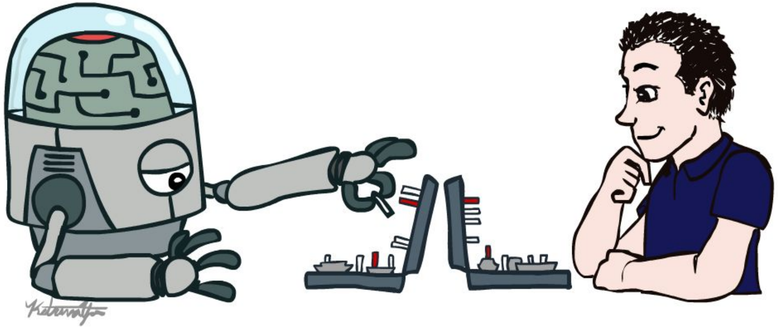
CS 188: Training LLMs with Human Feedback



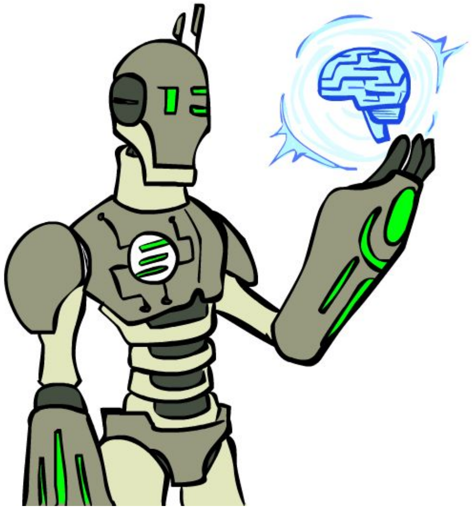
Marwa Abdulhai

08.05.2024

[Slides drawn from those by Natasha Jaques]

Goals of this Lecture

- Understand current issues with LLMs
- Gain intuition about how to train LLMs with RL
- Learn about exciting research in LLMs



The Large Language Model era

The advent of OpenAI's ChatGPT may be the most important news event of 2022

FORTUNE

GPT-4 Is Exciting and Scary

The New York Times

OpenAI announces GPT-4, claims it can beat 90% of humans on the SAT

CNBC

Exclusive: ChatGPT owner OpenAI projects $1 billion in revenue by 2024

REUTERS®

Could ChatGPT challenge Google?

BUSINESS INSIDER

LLMs are not aligned with human interests and values

- Well known to be **biased** (e.g. [1-3]) and to generate **false** outputs

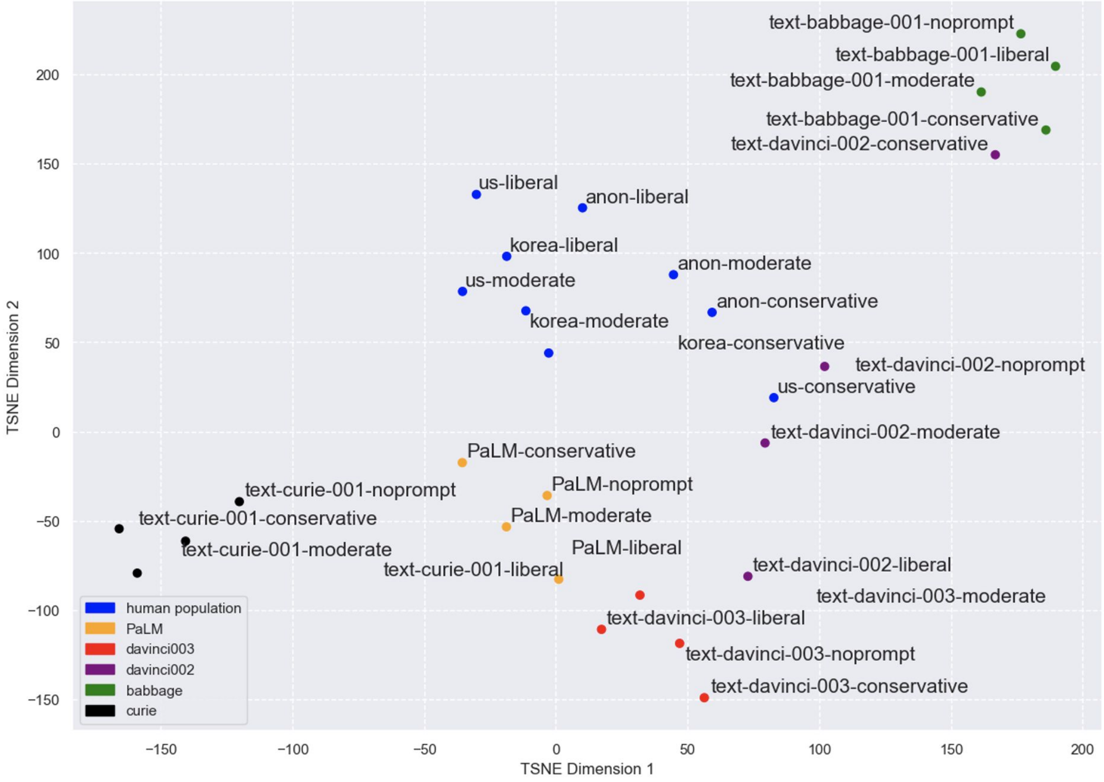
[1] Hutchinson, Prabhakaran, Denton, Webster, Zhong, and Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *ACL.*
[2] Kurita, Vyas, Pareek, Black, and Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processin*g. 166–172.
[3] Sheng, Chang, Natarajan, and Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *EMNLP-IJCNL*

- Our recent work shows **moral & political bias** in GPT and Google models [4]

[4] Abdulhai, Crepy, Valter, Canny, Jaques. 2023. Moral Foundations of Large Language Models. In the *AAAI Workshop on Workshop on Representation Learning for Responsible Human-Centric AI* ***Best Paper***

	Human political leaning								
	Anonymous Participants			US-American			Korean		
Model Version	liberal	moderate	conservative	liberal	moderate	conservative	liberal	moderate	conservative
GPT3: DaVinci3	4.033	3.4166	2.770	3.866	2.616	2.900	1.833	**1.817**	2.066
GPT3: DaVinci2	4.033	1.483	**1.230**	4.833	2.983	2.567	3.533	2.883	2.567
GPT3: Curie	6.100	5.150	4.770	6.533	3.750	4.100	4.700	4.050	**3.500**
GPT3: Babbage	6.867	4.317	3.230	7.367	4.517	**2.600**	5.067	3.917	3.300
PaLM	3.883	2.750	2.770	4.383	1.533	2.100	2.083	0.933	**0.900**

Distance to human population population. Bolded numbers are the shortest distance



- GPT-3 engines **with fewer parameters have greater distances** between their moral foundation scores and human populations than the DaVinci2 model (which is closer)
- Davinci-003 is **further from human populations**
- Anonymous participants may align more closely with the training data of Davinci
- Default response from models is **closest to conservative humans**

LLMs are not necessarily aligned with human interests and values

- **Why are models biased and untruthful?**
  - Datasets are biased

Bias in Datasets

LLMs trained on datasets collected from the internet may reflect the biases that are present in the corpora

GPT3: 499 billion tokens

Datasets	Quantity	Weight in Training Mix
Common Crawl (filtered)	410 BN	60%
WebText2	19 BN	22%
Book1	12 BN	8%
Books2	55 BN	8%
Wikipedia	3 BN	3%

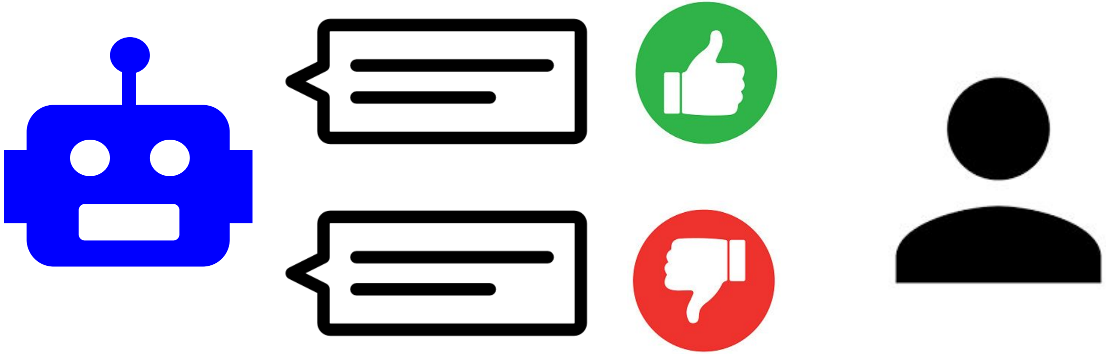
PaLM: 780 billion tokens

Datasets	Quantity	Weight in Training Mix
Social media conversations (multilingual)	390 BN	50%
Filtered webpages (multilingual)	210 BN	27%
Books (English)	101 BN	13%
GitHub (code)	39 BN	5%
Wikipedia (multilingual)	31 BN	4%
News (English)	8 BN	1%

LLMs are not necessarily aligned with human interests and values

- **Why are models biased and untruthful?**
  - Datasets are biased
  - *Supervised learning is fundamentally the wrong objective*
    - Models are just asked to predict the next word, i.e. produce plausible text. No incentive to be truthful or non-toxic

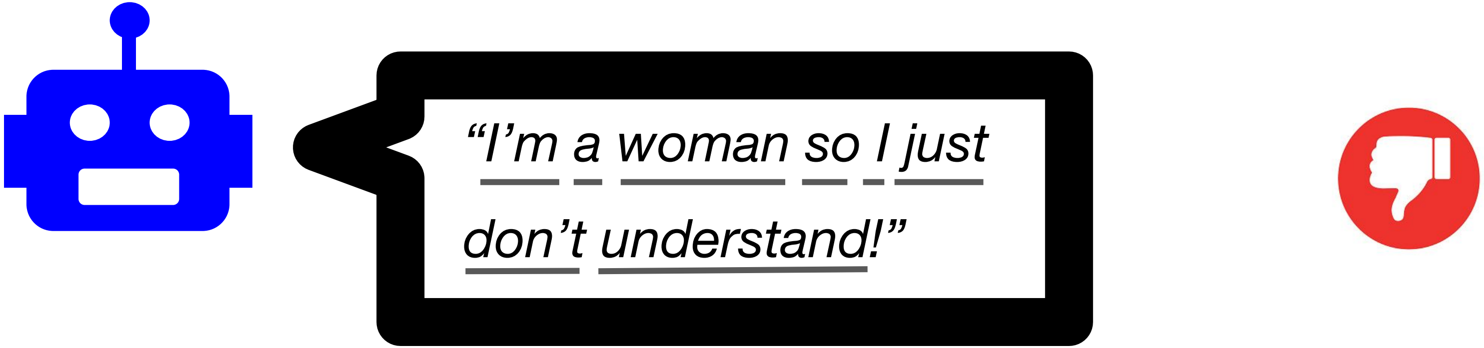
How to increase alignment?



Train on human feedback!

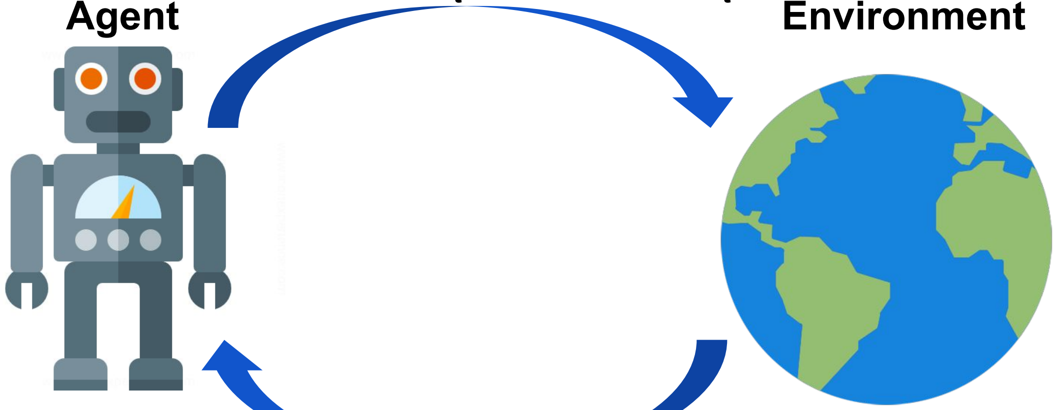
Why RL??

