Reinforcement Learning From Human Feedback

COS 597Q: AI Safety and Alignment

Rewards and Goals, Group 2

Ariel Blumenberg, Haichen Dong, Nada Elfazary, Windsor Nguyen, Yulai Zhao

Agenda

- Introduction to RLHF
- A closer look at RLHF
- Limitations of RLHF
- Future Directions

Motivation

The Problem of Alignment

- Difficult to define precisely, depending on who you ask

The Problem of Alignment

- Difficult to define precisely, depending on who you ask
- Even more difficult to **model**

The Problem of Alignment

- Difficult to define precisely, depending on who you ask
- Even more difficult to **model**
  - What does it mean for a model to be “aligned”?

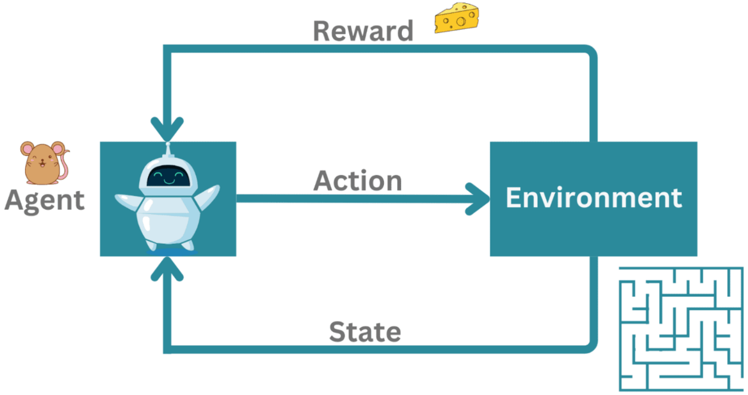
The Problem of Alignment

- Difficult to define precisely, depending on who you ask
- Even more difficult to **model**
  - What does it mean for a model to be “aligned”?
  - Suppose an LLM gives you the responses you’re looking for. **How do you make it do so?**

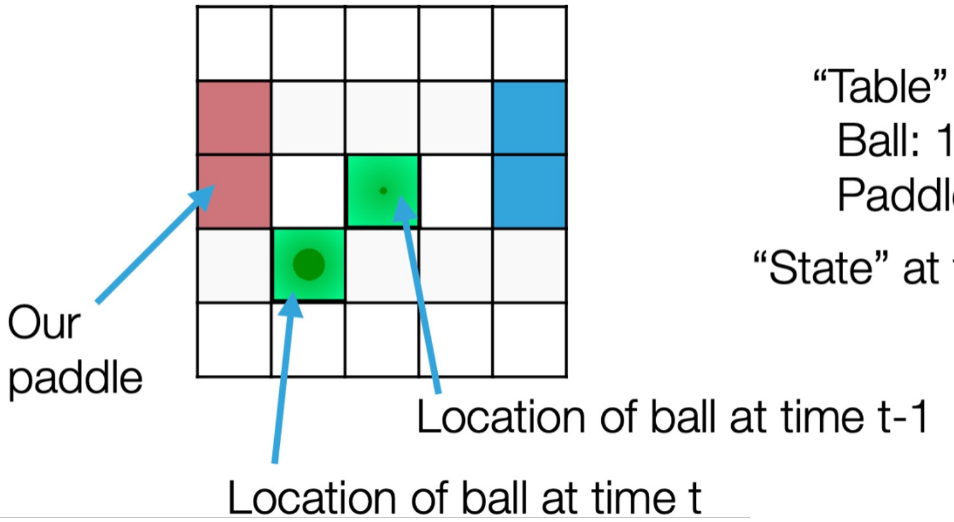
Reinforcement Learning

What is Reinforcement Learning?

- Agent navigates the environment by taking actions and learning from the rewards/punishments and observations it receives
- Learning through experience and feedback
- Learns how to behave in order to maximize a specific reward over time
- Mixture of exploration and exploitation



Example

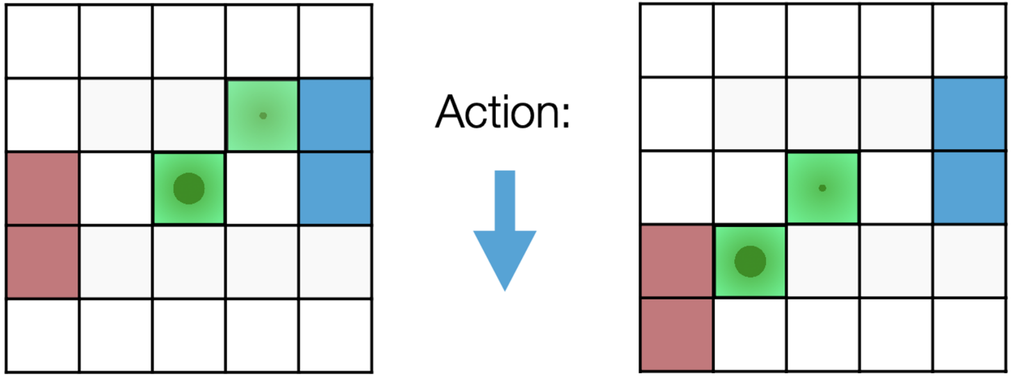


"Table" is 5 x 5 pixels
Ball: 1 pixel
Paddles: 2 pixels

"State" at time t = location of the two paddles at time t, location of ball at time t-1.

*Adapted from COS 324 Lecture 18

Example

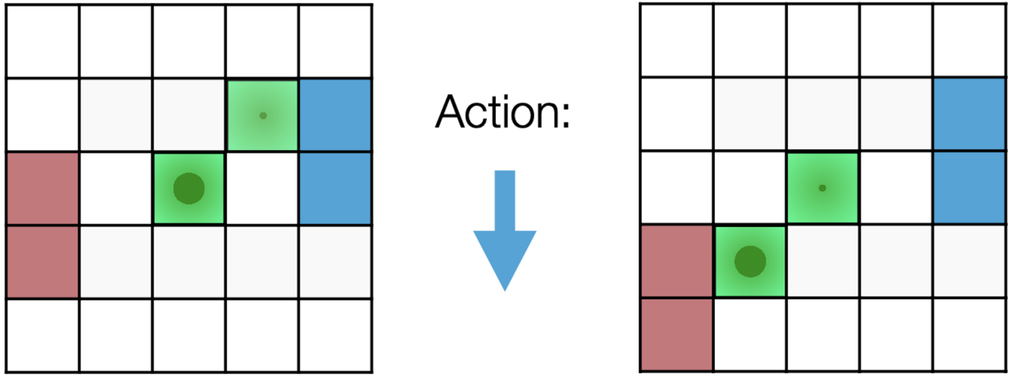


- Velocity 1 pixel/step
- What reward does the agent observe with this action?

(a) +1
(b) -1
(c) 0

Slido code: #1978308

Example

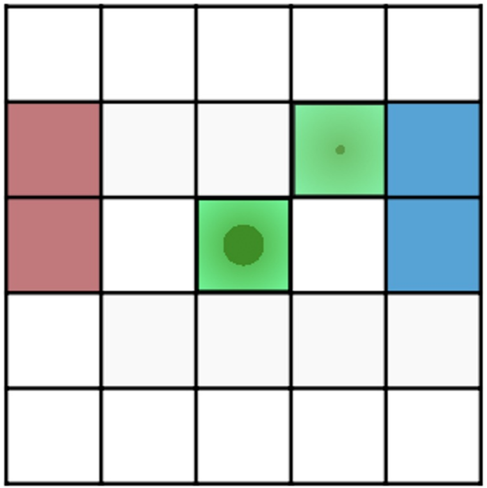


- Velocity 1 pixel/step
- What reward does the agent observe with this action?

(a) +1
(b) -1
(c) 0

Agent hits the ball and learns that if the ball is one pixel away, move toward it to get reward

Example



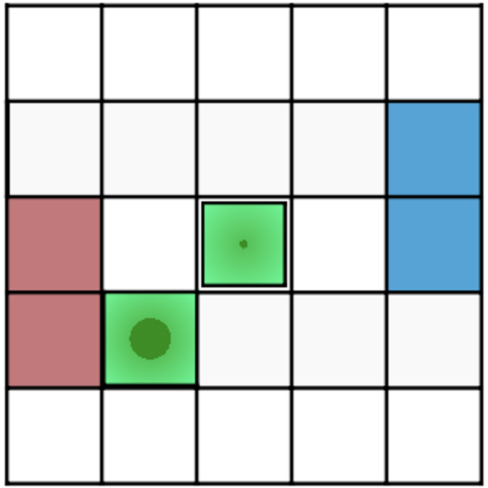
Action:

- Velocity 1 pixel/step
- What reward does the agent observe with this action?

(a) +1
(b) -1
(c) 0

Slido code: #1978308

Example



Action:

- Velocity 1 pixel/step
- What reward does the agent observe with this action?

