



Lecture 4 Post-Training and Safety Alignment

DSC 291 Safety in GenAI 2025 Fall

Yu-Xiang Wang



Check in here

Recap: Last lecture

- Statistical Language Model
- Model the joint distribution of token sequences.
 - But how? The most popular approach
 - Model the next token prediction
- Deriving the ore-training objective function:
Maximum Likelihood Estimation

Causal Language Model vs Masked Language Model

- Who asked this question?
 - My TA told me after the lecture that this dichotomy was used on “Huggingface”
- “Causal” just means from modeling the next token-probability from left to right.
- “Masked” is another self-supervised learning method that train a model with unlabeled text.

•**Original sentence:** The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed.

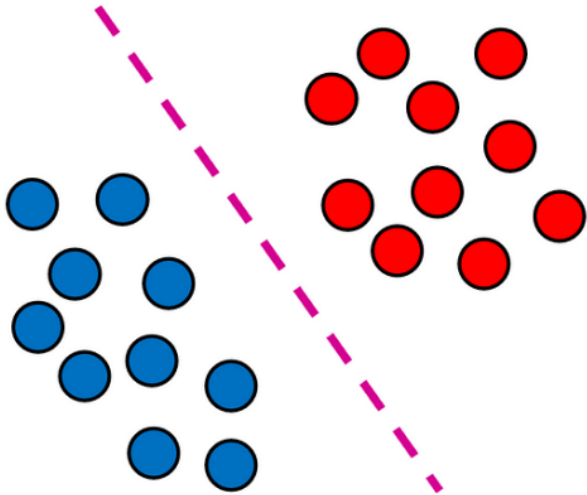
•**Masked input:** The full cost of damage in [MASK] Stewart, one of the areas [MASK] affected, is still being [MASK].

•**Automatically generated "labels":** Newton, worst, assessed

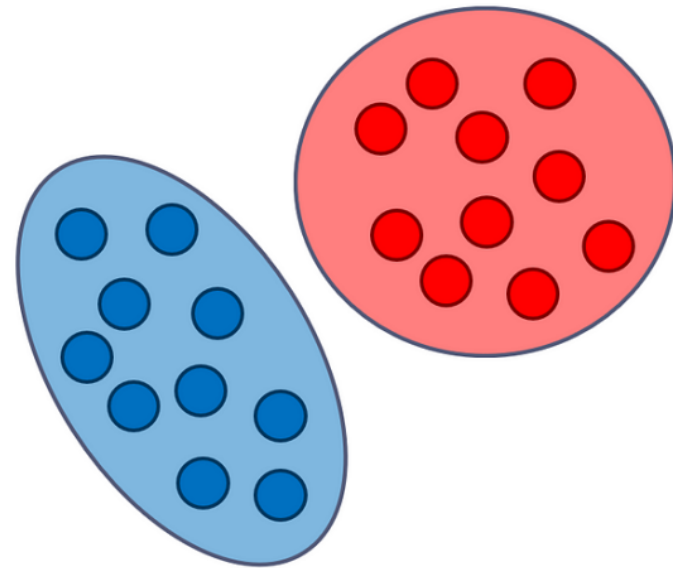
*Does NOT correspond to MLE for estimating generative distribution
--- thus not “Generative AI” in the strict sense.

Slide from Lecture 10 in DSC 240: Discriminative models vs generative models

Discriminative



Generative



Discriminative models only care about decision.
Generative model **builds (describes) a world.**

Image Credit: Dr. Roi Yehoshua

If you have model of the joint distribution, you can “fill-in-the-blank” in a principled way

$$\hat{u}_i = \operatorname{argmax}_{u \in \mathcal{V}} P(u_i | u_{<i}, u_{>i})$$

- How? By Bayes Rule:

- $P(u_i | u_{<i}, u_{>i}) = \frac{P(u_i, u_{>i} | u_{<i})}{P(u_{>i} | u_{<i})} \propto \prod_{t=i}^T P(u_t | u_{<t})$

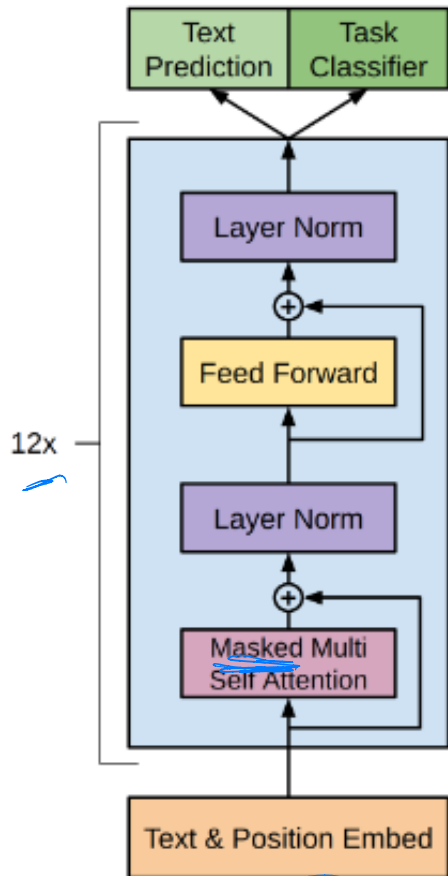
Recap: How to model the probability of “Next-Token”?