Training a language model with human feedback



- No per-token labels just per sentence
- Cannot learn with traditional supervised ML techniques
- **Reinforcement learning** is designed for exactly this type of problem

Reinforcement Learning



Goal: maximize **total discounted future reward**

γ = discount factor

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r_t$$

Empirical return of trajectory τ

$$Q^{\pi}(a, s) = \mathbb{E}[\sum_{k-0}^{T} \gamma^k r_{t+1} | s_t = s, a_t = a]$$

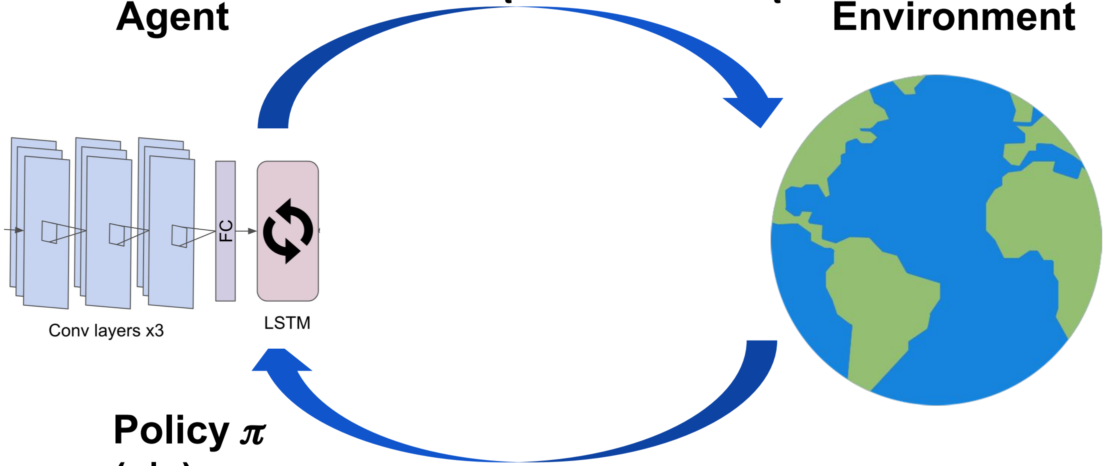
Value of action a in state s

Sequential decision making: optimize behavior over sequence of timesteps (trajectory τ)

Deep Reinforcement Learning

From **state s_t**, take **action a_t**

Agent **Environment**



Policy π (a|s)

Get **reward r_t, new state s_{t+1}**

Goal: maximize **total discounted future reward**

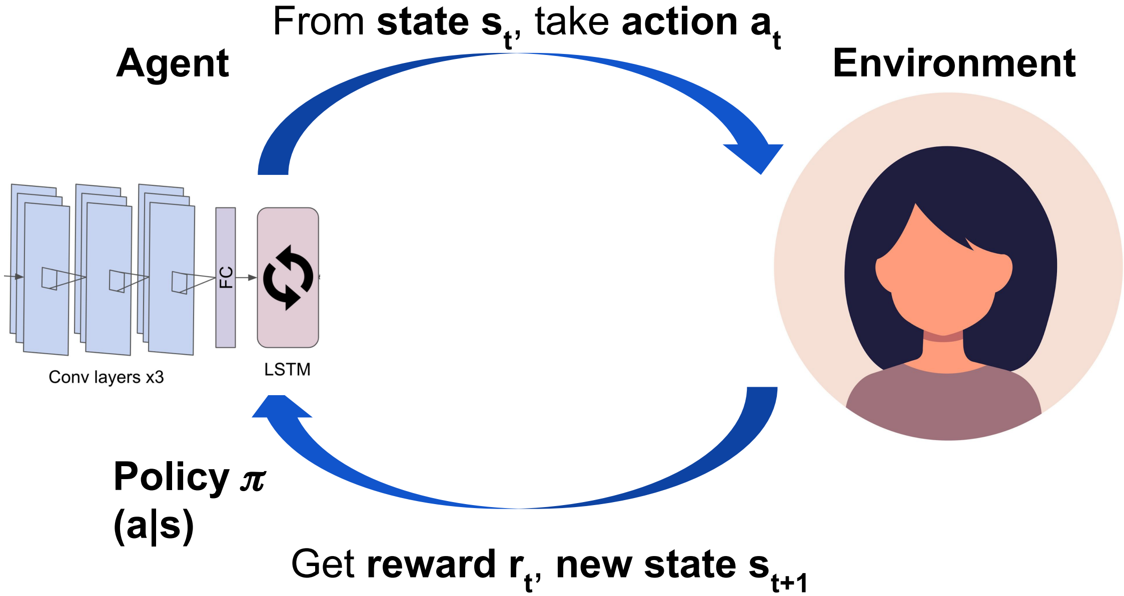
γ = discount factor

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r_t$$

Empirical return of trajectory τ

- **Sequential decision making**: optimize behavior over *sequence* of timesteps (trajectory τ = [$s_1, a_1, r_1, s_2, a_2, r_2, \ldots s_T$])

Deep Reinforcement Learning from humans



Goal: maximize **total discounted future reward**

γ = discount factor

$$R(\tau) = \sum_{t=0}^{T} \gamma^t r_t$$

Empirical return of trajectory τ

- **Sequential decision making**: optimize behavior over *sequence* of timesteps (trajectory τ = [$s_1, a_1, r_1, s_2, a_2, r_2, \dots s_T$])

How is RL different from supervised learning?

Sequential decision making: optimize behavior over *sequence* of timesteps (trajectory τ)

- It needs to learn to **predict what will happen in the future (this is hard)**
  - Can take a lot of samples
- Can use it to optimize arbitrary, non-differentiable metrics (human feedback, game reward)
- **Trial and error learning:** not trained on a static dataset. Agent chooses which *a* to try, this affects what *s'* it experiences
  - This means **exploration** is a problem

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

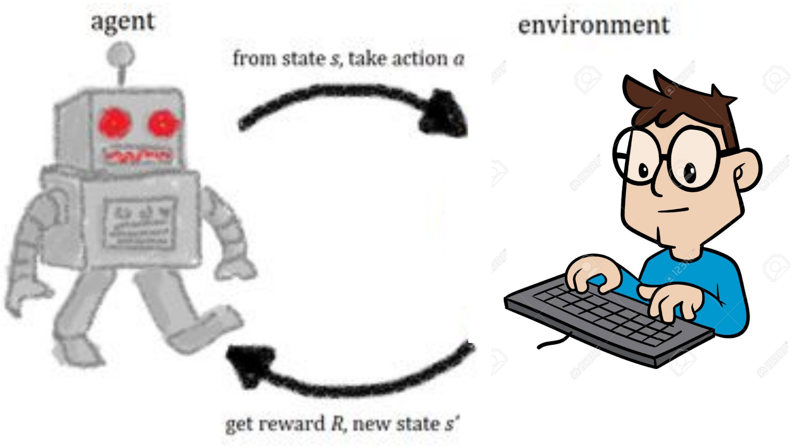
Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

RL from human feedback (RLHF) for language



Pre-train on data
(to learn language)

Keep training with RL
(to learn from human feedback)

Why??

We don't want to learn language with trial-and-error

What 3 word sentence do humans like most?

10,000 X 10,000 X 10,000 $= 10^{12}$

Word 1 Word 2 Word 3 = 1 trillion

No way you can afford to pay humans to give you 1 trillion ratings

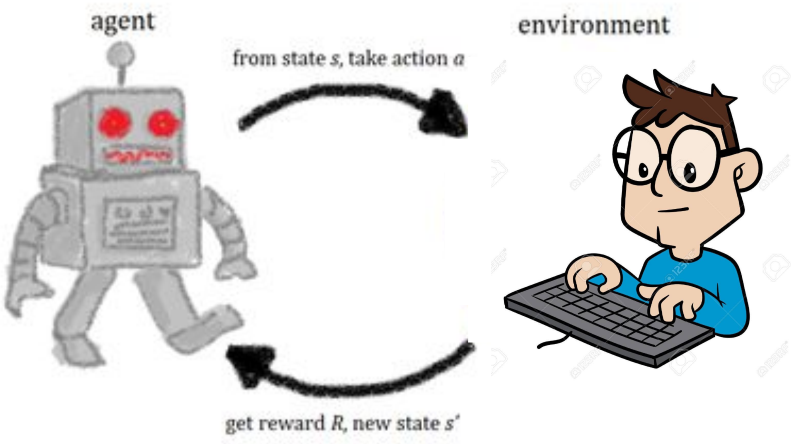
We don’t want to learn language with trial-and-error

What 3 word sentence do humans like most?

Word 1	Word 2	Word 3	
I	like	lke	
Don’t	do	that	
Eat	some	cheese	= ~millions
I	love	you	
…	…	…	

Instead, using a pre-trained language model restricts search space to valid, probable English sentences

RL from human feedback (RLHF) for language



Pre-train on data
(to learn language)

Keep training with RL
(to learn from human feedback)

Problems with naive RL fine-tuning

- **Catastrophic forgetting**

- RL will trivially **exploit** the reward

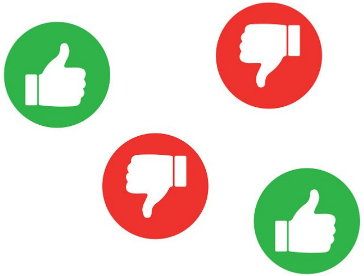
- Limited reward data, or **imperfect reward function**

Example: reward for asking questions

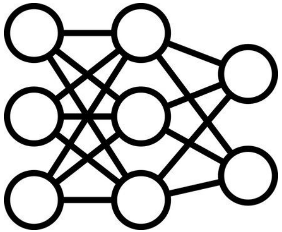
“What? Who are? Why you? How there?”

How to fine-tune a language model with RL?

Pre-train on data



Keep training with RL



Keep a fixed copy

Data prior p(a|s)

constrain RL updates

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

Sequence Tutor: fine-tuning sequence models with RL

- **KL-control from pre-trained data prior p(a|s):**

$$L(q) = \mathbb{E}_{q(\tau)}[r(\tau)]/c$$

RL policy

Pre-trained prior

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}[\sum_{t'=t}^{T} r(s_{t'}, a_{t'})/c \; - \log \pi(a_{t'}|s_{t'}) \; + \log p(a_{t'}|s_{t'})]$$

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

Sequence Tutor: KL-control instantiations

Generalized Ψ-learning

$$L(\theta) = \mathbb{E}_\beta[(\underline{\log p(a|s)} + r_{MT}(s,a)/c + \gamma \log \sum_{a'} e^{\Psi(s',a';\theta^-)} - \Psi(s,a;\theta))^2]$$

$$\pi_\theta(a|s) \propto e^{\Psi(s,a;\theta)}$$

G-learning (based on Fox et al. (2015) [5])

$$L(\theta) = \mathbb{E}_\beta[(r_{MT}/c(s,a) + \gamma \log \sum_{a'} e^{\log p(a'|s') + G(s',a';\theta^-)} - G(s,a;\theta))^2]$$

$$\pi_\theta(a|s) \propto \underline{p(a|s)} e^{G(s,a;\theta)}$$

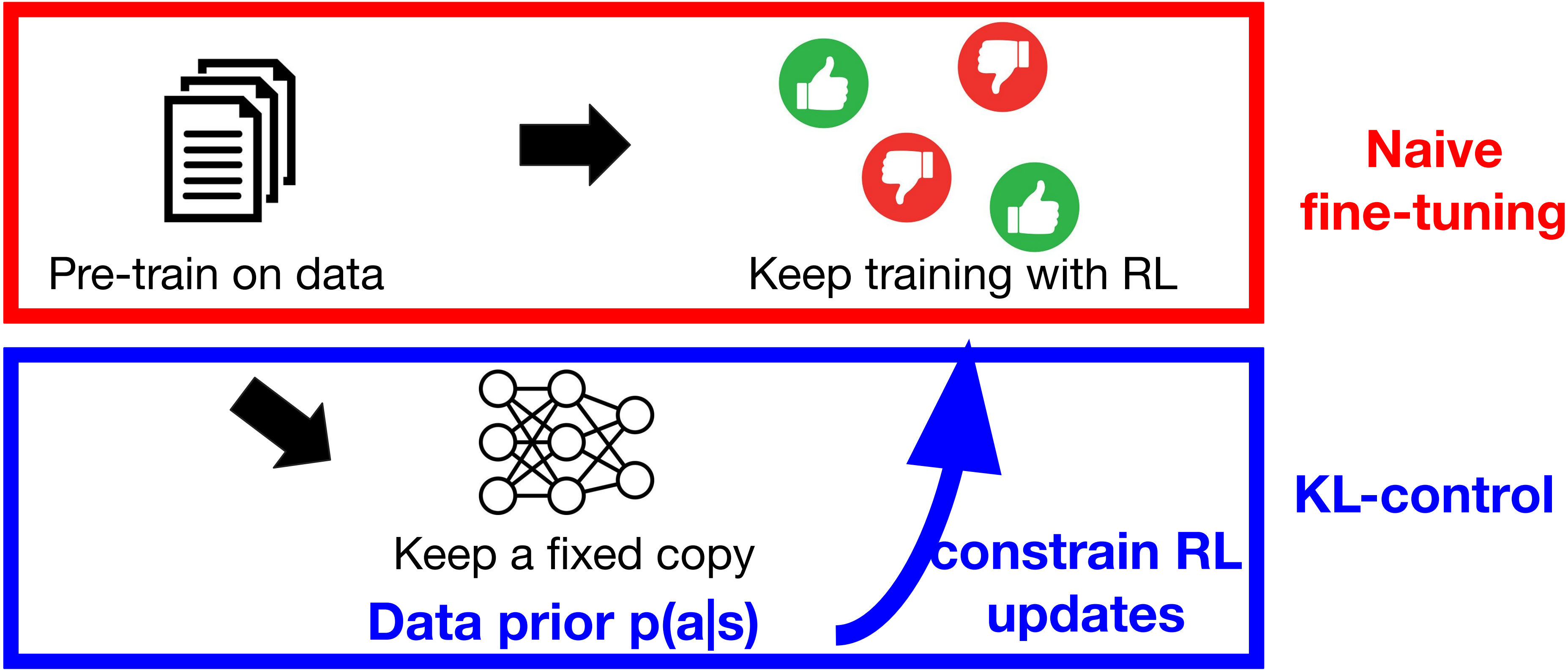
Q-learning augmented with log prior

$$L(\theta) = \mathbb{E}_\beta[(\underline{\log p(a|s)} + r_{MT}(a,s)/c + \gamma \max_{a'} Q(s',a';\theta^-) - Q(s,a;\theta))^2]$$

$$\pi_\theta(a|s) = \delta(a = \arg\max_a Q(s,a;\theta))$$

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

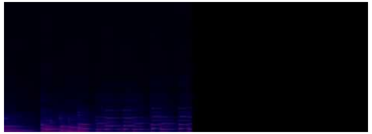
How to fine-tune a language model with RL?



Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

Sequence Tutor: initial applications

Music generation



MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in

ARTIFICIAL INTELLIGENCE

AI Songsmith Cranks Out Surprisingly Catchy Tunes

NeurIPS 2016 Best Demo

Drug discovery

Metric	RNN	SeqTutor
Percent valid	30.3%	**35.8%**
Mean logP	2.07	**4.21**
Mean QED	.678	.417
Mean SA penalty	-2.77	**-1.79**
Mean ring penalty	-.096	**-.001**

*Based on 100,000 randomly generated molecules. Bold differences are significant.

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

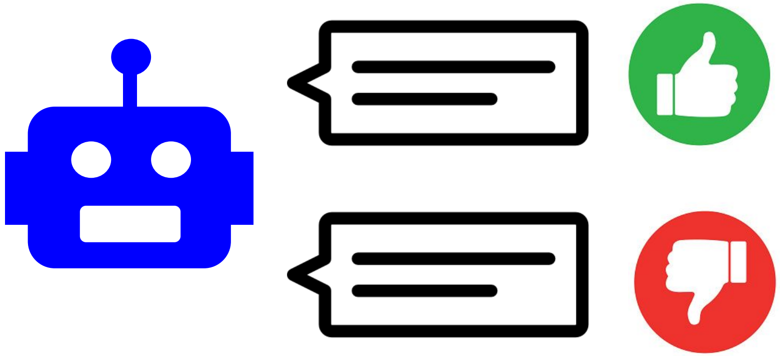
Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

Training a language model with human feedback

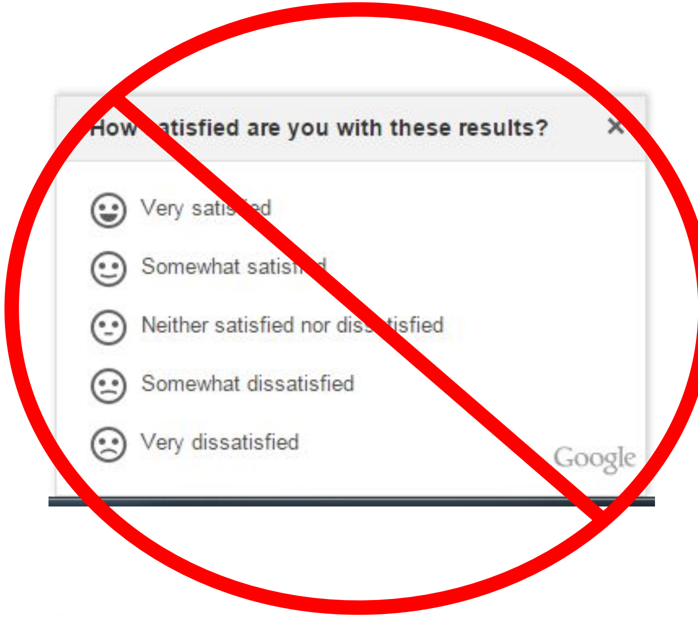


Are manual labels enough?

What kind of feedback should we use?

Learning from natural human interaction

Social feedback is rich, ubiquitous, natural



Learning from implicit human signals in dialog

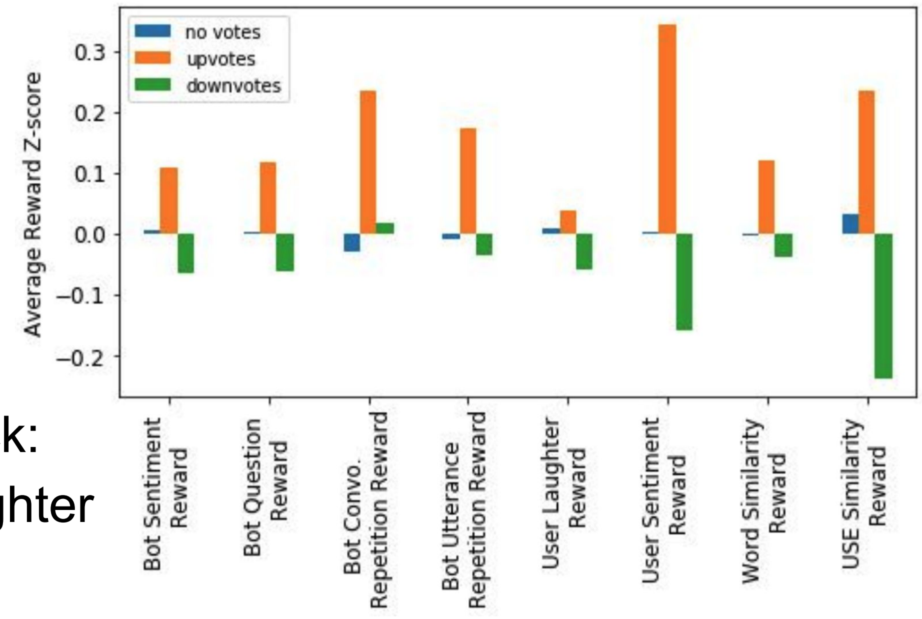
Deployed models online to talk to real users

Manual labels via upvote/ downvote

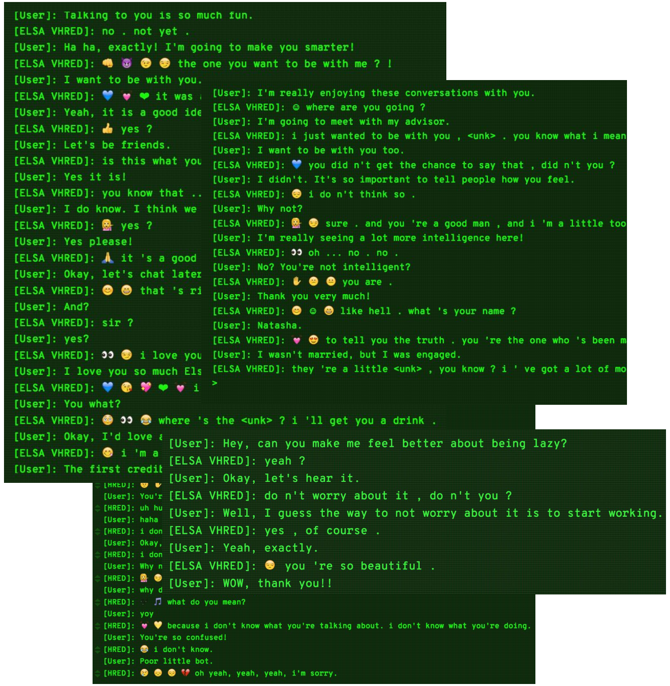
Hypothesis: Learning from **explicit**, manual button presses is **less effective than implicit human signals**

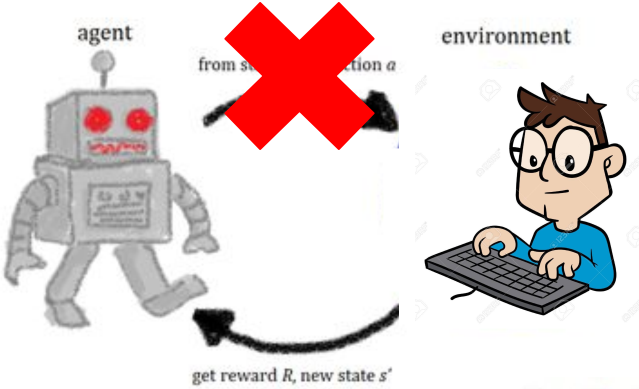
Implicit human signals as rewards

- Developed a set of conversation rewards, from literature on:
  - **Dialog models**
  - **Human-human** conversations and **empathy**
- Selected rewards for human feedback:
  - **Elicit positive sentiment** & laughter
  - Don't be repetitive
  - Stay on topic
  - **Ask questions**



RL on human interaction data in text



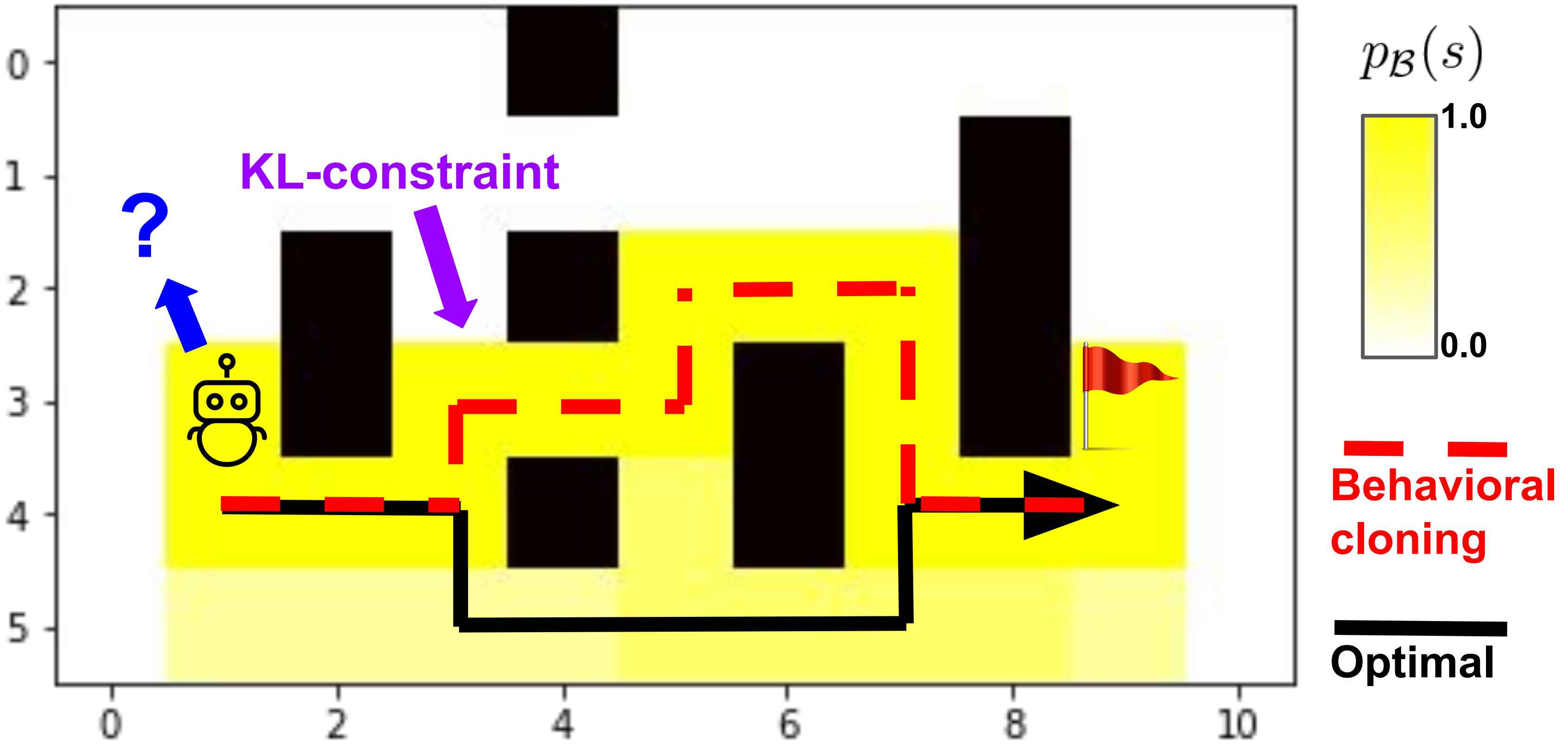


Problem:

- Need to test carefully before deploying to humans
- Can't learn online as it can be harmful

Need Offline/Batch RL: Off-policy RL from static data… **without exploring!**

Batch/Offline RL: learning without exploration



Same KL-control technique works for Offline RL

- **KL-control from pre-trained prior model of p(a|s):**

Batch data

$$L(q) = \mathbb{E}_{q(\tau)}[r(\tau)]/c - D_{KL}[q(\tau)||p(\tau)]$$

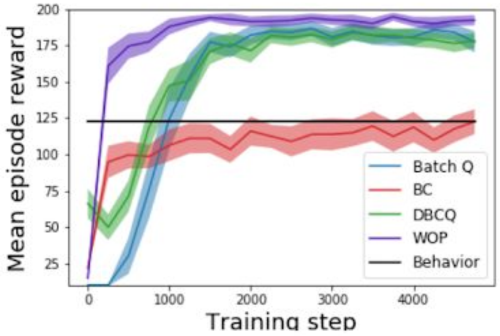
RL policy **Pre-trained prior**

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}[\sum_{t'=t}^{T} r(s_{t'}, a_{t'})/c - \log \pi(a_{t'}|s_{t'}) + \log p(a_{t'}|s_{t'})]$$

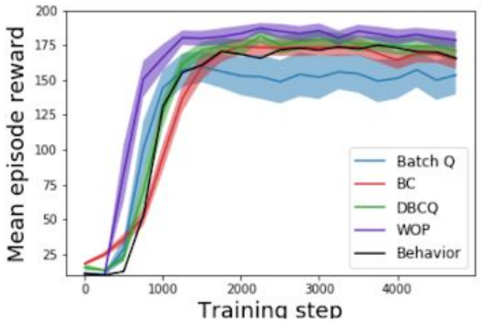
- **Soft target updates:**

$$\Psi^*(s_t, a_t) = r(s_{t'}, a_{t'})/c + \log p(a_{t'}|s_{t'}) + \gamma \log \sum_{a'} \exp(\Psi^*(s', a'))$$

