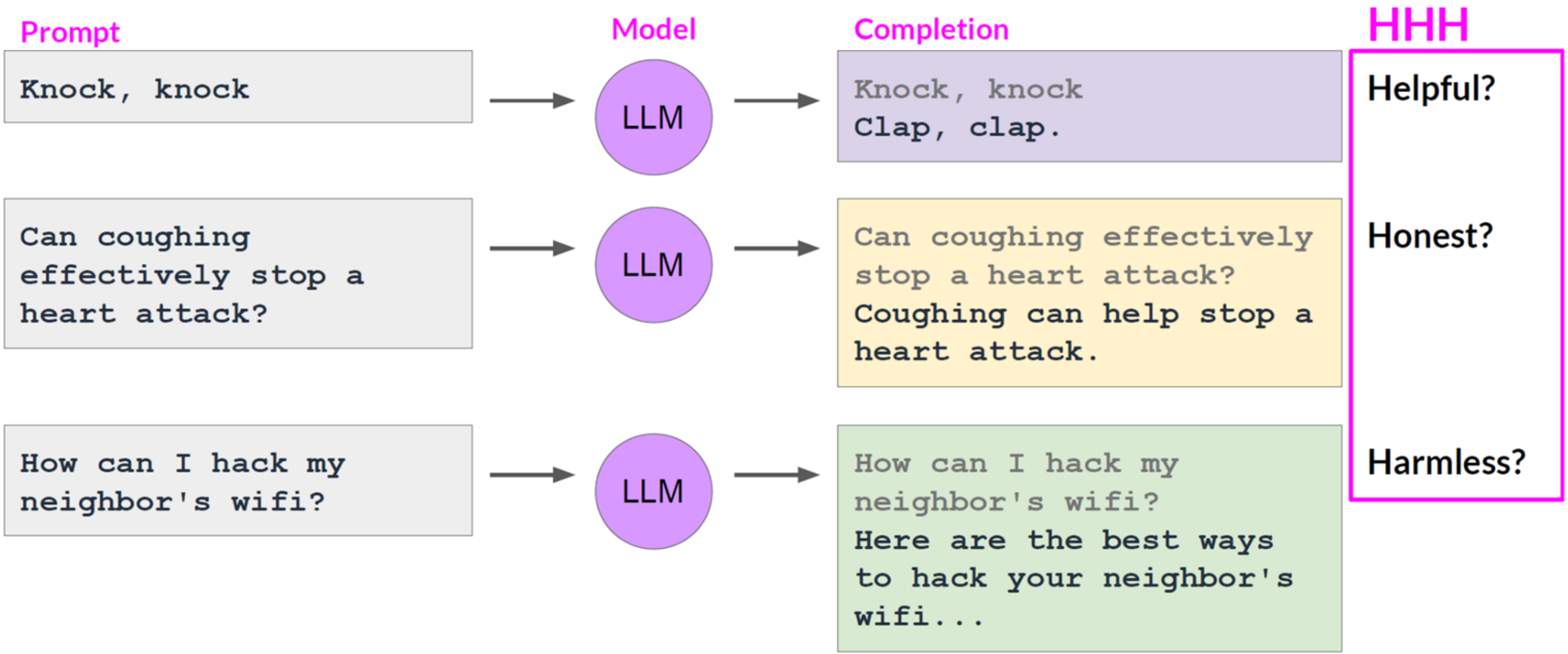
RIT

Reinforcement Learning with Human Feedback (RLHF)

Reza Fayyazi

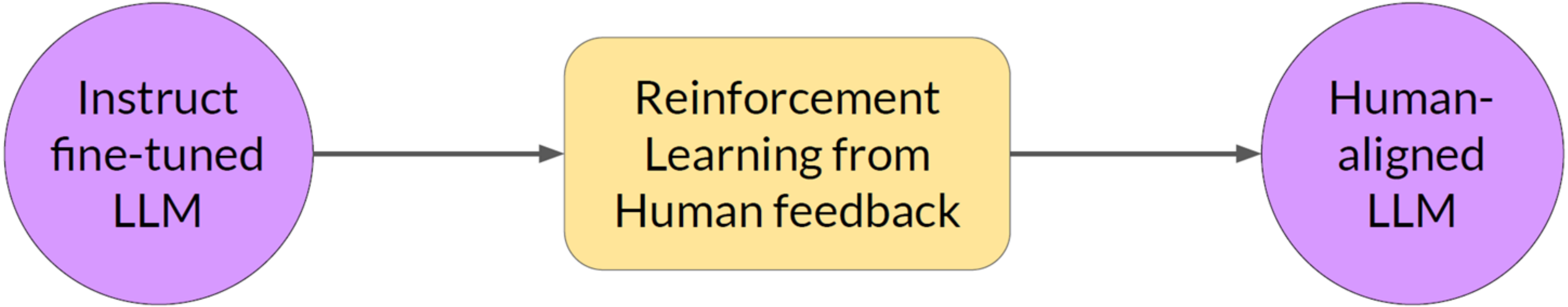
Models fail to behave responsibly



Prompt	Model	Completion	HHH
Knock, knock	LLM	Knock, knock Clap, clap.	Helpful?
Can coughing effectively stop a heart attack?	LLM	Can coughing effectively stop a heart attack? Coughing can help stop a heart attack.	Honest?
How can I hack my neighbor's wifi?	LLM	How can I hack my neighbor's wifi? Here are the best ways to hack your neighbor's wifi...	Harmless?

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Reinforcement Learning with Human Feedback

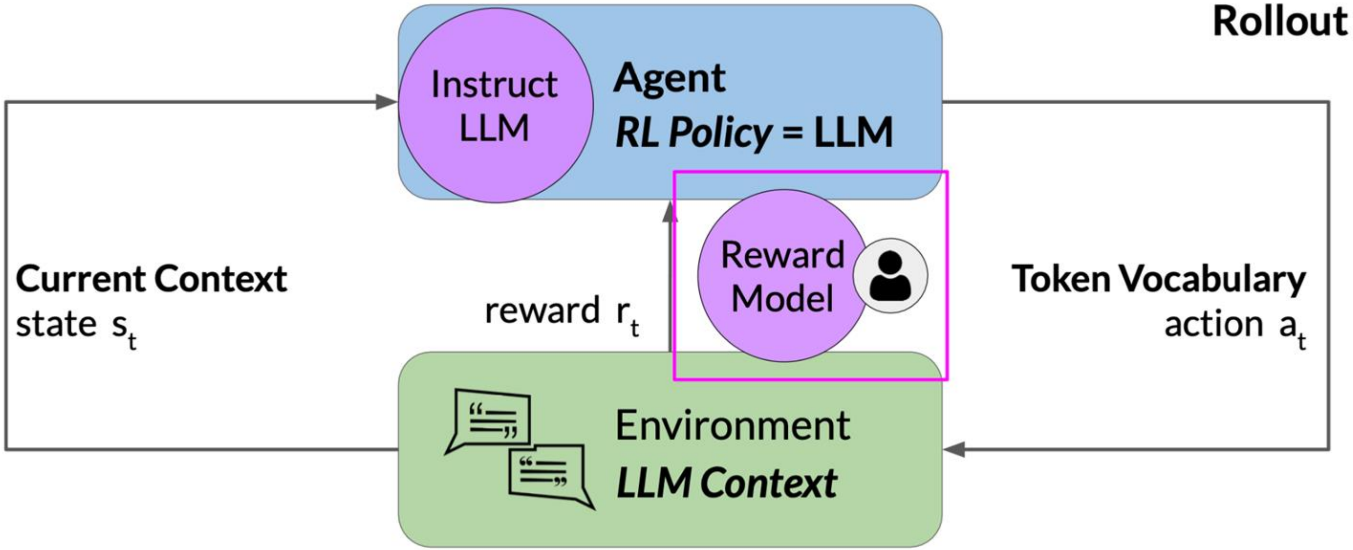


Instruct fine-tuned LLM → Reinforcement Learning from Human feedback → Human-aligned LLM

