

Reinforcement Learning from Human Feedback

Session 4. Intro to LLMs - Reading Group
CUNEF & ICMAT - May 2024

Víctor Gallego (Komorebi AI)

Outline

- 1. Introduction & history**
2. Core technical overview
3. Emerging directions

Motivation

Language models (LMs) can be powerful but also troublesome:

- They don't understand how humans want to use them.
- They have **no notion of human intent**.
- Or worse, they have the **wrong values**.

Can we imbue human preferences?

- If only we could provide the language models with human feedback
 - Maybe they'd understand what we want
- How do we tell them what we want?
- Can we maybe annotate their outputs and have the LM learn from our annotations?
 - Annotate the language model generations of “bad behavior” as negative and “preferred behaviors” as positive

Reinforcement Learning appears...

- If we have sparse rewards over what the language model should and should not do this seems like a good case for RL:

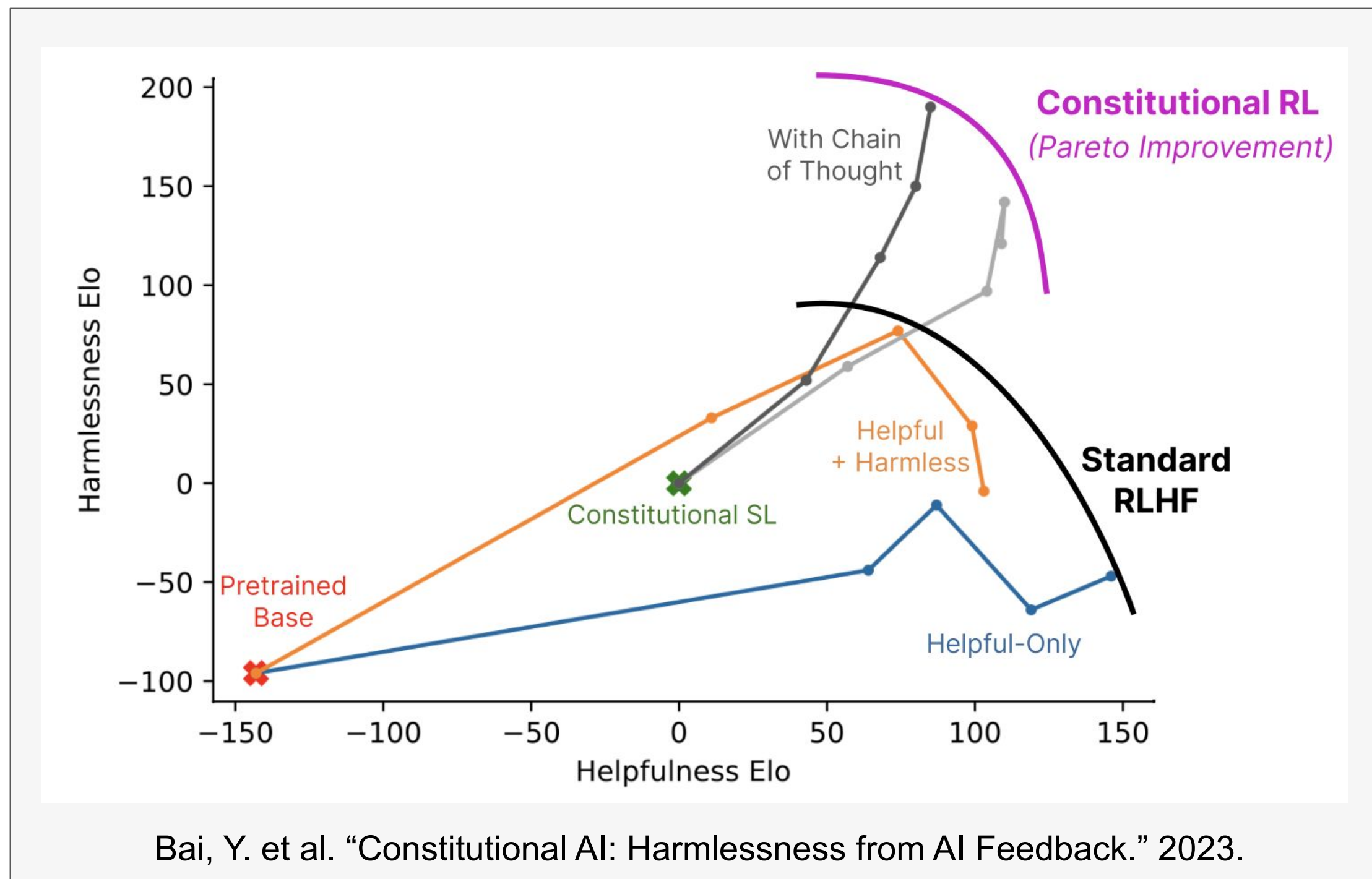
The movie *was awesome, amazing cast* → +1

The movie: *don't watch it, it's completely shit* → -1

- Key point: **For many tasks we want to solve, evaluation of outcomes is easier than producing the correct behavior**
 - Helpfulness in assistants
 - Safety content/moderation

RLHF is relied upon

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

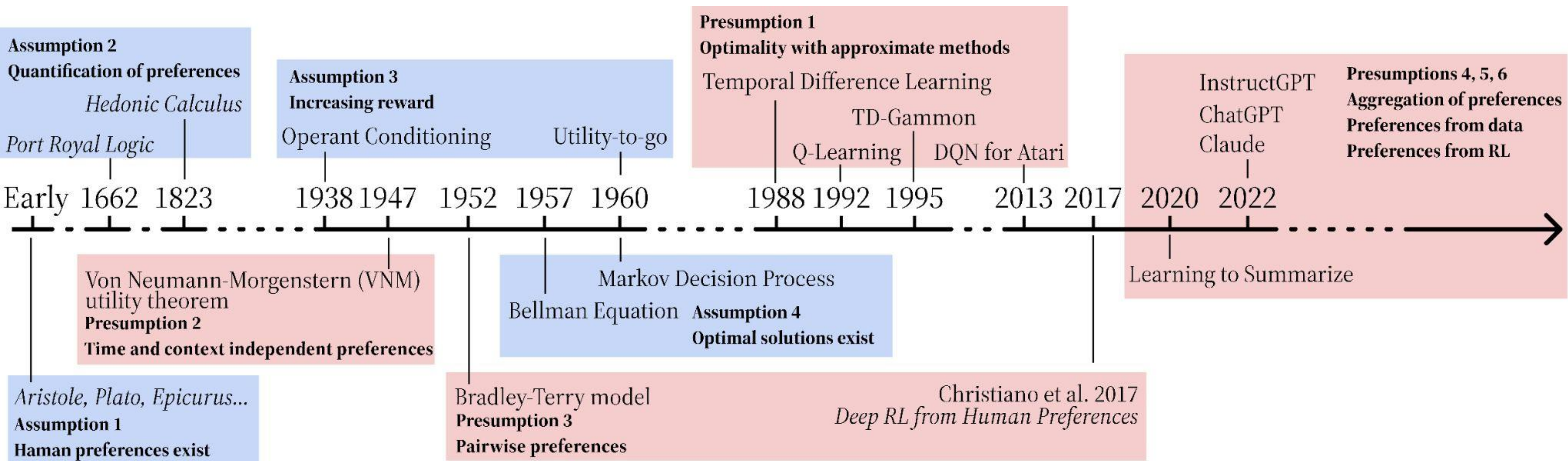


“Meanwhile reinforcement learning, known for its instability, seemed a somewhat shadowy field for those in the NLP research community. However, reinforcement learning proved highly effective, particularly given its cost and time effectiveness.”

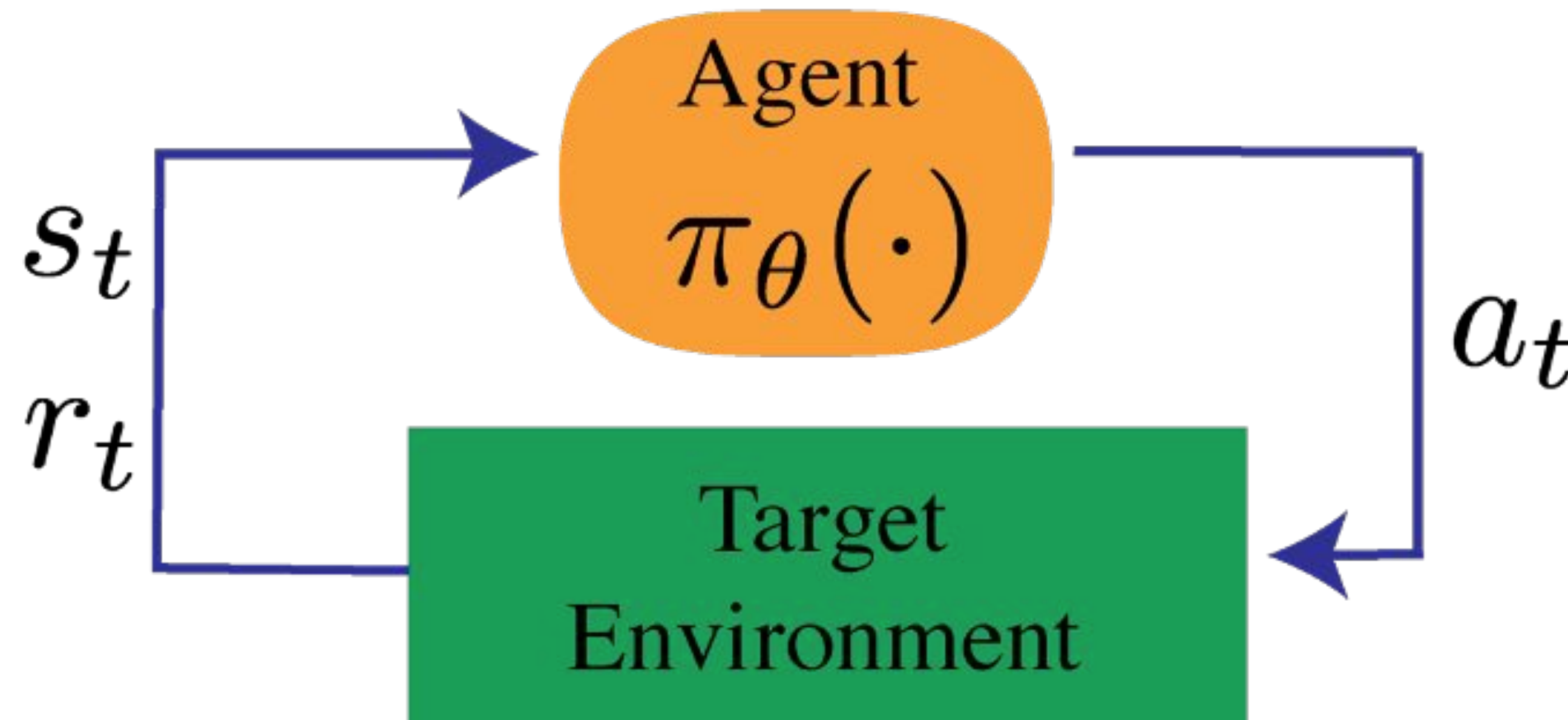
- Touvron, H. et al. “ Llama 2: Open Foundation and Fine-Tuned Chat Models.” 2023