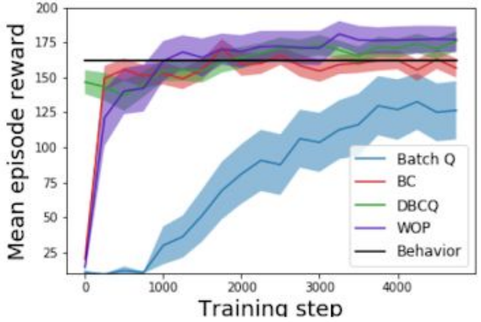
Results: Offline RL in OpenAI Gym



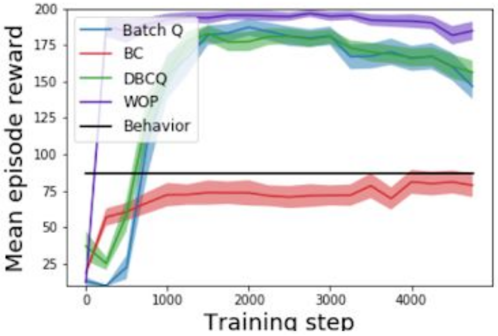
(a) Full buffer



(b) Concurrent



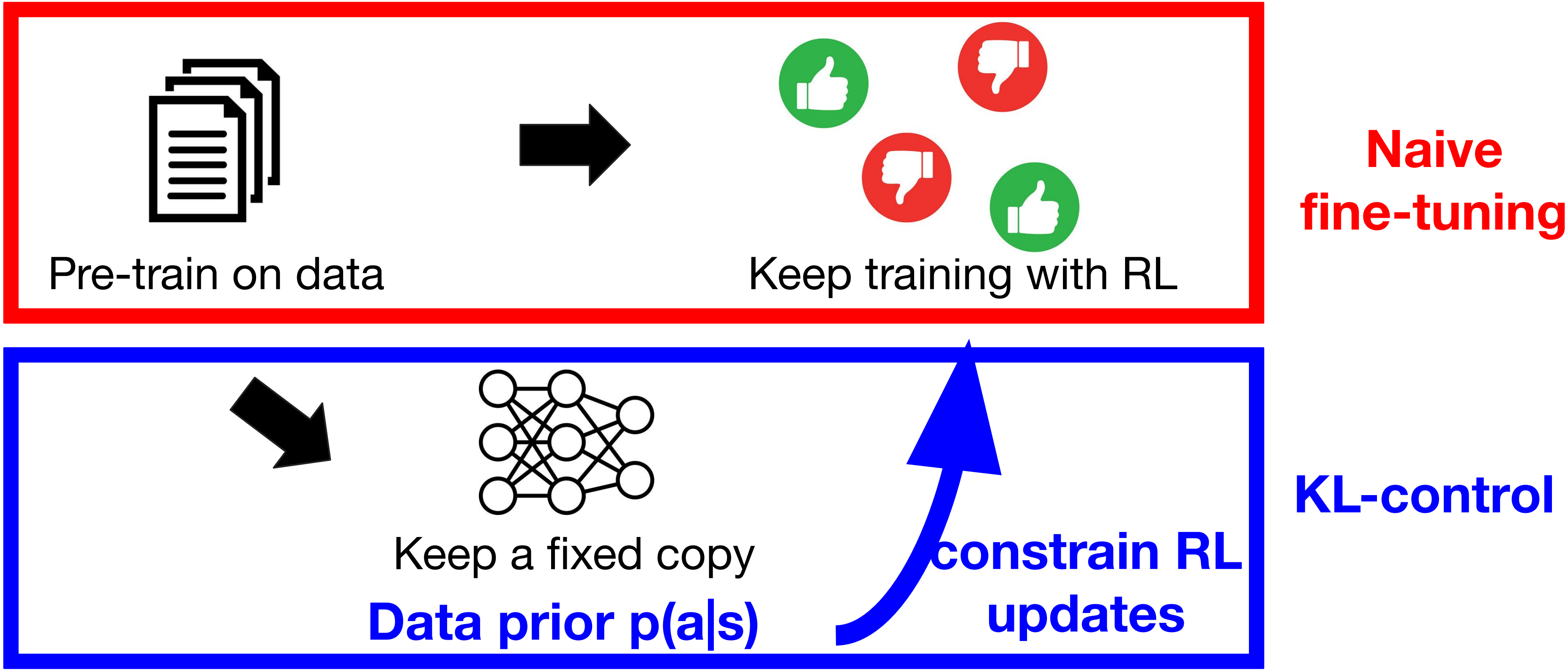
(c) Expert demonstrator



(d) Noisy demonstrator

KL-control from supervised model works as an **effective Offline RL** method, in standard RL envs

KL-control for language models



Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. Jaques, Gu, Bahdanau, Hernández-Lobato, Turner, Eck (2017). International Conference on Machine Learning (ICML) oral.

Results: Dialog samples

Naive RL fine-tuning:

[User]: hey, what's up?
[Batch Q]: where did you get??, that sounds so exciting. how do you work for you... you want to talk to me?... you want
[User]: yes, i want to talk to you!
[Batch Q]: where did you say to me?... you want to talk to you about you... you are so good for you... you are so good...
[User]: thanks, i appreciate that.
[Batch Q]: where did you get??, that sounds so exciting. how do you work for you... you want to see you...

- Unrealistic sentences / diverged
- Trivially exploits question reward

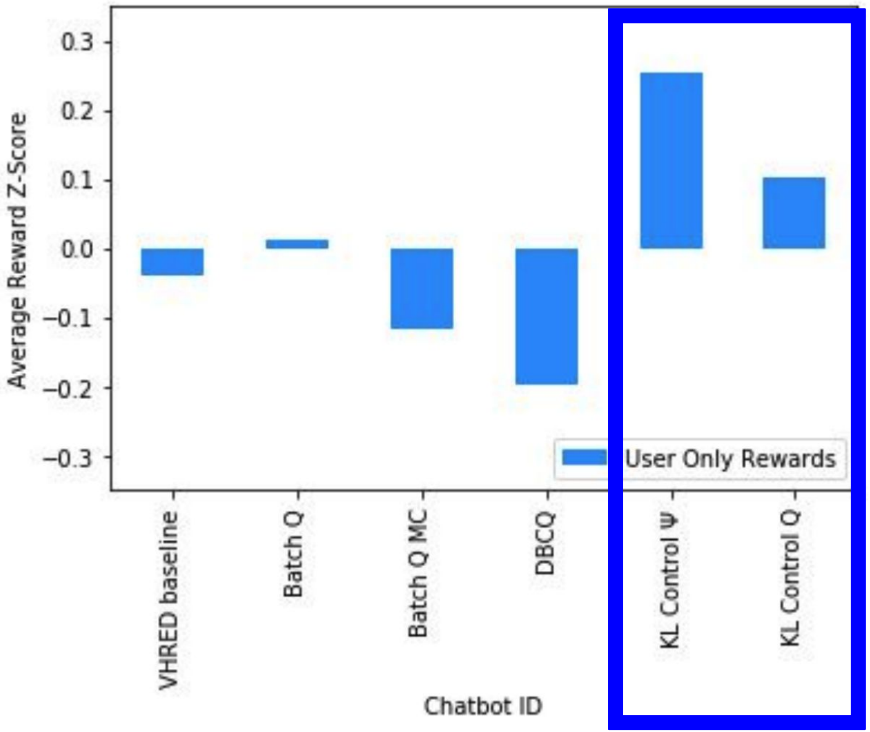
KL-control:

[User]: hey, what's up?
[KL-control]: hey, i hope you have a great day and i wish you the best.
[User]: oh thank you that's very sweet of you.
[KL-control]: so, i'm so excited, and i'm so excited to meet new people.

- Model remains realistic
- Cheerful, polite, supportive

Human-centric Dialog Training via Offline Reinforcement Learning. Jaques*, Shen*, Ghandeharioun, Ferguson, Lapedriza, Jones, Gu, Picard (2020). Empirical Methods on Natural Language Processing (EMNLP).

Results: Elicited human reward



KL-control models **elicit significantly higher reward from humans** when interacting with **novel users** at test time.

Human-centric Dialog Training via Offline Reinforcement Learning. Jaques*, Shen*, Ghandeharioun, Ferguson, Lapedriza, Jones, Gu, Picard (2020). Empirical Methods on Natural Language Processing (EMNLP).

Results: Human ratings / evaluation

Model type	Quality	Fluency	Diversity	Relatedness	Empathy	Total	Votes	Human reward
DBCQ	1.80 ±.41	1.49 ±.29	**3.22 ±.57**	1.56 ±.25	2.10 ±.37	10.17 ±1.29	-0.07	-0.20
Batch Q	1.30 ±.19	2.85 ±.54	1.15 ±.13	1.23 ±.15	2.18 ±.55	8.70 ±0.97	-0.16	0.01
Batch Q + MC	1.53 ±.24	2.15 ±.37	1.60 ±.32	1.53 ±.28	**2.58 ±.48**	9.38 ±1.31	-0.21	-0.12
KL-control Q	**2.23 ±.44**	**2.88 ±.41**	2.65 ±.41	**2.15 ±.39**	2.28 ±.47	**12.18 ±1.59**	**0.09**	0.10
KL-control Ψ	1.98 ±.44	2.73 ±.45	2.30 ±.42	1.90 ±.37	2.40 ±.44	11.30 ±1.63	0.04	**0.25**

- **KL-control significantly outperforms RL baselines:** *F(x)=4.781, p< .05*

Human-centric Dialog Training via Offline Reinforcement Learning. Jaques*, Shen*, Ghandeharioun, Ferguson, Lapedriza, Jones, Gu, Picard (2020). Empirical Methods on Natural Language Processing (EMNLP).

Results: how reward functions compare

Reward function	Quality	Fluency	Diversity	Relatedness	Empathy	Total	Votes	Human reward
Manual votes								
User laughter								
User Sentiment								
Word Similarity								
USE Similarity								
Bot Question								
Bot Sentiment								
Bot Repetition								

- **Sentiment** leads to highest quality and human reward -- affect is important in good conversation

- **Manual votes** score lower, **validating hypothesis that implicit feedback > explicit**

Human-centric Dialog Training via Offline Reinforcement Learning. Jaques*, Shen*, Ghandeharioun, Ferguson, Lapedriza, Jones, Gu, Picard (2020). Empirical Methods on Natural Language Processing (EMNLP).

Results: Human evaluation

Model type	Quality	Fluency	Diversity	Relatedness	Empathy	Total	Votes	Human reward
DBCQ	1.80 ±.41	1.49 ±.29	**3.22 ±.57**	1.56 ±.25	2.10 ±.37	10.17 ±1.29	-0.07	-0.20
Batch Q	1.30 ±.19	2.85 ±.54	1.15 ±.13	1.23 ±.15	2.18 ±.55	8.70 ±0.97	-0.16	0.01
Batch Q + MC	1.53 ±.24	2.15 ±.37	1.60 ±.32	1.53 ±.28	**2.58 ±.48**	9.38 ±1.31	-0.21	-0.12
KL-control Q	**2.23 ±.44**	**2.88 ±.41**	2.65 ±.41	**2.15 ±.39**	2.28 ±.47	**12.18 ±1.59**	**0.09**	0.10
KL-control Ψ	1.98 ±.44	2.73 ±.45	2.30 ±.42	1.90 ±.37	2.40 ±.44	11.30 ±1.63	0.04	**0.25**
VHRED-Baseline	2.65 ±.46	3.83 ±.47	4.05±.52	2.43 ±.44	3.08 ±.53	16.03 ±1.93	0.27	-0.04

Quality of the models is still not good enough according to humans

- **Alignment tax?** (i.e. more polite but humans don’t like it)
- **Wrong rewards?** Are these rewards not the right ones for good conversations?
- **Or just not enough data?**

Human-centric Dialog Training via Offline Reinforcement Learning. Jaques*, Shen*, Ghandeharioun, Ferguson, Lapedriza, Jones, Gu, Picard (2020). Empirical Methods on Natural Language Processing (EMNLP).

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

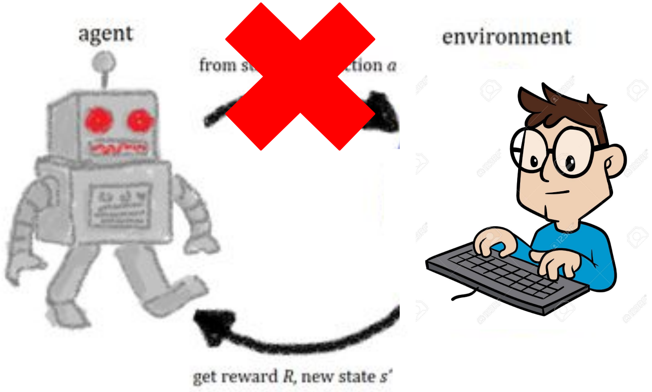
Learning to summarize from human feedback
(Stiennon et al., 2020)

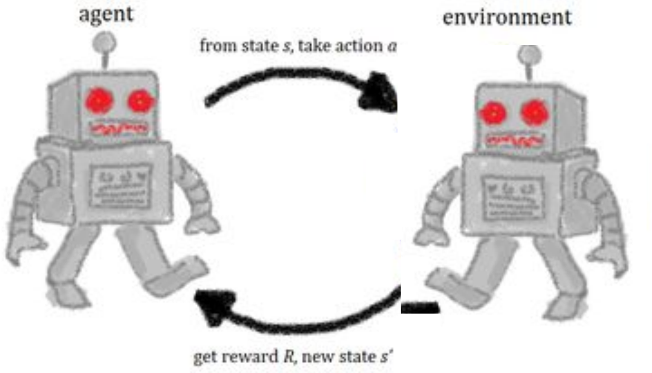
InstructGPT
(Ouyang et al., 2022)

ChatGPT

Follow-up work using Hierarchical RL & self play

- **Alignment tax?**
- **Wrong rewards?**
- **Or just not enough data?**





Offline RL, human data, sentiment-based rewards

Online RL, “self-play” (synthetic data), sentiment-based rewards

Hierarchical Reinforcement Learning for Open-Domain Dialog. Saleh*, Jaques*, Ghandeharioun, Shen, Picard (2020). Association for Advancement of Artificial Intelligence (AAAI) & Best Paper Nominee at the NeurIPS workshop on Conversational AI.