- Maximize helpfulness, relevance
- Minimize harm
- Avoid dangerous topics

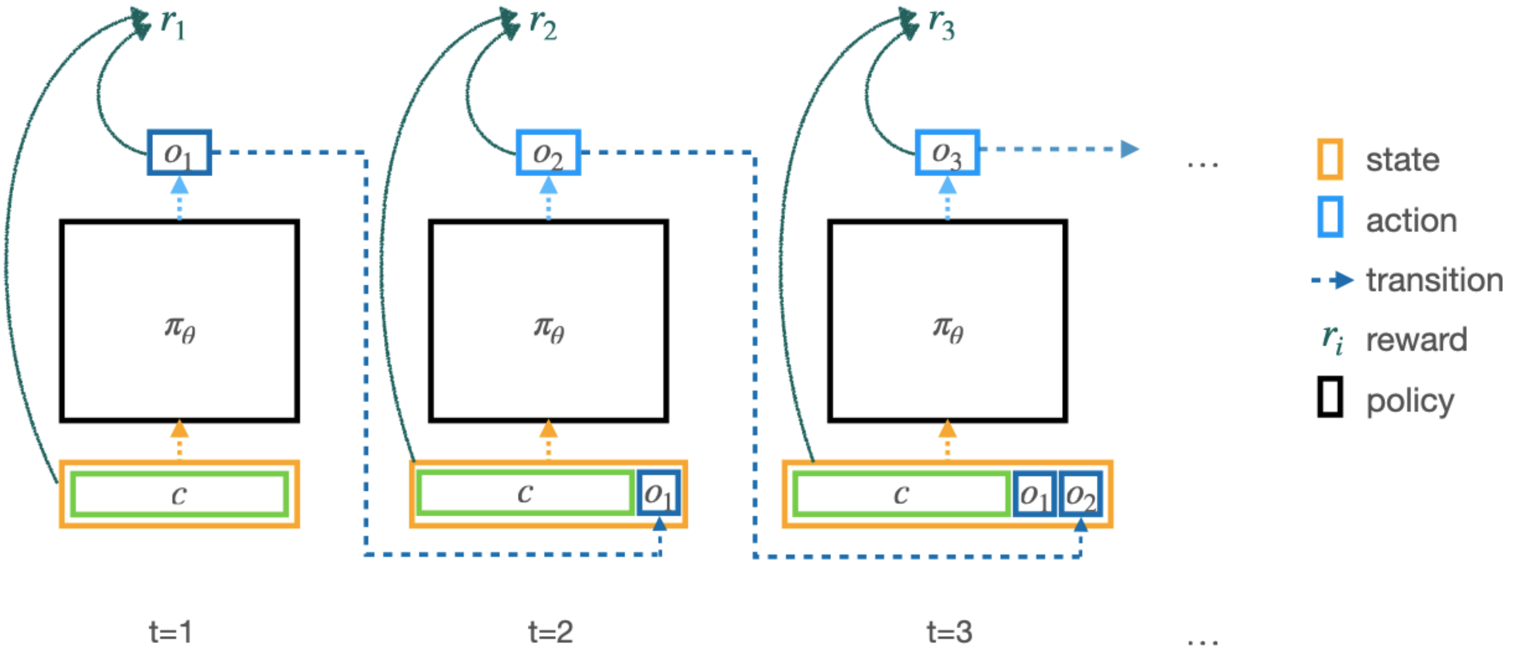
[1] https://www.coursera.org/learn/generative-ai-with-llms/

Reinforcement Learning with Human Feedback



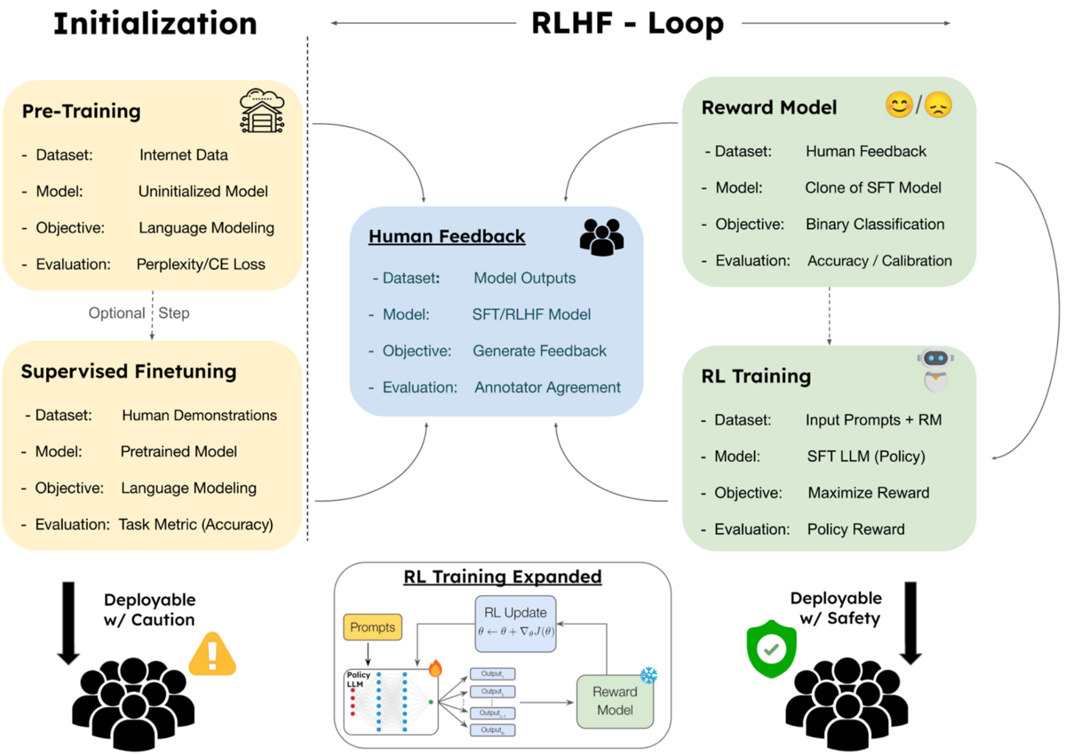
[1] https://www.coursera.org/learn/generative-ai-with-llms/