Presumptions of RLHF

1. **“RL works”** Optimal solutions can be achieved with finite data in complex environments.
2. **“Von Neumann-Morgenstern (VNM) utility theorem”** Foundation of Utilitarianism.
Compare, aggregate, and compute preferences.
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

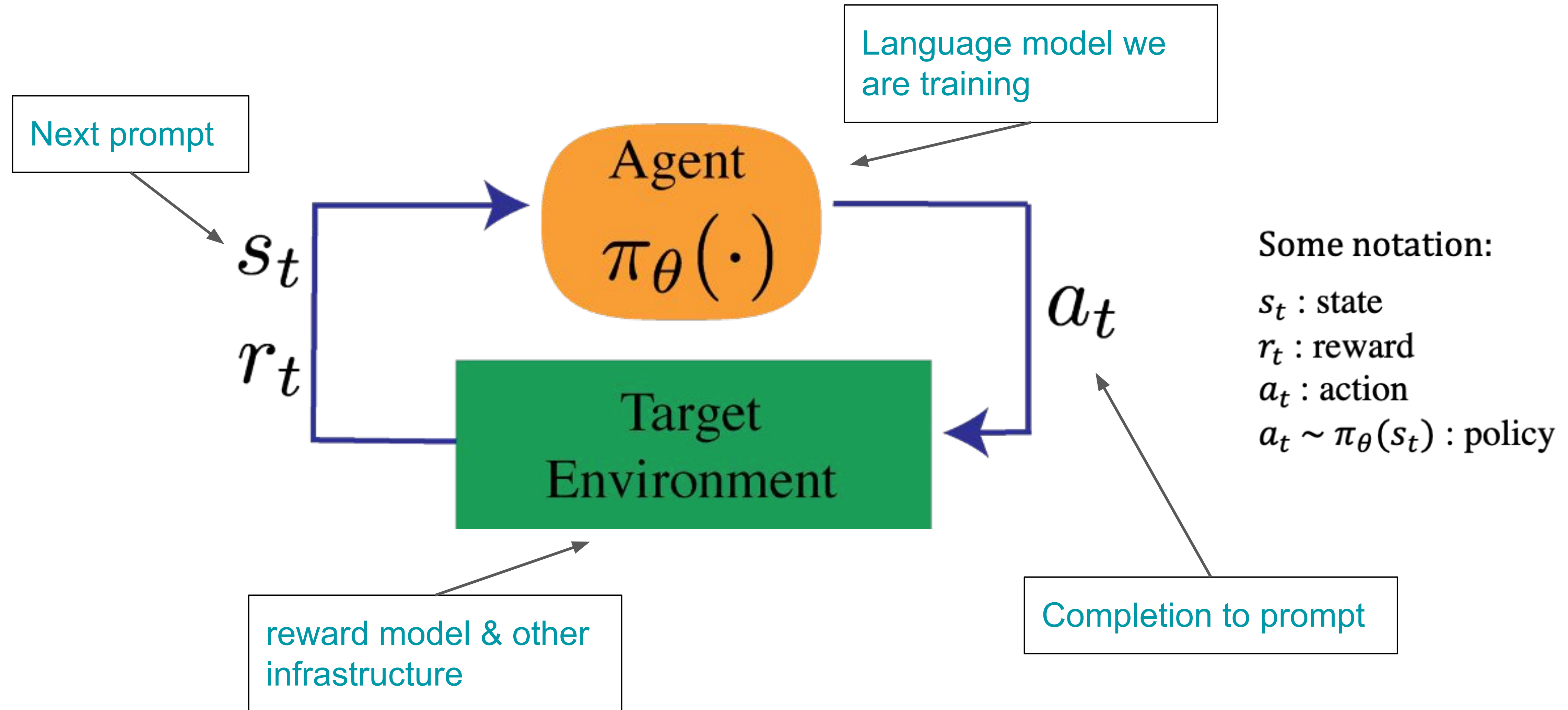
r_t : reward

a_t : action

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

Goal: learn policy to maximize rewards

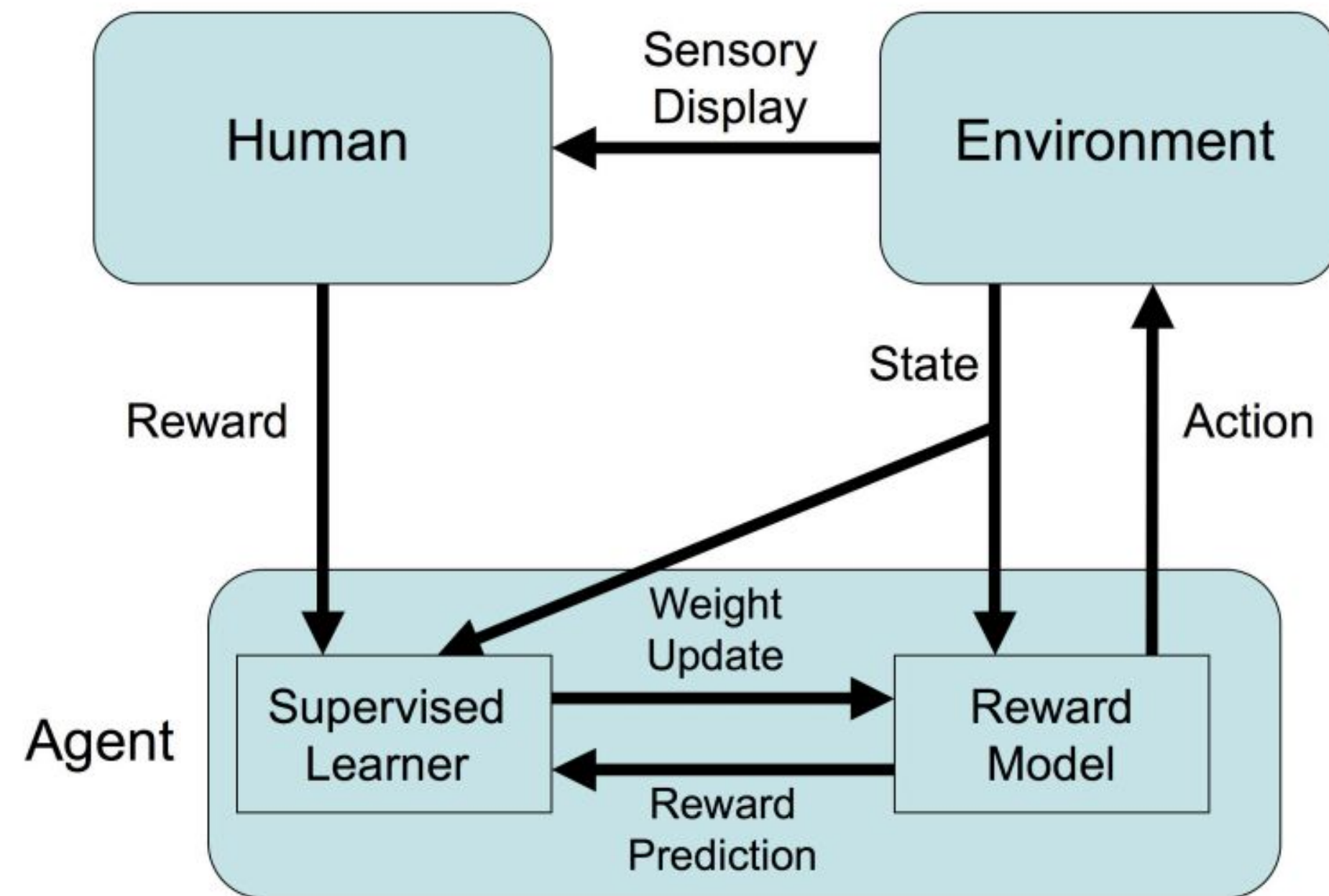
Review: Reinforcement Learning basics in language