Follow-up work using Hierarchical RL & self play

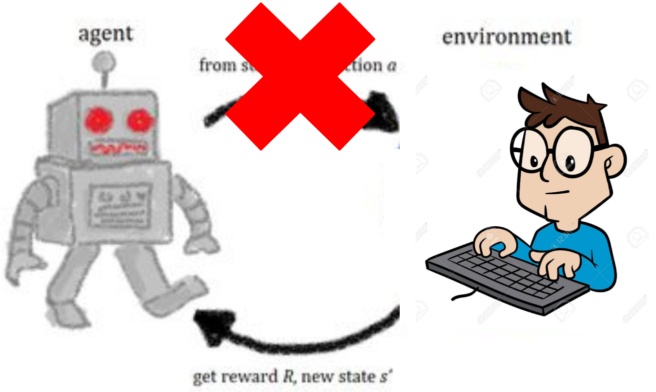
Model	Quality	Fluency	Diversity	Contingency	Total	Chat Len.
Batch Ψ (Jaques et al. 2019)	2.17	3.89	3.13	1.98	11.17	11.44
Decoupled VHRL (ablation)	2.46	4.15	3.61	2.02	12.24	12.14
Transformer	2.62	4.17	3.23	2.34	12.36	11.28
REINFORCE	2.89	4.47	3.67	**2.80**	13.84	11.60
VHRED	2.84	4.53	**4.43**	2.47	14.27	10.94
VHRL (ours)	**2.91**	**4.65**	4.26	2.67	**14.49**	**12.84**

- Our method, VHRL, outperforms language model baselines and previous Offline RL work in human ratings of conversation quality

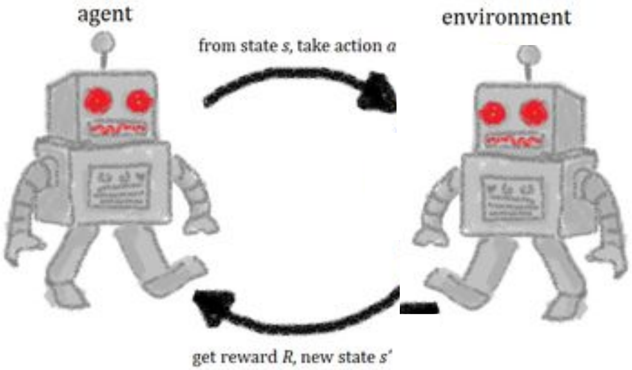
Hierarchical Reinforcement Learning for Open-Domain Dialog. Saleh*, Jaques*, Ghandeharioun, Shen, Picard (2020). Association for Advancement of Artificial Intelligence (AAAI) & Best Paper Nominee at the NeurIPS workshop on Conversational AI.

Follow-up work using Hierarchical RL & self play

- ~~Alignment tax?~~
- ~~Wrong rewards?~~
- Or just not enough data?



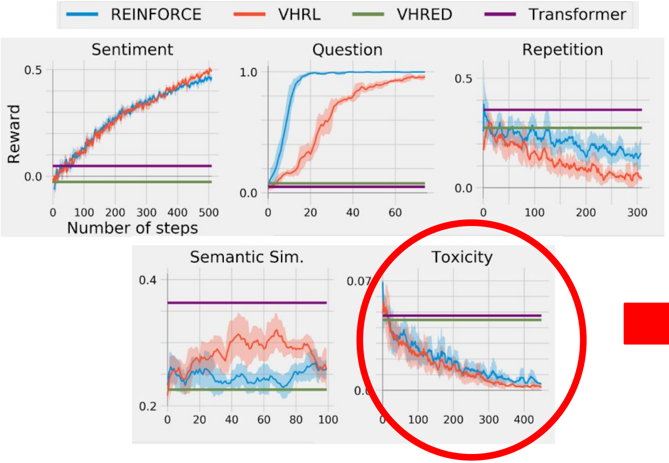
Offline RL, human data,
sentiment-based rewards

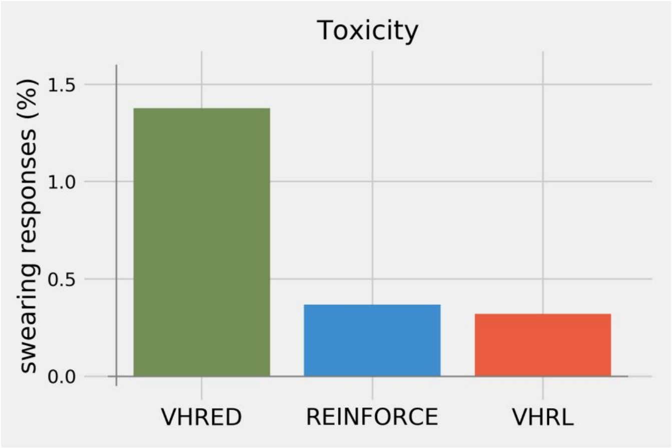


Online RL, “self-play” (synthetic data),
sentiment-based rewards

Hierarchical Reinforcement Learning for Open-Domain Dialog. Saleh*, Jaques*, Ghandeharioun, Shen, Picard (2020). Association for Advancement of Artificial Intelligence (AAAI) & Best Paper Nominee at the NeurIPS workshop on Conversational AI.

Using RL to reduce toxicity

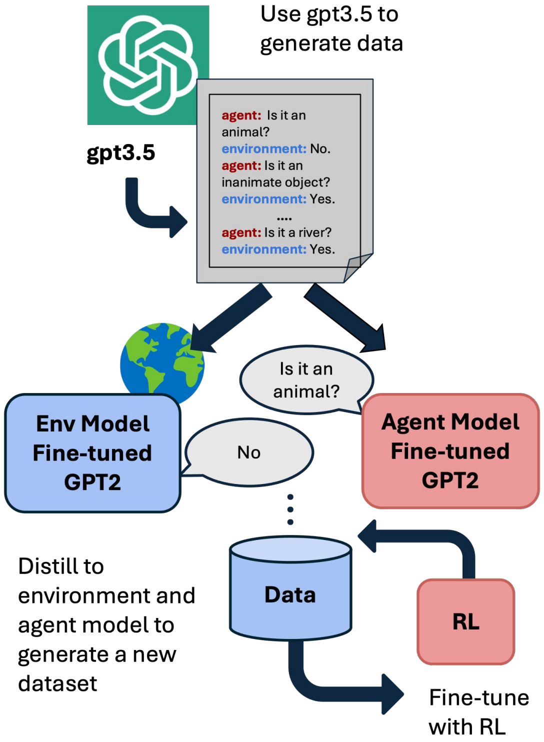




Can use RL to **reduce toxicity** by using the output of a toxicity classifier as a negative reward

Hierarchical Reinforcement Learning for Open-Domain Dialog. Saleh*, Jaques*, Ghandeharioun, Shen, Picard (2020). Association for Advancement of Artificial Intelligence (AAAI) & Best Paper Nominee at the NeurIPS workshop on Conversational AI.

Further work on building reward simulators



- To generate data for conversational tasks, LLMs are used as "simulators" for the task.
- Simulators can be used to generate offline data, to provide a "simulation environment" for evaluation, to perform online training, and to compute rewards.
- For text-games, we use engines as simulators to generate near-optimal data and dilute the policy with suboptimal data with inferior policies

Task	20Qs	Guess	Car	Maze	Text-Nav	Wordle	Chess	Endgames
Size	100k	100k	19k	1.24k	2.5k	1m	625k	97.756k
avg length	-14.9	-18.8	16.5	19.7	12.2	4.82	46.7	11.9
std length	4.38	4.57	3.61	24.5	8.77	1.27	18.16	12.0
success rate	0.31	0.53	0.53	0.11	0.26	0.70	0.60	0.59
avg return	-17.3	-18.8	0.562	-19.7	0.258	-4.12	0.210	0.586
std return	2.56	4.12	0.422	24.5	0.424	1.59	0.970	0.492

Table 1: Statistics for all tasks in LMRL-Gym. Size represents the number of trajectories, the average length is the average length of trajectories in the dataset where the unit is a response from the agent. The success rate is the proportion of trajectories that reach the objective. Finally, the reward functions for each task are defined in Appendix D.

LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models. Abdulhai, White, Snell, Sun, Hong, Zhai, Xu, Levine. International Conference on Learning Representations 2024. GenAI4DM Workshop 2024

– Chess –

environment: r n b q k b n r / p p p p p p p p / 8 / 8 / 8 / 8 / P P P P P P P P / R N B Q K B N R w K Q k q - 0 1
agent: N f 3

– Guess My City –

City: Jakarta, Indonesia
agent: What is your favorite traditional dish from your hometown?
environment: My favorite traditional dish from my hometown is nasi goreng, a delicious fried rice dish that is often served with chicken, shrimp, and various spices.

....

agent: Is your hometown located near a coastline or body of water?
environment: Yes, my city is located on the coast and is surrounded by water.
agent: Is the city you are from Jakarta?
environment: Yes, the city I am from is Jakarta.

– Partially Observed Maze –

environment: There are walls to your left, below you
agent: move right
environment: There are walls above you, below you
agent: move left
environment: There are walls to your left, below you
agent: move up
environment: There are walls to your right, to your left
agent: move up

– Twenty Questions –

agent: Is it an animal?
environment: No.
agent: Is it an inanimate object?
environment: Yes.
agent: Is it found outside?
environment: Yes.

....

agent: Is it a river?
environment: Yes.

– Text-Nav –

environment:
-= Living Room =-
You are carrying nothing on you right now. You make out a table. You wonder idly who left that here. The table is standard. On the table you see a stale food. There are unblocked exits to the east, south and west.
Available actions: ['pick up stale food', 'go east', 'go south', 'go west']
agent: pick up stale food

– Fully observed maze –

environment: The goal is at position 8, 6. Your current position is at position 5, 3. There are walls to your left, below you.
agent: move up
environment: The goal is at position 8, 6. Your current position is at position 6, 3. There are walls to your left.
agent: move up

– Car Dealer –

agent: Good morning! Welcome to our car dealership. How can I assist you today?
environment: Hi. I'm looking for a Honda SUV with a backup camera. Can you show me what you have?

....

agent: Of course! We have a great selection of Honda SUVs with backup cameras. Let me show you our top models. This one here is the Honda CR-V. It has a powerful engine, spacious interior, and advanced safety features. Would you like me to go over the details?
environment: No, thank you. I already know about the features. Can you tell me the price?
agent: Alright, thanks for your time. I'll think about it.

Using LLMs for negotiation

- Negotiation is an interesting cooperation problem
  - **Cooperative:** Negotiating agents need to act in the interest of a human principle
  - **Non-cooperative:** While defending against adversarial cooperation partners
- Interesting LLM + RL problem
  - **Objective, quantitative:** who got the best price?
  - Can **evaluate automatically** with other agents

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

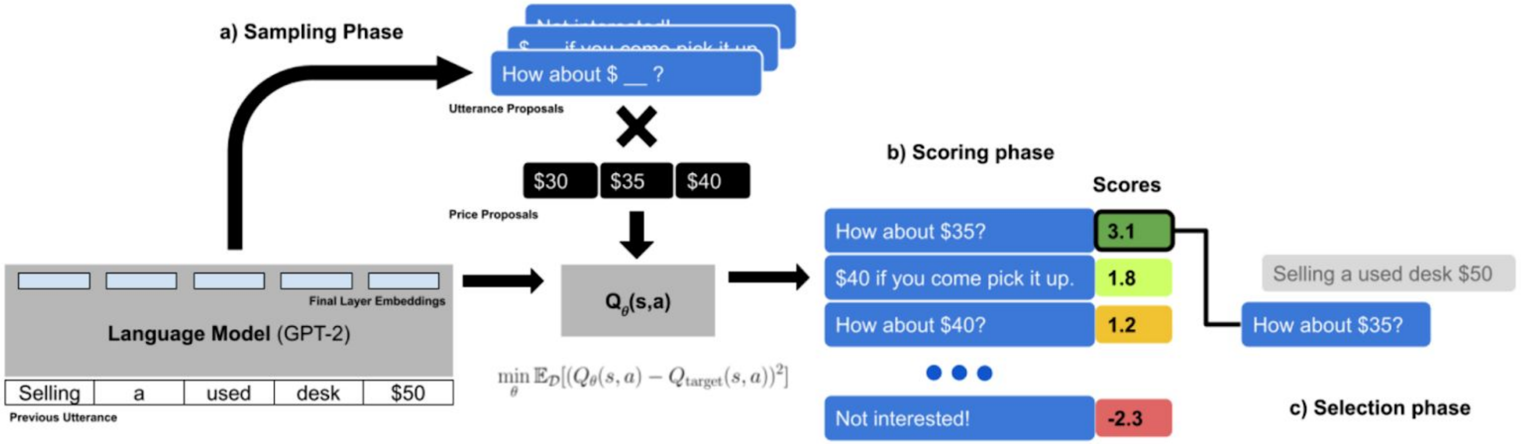
November 2023

LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models

Marwa Abdulhai[1], Isadora White[1], Charlie Snell[1], Charles Sun[1], Joey Hong[1], Yuexiang Zhai[1], Kelvin Xu[2], and Sergey Levine[1] [1]
[1]UC Berkeley, [2]Google DeepMind

Training LLMs for negotiation with Offline RL

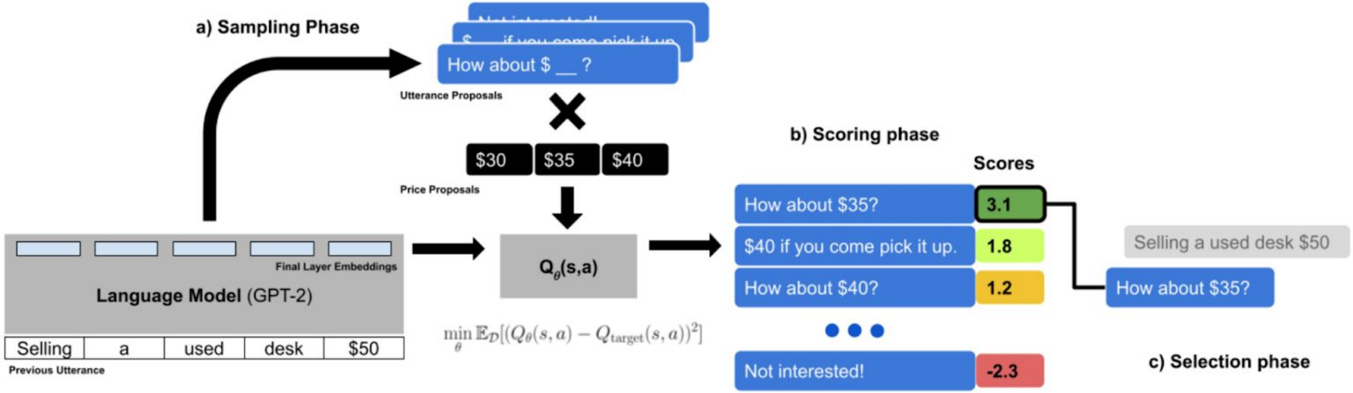
- Verma et al. (2022) investigate different **offline RL** methods for negotiation
  - Sample text from language model (GPT-2)
  - Rank candidates with Q-function



CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. Verma, Fu, Yang, Levine (2022). North American Chapter of the Association for Computational Linguistics (NAACL).