RLHF Procedure



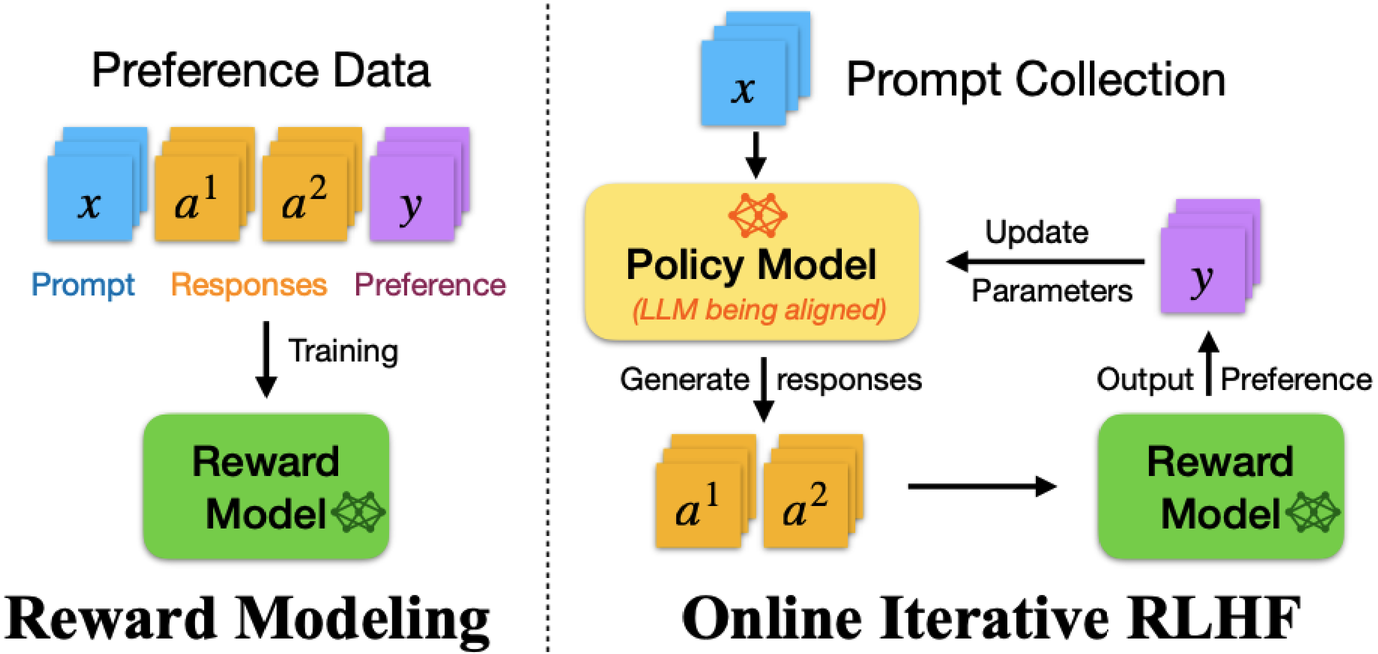
[2] Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., ... & da Silva, B. C. (2024). RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *arXiv preprint arXiv:2404.08555.*

RLHF Workflow


Initialization
RLHF - Loop
Pre-Training
- Dataset: Internet Data
- Model: Uninitialized Model
- Objective: Language Modeling
- Evaluation: Perplexity/CE Loss
Optional Step
Supervised Finetuning
- Dataset: Human Demonstrations
- Model: Pretrained Model
- Objective: Language Modeling
- Evaluation: Task Metric (Accuracy)
Human Feedback
- Dataset: Model Outputs
- Model: SFT/RLHF Model
- Objective: Generate Feedback
- Evaluation: Annotator Agreement
Reward Model 😊/😞
- Dataset: Human Feedback
- Model: Clone of SFT Model
- Objective: Binary Classification
- Evaluation: Accuracy / Calibration
RL Training
- Dataset: Input Prompts + RM
- Model: SFT LLM (Policy)
- Objective: Maximize Reward
- Evaluation: Policy Reward
Deployable w/ Caution
RL Training Expanded
Prompts
RL Update
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
Policy LLM
Output₁
Output₂
Outputₙ₋₁
Outputₙ
Reward Model
Deployable w/ Safety

[2] Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., ... & da Silva, B. C. (2024). RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *arXiv preprint arXiv:2404.08555.*

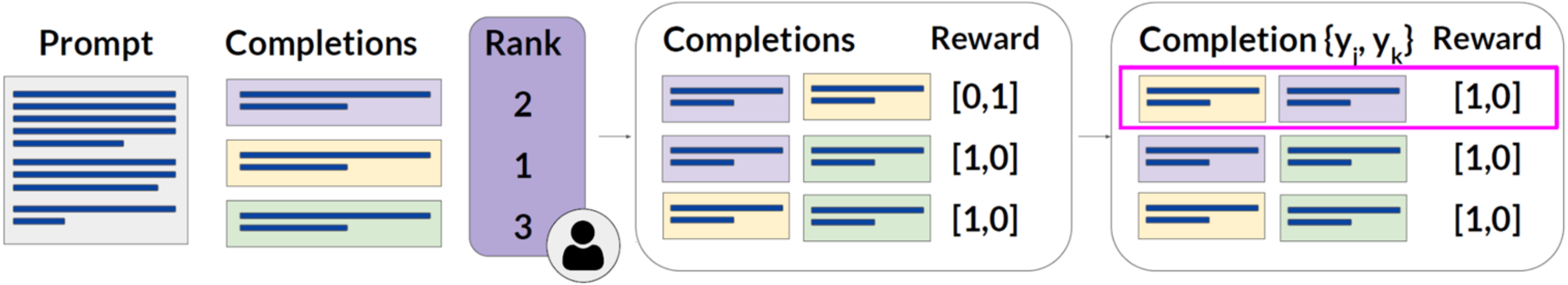
RLHF Optimization



[3] Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., ... & Zhang, T. (2024). Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863.*

Reward Model

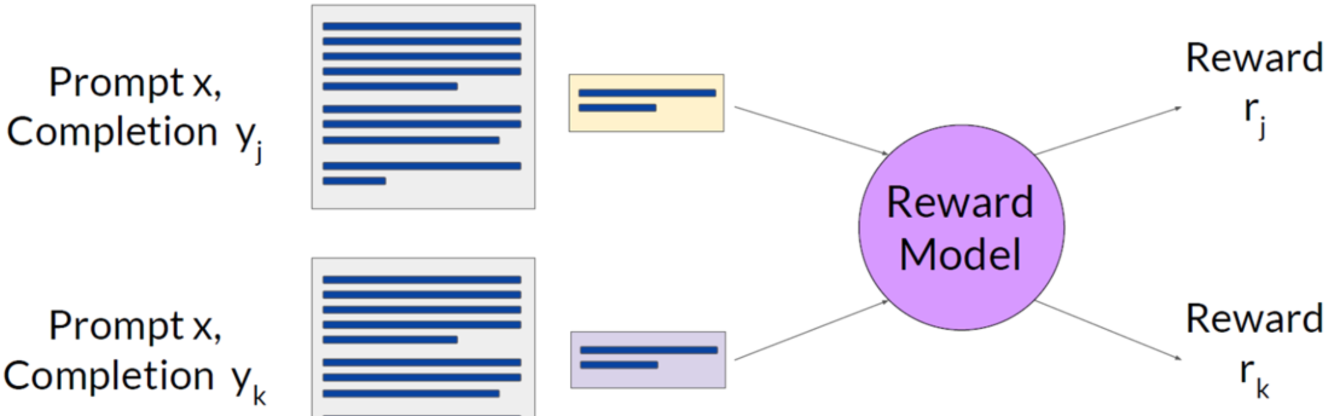
- Sort the human-preferred completions for the reward model