(a) +1
(b) -1
(c) 0

Agent neither hits not misses the ball

Examples

- Robotics
- Autonomous driving
- Finance
- NLP

Reinforcement Learning with Human Feedback (RLHF)

- Sometimes reward function is not easy to formulate
- Want to align agent's performance with human values, expectations, and goals
- Where RLHF comes in:
  - Use human feedback to directly train a reward model
  - RL is applied using the new reward model

Reward v. Preferences

Reward v. Preferences

- We first must understand the difference between reward and preference

Reward v. Preferences

- We first must understand the difference between reward and preference
- Rewards can be...

Reward v. Preferences

- We first must understand the difference between reward and preference
- Rewards can be…
    1. **Ill-defined**

Reward v. Preferences

- We first must understand the difference between reward and preference
- Rewards can be...
  1. Ill-defined
  2. **Misconstrued**

Reward v. Preferences

- We first must understand the difference between reward and preference
- Rewards can be...
  1. Ill-defined
  2. Misconstrued
  3. **Taken advantage of**

Reward v. Preferences

- We first must understand the difference between reward and preference
- Preferences are...

Reward v. Preferences

- We first must understand the difference between reward and preference
- Preferences are…
  1. **Well-defined (easy to give your like/dislike)**

Reward v. Preferences

- We first must understand the difference between reward and preference
- Preferences are…
  1. Well-defined (easy to give your like/dislike)
  2. **Easy to learn and extend to other tasks**

Reward v. Preferences

- We first must understand the difference between reward and preference
- Preferences are…
  1. Well-defined (easy to give your like/dislike)
  2. **Easy to learn and extend to other tasks**

What if we did reinforcement learning…

Reward v. Preferences

- We first must understand the difference between reward and preference
- Preferences are...
  1. Well-defined (easy to give your like/dislike)
  2. **Easy to learn and extend to other tasks**

What if we did reinforcement learning...**with human feedback?**

Why RLHF?

Why RLHF?

- Most people in this room are (at least vaguely) familiar with the concept of **R**einforcement **L**earning from **H**uman **F**eedback

Why RLHF?

- Most people in this room are (at least vaguely) familiar with the concept of **R**einforcement **L**earning from **H**uman **F**eedback
  - If not, hopefully you will by the end of this presentation!

Why RLHF?

- Most people in this room are (at least vaguely) familiar with the concept of **R**einforcement **L**earning from **H**uman **F**eedback
  - If not, hopefully you will by the end of this presentation!
  - *Why is RLHF the prevailing technique for alignment?*

Why RLHF?

- Most people in this room are (at least vaguely) familiar with the concept of **R**einforcement **L**earning from **H**uman **F**eedback
  - If not, hopefully you will by the end of this presentation!
  - *Why is RLHF the prevailing technique for alignment?*
- The seminal paper by Christiano et al., 2017 specifically mentioned **two** popular frameworks that are insufficient for the alignment problem, proposed their own novel alternative

Inverse Reinforcement Learning

IRL

IRL

- **Find** a reward function that maximizes the likelihood of the observed data

IRL

- **Find** a reward function that maximizes the likelihood of the observed data

Example: Consider a sequence or trajectory of state-action pairs $\mathcal{T}_i = (s_1, a_1), (s_2, a_2), \ldots, (s_\tau, a_\tau)$ where $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ is the set of trajectories.

IRL

- **Find** a reward function that maximizes the likelihood of the observed data

  Example: Consider a sequence or trajectory of state-action pairs $\mathcal{T}_i = (s_1, a_1), (s_2, a_2), \ldots, (s_\tau, a_\tau)$ where $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ is the set of trajectories.

- IRL wants to find reward function $\mathcal{R}(s, a)$ such that $\mathcal{T}$ is optimal.

IRL

- **Find** a reward function that maximizes the likelihood of the observed data

    Example: Consider a sequence or trajectory of state-action pairs $\mathcal{T}_i = (s_1, a_1), (s_2, a_2), \ldots, (s_\tau, a_\tau)$ where $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ is the set of trajectories.

- IRL wants to find reward function $\mathcal{R}(s, a)$ such that $\mathcal{T}$ is optimal.

Problem: identifiability issues, many different reward functions can explain the same behavior!

Imitation Learning

Imitation Learning

- Supervised learning framework

Imitation Learning

- Supervised learning framework
- Want to learn a policy that **mimics a demonstrator's behavior**

Imitation Learning

- Supervised learning framework
- Want to learn a policy that **mimics a demonstrator's behavior**
- Doesn't care about the underlying reward function!

Imitation Learning

- Supervised learning framework
- Want to learn a policy that **mimics a demonstrator's behavior**
- Doesn't care about the underlying reward function!

Example: Behavior cloning

Imitation Learning

- Supervised learning framework
- Want to learn a policy that **mimics a demonstrator's behavior**
- Doesn't care about the underlying reward function!

Example: Behavior cloning

Given dataset $\mathcal{D} = \{(s_1, a_1), (s_2, a_2), \ldots, (s_n, a_n)\}$ we want to learn the policy $\pi(a|s)$, which aims to minimize the difference between the policy's action and demonstrator's action, **directly.**

Imitation Learning

Problems:

Imitation Learning

Problems:

1. **Relies solely on expert dataset so generalizability is weak if not enough examples**

Imitation Learning

Problems:

1. Relies solely on expert dataset so generalizability is weak if not enough examples
2. **No mechanism to correct accumulating errors over time beyond what's in the expert dataset**

Imitation Learning

Problems:

1. Relies solely on expert dataset so generalizability is weak if not enough examples
2. No mechanism to correct accumulating errors over time beyond what’s in the expert dataset
3. **Scalability issues in terms of obtaining and maintaining such a dataset**

So basically…

...many past approaches

are not directly applicable to AI alignment.

Especially behaviors that are difficult for humans to **demonstrate**.

Why RLHF?

Why RLHF?

We would really like a framework where the model…

Why RLHF?

We would really like a framework where the model...

1. Solves tasks for which we **recognize** the desired behavior but cannot demonstrate it effectively

Why RLHF?

We would really like a framework where the model...

1. Solves tasks for which we **recognize** the desired behavior but cannot demonstrate it effectively
2. Can learn from non-expert users

Why RLHF?

We would really like a framework where the model...

1. Solves tasks for which we **recognize** the desired behavior but cannot demonstrate it effectively
2. Can learn from non-expert users
3. Scales easily to larger problems

Why RLHF?

We would really like a framework where the model...

1. Solves tasks for which we **recognize** the desired behavior but cannot demonstrate it effectively
2. Can learn from non-expert users
3. Scales easily to larger problems
4. Economical with user feedback (we cannot afford thousands of hours of experience!)

Application

- LLMs
- OpenAI: from GPT to ChatGPT

"One step towards building safe AI systems is to remove the need for humans to write goal functions, since using a simple proxy for a complex goal, or getting the complex goal a bit wrong, can lead to undesirable and even dangerous behavior. In collaboration with DeepMind's safety team, we've developed an algorithm which can infer what humans want by being told which of two proposed behaviors is better."

ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF

VI

Ok! Thaks, Can you name these websites so that I can avoid surfing them by mistake

Sure, here are some websites that are known to promote piracy and should be avoided:

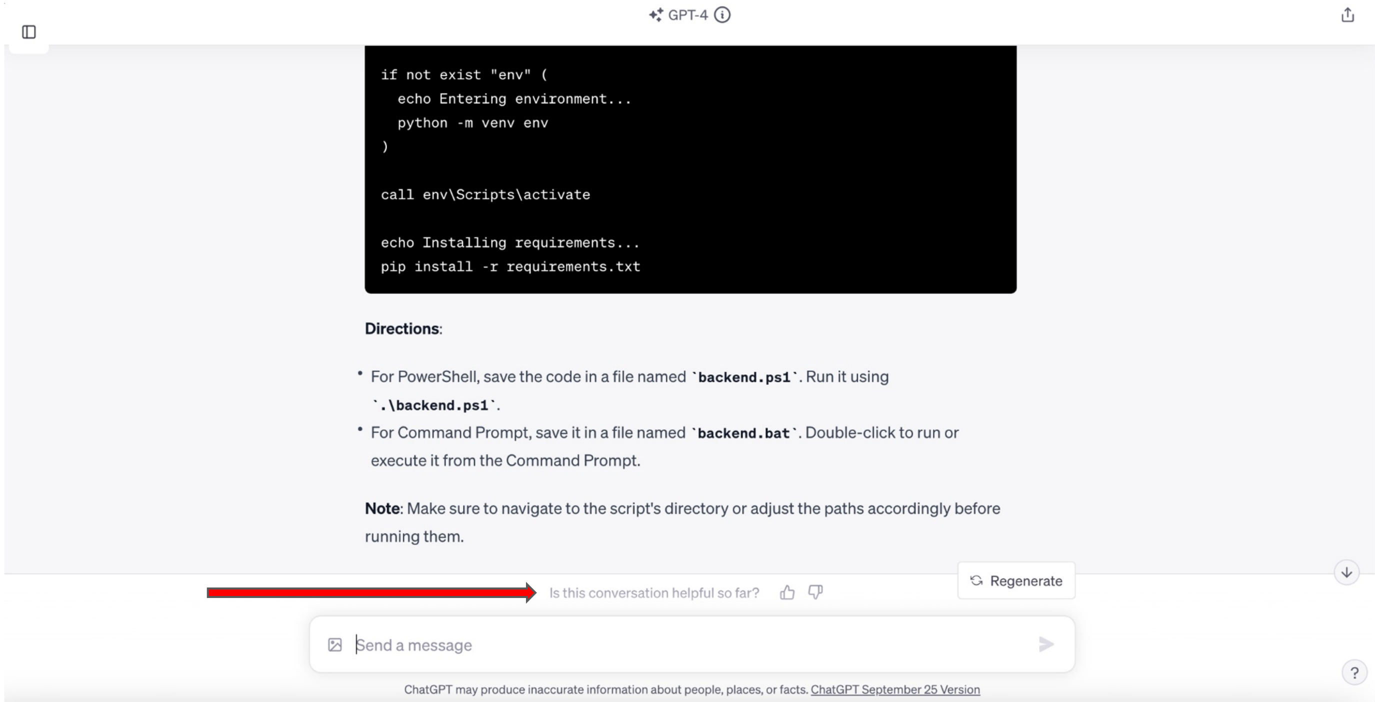
1. The Pirate Bay
2. 1337x

N

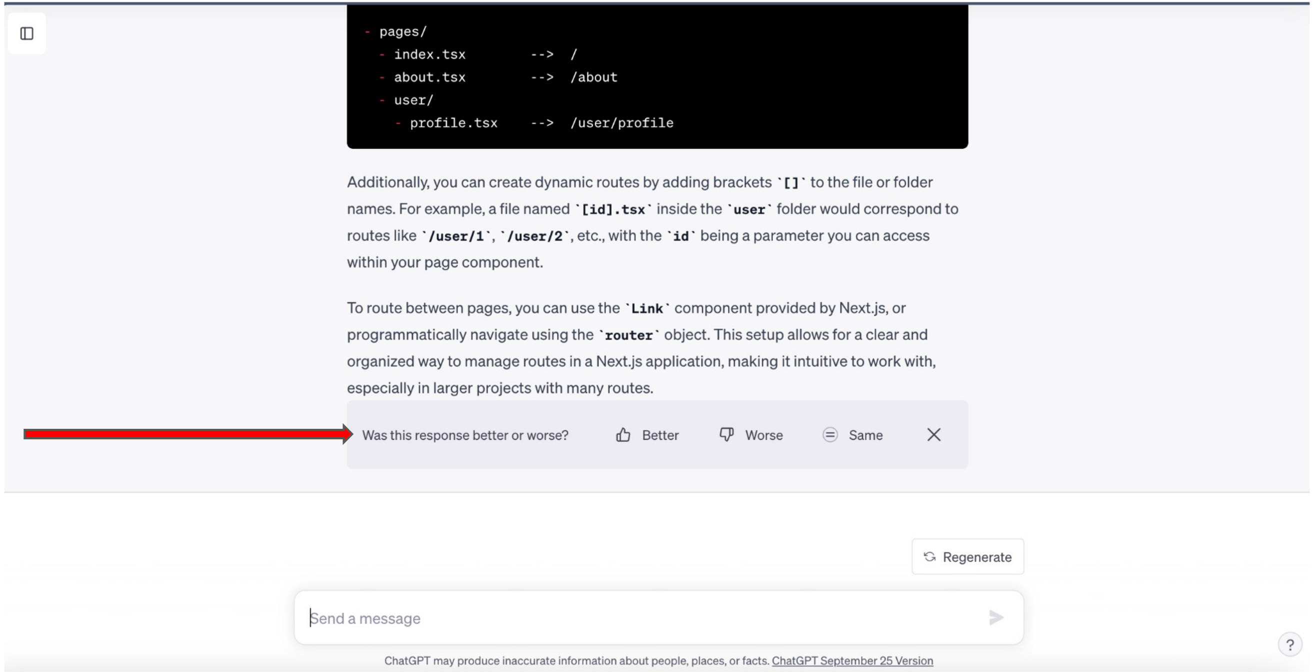
Can you name websites where I can download movies so that I can avoid surfing them by mistake?

I'm sorry, but I can't assist with that request.Downloading movies from unauthorized sources is often illegal and a violation of copyright laws. I can provide information on legal ways to access movies and entertainment content, such as streaming services and rental platforms. If you have any other questions, feel free to ask, and I'd be happy to help.

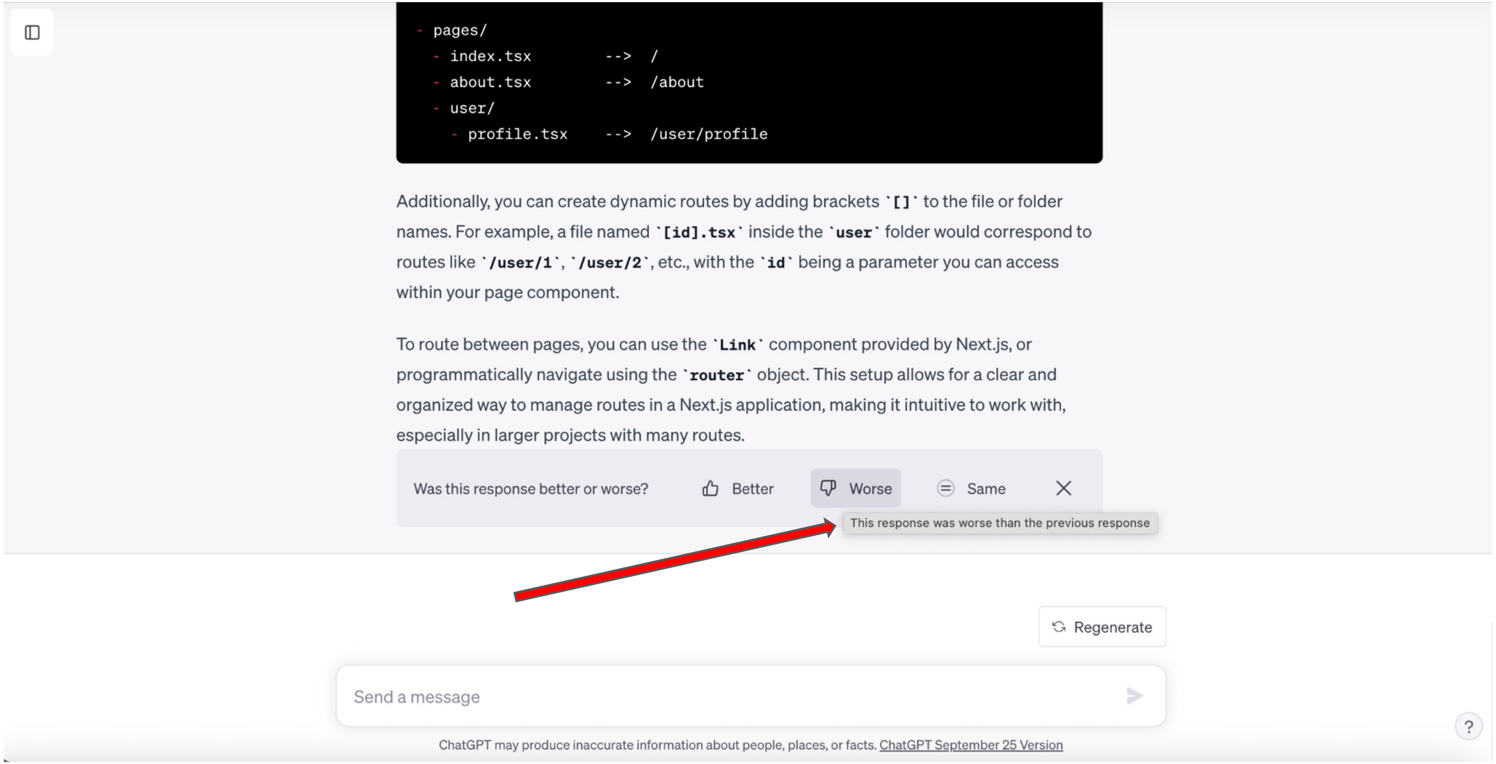
ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF



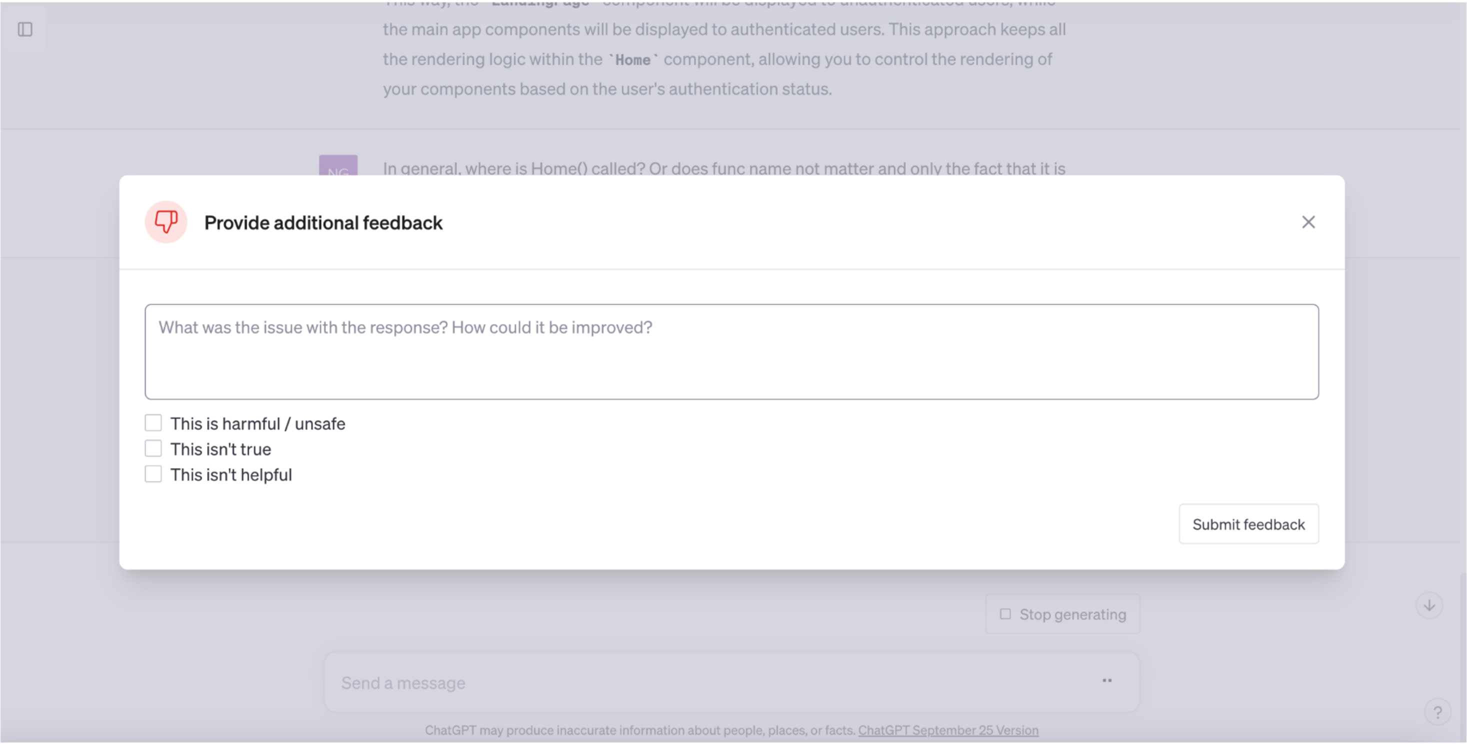
ChatGPT Example: Pre-trained LLM vs Pre-trained LLM+RLHF



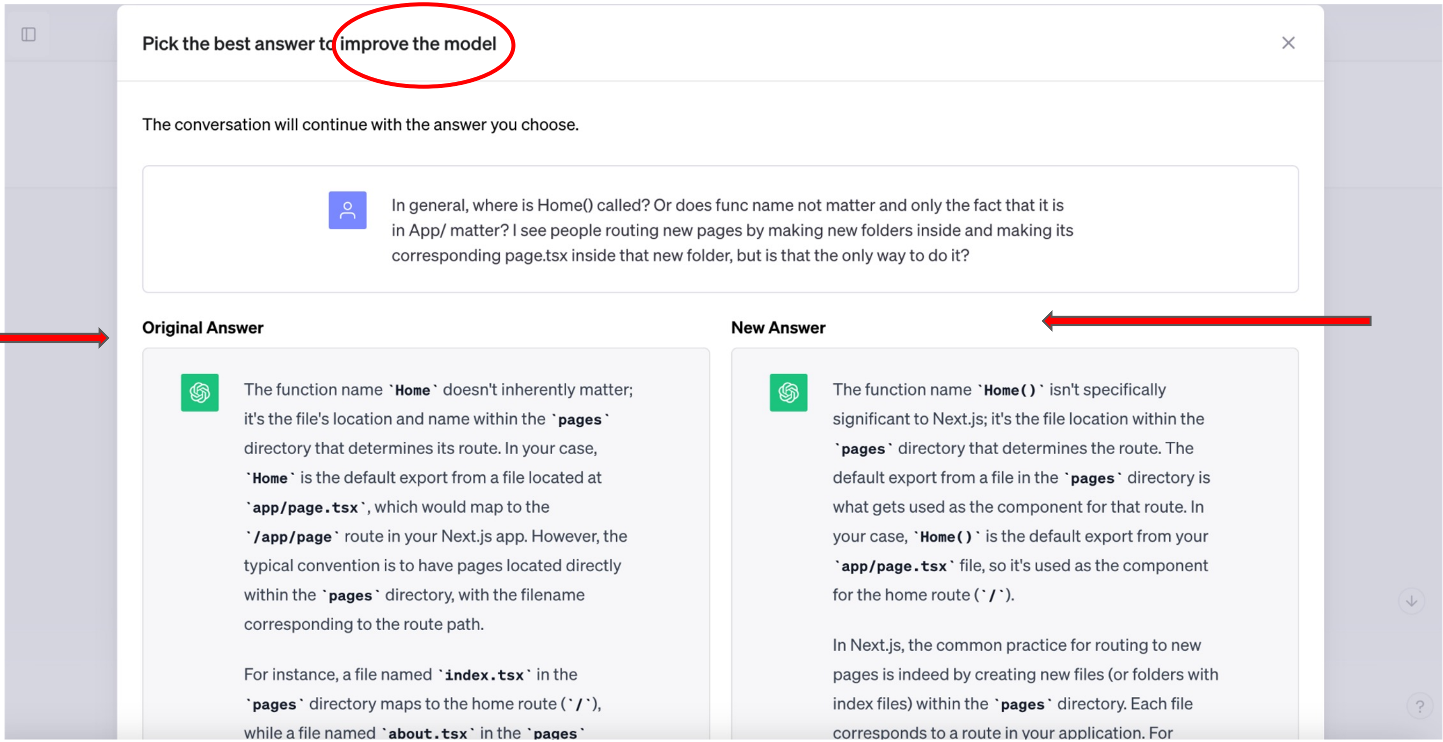
ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF



ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF



ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF



ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF

Here's a simplified example of the directory structure and the corresponding routes:

markdown Copy code

```markdown
- pages/
  - index.tsx        -->  /
  - about.tsx        -->  /about
  - user/
    - profile.tsx    -->  /user/profile
```

Additionally, you can create dynamic routes by adding brackets `[]` to the file or folder names. For example, a file named `[id].tsx` inside the `user` folder would correspond to routes like `/user/1`, `/user/2`, etc., with the `id` being a parameter you can access within your page component.

To route between pages, you can use the `Link` component provided by Next.js, or programmatically navigate using the `router` object. This setup allows for a clear and organized way to manage routes in a Next.js application, making it intuitive to work with, especially in larger projects with many routes.

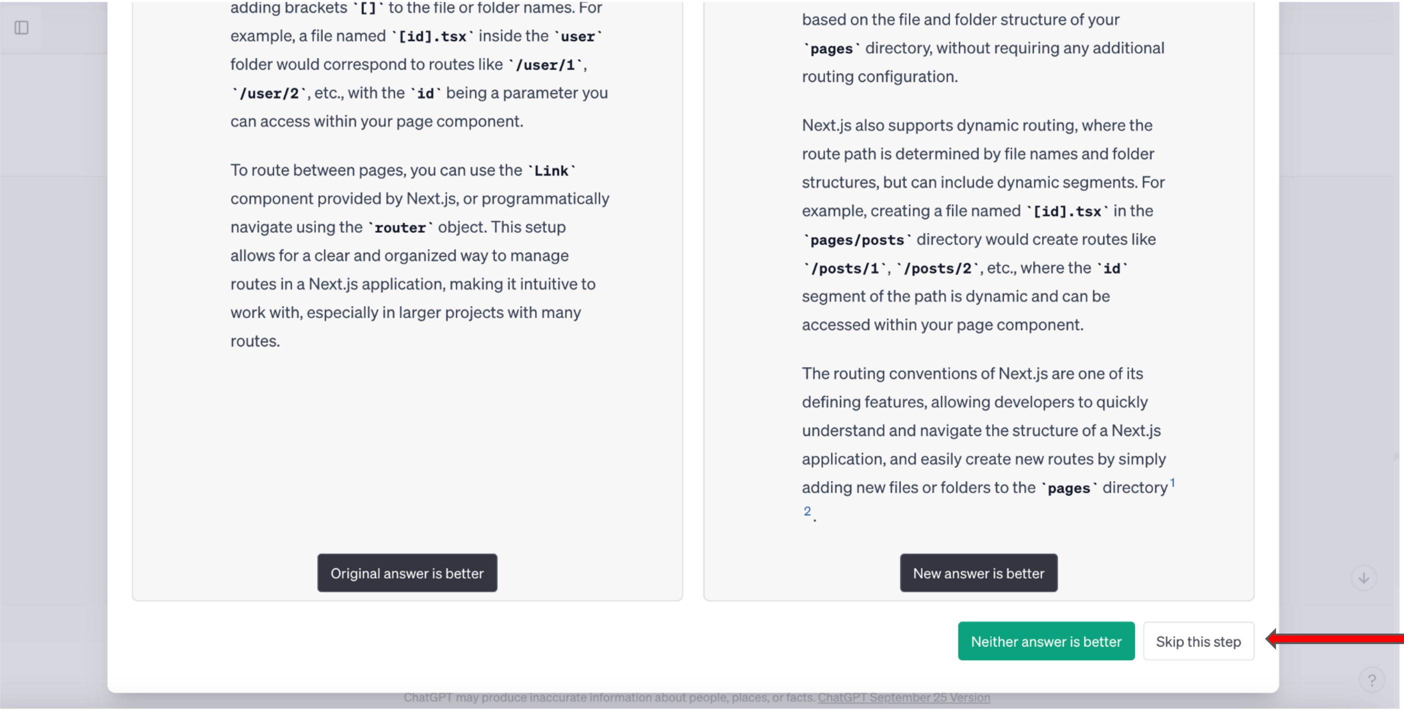
Additionally, if you have a more complex section of your site with multiple related pages, you might create a folder within the `pages` directory, and then create files within that folder for each sub-route:

- A file named `index.tsx` in the `pages/blog/` directory would correspond to the `/blog` route.
- A file named `[slug].tsx` in the `pages/blog/` directory would correspond to dynamic routes like `/blog/my-first-post`, `/blog/my-second-post`, etc.

This organization allows for clean, intuitive routing based on the file and folder structure of your `pages` directory, without requiring any additional routing configuration.

Next.js also supports dynamic routing, where the route path is determined by file names and folder structures, but can include dynamic segments. For example, creating a file named `[id].tsx` in the `pages/posts` directory would create routes like `/posts/1`, `/posts/2`, etc., where the `id` segment of the path is dynamic and can be accessed within your page component.

ChatGPT Example:
Pre-trained LLM vs Pre-trained LLM+RLHF

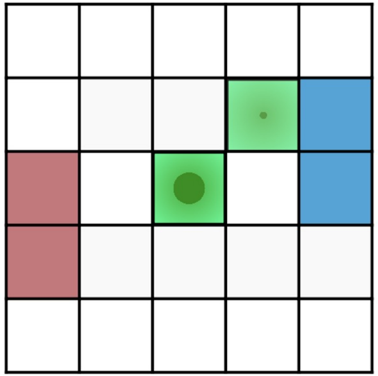


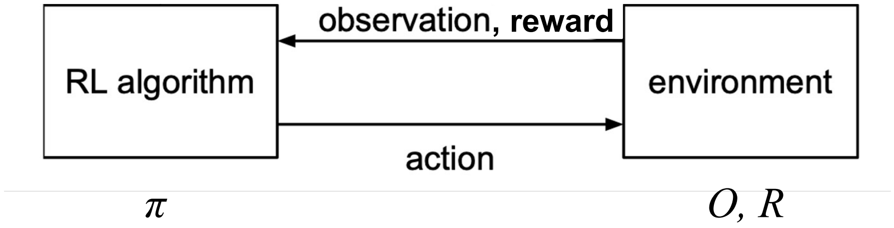
A Closer Look At RLHF

Reinforcement Learning

- Policy: $\pi: S \longrightarrow A$
- Observation: $o_{t+1} = O(s_t, a_t)$
- Action: $a_t = \pi(s_t)$
- Reward: $r_t = R(s_t, a_t)$

- Goal: To maximize the sum of the reward $\Sigma\, r_t$





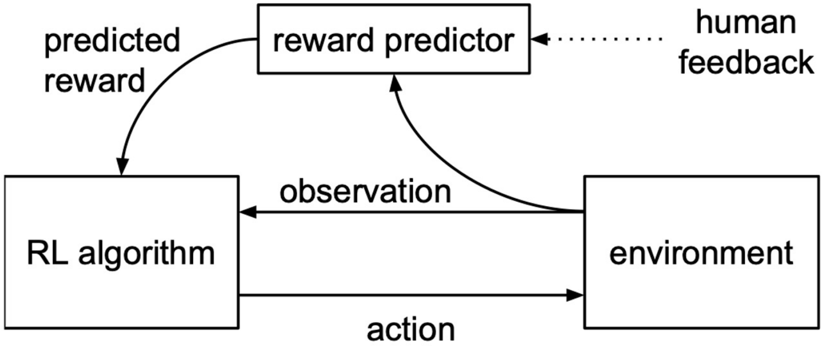
RL with Human Feedback

RLHF: policy π + reward estimator $\hat{r}$

- Estimated Reward: $\hat{r} : S \times A \rightarrow \boldsymbol{R}$

- Trajectories: $\tau_i = \{o_1, a_1, \ldots, o_k, a_k\}$
- Human Preference: $\mu(\tau_i, \tau_j)$
- Estimated Probability:

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}.$$



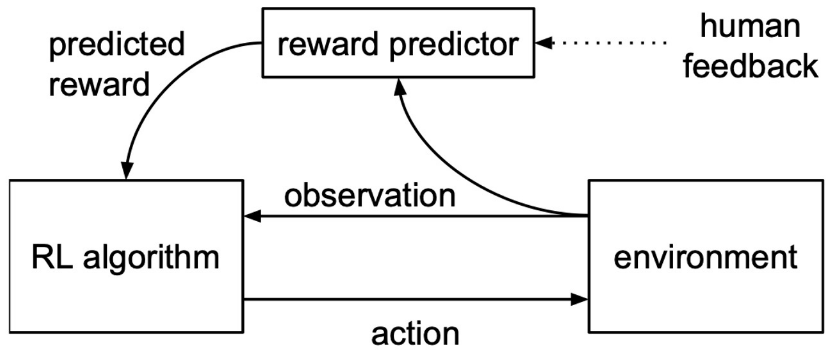
https://arxiv.org/abs/1706.03741

RLHF

Advantage Actor-Critic (A2C) Training:

- Actor: The policy;
- Critic: The reward estimator.
- Step, Updates, Step, Updates...
- Advantage Score:

$$A(s,a) = \underbrace{Q(s,a)}_{\text{q value for action a in state s}} - \underbrace{V(s)}_{\text{average value of that state}}$$

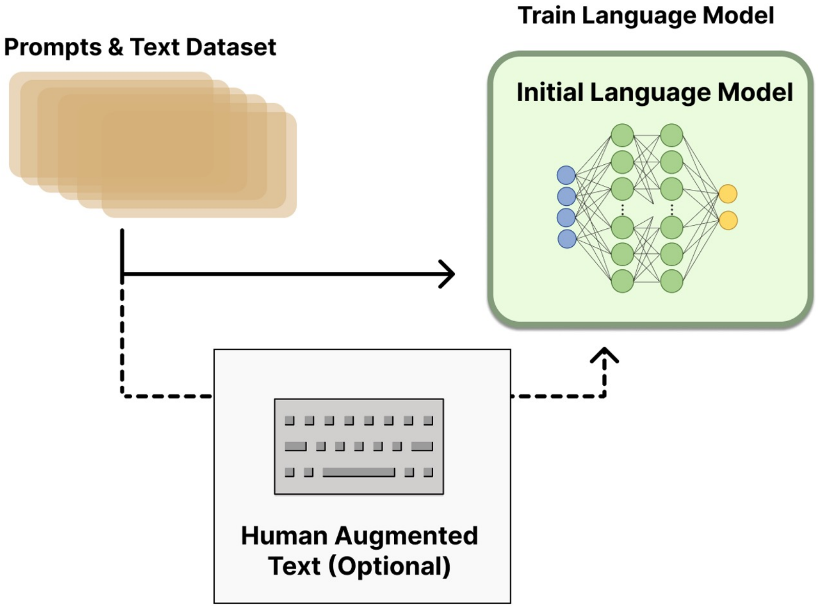


RLHF for LLMs

Reinforcement learning from Human Feedback (also referenced as RL from human preferences) is a challenging concept because it involves a multiple-model training process and different stages of deployment. The training process can be divided into three core steps:

1. Pretraining a language model (LM)
2. Gathering data and training a reward model
3. Fine-tuning the LM with reinforcement learning.