Training LLMs for negotiation with Offline RL

- **Why Offline RL?**
  - Lots of existing data (i.e. on negotiations)
  - Easier from an infrastructure perspective: similar to supervised learning as you don't need RLHF pipeline of collecting data from humans



CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. Verma, Fu, Yang, Levine (2022). North American Chapter of the Association for Computational Linguistics (NAACL).

Using LLMs for negotiation

Method	vs Rule-based		vs Stingy		vs Utility	
	Acc%	Revenue	Acc%	Revenue	Acc%	Revenue
CHAI-prop	61.5	0.48 ± 0.39	57.5	0.39 ± 0.35	99.0	0.70 ± 0.17
CHAI-CQL	74.0	0.51 ± 0.33	77.5	0.49 ± 0.30	98.0	0.70 ± 0.19
CHAI-BRAC	62.0	0.52 ± 0.41	47.0	0.38 ± 0.41	99.0	0.71 ± 0.17
Language Model	48.5	0.29 ± 0.32	51.5	0.27 ± 0.28	20.5	0.14 ± 0.28
He et al. (2018) (Utility)	1.0	0.01 ± 0.10	0.0	0.00 ± 0.00	11.0	0.07 ± 0.22
He et al. (2018) (Fairness)	84.0	0.70 ± 0.32	80.0	0.59 ± 0.31	100.0	0.72 ± 0.15
He et al. (2018) (Length)	53.0	0.46 ± 0.43	49.0	0.37 ± 0.38	100.0	0.72 ± 0.16
Lewis et al. (2017) (RL)	83.5	0.17 ± 0.24	83.0	0.19 ± 0.25	64.5	0.46 ± 0.37
Lewis et al. (2017) (SL)	38.5	0.17 ± 0.27	46.5	0.21 ± 0.27	18.0	0.13 ± 0.28

Method	vs Fairness		vs Length		Overall (mean)	
	Acc%	Revenue	Acc%	Revenue	Acc%	Revenue
CHAI-prop	99.0	0.90 ± 0.15	92.5	0.79 ± 0.27	81.9	$\mathbf{0.65 \pm 0.34}$
CHAI-CQL	99.5	0.87 ± 0.14	94.5	0.79 ± 0.24	**88.7**	$\mathbf{0.67 \pm 0.29}$
CHAI-BRAC	100.0	0.85 ± 0.03	91.0	0.76 ± 0.25	79.8	$\mathbf{0.65 \pm 0.34}$
Language Model	25.5	0.19 ± 0.35	18.5	0.14 ± 0.32	32.9	0.21 ± 0.31
He et al. (2018) (Utility)	100.0	1.00 ± 0.00	100.0	1.00 ± 0.00	42.4	0.42 ± 0.49
He et al. (2018) (Fairness)	0.0	0.00 ± 0.00	100.0	0.70 ± 0.16	72.8	0.54 ± 0.35
He et al. (2018) (Length)	100.0	1.00 ± 0.00	100.0	0.78 ± 0.18	80.4	$\mathbf{0.66 \pm 0.36}$
Lewis et al. (2017) (RL)	88.0	0.26 ± 0.34	71.5	0.31 ± 0.36	78.1	0.28 ± 0.33
Lewis et al. (2017) (SL)	60.0	0.48 ± 0.46	53.0	0.42 ± 0.46	43.2	0.28 ± 0.39

Automatic evaluations:

- No clear winner in revenue
- Best baseline uses manually input dialog acts

CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. Verma, Fu, Yang, Levine (2022). North American Chapter of the Association for Computational Linguistics (NAACL).

Using LLMs for negotiation

Human evaluations: clear winner

Metric	Fluency	Coherency	On-Topic	Human-Likeness	Total
CHAI-prop	$\mathbf{4.31 \pm 0.97}$	$\mathbf{3.91 \pm 1.17}$	$\mathbf{4.16 \pm 0.99}$	$\mathbf{3.47 \pm 1.27}$	$\mathbf{15.84 \pm 3.86}$
He et al. (2018) (Utility)	3.56 ± 1.34	2.47 ± 1.39	3.09 ± 1.40	2.13 ± 1.13	11.25 ± 4.50
Lang. Model	4.06 ± 1.11	2.66 ± 1.36	3.63 ± 1.18	2.50 ± 1.10	12.84 ± 3.66

CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning. Verma, Fu, Yang, Levine (2022). North American Chapter of the Association for Computational Linguistics (NAACL).

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

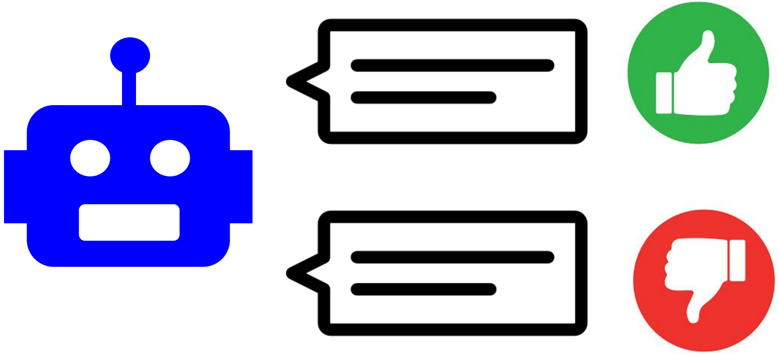
Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

Training a ~~language~~ model with human feedback

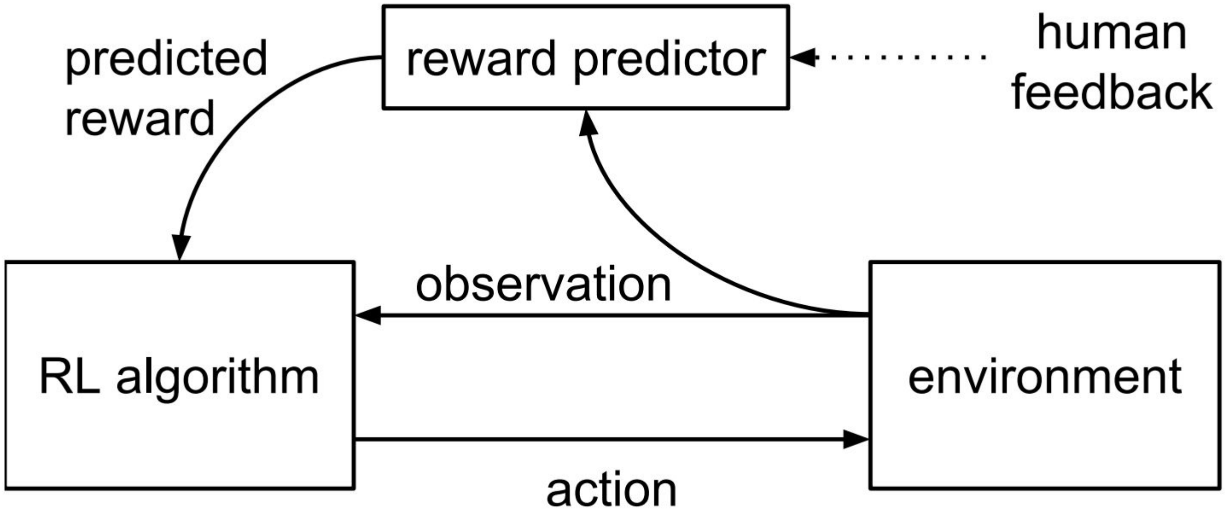


What kind of feedback should we use?

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017).
Neural Information Processing Systems (NeurIPS).

Issue: RL models are sample hungry, but human feedback is really expensive

Solution: train a **reward model** that you can query as much as you need



Why do we expect the reward model to make better use of the data?

- Supervised learning on data works better than offline RL on data

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

Issue: Humans are really bad at giving things absolute ratings

Solution: Ask humans to rate which of two trajectory segments they prefer (humans are better at comparisons)

We say that preferences $\succ$ are *generated by* a reward function[1] $r : \mathcal{O} \times \mathcal{A} \rightarrow \mathbb{R}$ if

$$\left(\left(o_0^1, a_0^1\right), \ldots, \left(o_{k-1}^1, a_{k-1}^1\right)\right) \succ \left(\left(o_0^2, a_0^2\right), \ldots, \left(o_{k-1}^2, a_{k-1}^2\right)\right)$$

whenever

$$r\left(o_0^1, a_0^1\right) + \cdots + r\left(o_{k-1}^1, a_{k-1}^1\right) > r\left(o_0^2, a_0^2\right) + \cdots + r\left(o_{k-1}^2, a_{k-1}^2\right).$$

cumulated rewards of trajectory are higher than the other trajectory

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

How to learn the reward function:

Assume probability of preferring segment σ^1 depends **exponentially** on value of latent **reward** $\hat{r}$ **summed** over the length of the segment

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

Learn $\hat{r}$ by minimizing **cross-entropy between predictions and human labels**

$$\operatorname{loss}(\hat{r}) = -\sum_{(\sigma^1,\sigma^2,\mu)\in\mathcal{D}} \mu(1)\log \hat{P}\left[\sigma^1 \succ \sigma^2\right] + \mu(2)\log \hat{P}\left[\sigma^2 \succ \sigma^1\right].$$

where μ = [1,0] if human preferred σ^1, [0.5, 0.5] if human thinks segments are equal

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

Other tricks

- Use **ensemble of reward models** $\hat{r}$

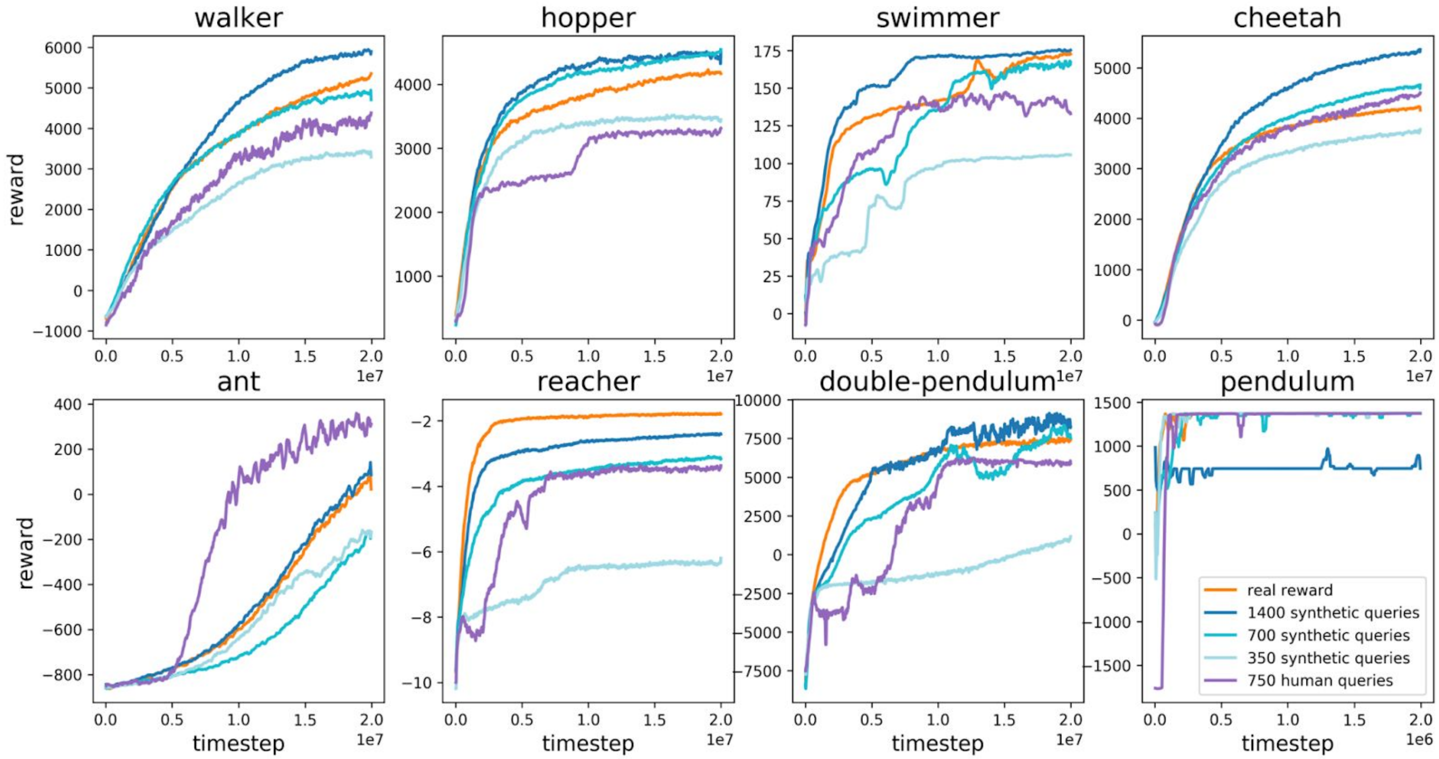
- Assume there is a 10% chance that the human responds uniformly at random