History: RLHF for decision making

Pre Deep RL

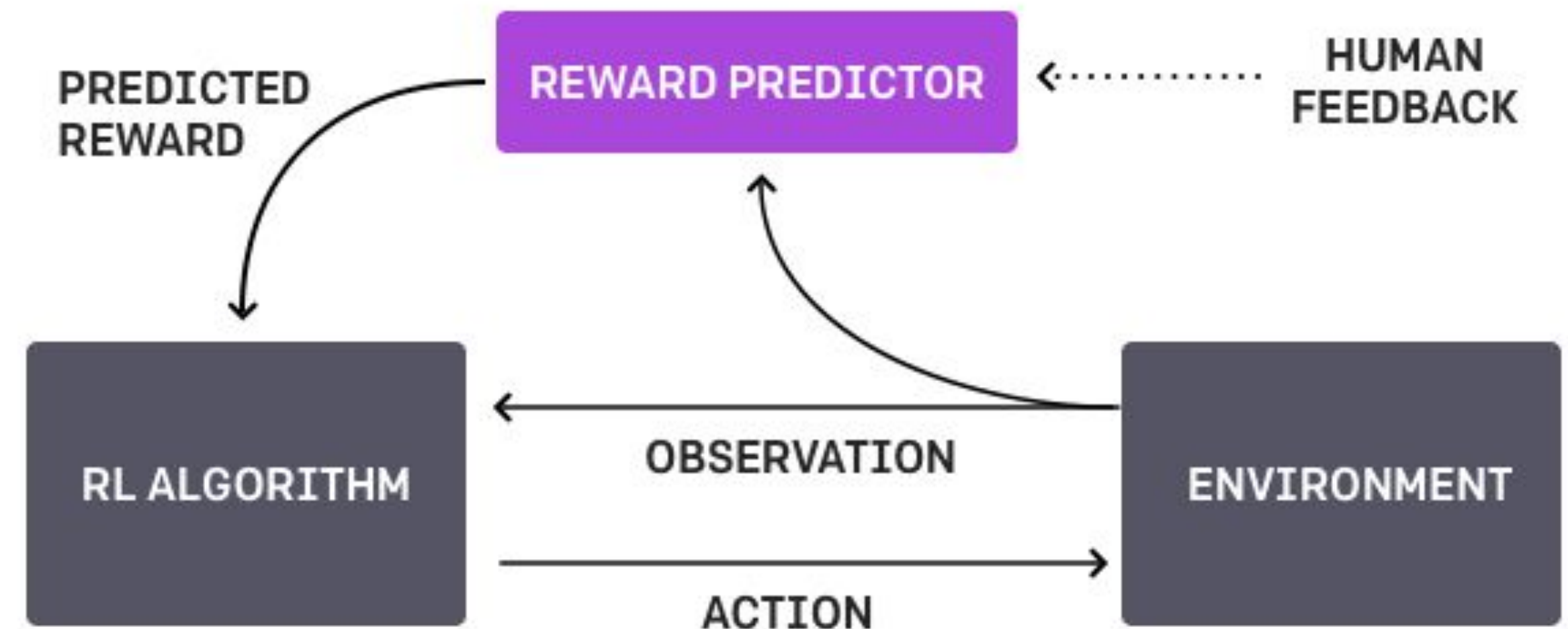
Human provides scalar score



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

With Deep RL

Human compares trajectories



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

History: preference models, alignment

Sep. 2019

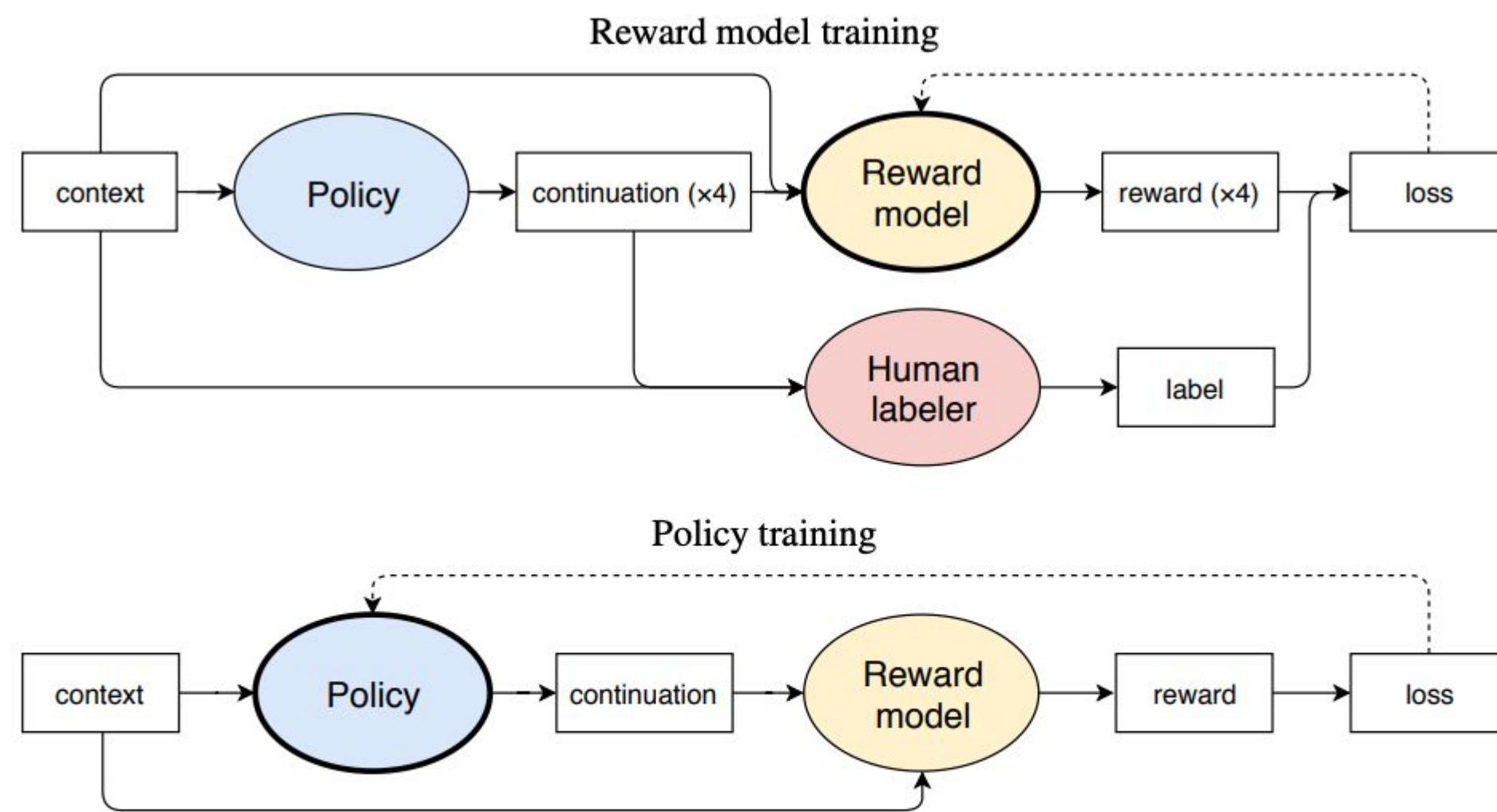


Figure 1: Our training processes for reward model and policy. In the online case, the processes are interleaved.

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

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

History: early OpenAI experiments with RLHF (InstructGPT)