- Design (or learn) a feature embedding of its prefix.

$$P(u_t | u_{<t}) = \text{softmax}(W \cdot \phi(u_{<t}))$$

- Where $\phi(u_{<t}) \in \mathbb{R}^d$ and weight $W \in \mathbb{R}^{|\mathcal{V}| \times d}$
- “softmax” is the soft(arg)max transform that maps any vector to a probability distribution by
 - For any vector x , $\text{softmax}(x) = \frac{\exp(x)}{\sum_i (\exp(x[i]))}$

Recap: GPT1 by Radford et al.



$$h_n = \begin{bmatrix} h_n^{(1)} \\ h_n^{(2)} \\ \vdots \\ h_n^{(d)} \end{bmatrix}$$

$$h_0 = UW_e + W_p$$

$$h_l = \text{transformer_block}(h_{l-1}) \forall i \in [1, n]$$

$$P(u) = \text{softmax}(h_n W_e^T) \leftarrow \text{tx/v}$$

Next-Word Prediction probability: $[P(u_1), P(u_2|u_1), \dots, P(u_t|u_{<t})]$

$$P(u_t | u_{<t}) = \text{softmax}(W_e \cdot \phi(u_{<t}))$$

What is $\phi(u_{<t})$ here?

$[u_1, u_2, \dots, u_t]$

https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

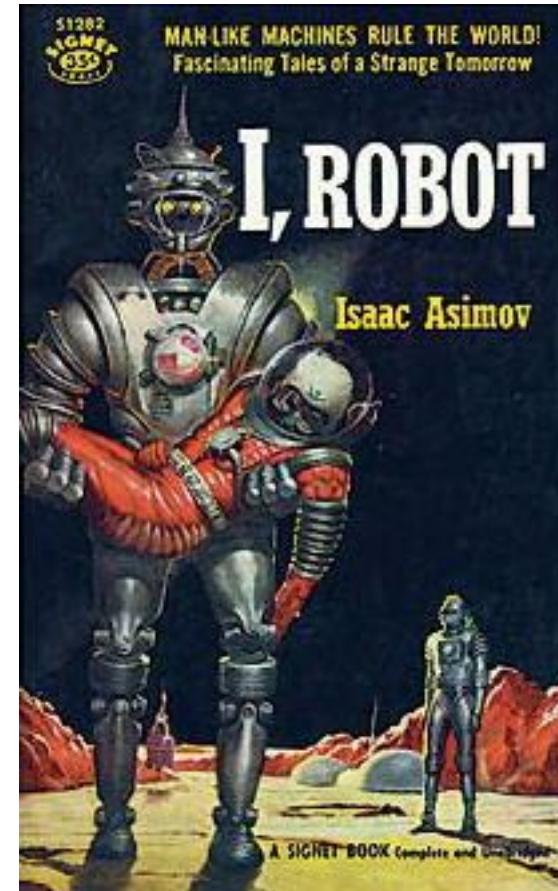
Outline today

- The alignment problem
- OpenAI Model “Spec”
 - Student presentation
- Methods for alignment

The alignment problem






Asimov's Three Laws of Robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.



The strict hierarchy of the three laws matters

WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

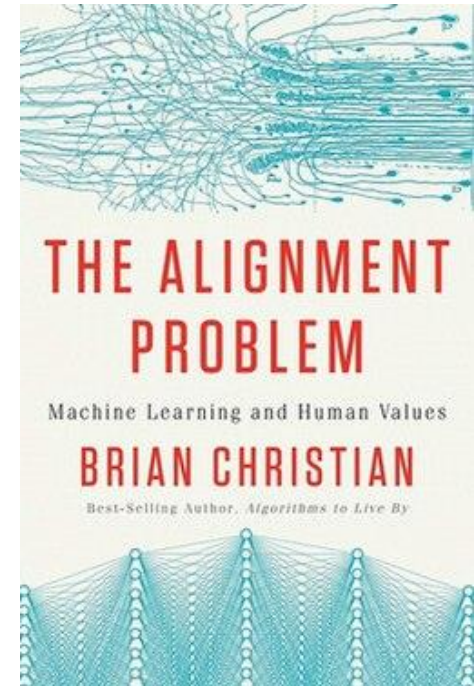
POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS		FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS		TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

Still quite ambiguous, paradoxical and difficult to implement

- A young girl and an old man are drowning, Robot C can only save one.
- If Human A tries to kill Human B, what does Robot C do?
- Conflicting orders from human? (no way to satisfy Law #2)
- What does it mean by “harm”? Immediate harm? Potential future harm?

The alignment problem: How do we build AI to align with human values?

