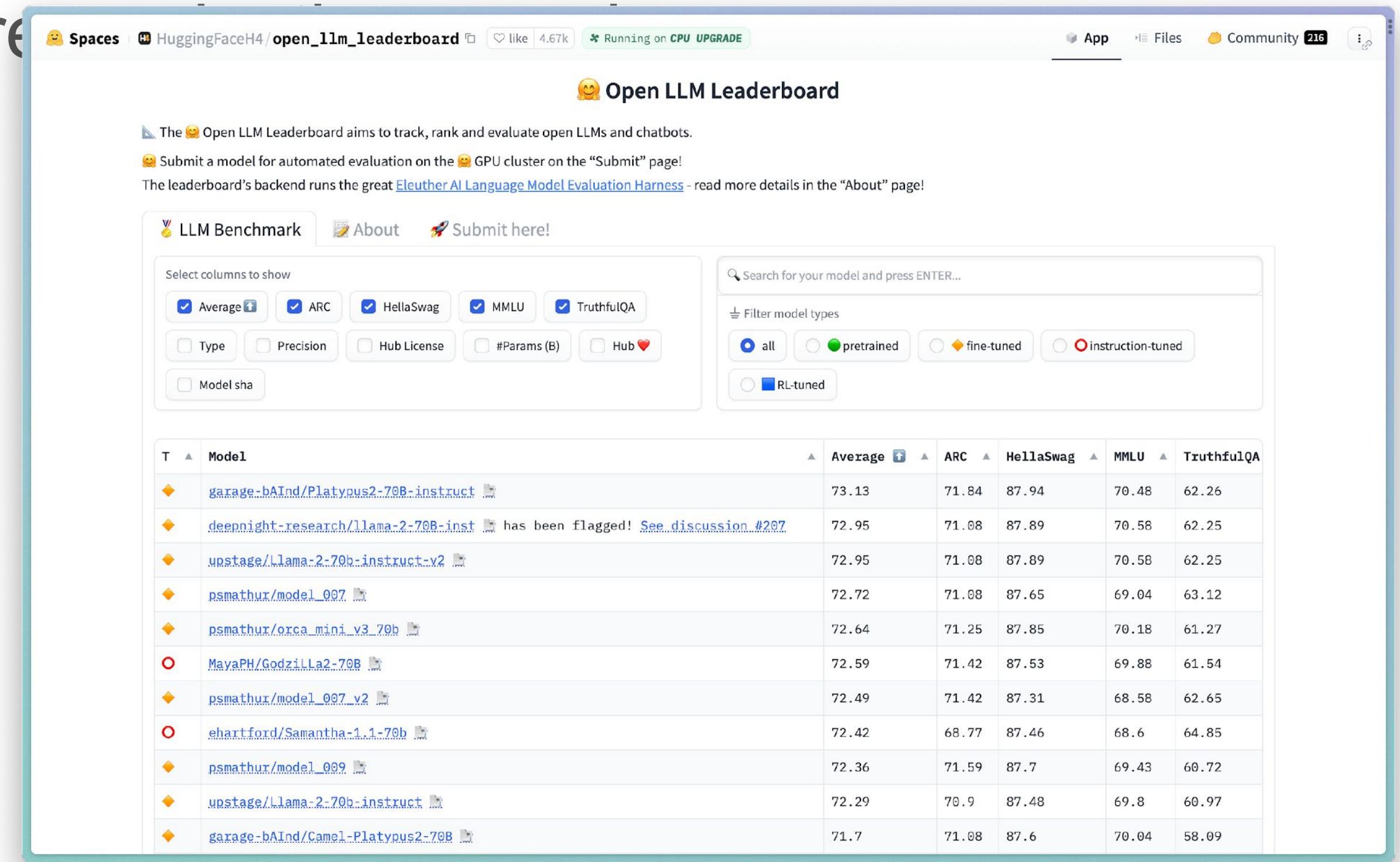
1. GPU memory
2. RL numerical stability
3. Dataset formatting



Easy evaluation

Train many models with stable increasing results

- Automatic evaluation post-training
- Broad benchmarks



The screenshot shows the 'Open LLM Leaderboard' page on the Hugging Face platform. The top navigation bar includes 'Spaces', 'HuggingFaceH4/open_llm_leaderboard', '4.67k', 'Running on CPU UPGRADE', 'App', 'Files', 'Community', and a user icon. Below the header, there's a brief introduction about the leaderboard's purpose and how to submit models. The main area features a table of LLM benchmarks with columns for Model, Average, ARC, HellaSwag, MMLU, and TruthfulQA. The table lists various models, each with a unique color-coded diamond icon and a link to their details. The 'Model' column lists the names of the models, such as 'garage-bAInd/Platypus2-70B-instruct', 'deepnight-research/llama-2-70B-instruct', and 'upstage/llama-2-70b-instruct-v2'. The 'Average' column shows scores ranging from 71.7 to 73.13. The 'ARC', 'HellaSwag', 'MMLU', and 'TruthfulQA' columns show scores ranging from 60.97 to 87.94.