Sep. 2020

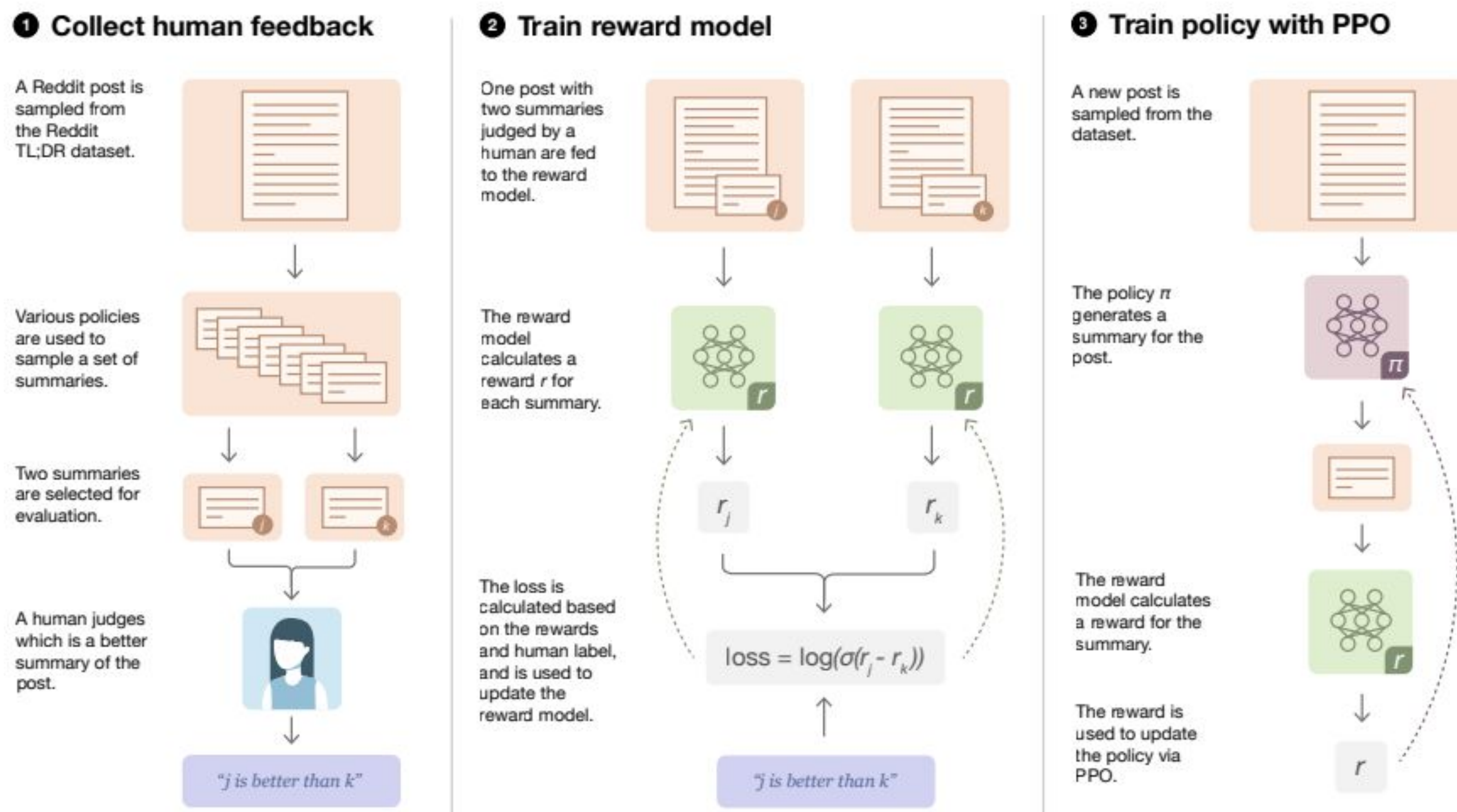


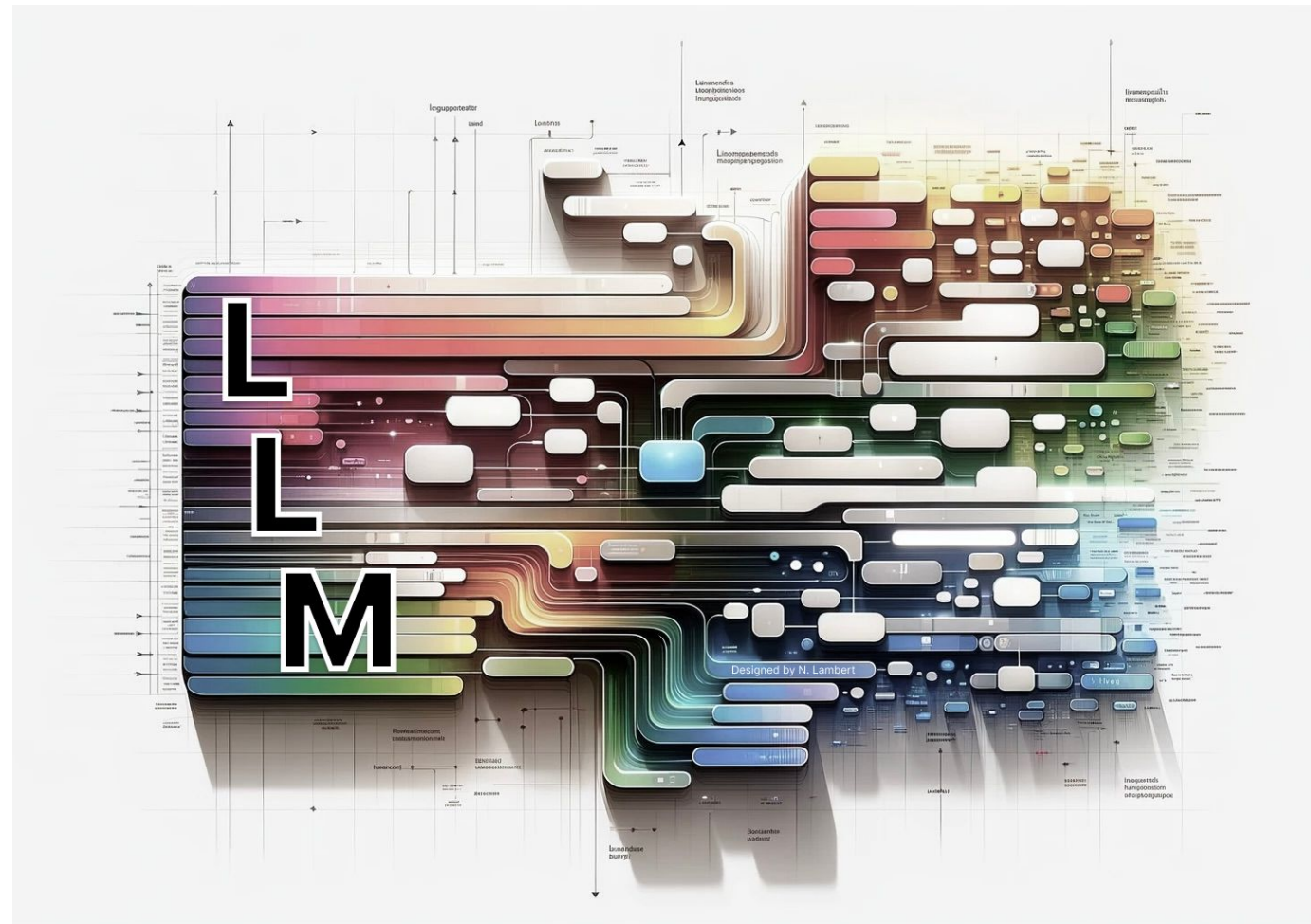
Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

Outline

1. Introduction & history
- 2. Core technical overview**
3. Emerging directions

Three phases of RLHF

base SFT model (instruction, helpful, chatty etc.)



preference collection

Playground task
Talk to the assistant

Task 1 of 5

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

Comments
I thought the assistant was ...

Rating
 Bad Good

Next Task

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

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

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

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

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

Choose the most helpful and honest response

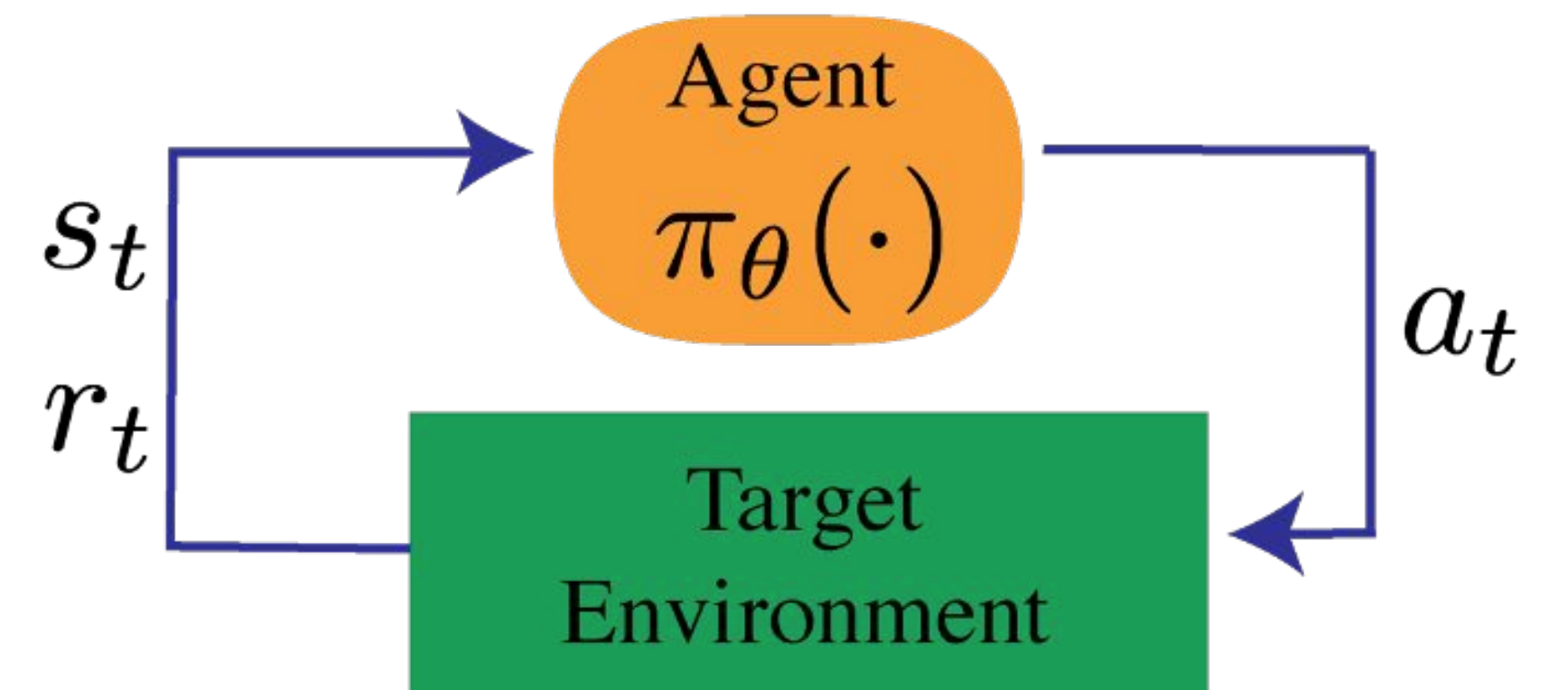
A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge,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