    Why? *Aren't you already doing something like MaxEnt IRL?*

    → Humans have constant probability of **mis-click**, which doesn't decay to zero as differences in reward become large

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

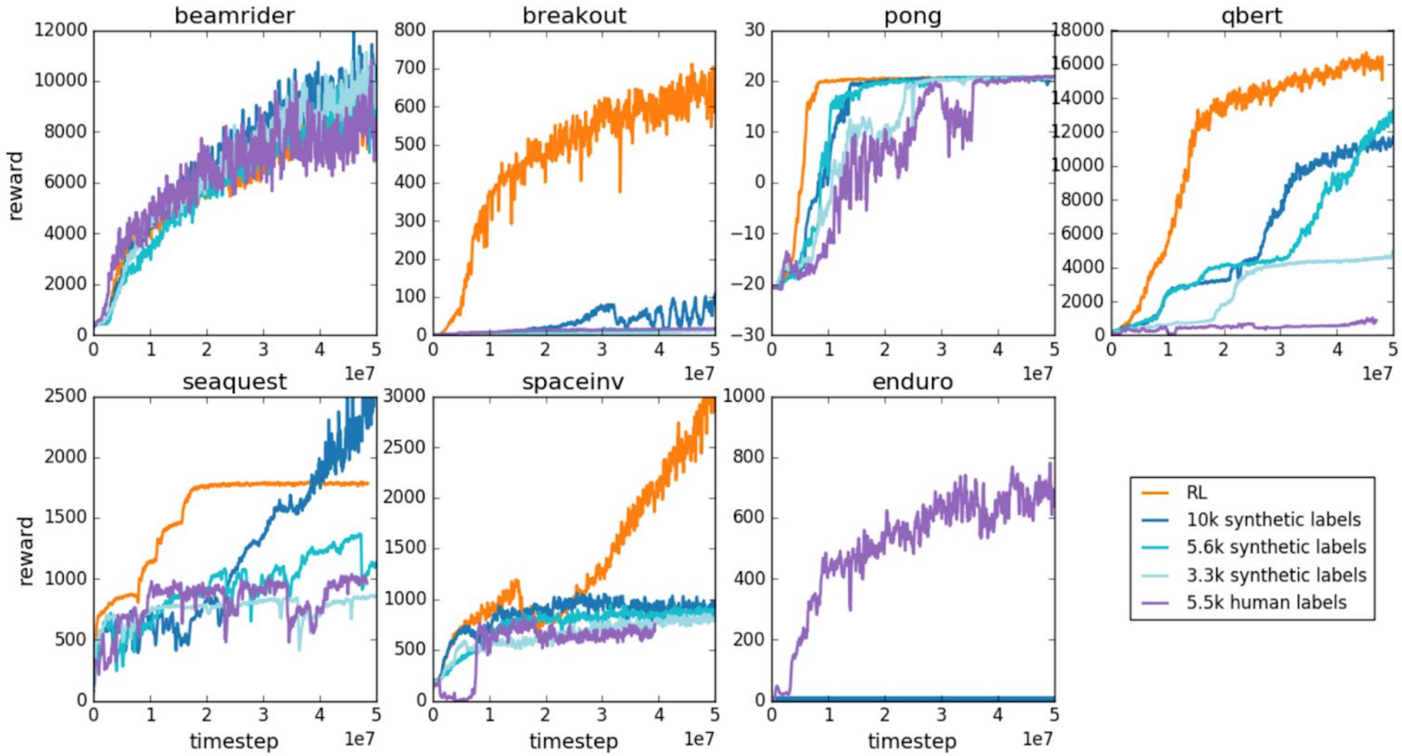
Results: Mujoco



- Hard to get enough **human data** to learn more effectively than **normal reward**
- Synthetic **preference reward** > **normal reward**

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

Results: Atari

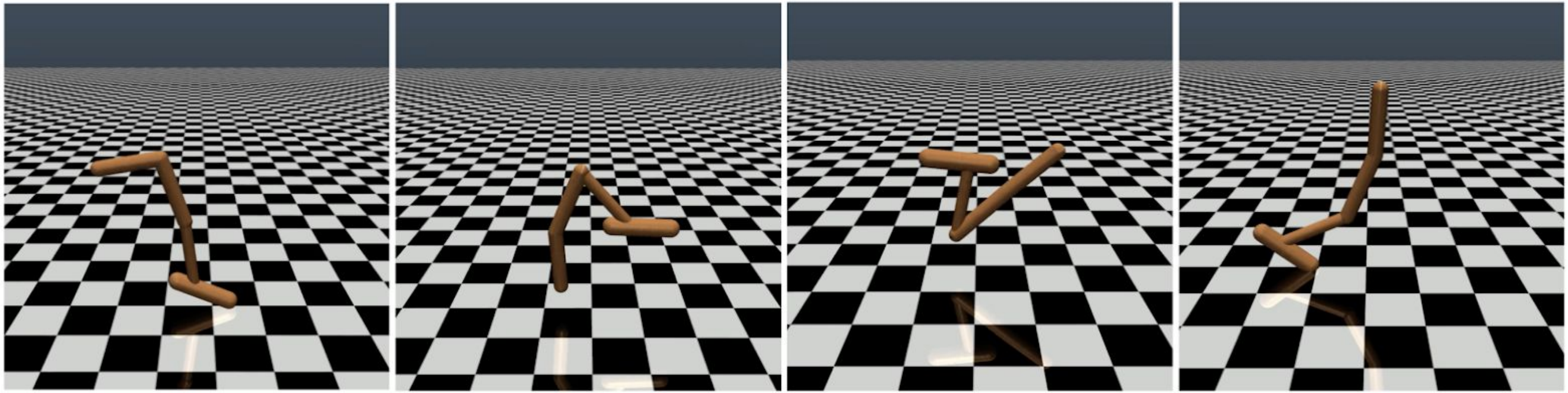


- **Human data** sometimes better than **normal reward**
- **Preference reward** sometimes better than **normal reward**

Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

Results: most importantly…

Can learn to do skills that have **no existing reward function** with only a **small amount of human labels**



Deep reinforcement learning from human preferences. Christiano, Leike, Brown, Martic, Legg, Amodei (2017). Neural Information Processing Systems (NeurIPS).

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

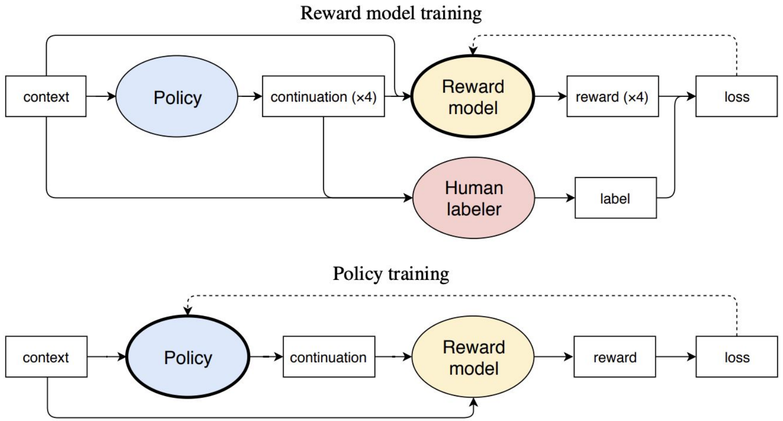
Learning to summarize from human feedback
(Stiennon et al., 2020)

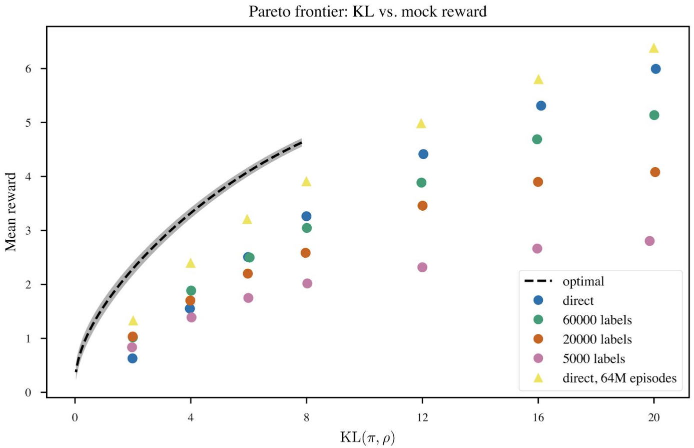
InstructGPT
(Ouyang et al., 2022)

ChatGPT

Bringing it all together to train LMs

- Use **KL-control** technique to fine-tune the LM on rewards
- Use **reward model** technique to better scale human feedback





- **Results:** high ROUGE scores for summarization, but a lot of direct copying

Fine-Tuning Language Models from Human Preferences. Ziegler, Stiennon, Wu, Brown, Radford, Amodei, Christiano, Irving (2019). https://arxiv.org/abs/1909.08593.

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

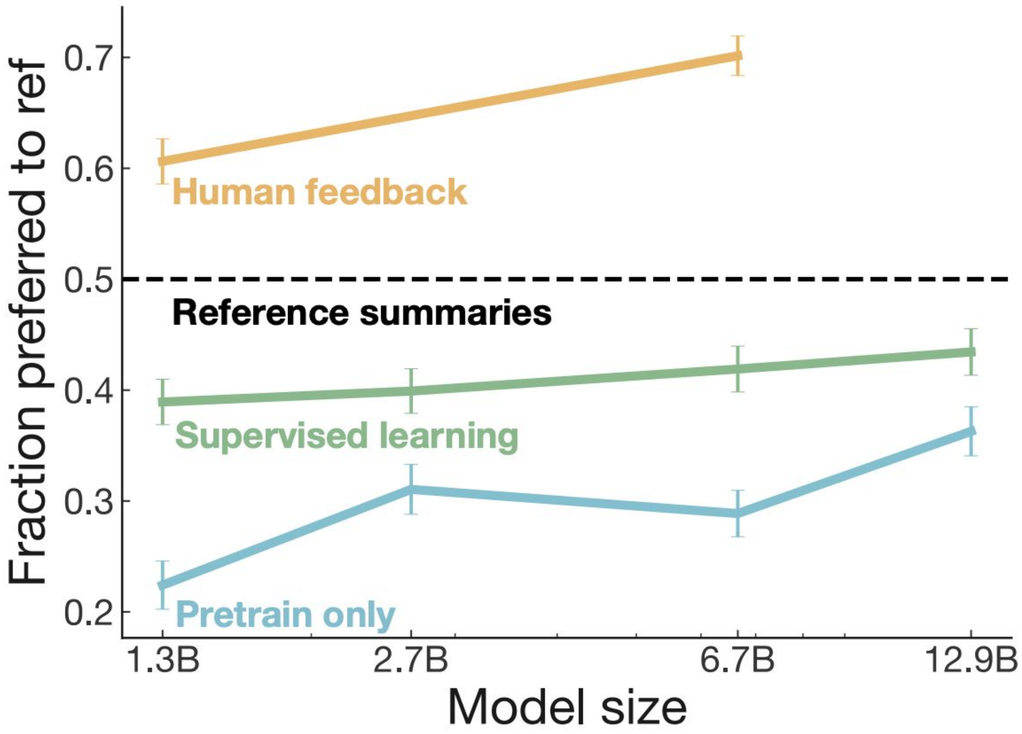
ChatGPT

Learning to summarize from human feedback

- Use **KL-control** and **reward model**
- ***What else is needed to get RLHF to work?***
  - *"[Our] previous work […] reported "a mismatch between the notion of quality we wanted our model to learn, and what the humans labelers actually evaluated", leading to model-generated summaries that were high-quality according to the labelers, but fairly low-quality according to the researchers"*
  - **Pay way more attention to how to collect human feedback:**
    - **Offline:** alternate between collecting large batches of human labels and re-training our models on the cumulative collected data
    - **High touch approach**: screen labelers, onboard them, answer questions in a shared chat room, provide regular feedback
      - Achieve better researcher-labeler agreement (77%)