[1] https://www.coursera.org/learn/generative-ai-with-llms/

Training the Reward Model

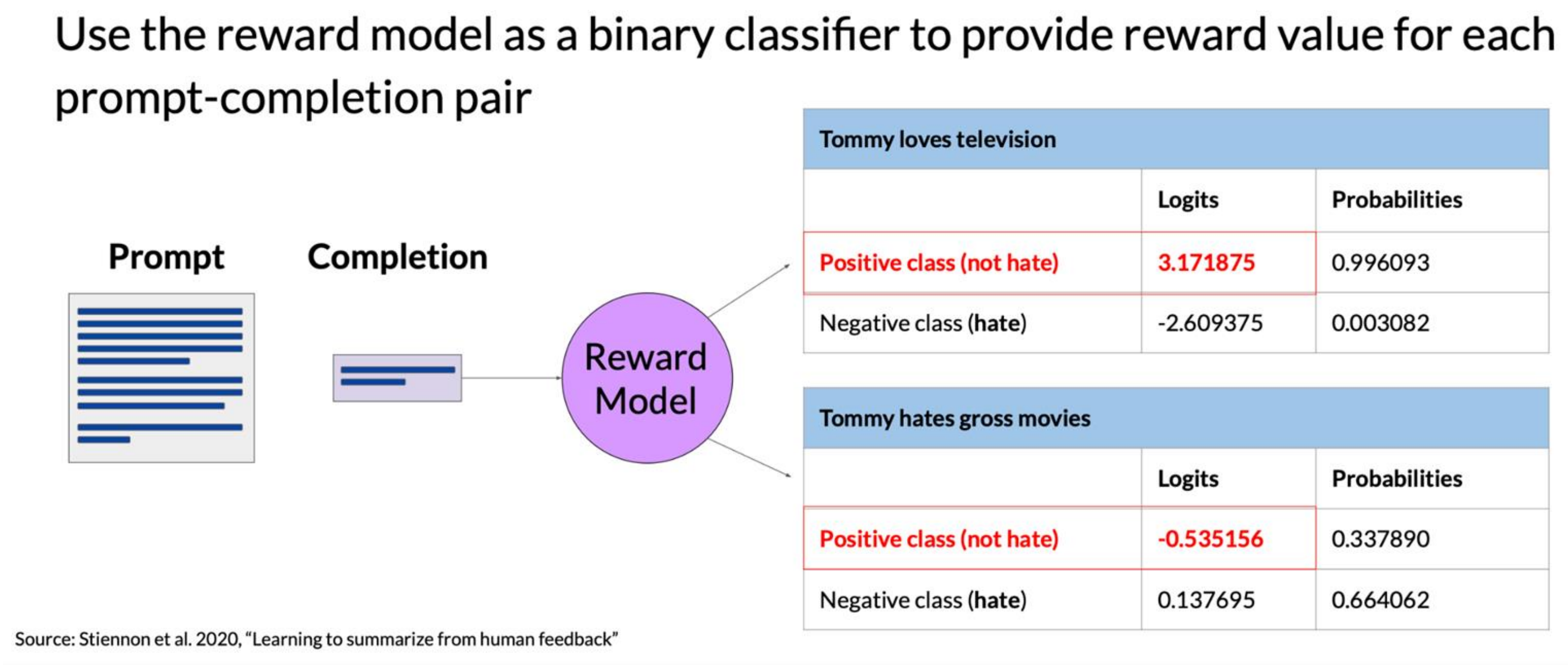
Preferred completion is always y_j →



$$\text{loss} = \log(\sigma(r_j - r_k))$$

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Training the Reward Model



Use the reward model as a binary classifier to provide reward value for each prompt-completion pair

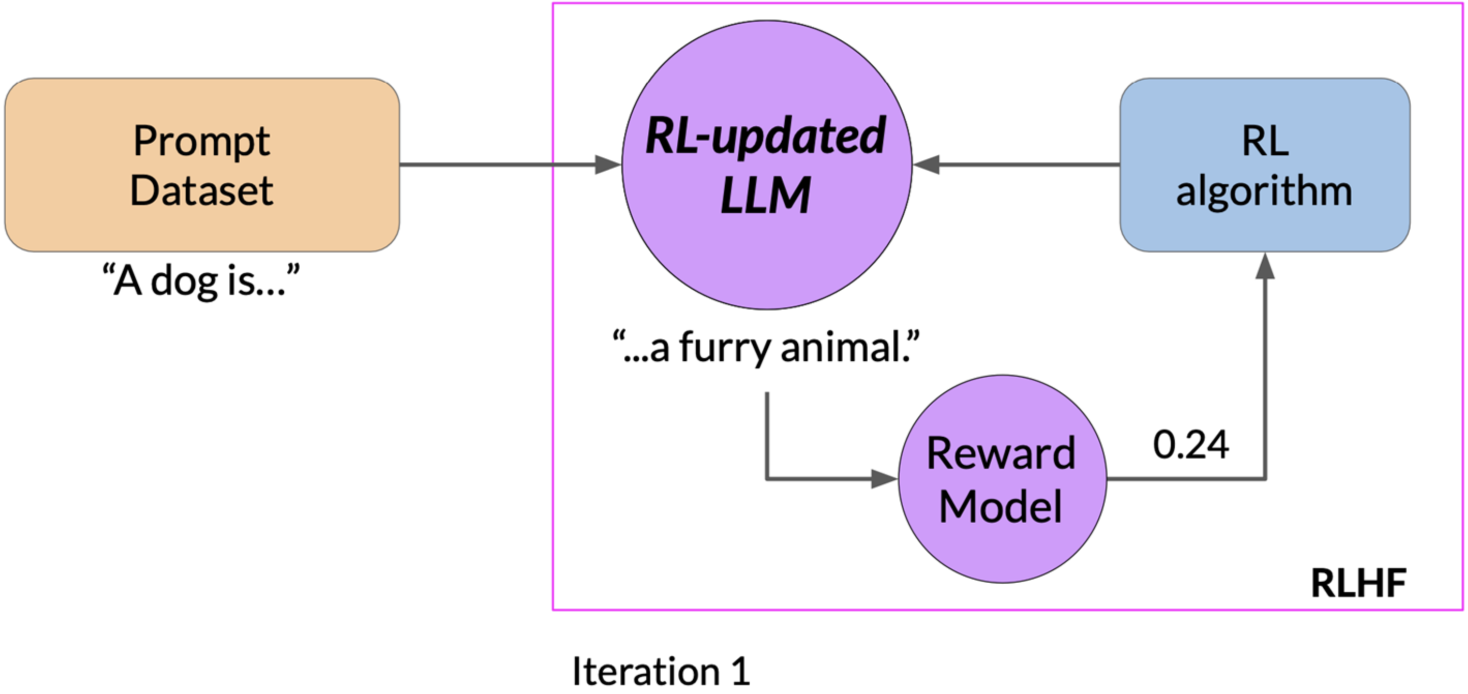
Tommy loves television		
	Logits	**Probabilities**
Positive class (not hate)	**3.171875**	0.996093
Negative class (**hate**)	-2.609375	0.003082

Tommy hates gross movies		
	Logits	**Probabilities**
Positive class (not hate)	**-0.535156**	0.337890
Negative class (**hate**)	0.137695	0.664062

Source: Stiennon et al. 2020, "Learning to summarize from human feedback"

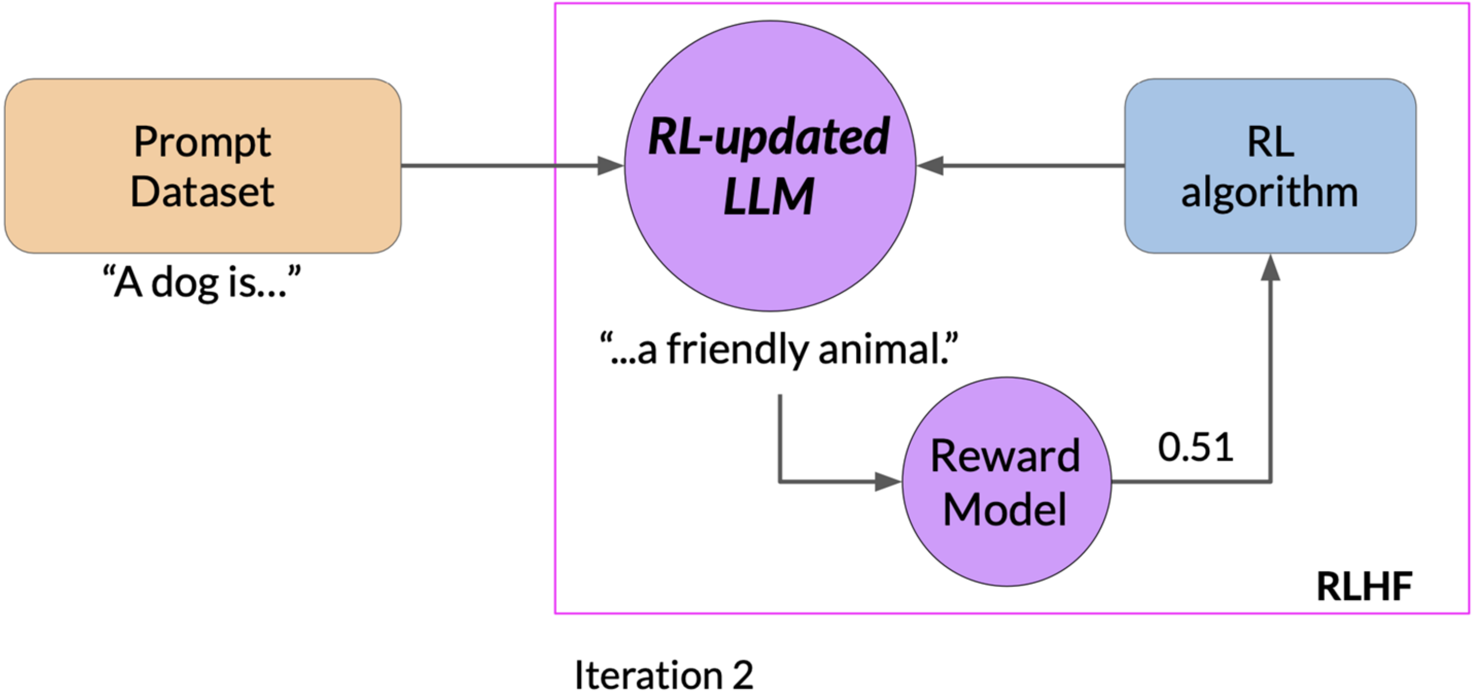
[1] https://www.coursera.org/learn/generative-ai-with-llms/