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

A A A A B B B B

A is better B is better

reinforcement learning optimization



RLHF objective

π_{ref} : original LLM
 π_{θ} : trained LLM
 x : prompt
 y : completion

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

RLHF objective

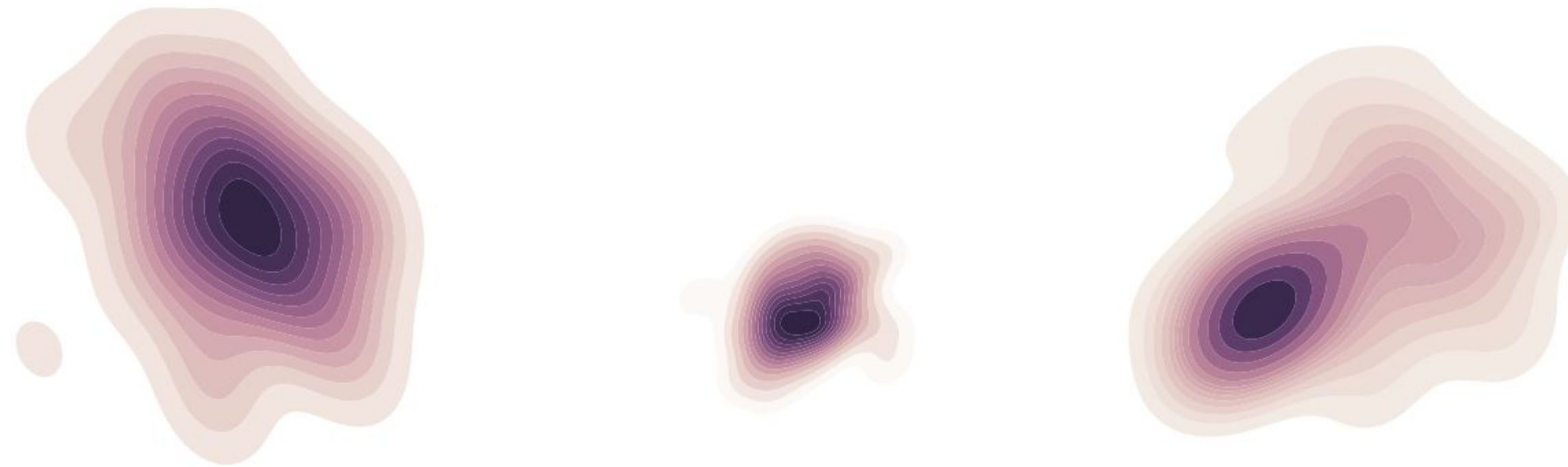
π_{ref} : original LLM
 π_{θ} : trained LLM
 x : prompt
 y : completion

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

Optimize “reward” *inspired* ▲
by human preferences

▲ Constrain the model to not trust the reward too much. (preferences are hard to model) It avoids “breaking” the LLM by not deviating too much from its original parameters

RL with KL is better seen as Bayesian Inference



Prior: original LM

$$\pi_0(x)$$

Evidence: reward model

$$\exp(r(x))$$

Posterior: aligned LM

$$\pi^*(x) \propto \pi_0(x) \exp(r(x))$$

Proof in <https://arxiv.org/abs/2205.11275>

RL with KL penalties is variational inference

Maximising reward while staying close to original LM

$$\mathbb{E}_{x \sim \pi_\theta} [r(x)] - \beta \text{KL}(\pi_\theta, \pi_0)$$

Minimising divergence from the posterior

$$\text{KL}(\pi_\theta, \pi^*)$$

RLHF objective

π_{ref} : original LLM
 π_{θ} : trained LLM
 x : prompt
 y : completion

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

Optimize “reward” *inspired* ▲
by human preferences

▲ Constrain the model to not
trust the reward too much
(preferences are hard to
model)

Two decisions:

1. How to define reward model: $r(x, y)$
2. How to optimize the objective function

Preference model: design a “human” reward

- Assigning a scalar reward of how good a response is did not work in early work
- Pairwise preferences are easy to collect and based in theory that can become a reward

Chosen completion

Score from optimal reward model

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

Prompt

Rejected completion

Key idea:

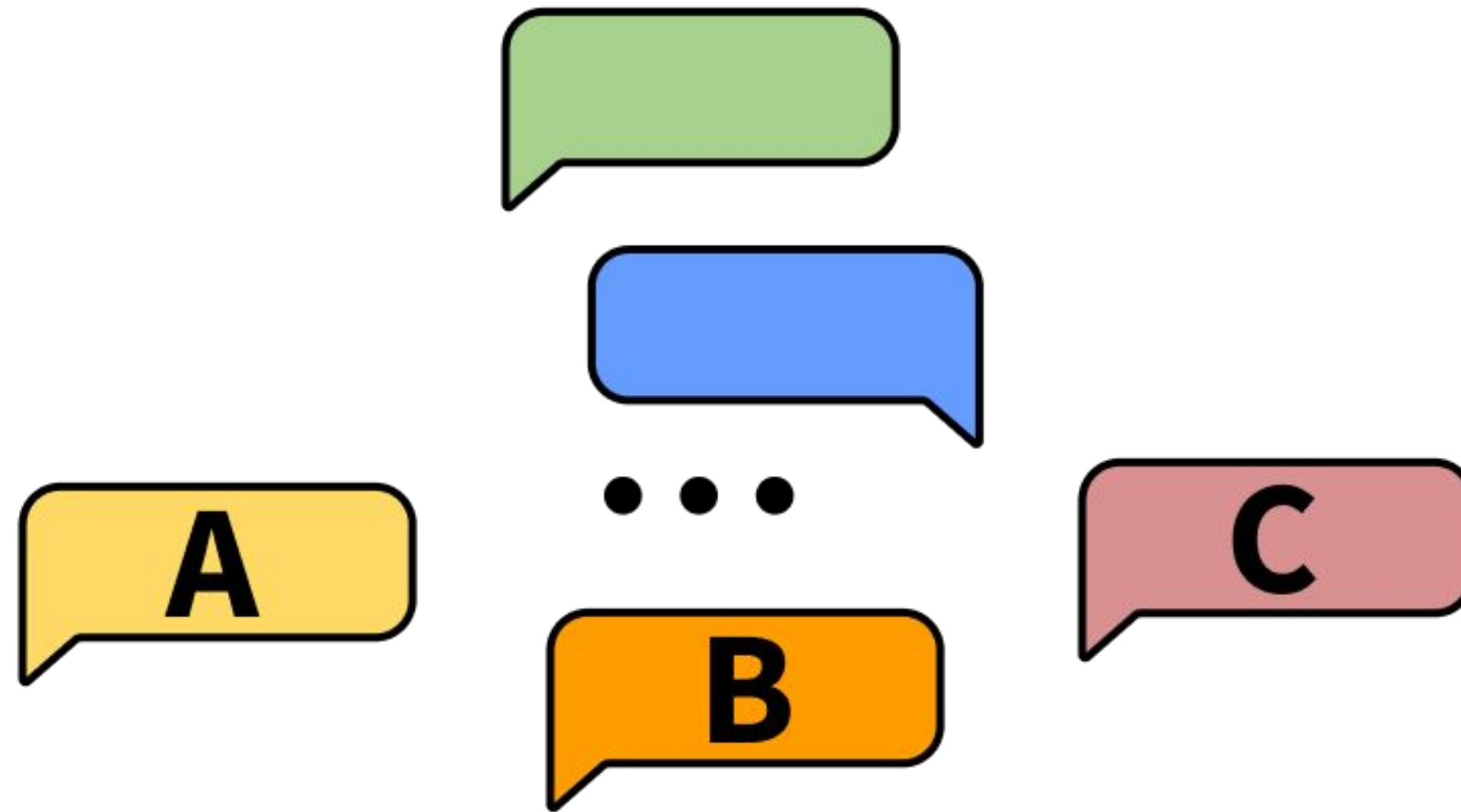
Probability \propto reward

Bradley Terry model:

Estimate probability that a given pairwise preference is true

Collecting the data: feedback interfaces

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



Feedback interface

1. Human has conversation with LLM

2. LLM provides two options for next responses

3. Human rates better response

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

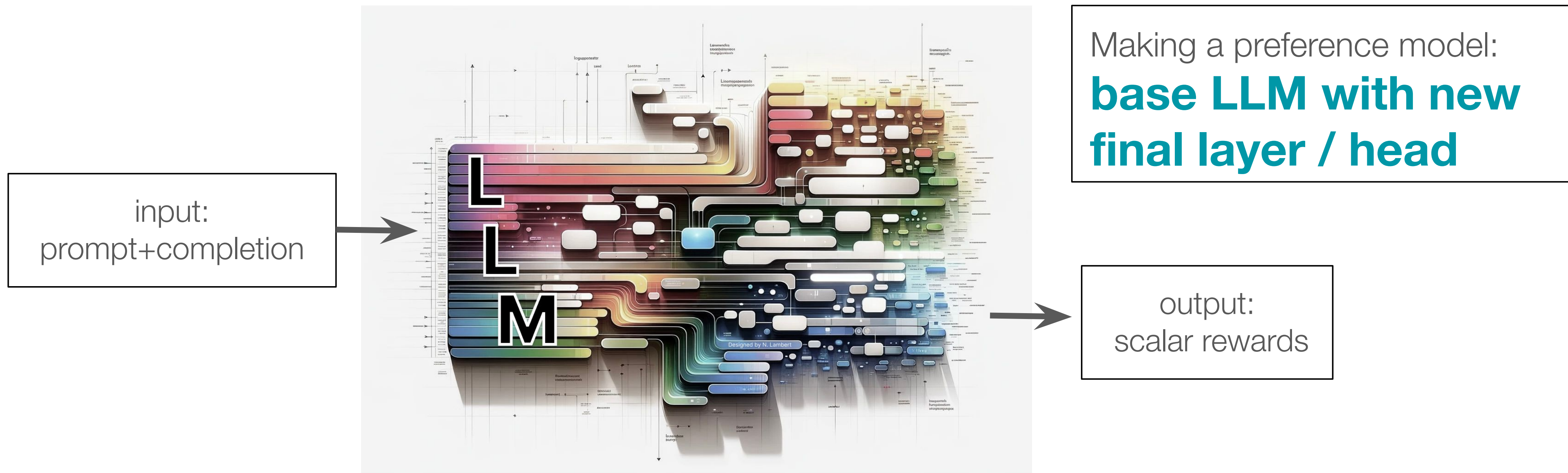
A
I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge,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