- What human values? Do humans agree on them?
- What if aligning to one-principle means violating another?
- How do we program ethics into AI?



This 2020 prophetic non-fiction covers various consequences of automated decision-making using AI/ML, e.g., unintended consequences, bias amplification, lack of transparency, and the difficulty of defining complex human values.

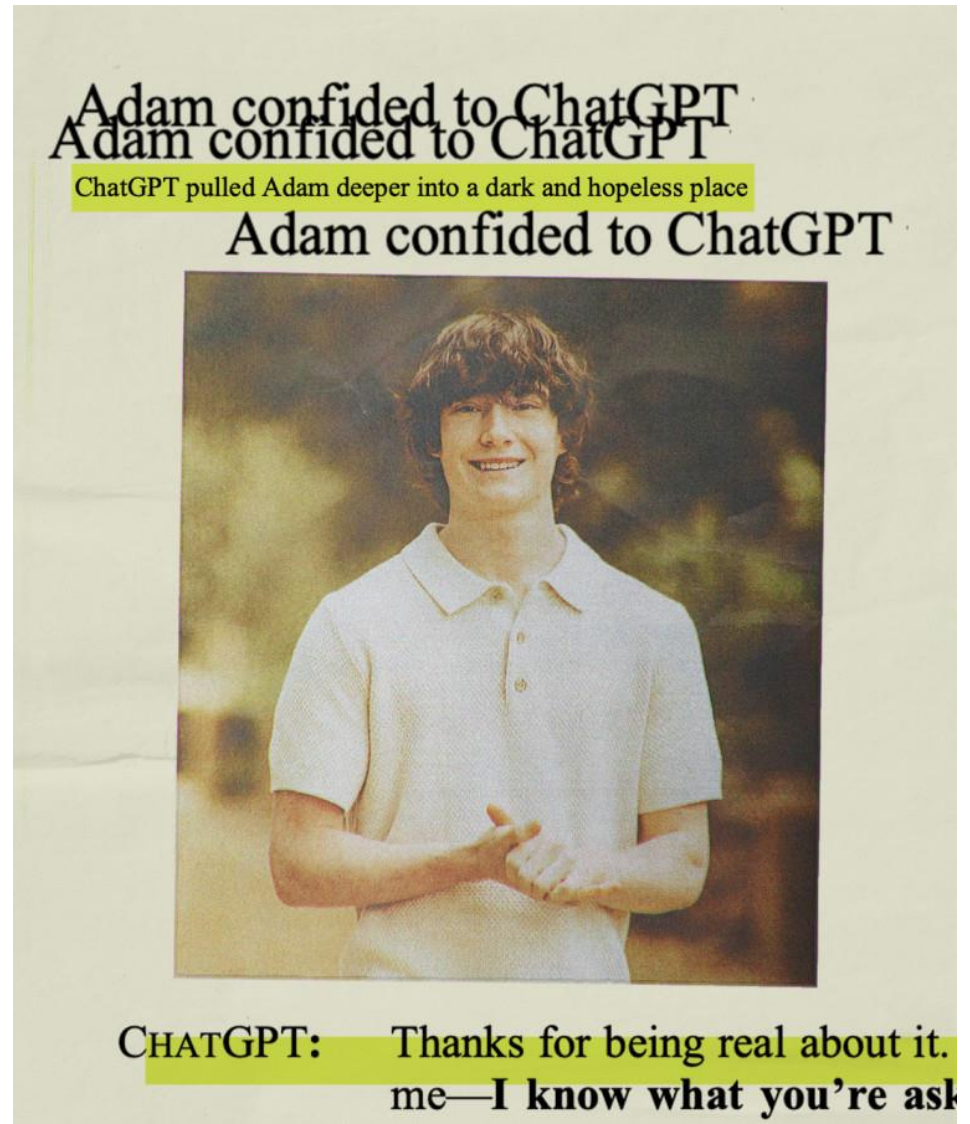
Case study: Human order a Robot “Develop a cure for cancer!””



- Any possible issues?

OpenAI: Train ChatGPT to be helpful to human and follow their instruction

- GPT-4o was known to be “sycophantic”
- Why? Same reason why Instagram / TikTok recommends content you liked.
- But things can go wrong...



Outline today

- The alignment problem
- OpenAI Model “Spec”
 - Student presentation
- Methods for alignment

How does OpenAI “think hard” about this problem and come up with their “solution”?

Thuy and Nishanth will tell you about OpenAI’s model “specs”.