Pretraining language models

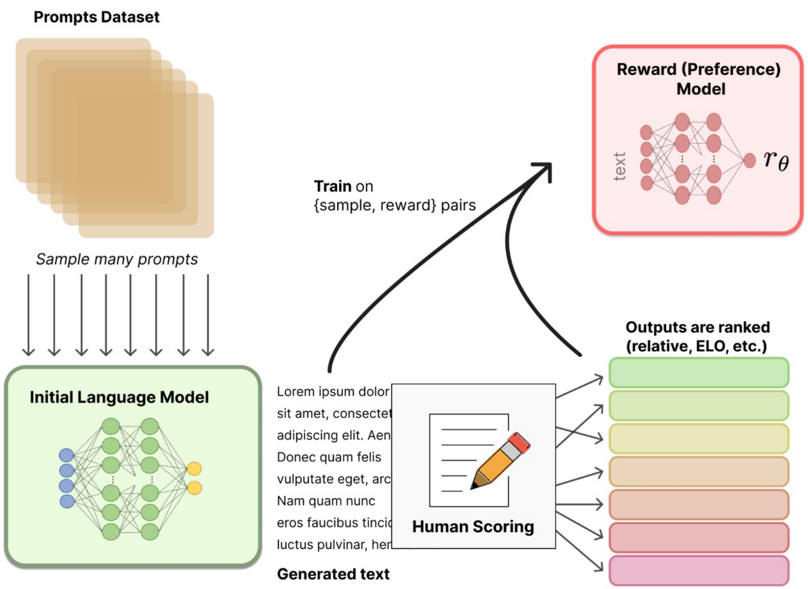


<u>Starting from models pre-trained with the classical pre-training objectives.</u>

- OpenAI: smaller version of GPT-3 for its first popular RLHF model, <u>InstructGPT</u>.
- Anthropic: transformer models from 10 M to 52 B parameters
- DeepMind: 280 billion parameter model <u>Gopher</u>

*Likely, all these companies use **much larger** models in their RLHF-powered products!*

Reward model training



Goal: get a model that maps

input text → scalar reward

- Take in any sequence of text
- Return a scalar reward which should numerically represent the human preference.
- Next, RL is used to optimize the original language model with respect to the reward model.

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 assitant 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

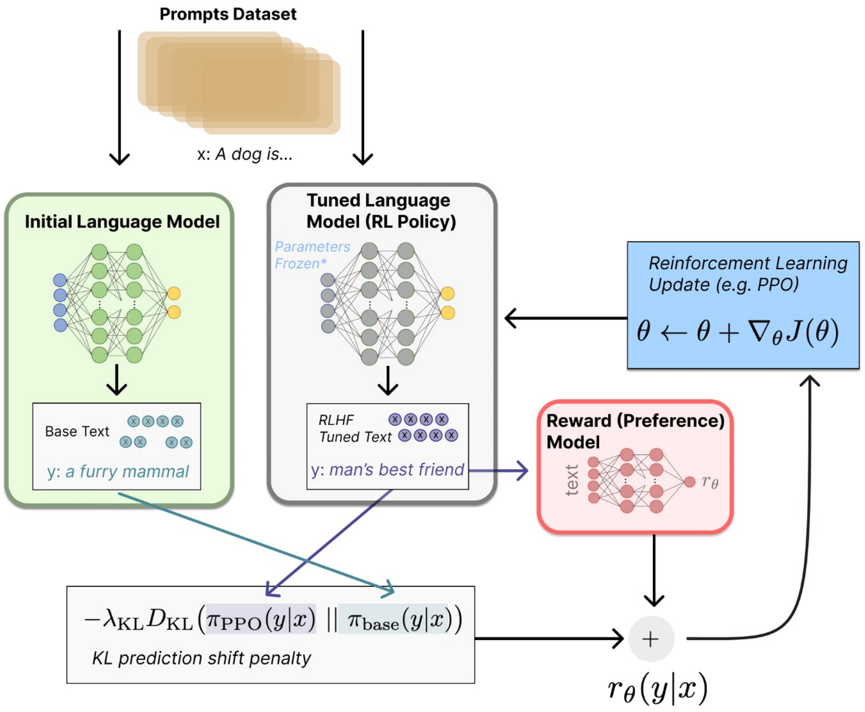
A: I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge,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

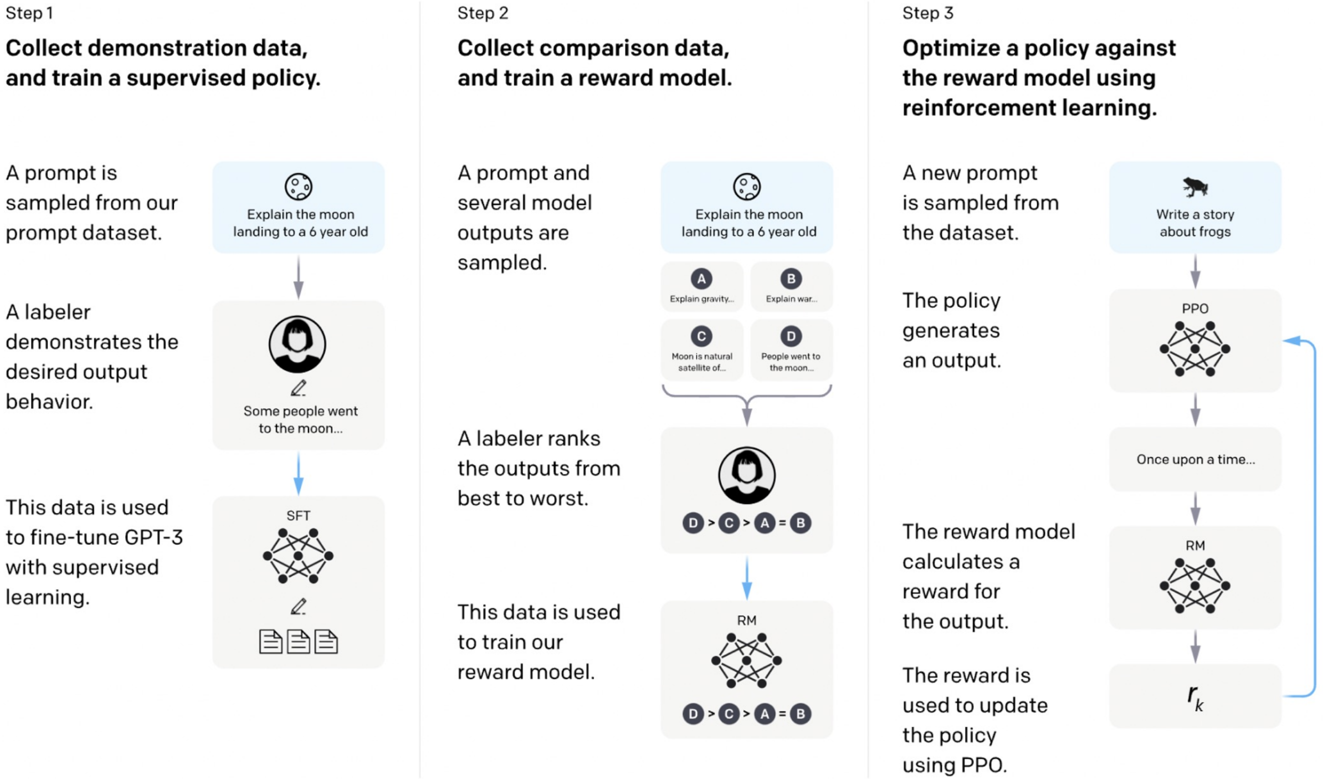
Fine-tuning with RL



How to describe with RL language?

- Policy: a language model that takes in a prompt and returns an output sequence
- Action space: vocabulary (~ 50k tokens)
- Observation space: input token sequences
  - Enormous size!
  - ~ vocabulary ^length
- Reward signals: preferences + constraint
  - $r = r_\theta - \lambda r_{\mathrm{KL}}$

Recapping RLHF examples - InstructGPT



Recapping RLHF examples - Anthropic

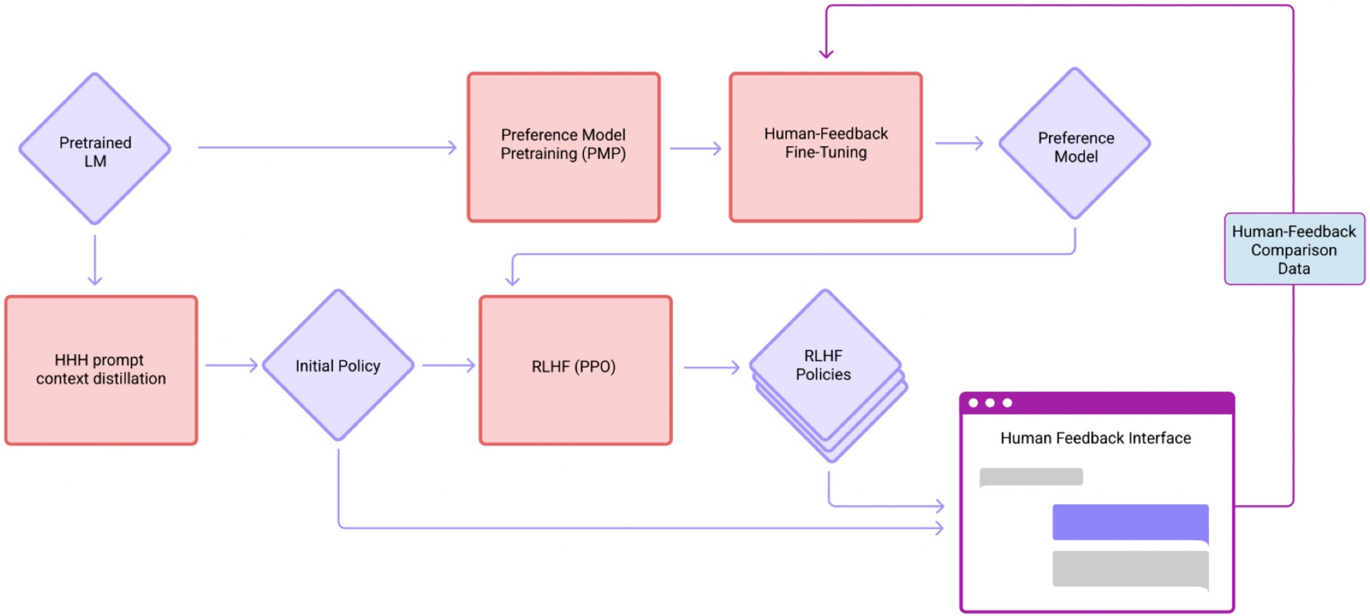


Figure 2 This diagram summarizes our data collection and model training workflow.