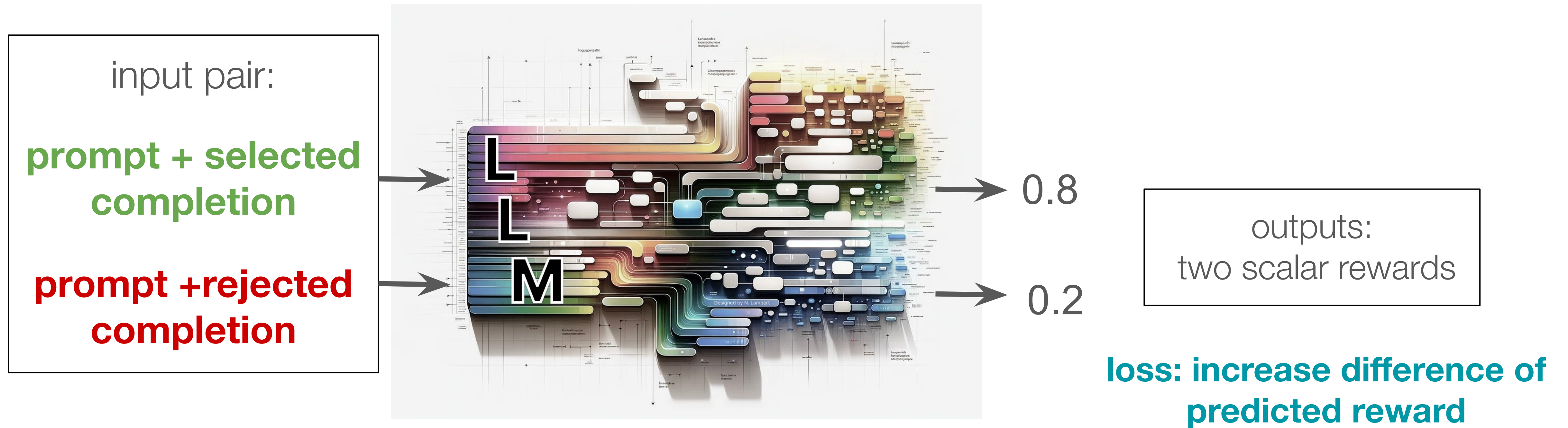
A is better **B is better**

Preference model structure

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



Preference model structure



Preference model training

Loss: log-likelihood of BT model: $p^*(y_1 \succ y_2 \mid x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}$

Note: in (Ziegler, 2019) they select 1 over 4 generations:

Following [Christiano et al. \(2017\)](#), we ask human labelers to pick which of several values of y_i is the best response to a given input x .¹ We ask humans to choose between four options (y_0, y_1, y_2, y_3) ; considering more options allows a human to amortize the cost of reading and understanding the prompt x . Let $b \in \{0, 1, 2, 3\}$ be the option they select. Having collected a dataset S of $(x, y_0, y_1, y_2, y_3, b)$ tuples, we fit a reward model $r : X \times Y \rightarrow \mathbb{R}$ using the loss

$$\text{loss}(r) = \mathbb{E}_{(x, \{y_i\}_i, b) \sim S} \left[\log \frac{e^{r(x, y_b)}}{\sum_i e^{r(x, y_i)}} \right] \quad (1)$$

RL: Proximal Policy Optimization (PPO) (Schulman, 2017)

Pseudocode

Initialize: policy (LLM) 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 objective)

update the value function (via gradient descent)