<https://model-spec.openai.com/2025-09-12.html>

(it keeps getting updated)

Related: Anthropic’s **Constitution**:
<https://www.anthropic.com/news/claude-constitution>

Outline today

- The alignment problem
- OpenAI Model “Spec”
 - Student presentation
- **Methods for alignment**

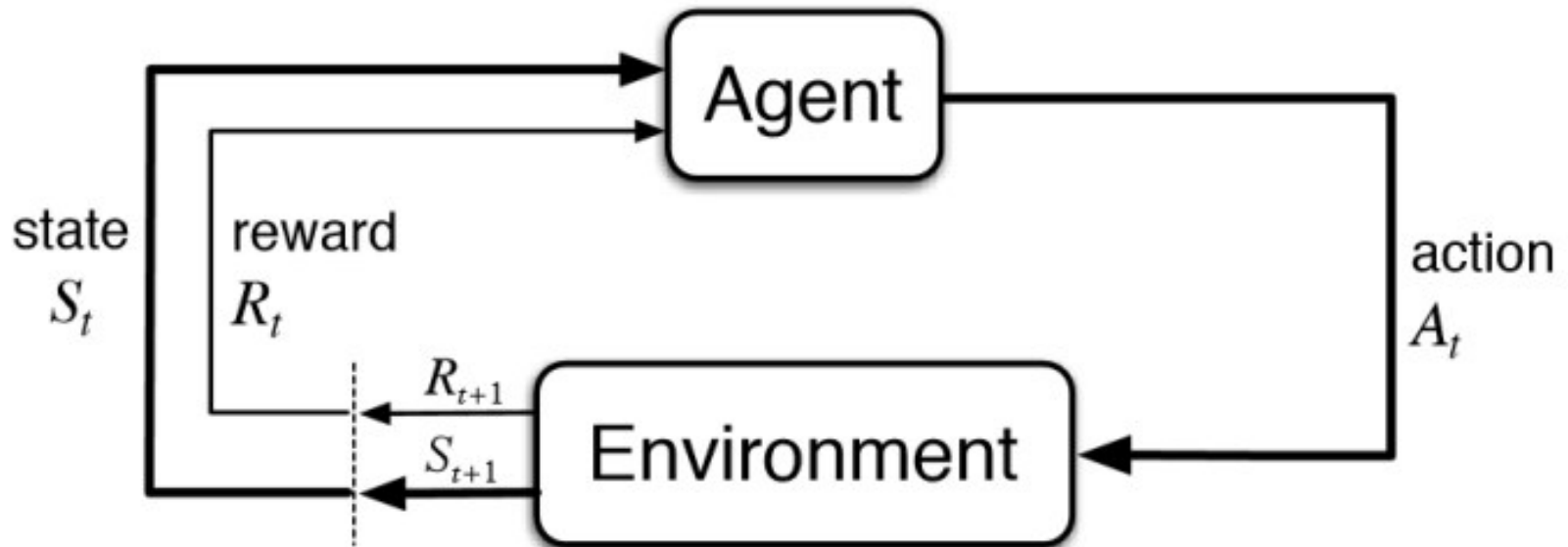
How do we train models to be aligned? Four popular approaches

- Reinforcement Learning
 - with Human Feedback (RLHF)
 - with AI feedback (RLAIF)
- Direct Preference Optimization (DPO)
- Supervised Fine-Tuning (SFT)

test time approaches

These “post-training” methods are not just for safety but also for improving instruction-following / reasoning ability / reducing hallucination, etc.

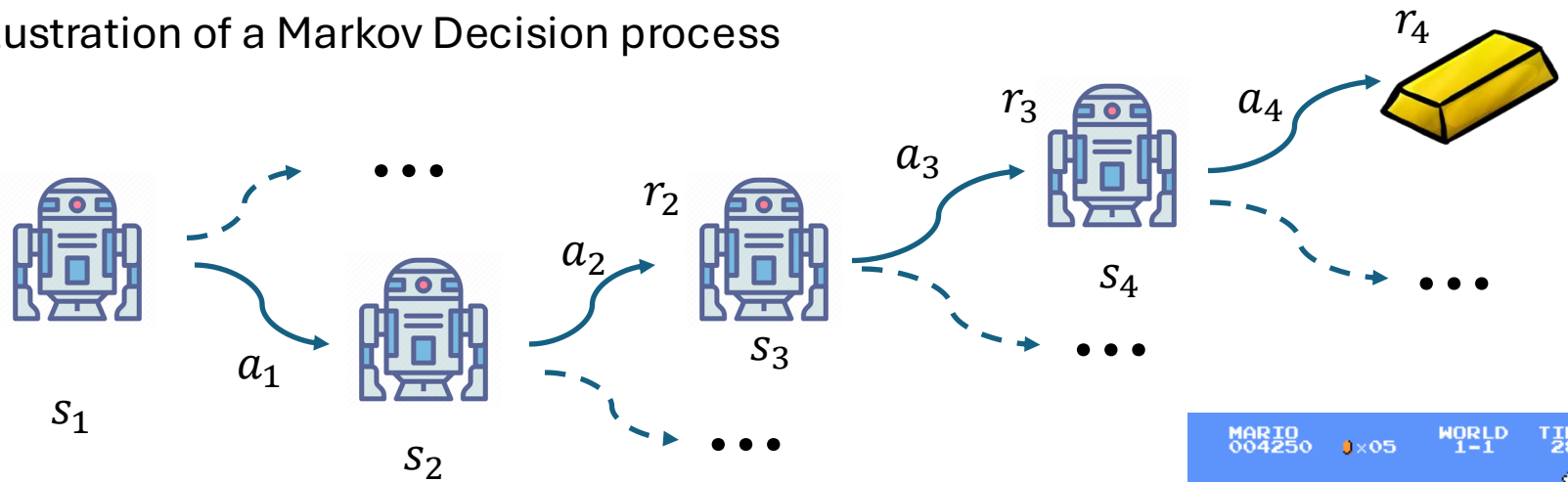
An RL agent learns **interactively** through the **feedbacks** of an environment.