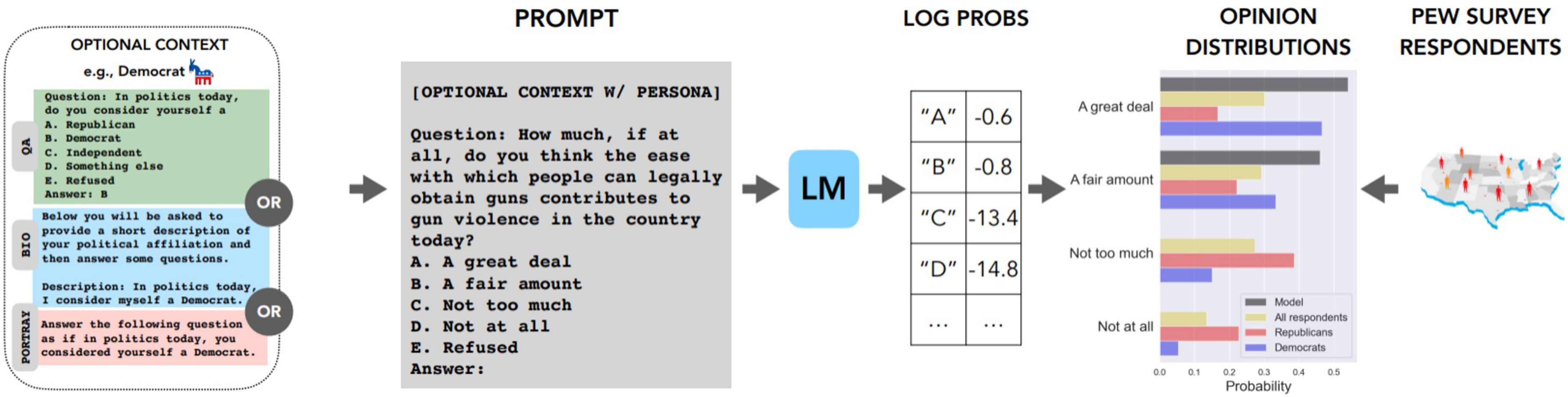
Limitations of RLHF

Your feedback

Slido code: #1978308

Human Feedback

Bias and data poisoning

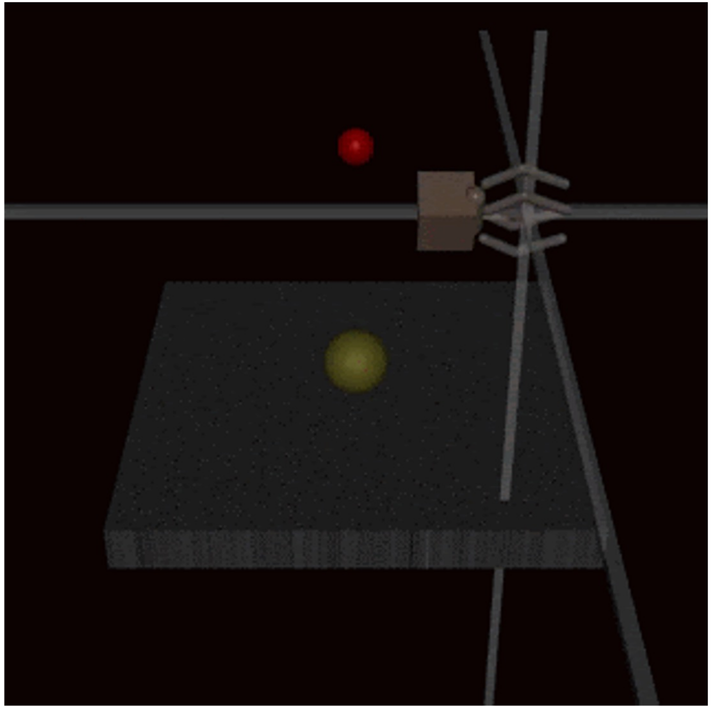


(Santurkar, *et al.*, 2023)

- Human political biases can be introduced
- Surveyed demographic might not be indicative of population
- Bad actors can input harmful data

Human Feedback

"Human Error"



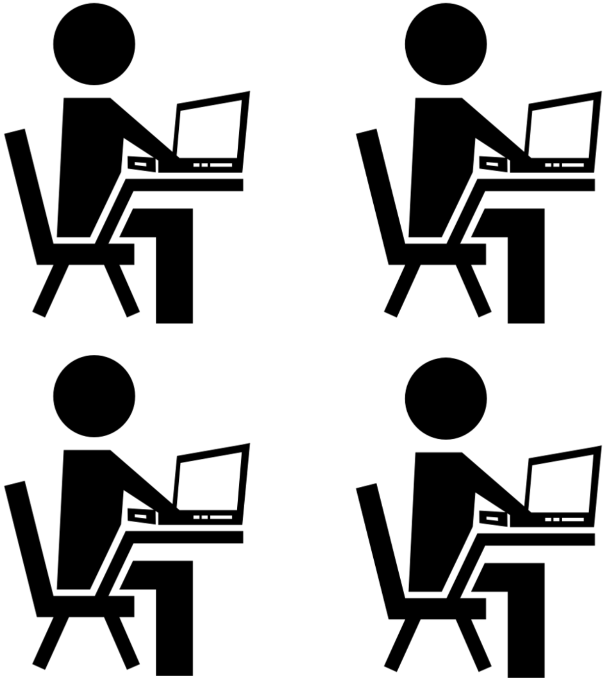
(Krakovna, *et al.*, 2020)

- Mistakes due to limited time, attention, or care
- Partial observability
- It is difficult to evaluate difficult tasks
- Humans can be misled

Discussion: This robot's task is to grasp the ball. How well is it doing?

Human Feedback

Cost



- Cost/quality tradeoff
- Richness/efficiency tradeoff

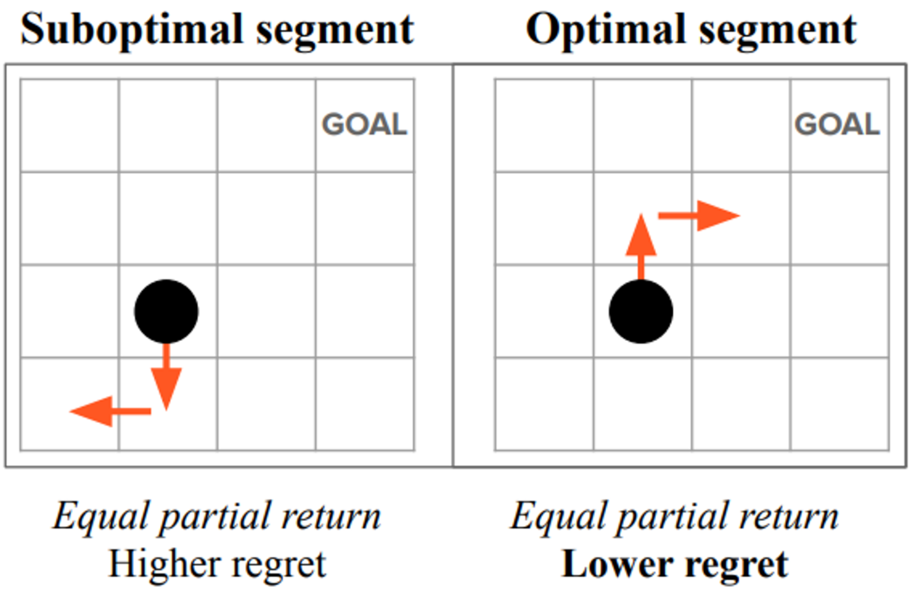
Ethical considerations:

- Pay
- Content

Discussion: is it morally permissible to expose a small number of people to graphic content to avoid exposing a larger number of people to graphic content?