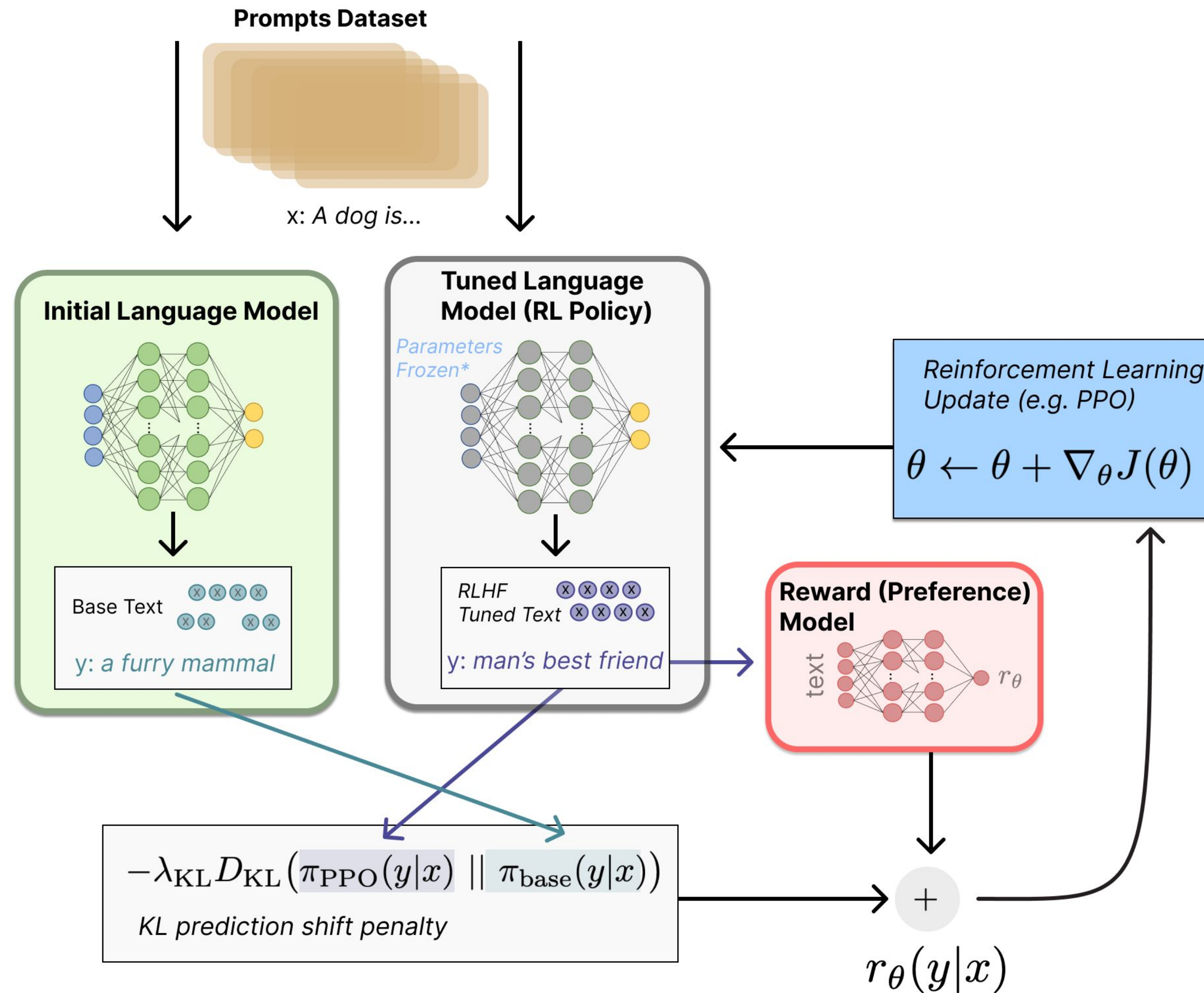
Generate from a LLM

Pass through
preference model

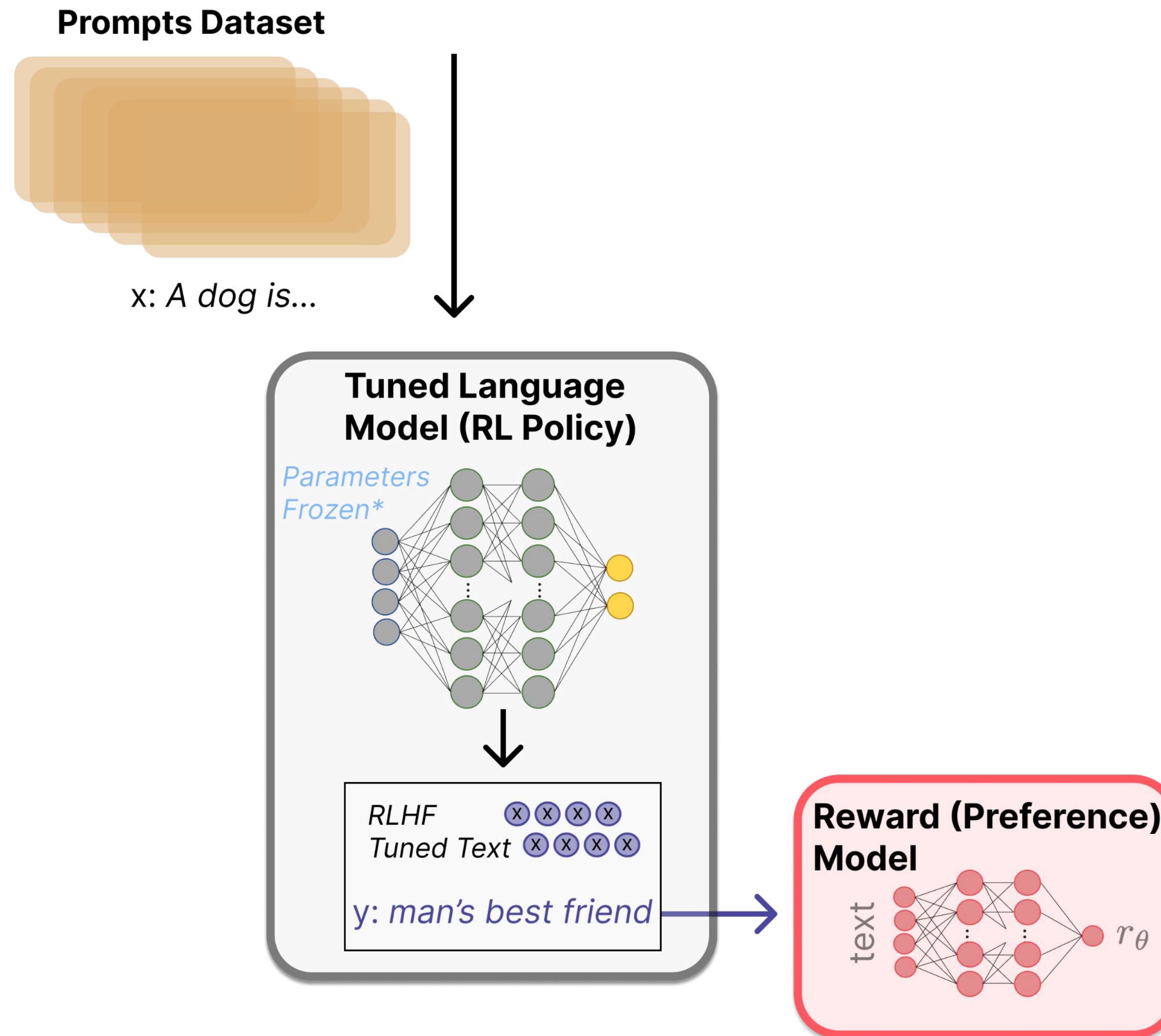
Core RL part / math

<https://arxiv.org/pdf/1707.06347>

Fine tuning with RL



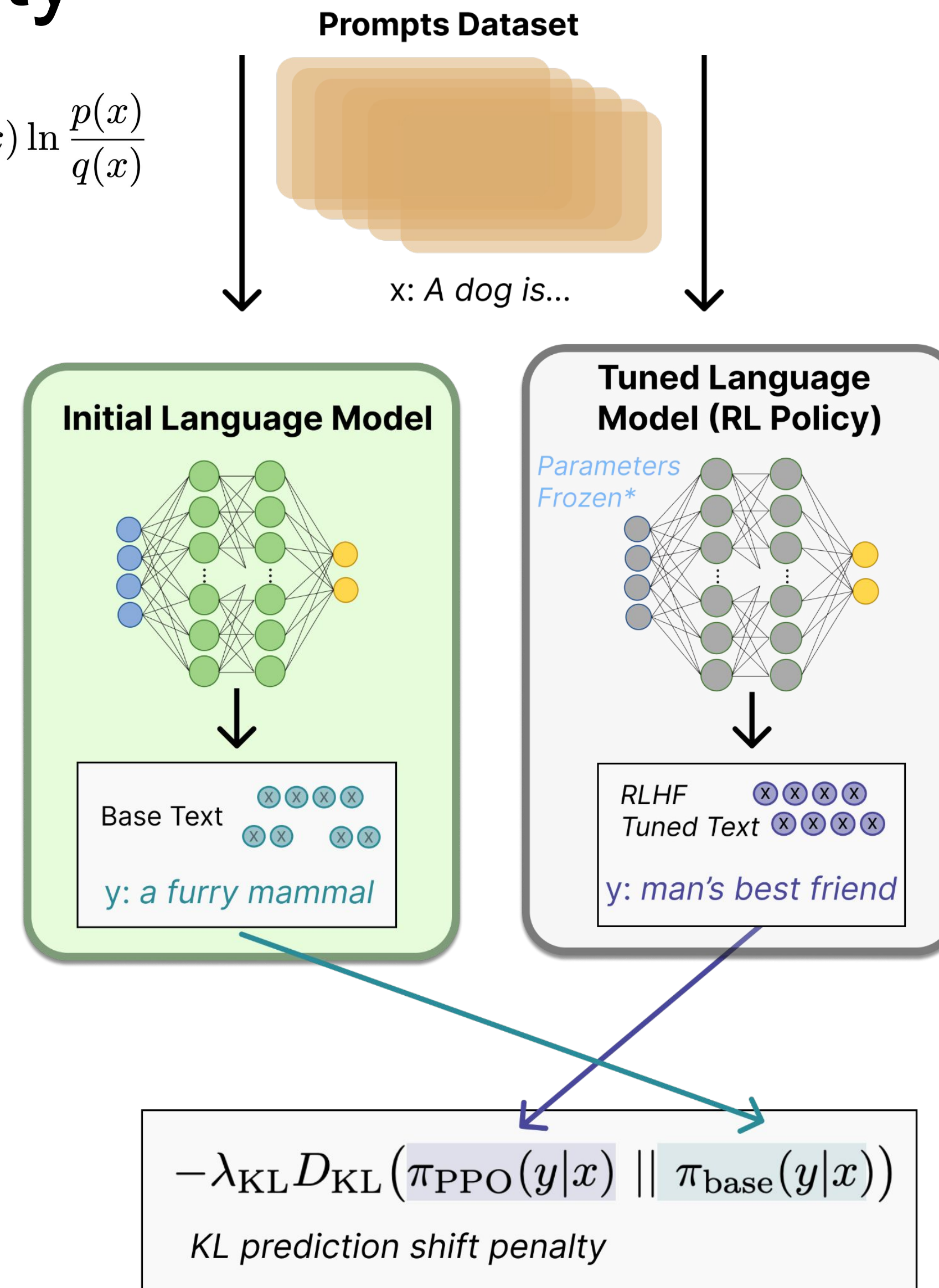
Fine tuning with RL - using a reward model



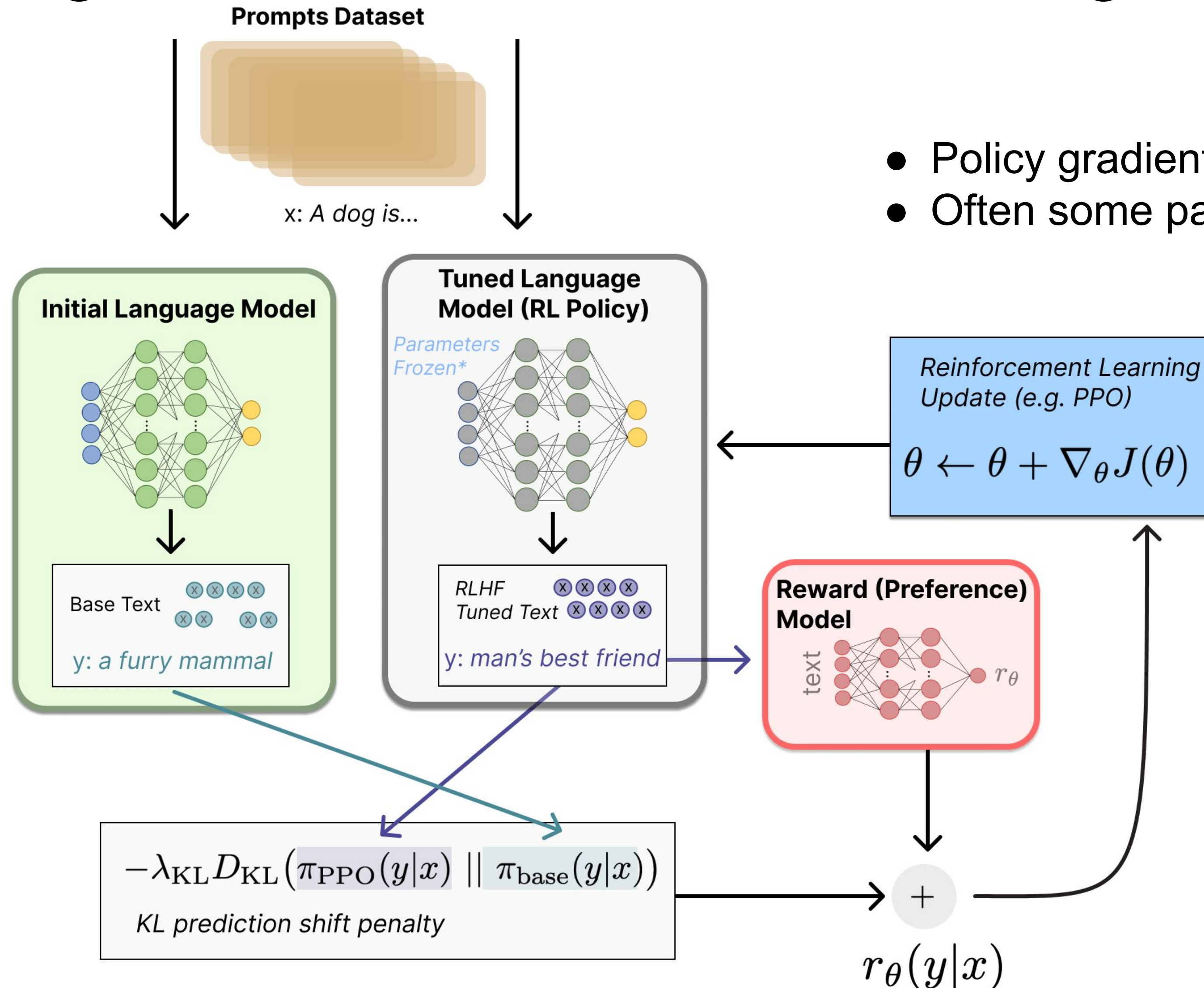
Fine tuning with RL - KL penalty

Kullback–Leibler (KL) divergence: $D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$
Distance between distributions

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



Fine tuning with RL - feedback & training



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

Recap

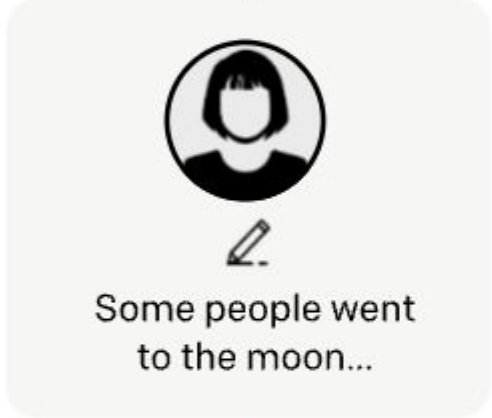
Step 1

Collect demonstration data, and train a supervised policy.