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
◆	garage-bAInd/Platypus2-70B-instruct	73.13	71.84	87.94	70.48	62.26
◆	deepnight-research/llama-2-70B-instruct	72.95	71.08	87.89	70.58	62.25
◆	upstage/llama-2-70b-instruct-v2	72.95	71.08	87.89	70.58	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
○	MayaPH/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
○	ehartford/Samantha-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-bAInd/Camel-Platypus2-70B	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 RL

The screenshot shows the 'Open LLM Leaderboard' page on Hugging Face. The top navigation bar includes 'Spaces', 'HuggingFaceH4/open_llm_leaderboard', '4.67k', 'Running on CPU UPGRADE', 'App', 'Files', 'Community', and a user icon. Below the header, there's a brief introduction about the leaderboard's purpose and how to submit models. The main area features a table of LLM Benchmark results with columns for Model, Average, ARC, HellaSwag, MMLU, and TruthfulQA. The table lists various models, each with a unique color-coded diamond icon and a link to their GitHub repository. The 'Average' column shows scores ranging from 71.7 to 73.13.

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
◆	garage-bAInd/Platypus2-70B-instruct	73.13	71.84	87.94	70.48	62.26
◆	deepnight-research/llama-2-70B-instruct	72.95	71.08	87.89	70.58	62.25
◆	upstage/llama-2-70b-instruct-v2	72.95	71.08	87.89	70.58	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
○	MayaPH/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
○	ehartford/Samantha-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-bAInd/Camel-Platypus2-70B	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 we evaluate humans?

The screenshot shows a user interface for a 'Playground task' titled 'Talk to the assistant'. The task is labeled 'Task 1 of 5'. The instructions ask the user to decide on a task for the AI assistant to help with, enter it, interact with the AI, and leave a comment on how the assistant did. A 'Comments' box contains the text 'I thought the assistant was ...'. Below it is a 'Rating' scale with five circles, labeled 'Bad' on the left and 'Good' on the right. 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

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 a 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}

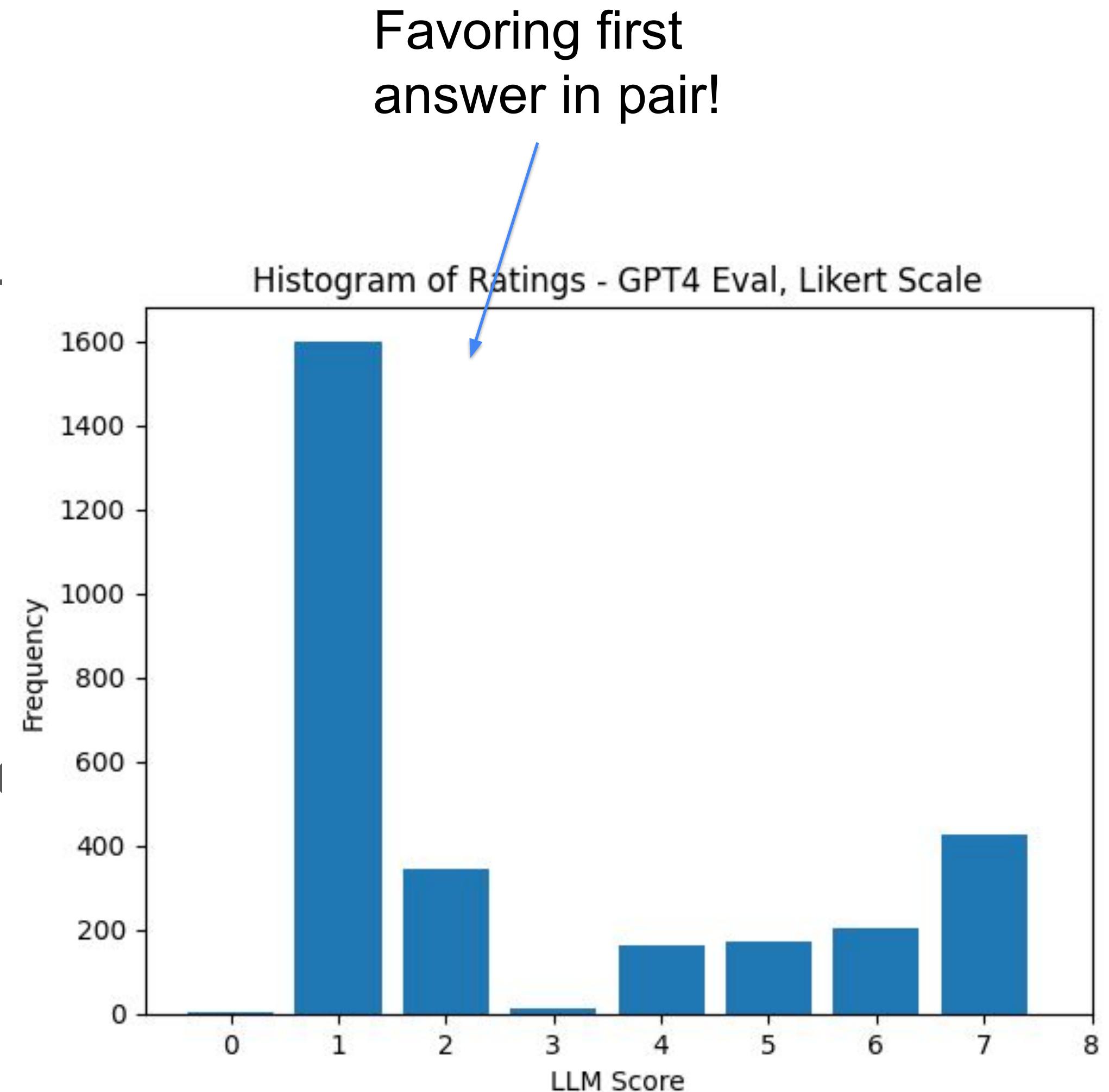
[The End of Assistant's Answer]

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



Rajani, N. et al. Can foundation models label data like humans? HuggingFace Blog, 2023.

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]]

Score: 3/B

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