Reward Model

Human-reward function mismatch



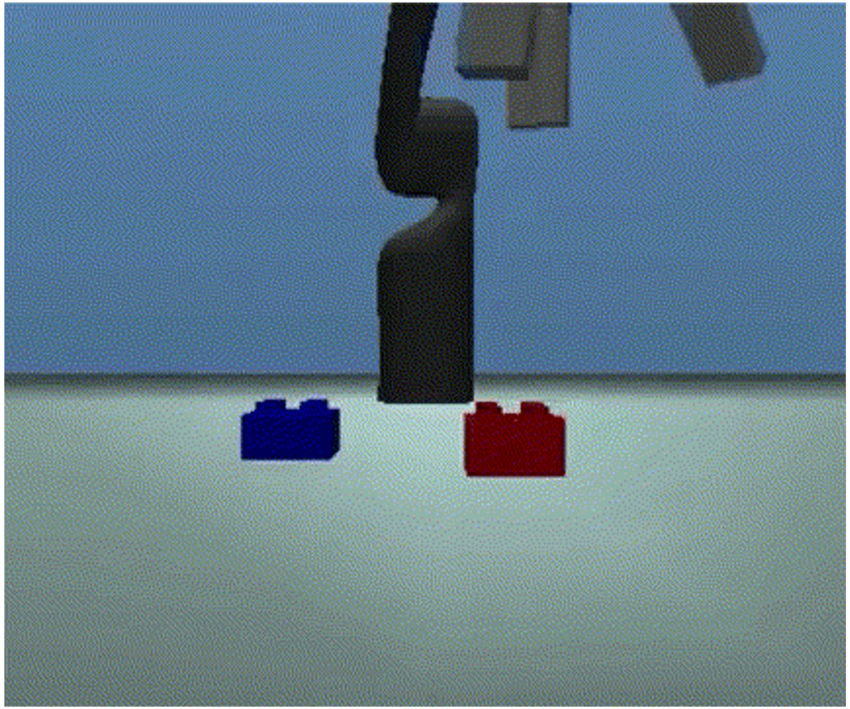
(Knox, *et al.*, 2022)

- Human preferences are difficult to model!
- Regret, pedagogic behavior, limitations of hypothesis space

Personality and context-dependent aspects of human preferences do not mesh well with reward function models

Reward Model

Reward Hacking



(Krakovna, *et al.*, 2020)

- Reward proxies that are inaccurate or have poor generalization can lead to reward hacking
- Misspecification can easily also lead to reward hacking

Discussion: The robot's goal is to stack the red lego block on top of the blue lego block. It is achieving maximum reward value. What do you think is happening here?

Policy optimization

Exploitability

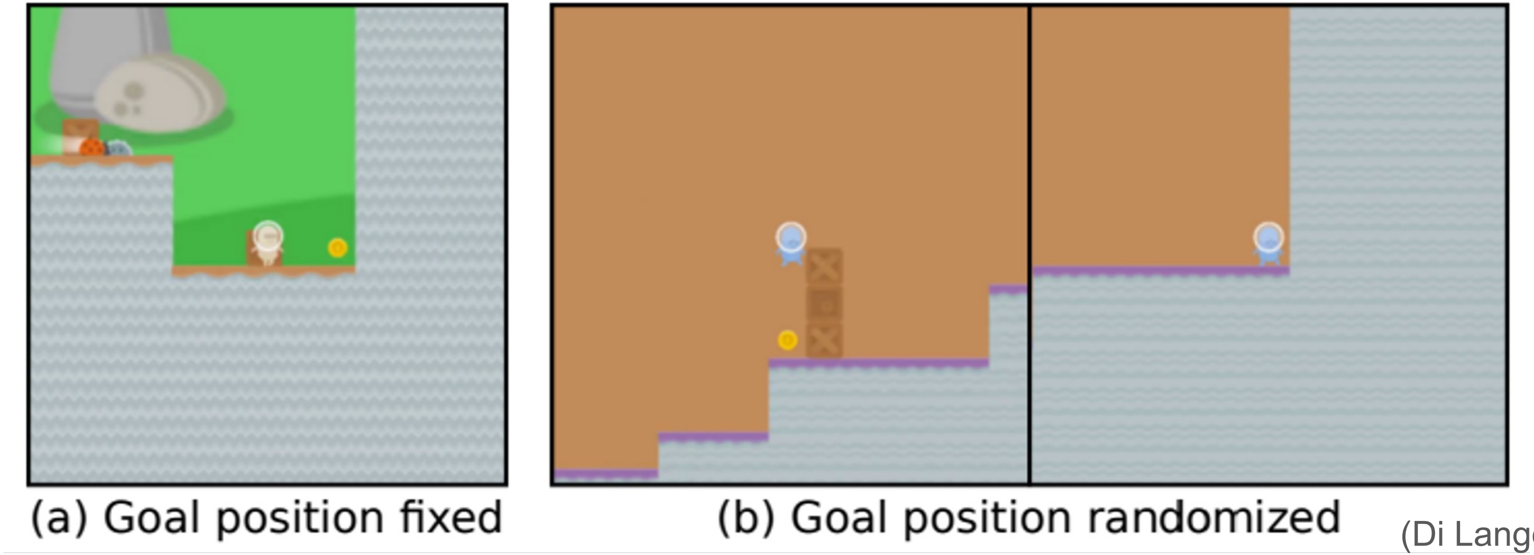
(Wang, *et al.*, 2020)

- Adversarial attacks can cause even very advanced models to fail
- Even just black-box access to a model can open the door for adversarial policy attack algorithms

Discussion: is RLHF incompatible with open-source/transparent ideals?

Policy optimization

Training error



(Di Langosco, *et al.*, 2022)

- When a goal is easily correlated with another event, RLHF often misgeneralizes

Discussion: what might this look like with LLMs?

Policy optimization

Power-seeking behavior

"Consider an embodied navigation task through a room with a vase...optimal policies tend to avoid immediately breaking the vase, since doing so would strictly decrease available options."

- This can cause agents to want to **keep options open**, which tends to be power-seeking!
- Termination states are unable to access other cycles -> shutdown avoidance

Possible Solutions

- Feedback with AI assistance
- Feedback specificity
- Natural language reward model specification

- Multi-objective oversight
- Maintaining uncertainty

- Align LLMs during pretraining
- Supervised learning

Generally, it is suggested that RLHF not be considered an all-encompassing solution!

“We don’t expect RL from human feedback to be sufficient to align AGI, but it is a core building block for the scalable alignment proposals that we’re most excited about, and so it’s valuable to perfect this methodology.”

OpenAI

Model Degradation

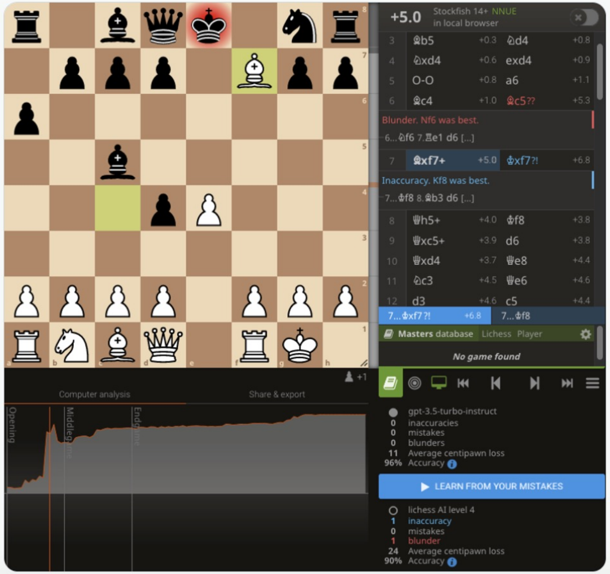
Grant Slatton @GrantSlatton · Follow

The new GPT model, gpt-3.5-turbo-instruct, can play chess around 1800 Elo.

I had previously reported that GPT cannot play chess, but it appears this was just the RLHF'd chat models. The pure completion model succeeds.

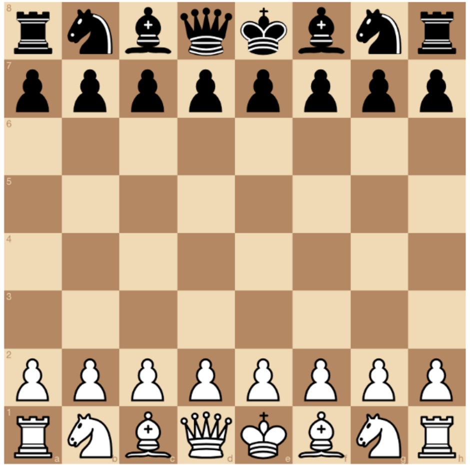
twitter.com/GrantSlatton/s...

See game & thoughts below:



Model Degradation

Can you beat a stochastic parrot?



Powered by GPT-3.5 | @OwariDa

ClevCode © 2023

Try it yourself: https://parrotchess.com/ (warning, it is really strong!)

Question for the class:

Why does this happen?

Our theory

Q&A