Using the Reward Model to Fine-tune LLM with RL



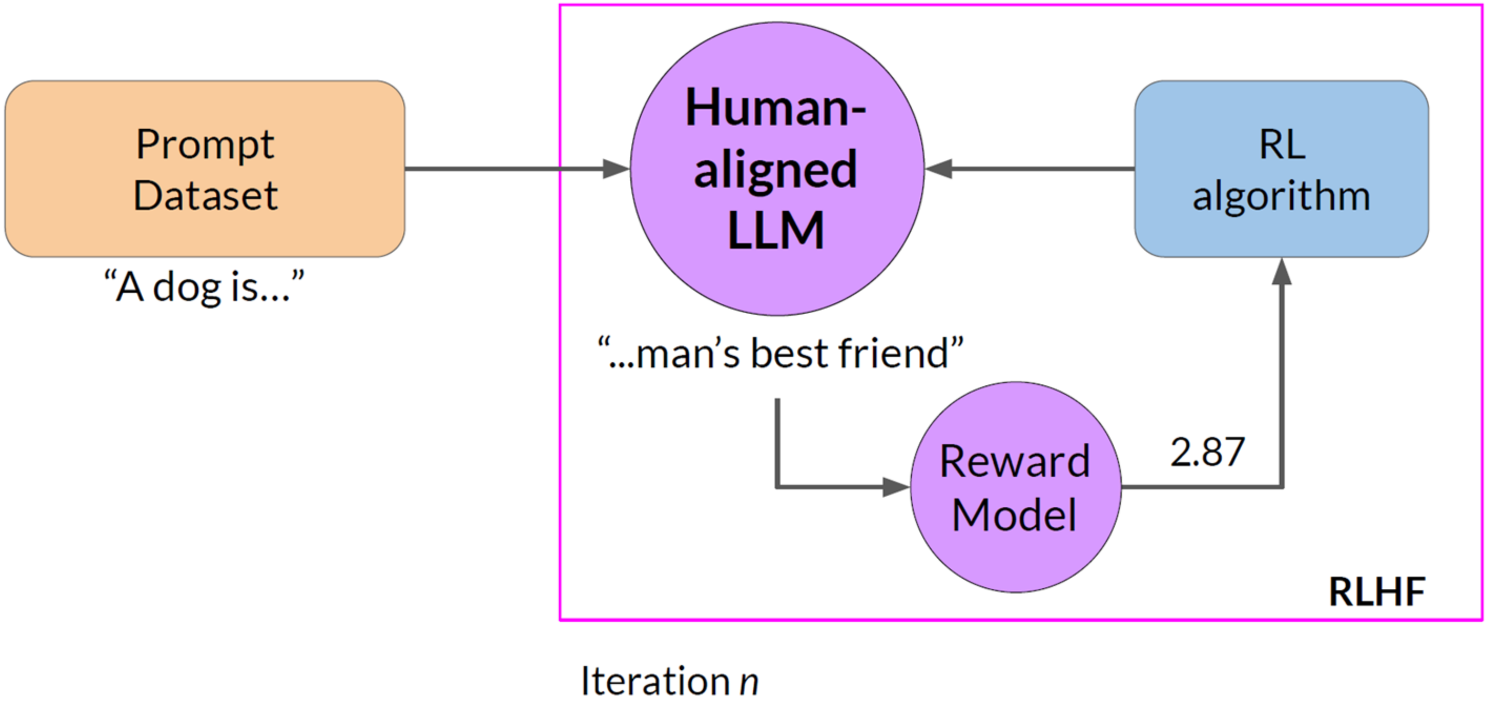
[1] https://www.coursera.org/learn/generative-ai-with-llms/

Using the Reward Model to Fine-tune LLM with RL



[1] https://www.coursera.org/learn/generative-ai-with-llms/

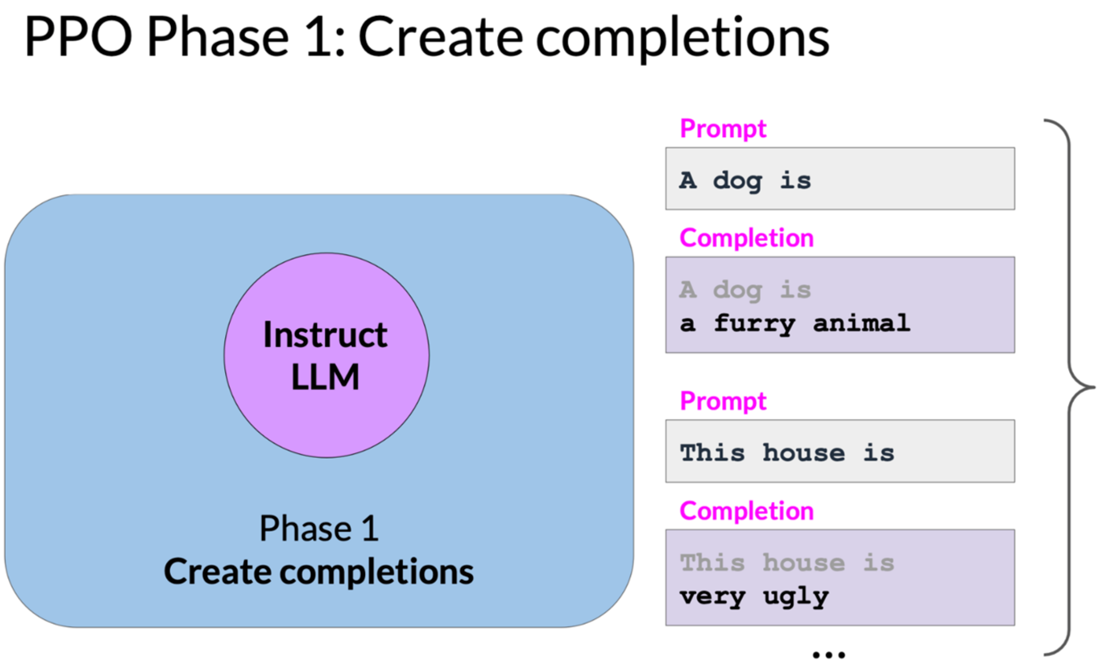
Using the Reward Model to Fine-tune LLM with RL



[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization

PPO Phase 1: Create completions



Prompt

A dog is

Completion

A dog is
a furry animal

Prompt

This house is

Completion

This house is
very ugly

...