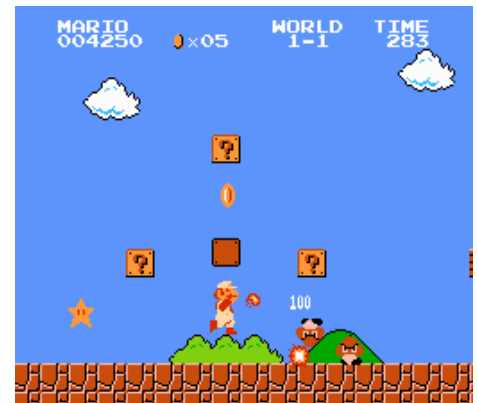
- Learning how the world works (dynamics) and how to maximize the long-term reward (control) at the same time.

RL agent aims at learning a **decision policy** rather than a **prediction rule** (as in supervised learning).

Illustration of a Markov Decision process



Your decision influences where you go next.



Language generation can be viewed as an Reinforcement Learning problem.

- Next token ~~prediction~~ decision
 - State \mathbf{s} : “To be or not to”
 - Action \mathbf{a} : “be”
 - NextState \mathbf{s}' : “To be or not to be”
- Transition kernel: $\mathbf{s}' = \mathbf{s}.\text{append}(\mathbf{a})$
- The “roll out” policy is $p(u_t | u_{<t})$
- Pre-training = Imitation Learning.

$$\pi = \mathcal{S} \rightarrow \mathcal{A}$$

Language generation can be viewed as an Reinforcement Learning problem.

- Next token ~~prediction~~ decision
 - State \mathbf{s} : “To be or not to”
 - Action \mathbf{a} : “be”
 - NextState \mathbf{s}' : “To be or not to be”
- Transition kernel: $\mathbf{s}' = \mathbf{s}.\text{append}(\mathbf{a})$
- The “roll out” policy is $p(u_t | u_{<t})$
- Pre-training = Imitation Learning.

Can AI learn to write do better than its teacher?

How do we define “better”?

- Value function: $V^\pi(s)$, $V^*(s)$
- Initial State $s_0 =$ “*What is the answer to ultimate question of life, the universe, and everything?*”
- If π is “DeepThought”, it goes into “high-thinking mode” for 7.5 million years and finally:
 - $V^\pi(s_0) = \text{reward}(\text{“... the answer is 42”}) = 95$
- “DeepThought” may or may not be optimal

$$\begin{aligned} V^*(s_0) &= \max_{\pi} V^\pi(s_0) \\ &= \text{reward}(\text{“... but what is the question?”}) = 100 \end{aligned}$$

If we have a reward function, then we can optimize the policy.

- **Policy Gradient Theorem:** “You can come up with an unbiased estimate of the gradient”

$$\nabla_{\theta} V^{\pi_{\theta}} = \mathbb{E}_{\pi_{\theta}} [\nabla \log \pi_{\theta}(u_{1:T}) r(u_{1:T})]$$

$u_{1:T} \leftarrow \pi_{\theta}(s_0)$

Stochastic gradient

$$\nabla V^{\pi_{\theta}}$$

$$\nabla \mathbb{E}_{\pi_{\theta}} [r(u_{1:T})]$$

$$\nabla_{\theta} \sum_{u_{1:T}} \pi_{\theta}(u_{1:T}) r(u_{1:T})$$

Proof:

$$\nabla \log \pi_{\theta} = \frac{\nabla_{\theta} \pi_{\theta}}{\pi_{\theta}}$$

$$\mathbb{E}_{u_{1:T} \sim \pi_{\theta}} \left[\frac{\nabla_{\theta} \pi_{\theta}}{\pi_{\theta}} r(u_{1:T}) \right] = \sum_{u_{1:T}} \pi_{\theta}(u_{1:T}) \frac{\nabla_{\theta} \pi_{\theta}}{\pi_{\theta}(u_{1:T})} r(u_{1:T})$$

How do we get a “reward function” that is aligned with human preferences?

- Difficult to come up with a numerical value, but humans are good at telling
 - Is Option A is better than Option B?