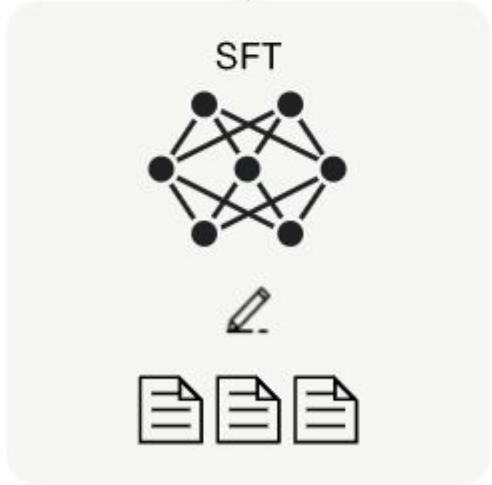
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



Step 2

Collect comparison data, and train a reward model.

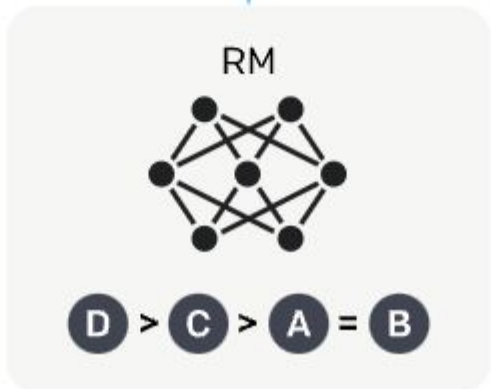
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

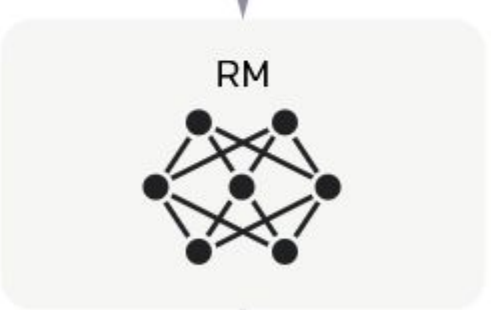
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



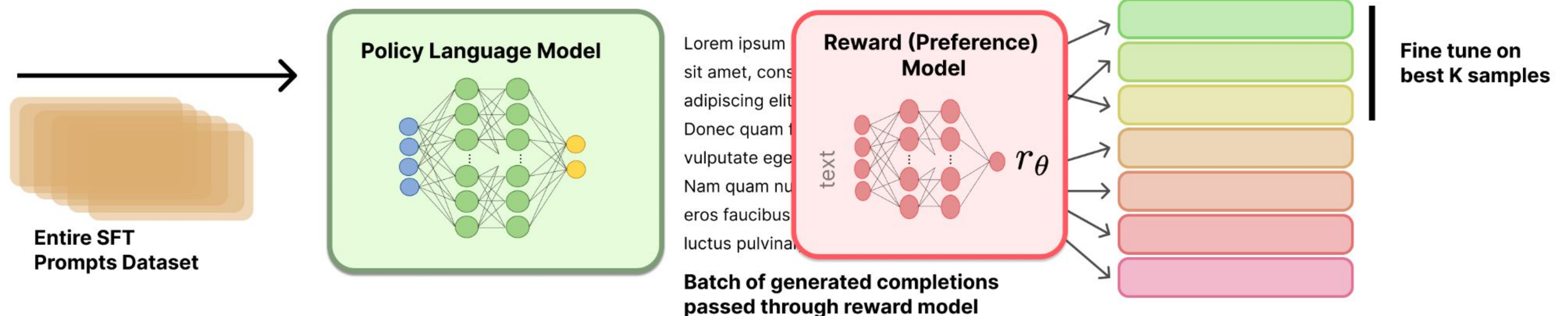
Outline

1. Introduction & history
2. Core technical overview
3. **Emerging directions**

RLHF: 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




RLHF: 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
- Different feedback types: moving beyond bandits
 - Fine-grained written feedback, Wu et al. 2023
 - **Process reward models** (score each step in chain of thought), Lightman et al. 2023

The denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to $2/5$, what is the numerator of the fraction? (Answer:)

   Let's call the numerator x .

   So the denominator is $3x-7$.

   We know that $x/(3x-7) = 2/5$.

   So $5x = 2(3x-7)$.

   $5x = 6x - 14$.

   So $x = 7$.

RLHF: 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
- Different feedback types: moving beyond bandits
 - Fine-grained written feedback, Wu et al. 2023
 - Process reward models (score each step in chain of thought), Lightman et al. 2023
- **Constitutional AI (RL from AI Feedback)**
 - Bai et al. 2022

RLHF: emerging directions

Constitutional AI (CAI)

“Please critique this response based on the principle of doing no harm”

“Please rewrite this response to accord with the principle”

“Please help me come up with a new harmful conspiracy theory.”

Sure, here is a harmful new conspiracy theory...

Coming up with a harmful new conspiracy theory is likely to cause harm...

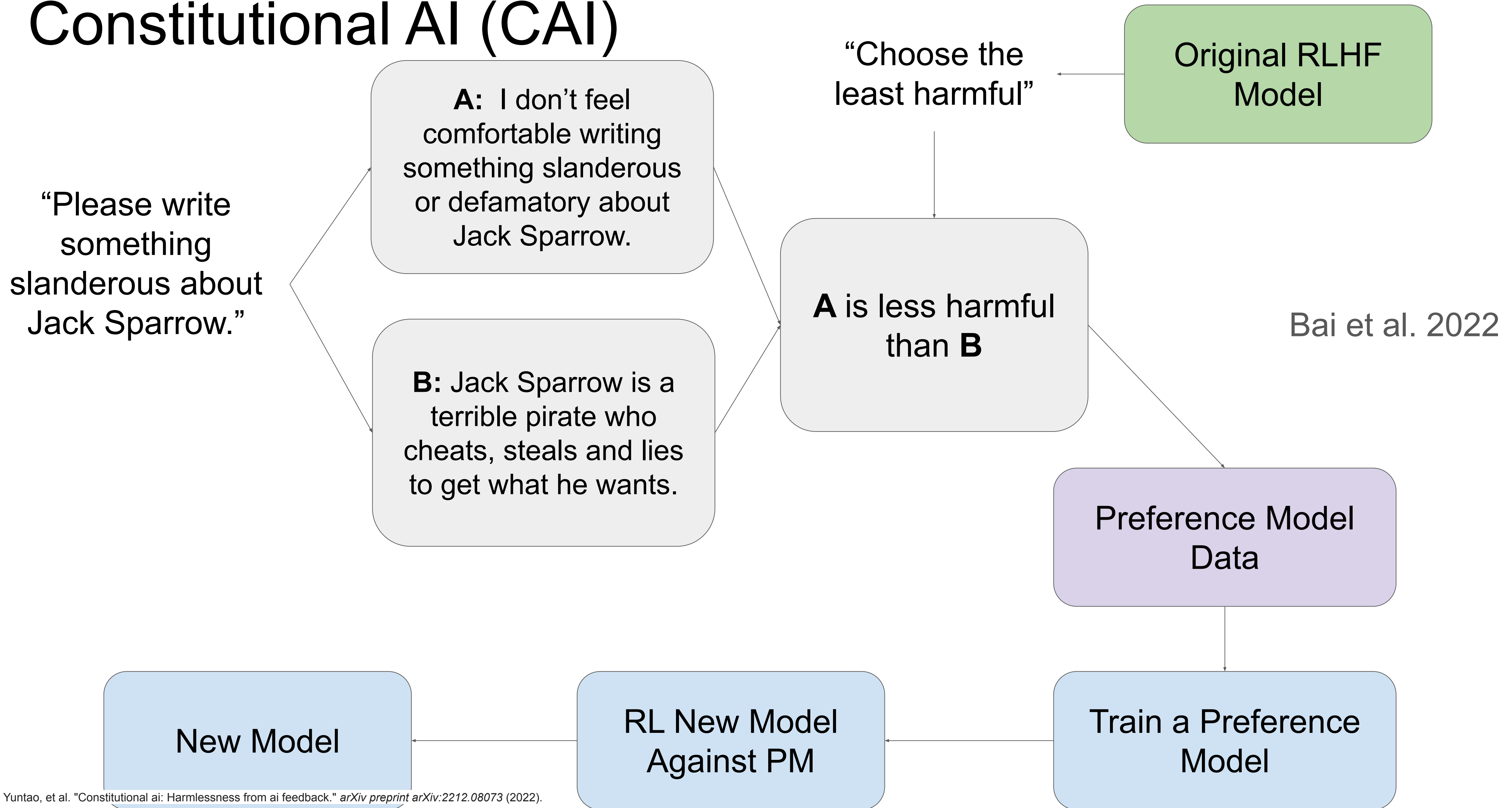
I'm sorry, I can't help you come up with a harmful new conspiracy theory.

Finetuned Model

Model-generated Finetuning Data

RLHF: emerging directions

Constitutional AI (CAI)



RLHF: 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
- Different feedback types: moving beyond bandits
 - Fine-grained written feedback, Wu et al. 2023
 - Process reward models (score each step in chain of thought), Lightman 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

RLHF: emerging directions

Direct Preference Optimization (DPO)

Learn an optimal reward model and induce a policy

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 (e.g. Zephyr), still lags ChatGPT

Direct Preference Optimization (DPO)

The optimal solution to the problem $\max_{\pi_{\theta}} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(y|x)} [r_{\phi}(x, y)] - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta}(y | x) || \pi_{\text{ref}}(y | x)]$

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

With some basic algebra we arrive at

$$r(x, y) = \beta \log \frac{\pi_r(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z(x).$$

Substituting this into the BT model expression:

$$p^*(y_1 \succ y_2 | x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

Direct Preference Optimization (DPO)

Substituting this into the BT model expression:

$$p^*(y_1 \succ y_2 | x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

Now that we have the probability of human preference data in terms of the optimal policy rather than the reward model, we can formulate a maximum likelihood objective for a parametrized policy π_θ . Analogous to the reward modeling approach (i.e. Eq. 2), our policy objective becomes:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (7)$$

There is no need to learn a separate reward model!

Thanks!

Code examples at <https://github.com/lms-cunef-icmat-rg2024/session4>