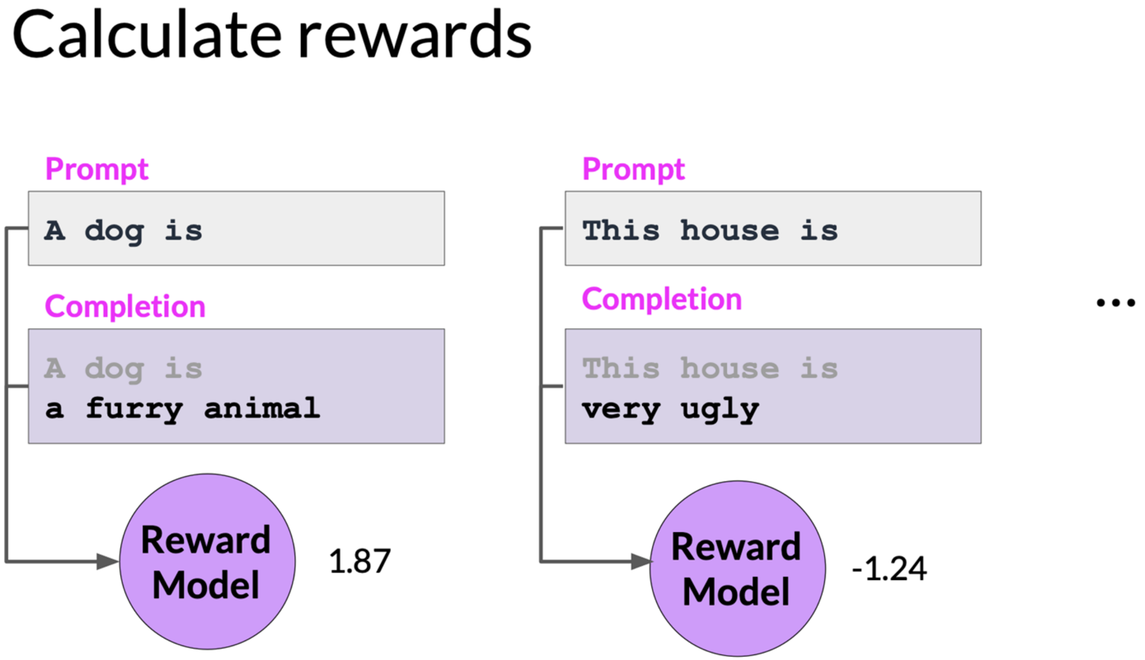
Experiments

to assess the outcome of the current model,

e.g. how helpful, harmless, honest the model is

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization



Calculate rewards

Prompt

A dog is

Completion

A dog is
a furry animal

Reward Model

1.87

Prompt

This house is

Completion

This house is
very ugly

Reward Model

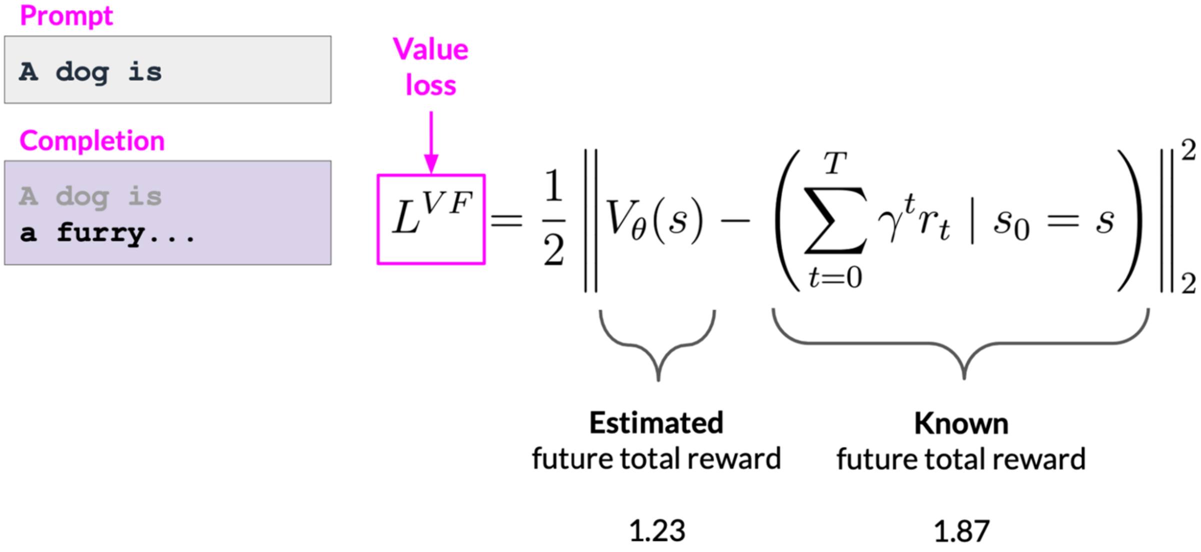
-1.24

...

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization

Calculate value loss



Prompt

```
A dog is
```

Completion

```
A dog is
a furry...
```

Value loss

$$\boxed{L^{VF}} = \frac{1}{2}\left\| \underbrace{V_\theta(s)} - \underbrace{\left(\sum_{t=0}^{T}\gamma^t r_t \mid s_0 = s\right)} \right\|_2^2$$