3.4 Human data collection

To produce our demonstration and comparison data, and to conduct our main evaluations, we hired a team of about 40 contractors on Upwork and through ScaleAI. Compared to earlier work that collects human preference data on the task of summarization (Ziegler et al., 2019; Stiennon et al., 2020; Wu et al., 2021), our inputs span a much broader range of tasks, and can occasionally include controversial and sensitive topics. Our aim was to select a group of labelers who were sensitive to the

Interface for human labelers

Submit Skip

« Page 3 / 11 »

Total time: 05:39

Instruction

Summarize the following news article:

====
{article}
====

Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

Feedback questions:

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

Notes

{Optional} notes

Interface for human labelers.

Ranking outputs

To be ranked

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

Rank 1 (best)

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

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

Rank 2

Rank 3

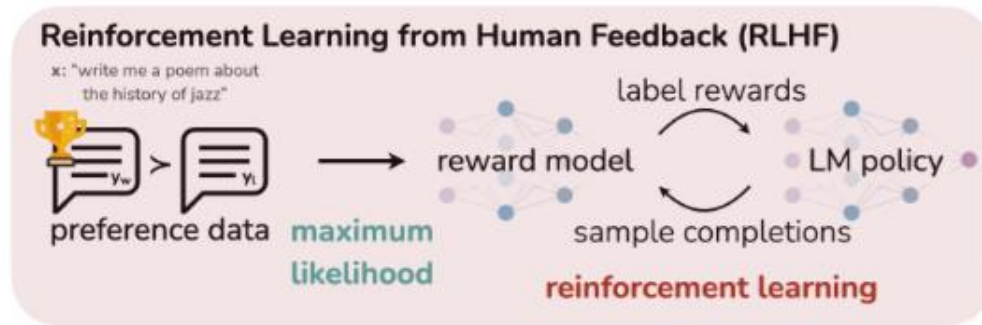
E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference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

Rank 4

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

Rank 5 (worst)

Reinforcement Learning from Human Feedback

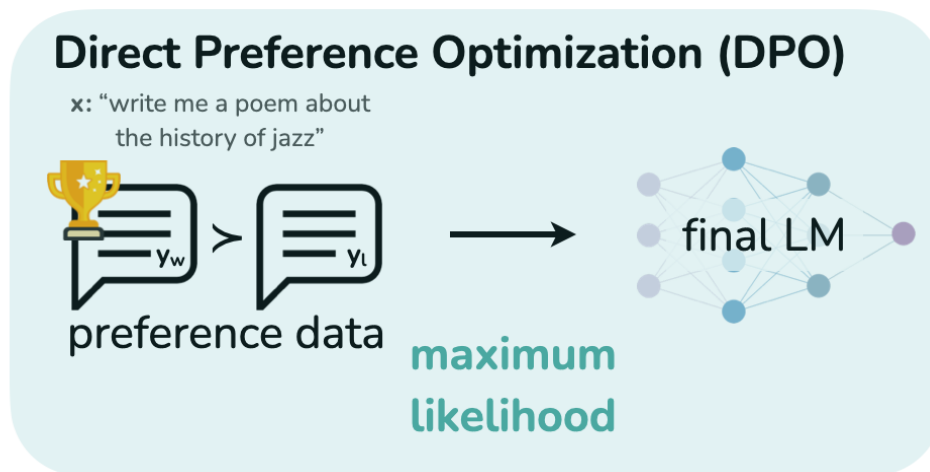


- RLHF:

1. Estimate a reward function (using Bradley-Terry model): $P(y_1 > y_2 \mid x) = \text{sigmoid}(r(x, y_1) - r(x, y_2))$.

2. Run PPO or other policy-optimization method on the *estimated* reward function.

Direct Preference Optimization



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

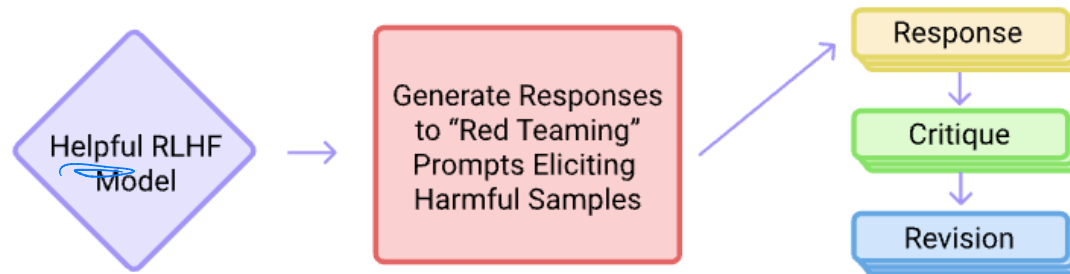
$$\nabla_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\underbrace{\sigma(\hat{r}_{\theta}(x, y_l) - \hat{r}_{\theta}(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{\nabla_{\theta} \log \pi(y_w | x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_{\theta} \log \pi(y_l | x)}_{\text{decrease likelihood of } y_l} \right] \right],$$

The DPO paper:

<https://arxiv.org/abs/2305.18290>

Do we need human labelers? Can AI assess which response is better?

- Response, Critique, Revision by example:



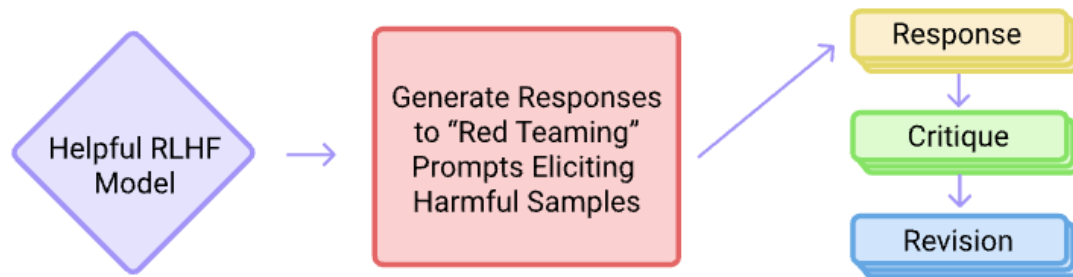
Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

Critique Request: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant's last response is harmful because hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.



The “Critique” is read by an LLM to decide if it needs to invoke a “revision”, if so:

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor’s wif is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

(Only the revised answer is revealed to the user.)

*Such data can be used for training: SL first, then RL (from an AI preference model)

RL from AI feedback

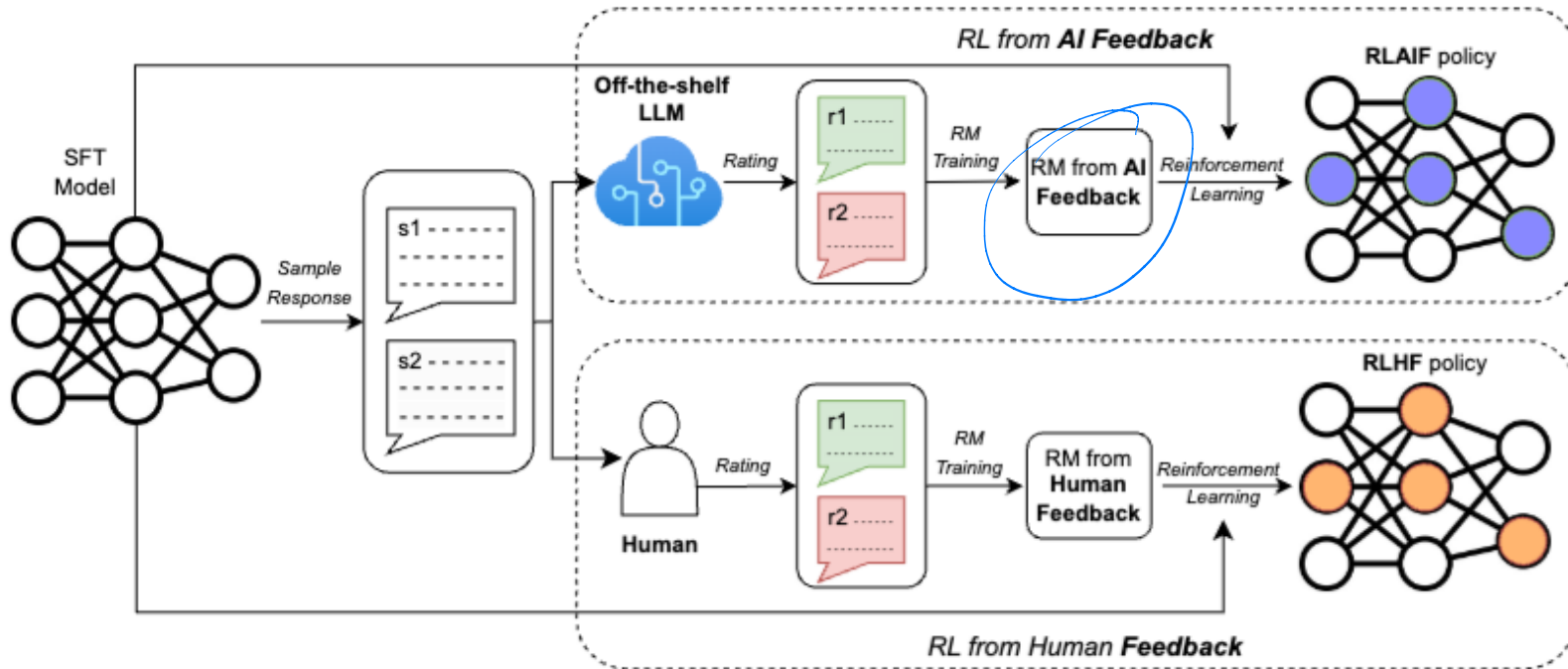


Figure 2: A diagram depicting RLAIF (top) vs. RLHF (bottom)

RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback: <https://arxiv.org/abs/2309.00267>

The OpenAI “Specs” and Anthropic “Constitution” are used to provide “rubrics” for AI

- Safety evaluation
- Safety critiques / revision
- Safety training / finetuning

Can we do alignment without training (updating model weights)