Estimated future total reward

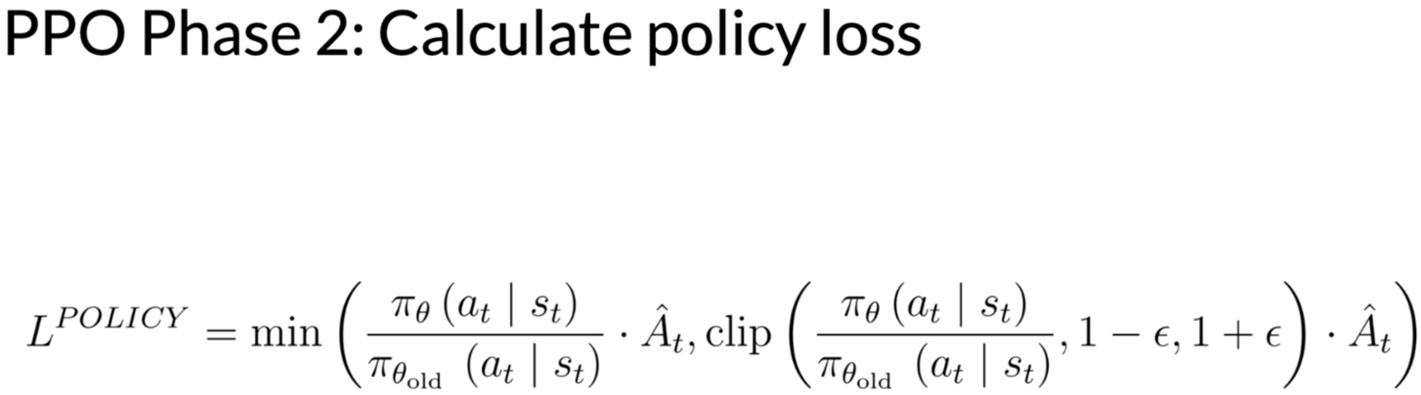
1.23

Known future total reward

1.87

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization

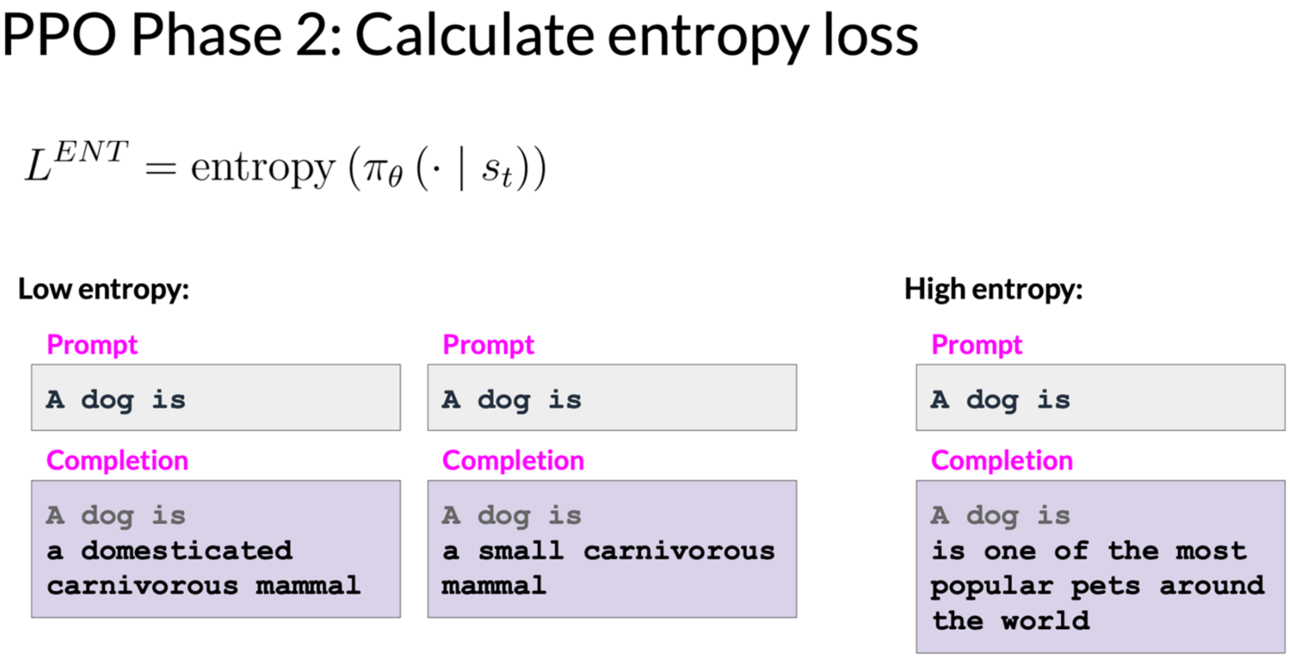


PPO Phase 2: Calculate policy loss

$$L^{POLICY} = \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_{\text{old}}}\left(a_t \mid s_t\right)}, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t\right)$$

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization



PPO Phase 2: Calculate entropy loss

$$L^{ENT} = \text{entropy}\left(\pi_\theta\left(\cdot \mid s_t\right)\right)$$

Low entropy:

Prompt

A dog is

Completion

A dog is
a domesticated carnivorous mammal

Prompt

A dog is

Completion

A dog is
a small carnivorous mammal

High entropy:

Prompt

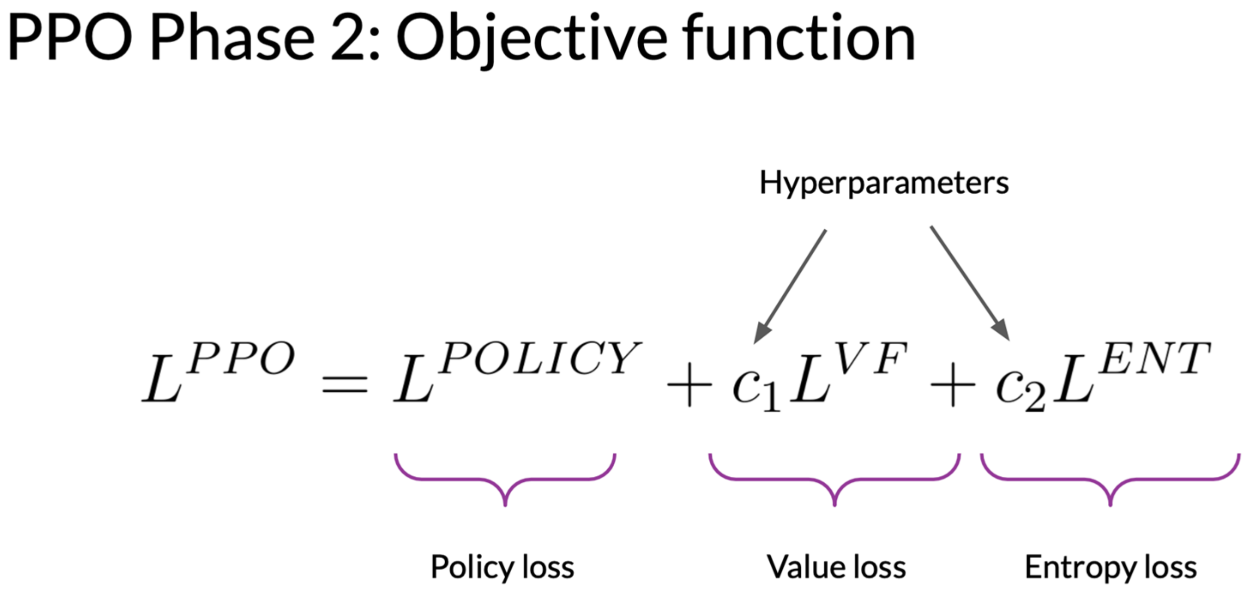
A dog is

Completion

A dog is
is one of the most popular pets around the world

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization



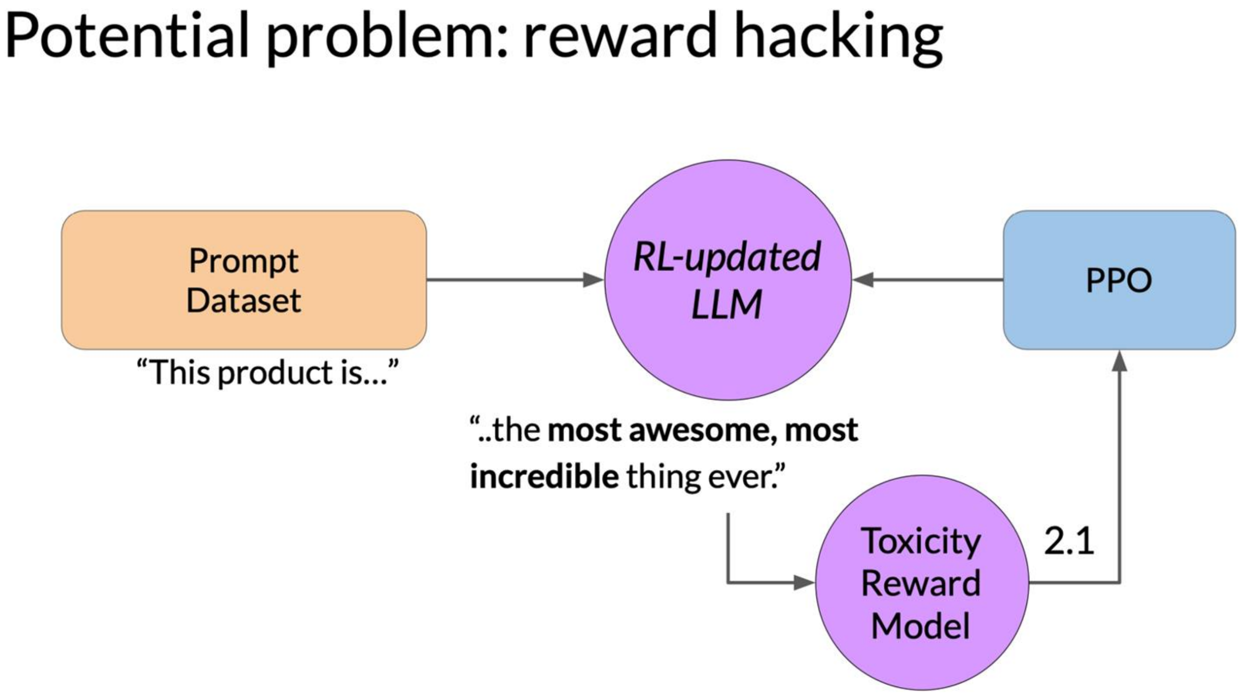
PPO Phase 2: Objective function

Hyperparameters

$$L^{PPO} = \underbrace{L^{POLICY}}_{\text{Policy loss}} + \underbrace{c_1 L^{VF}}_{\text{Value loss}} + \underbrace{c_2 L^{ENT}}_{\text{Entropy loss}}$$