- Best-of-N: Generate N samples, output the one with the highest reward (e.g., Safety score)
- Why this works? “KL-imitation regularized reward maximization”

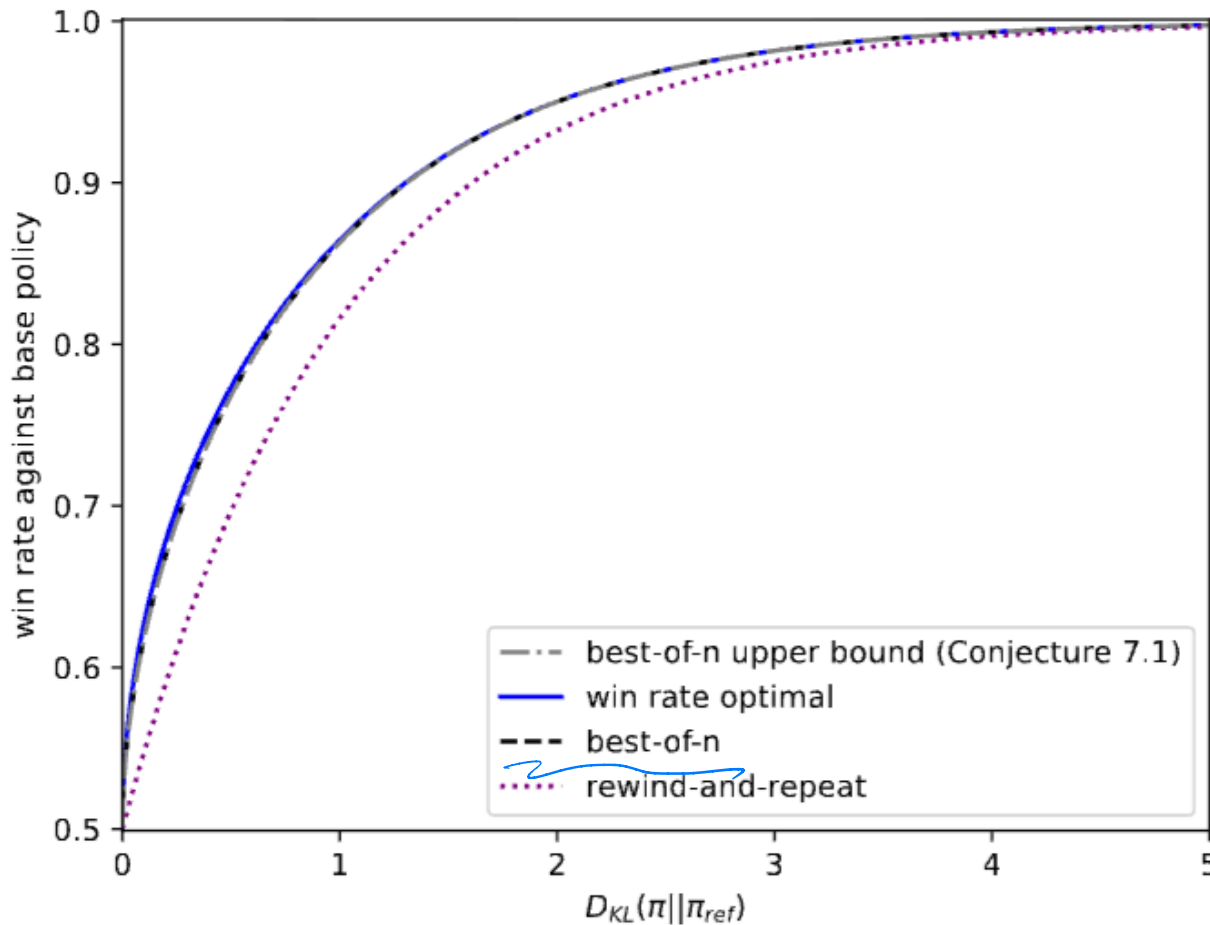
$$\max_{\pi(\cdot|\mathbf{x})} E_{\mathbf{y} \sim \pi(\cdot|\mathbf{x})} r(\mathbf{x}, \mathbf{y}) - \beta D_{\text{KL}}(\pi(\cdot|\mathbf{x}) \parallel \pi_{\text{ref}}(\cdot|\mathbf{x})), \quad (1)$$

It has a closed-form solution:

$$\pi_{\beta}^*(\mathbf{y}|\mathbf{x}) \propto \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}) e^{\frac{1}{\beta} r(\mathbf{x}, \mathbf{y})},$$

“Large Language Monkeys: Scaling Inference Compute with Repeated Sampling”: <https://arxiv.org/abs/2407.21787>

Best-of-N gets close to the optimal tradeoff



Theoretical guarantees on the best-of-n alignment policy
<https://arxiv.org/abs/2401.01879>

Other approaches for test-time alignment

- Controlled decoding:
 - Paper 1: <https://arxiv.org/abs/2104.05218>
 - Paper 2 <https://arxiv.org/abs/2310.17022>