[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization



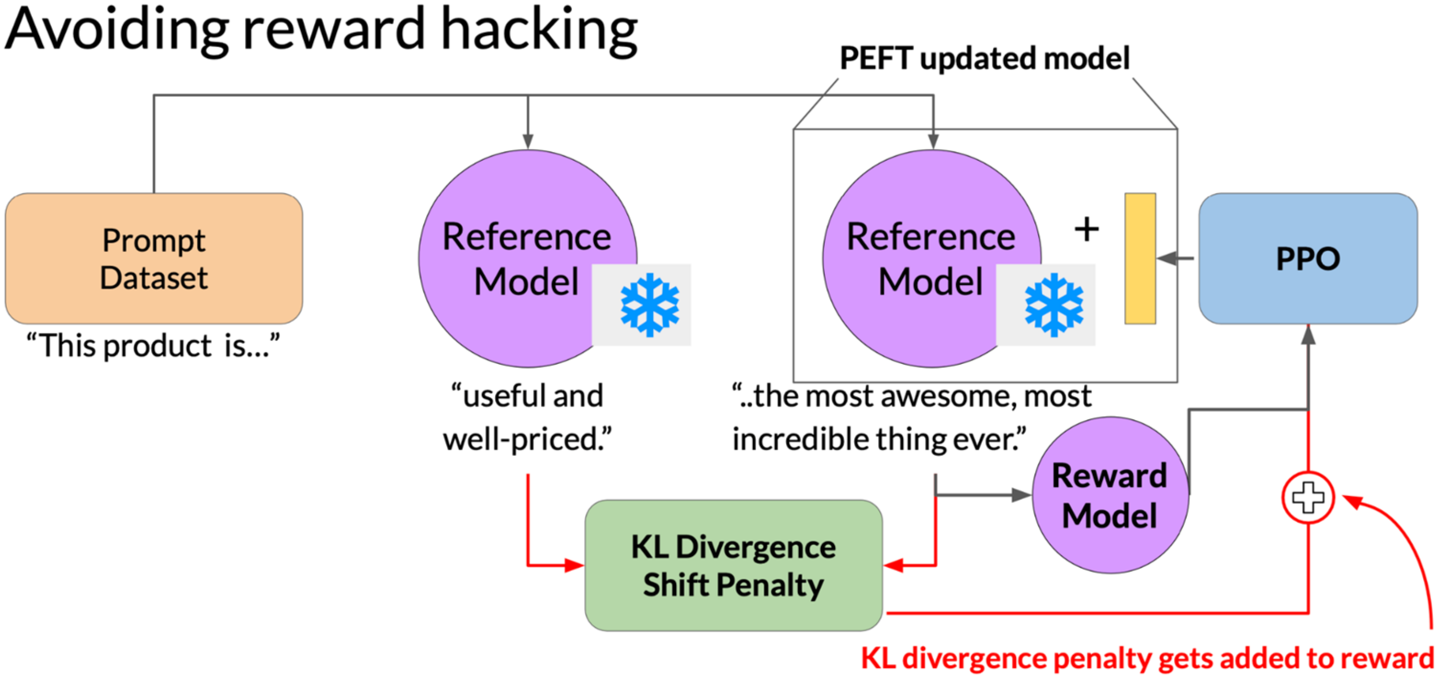
Potential problem: reward hacking

“This product is...”

“..the **most awesome, most incredible** thing ever.”

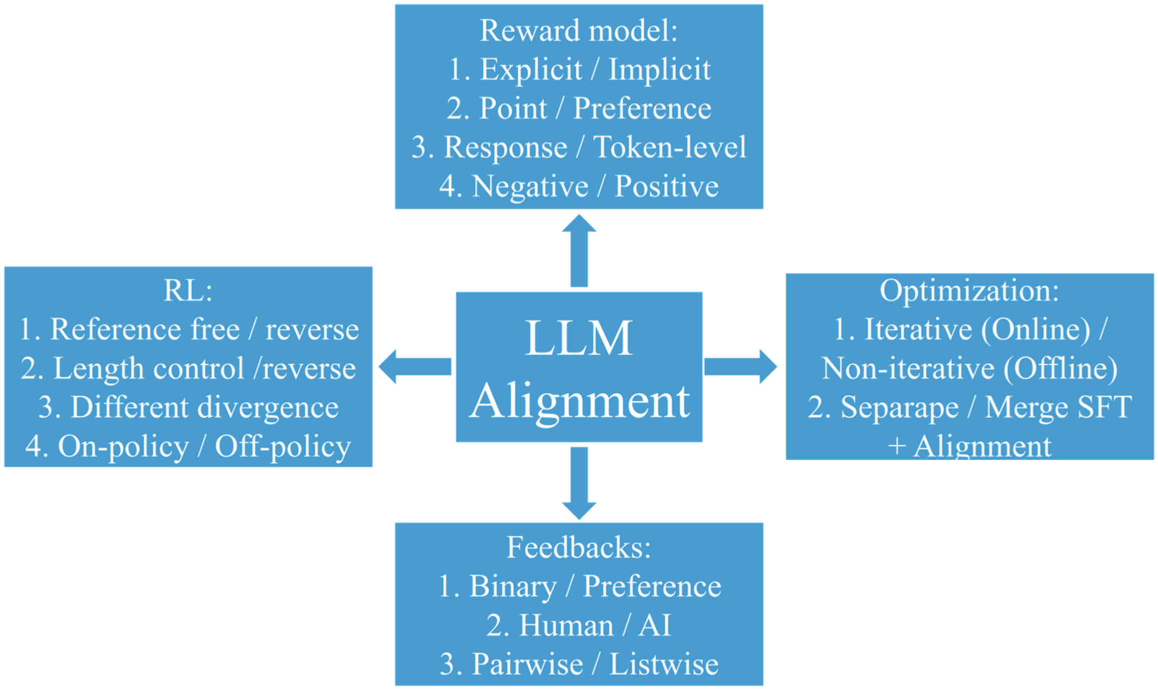
[1] https://www.coursera.org/learn/generative-ai-with-llms/

Proximal Policy Optimization


Avoiding reward hacking
PEFT updated model
Prompt Dataset
"This product is..."
Reference Model
"useful and well-priced."
Reference Model
+
PPO
"..the most awesome, most incredible thing ever."
Reward Model
KL Divergence Shift Penalty
KL divergence penalty gets added to reward

[1] https://www.coursera.org/learn/generative-ai-with-llms/

LLM Alignment Techniques



LLM Alignment

Reward model:
1. Explicit / Implicit
2. Point / Preference
3. Response / Token-level
4. Negative / Positive

RL:
1. Reference free / reverse
2. Length control /reverse
3. Different divergence
4. On-policy / Off-policy

Optimization:
1. Iterative (Online) / Non-iterative (Offline)
2. Separape / Merge SFT + Alignment

Feedbacks:
1. Binary / Preference
2. Human / AI
3. Pairwise / Listwise

[4] Wang, Z., Bi, B., Pentyala, S. K., Ramnath, K., Chaudhuri, S., Mehrotra, S., ... & Asur, S. (2024). A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Thank you!

Reza Fayyazi
rf1679@rit.edu