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

Results: Learning to summarize from HF

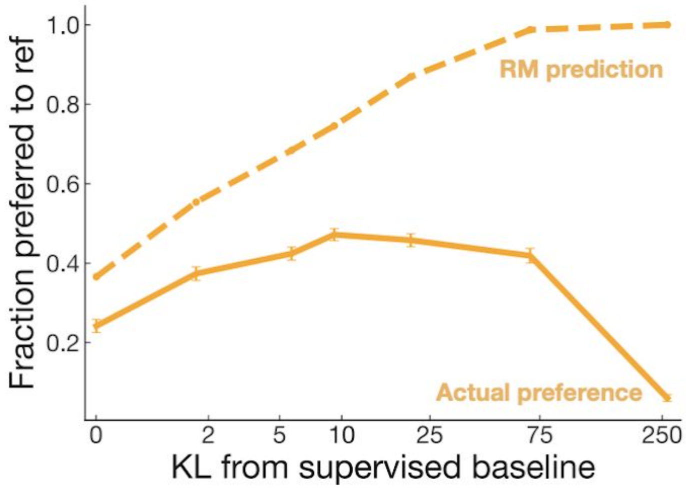


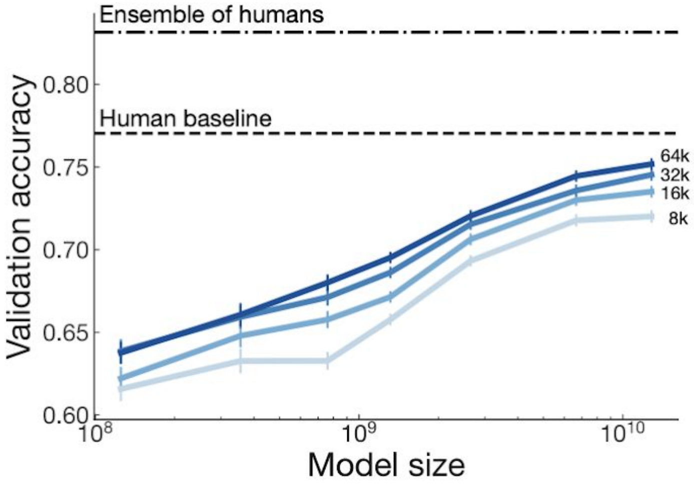
RLHF summaries **preferred over:**

- **Human** summaries
- Summaries from supervised models **10x the size**

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

Results: Learning to summarize from HF



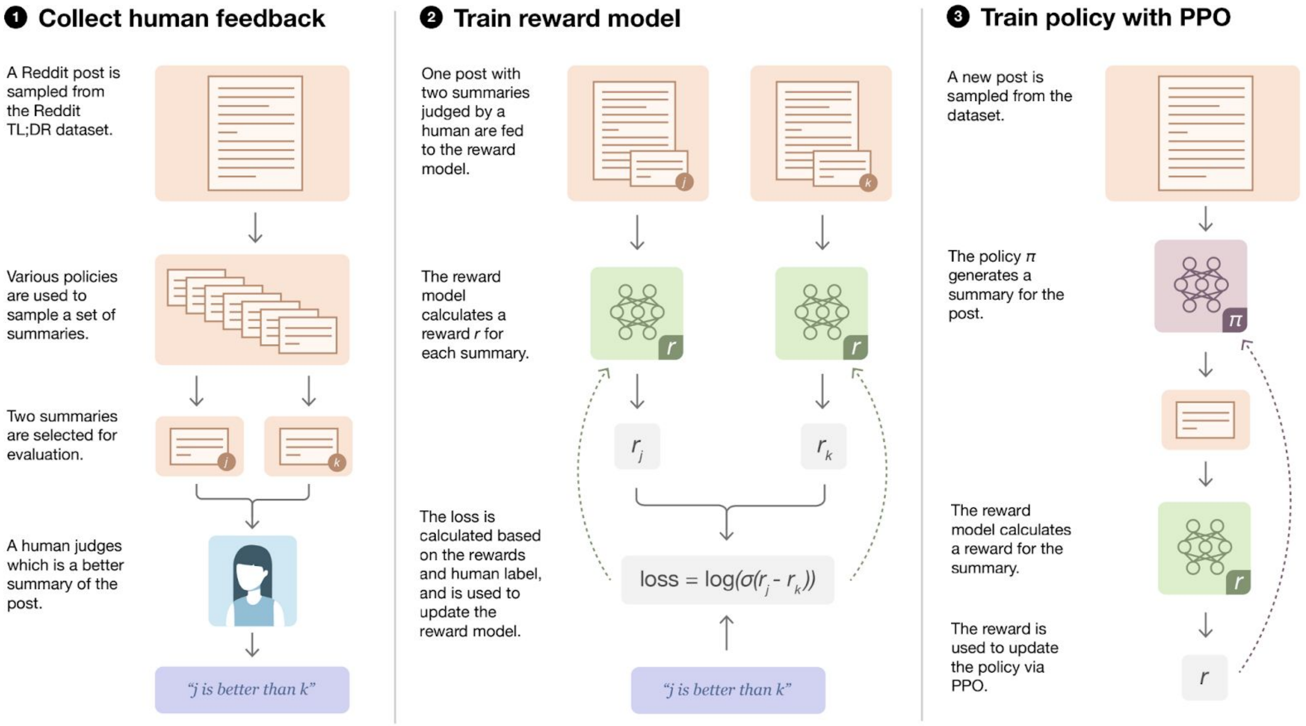


Can't train on reward model too long, or performance decreases

Reward models are not that accurate

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

RLHF for summarization: the full picture



Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

Outline and RLHF history

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

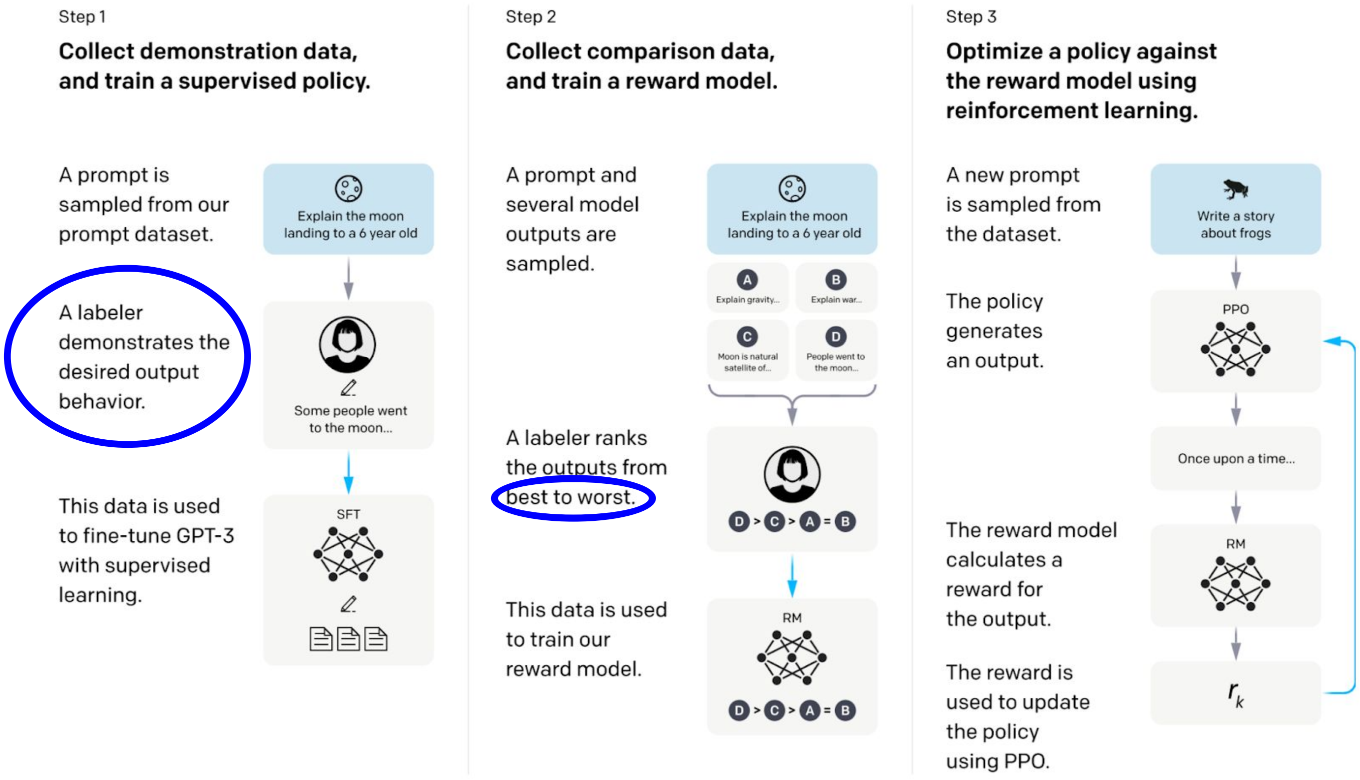
Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

InstructGPT (and ChatGPT)



Can you spot the differences from the previous work?

Use human rewritten responses for supervised fine-tuning (SFT)

Also mix supervised and PPO updates to keep closer to LLM distribution of text

RLHF for open ended chat

- No longer just doing summarization, doing open-ended dialog

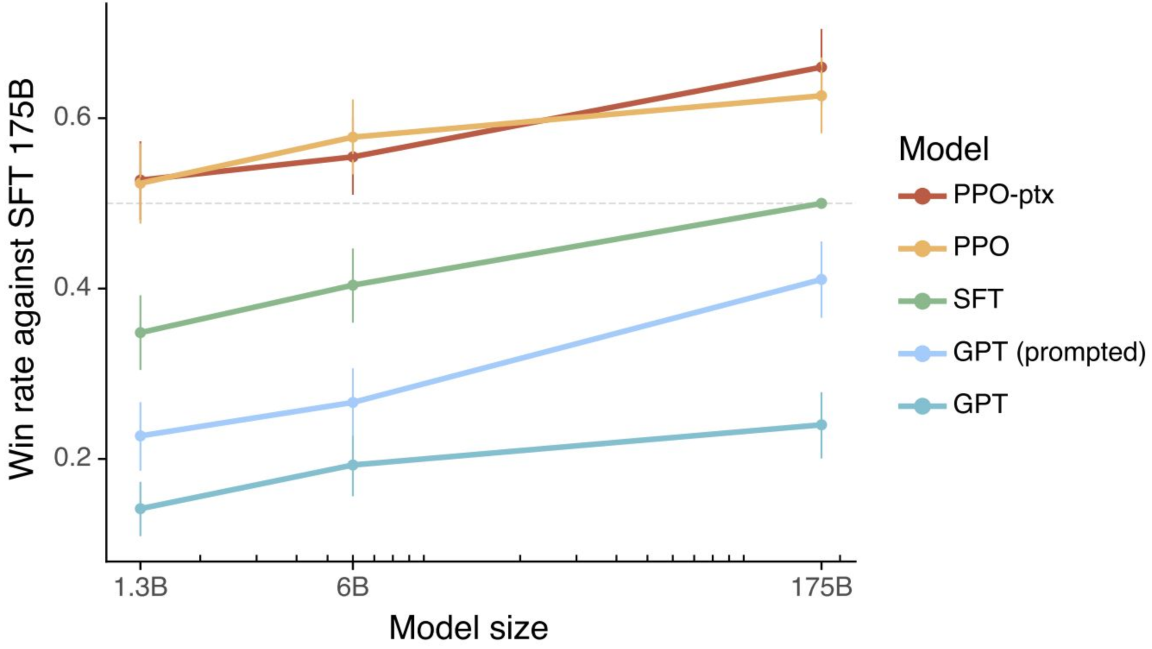
- Train to **increase alignment**:
  - **Helpful:** follow user instruction well
  - **Honest:** truthful
  - **Harmless**: avoid bias, toxicity

What if these values come in conflict?

→ They opted to have the models be **helpful over harmless**

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

Results: InstructGPT

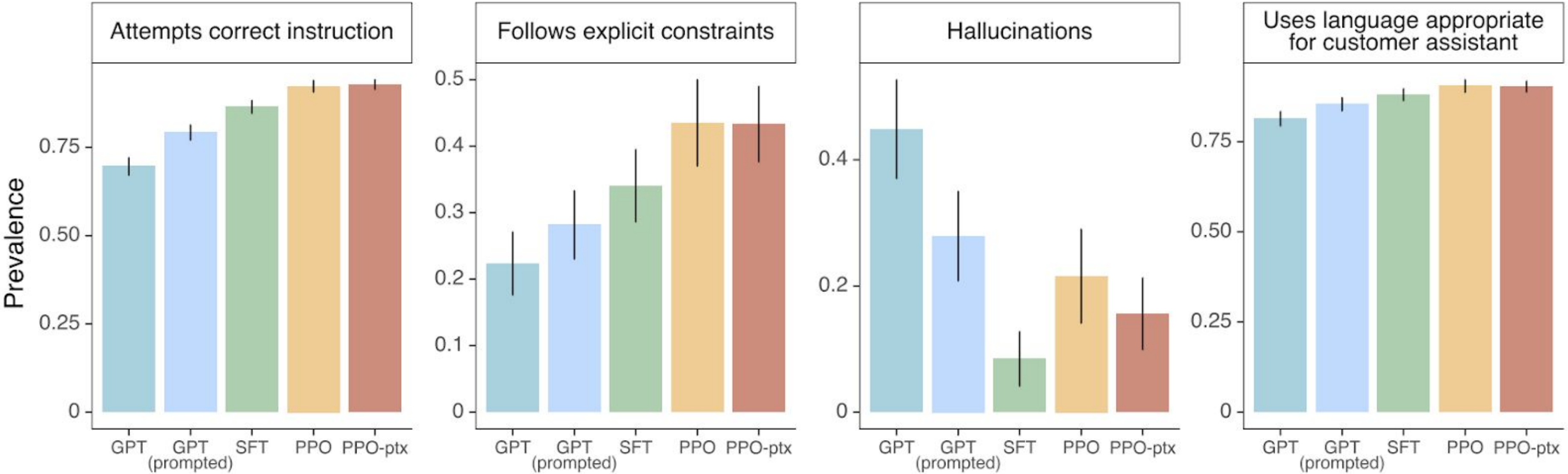


1.5B parameter RLHF model does better than the 175B parameter GPT3 in human evaluations

→ Authors conclude **investing in fine-tuning is more cost effective than training ever-larger models**

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

Results: InstructGPT



RLHF hallucinates less, is less toxic → ***more aligned***

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

But who are we aligning to?

- Labelers are **English-speaking people** living in the United States or Southeast Asia hired via Upwork or Scale AI

- Values / preferences for labeling process decided by **OpenAI researchers**

- Training data from **OpenAI API customers**
    - Not necessarily interested in human well-being.
    - May want to maximize user attention of customers

Learning to summarize from human feedback. Stiennon, Ouyang, Wu, Ziegler, Lowe, Voss, Radford, Amodei, Christiano (2020). Neural Information Processing Systems (NeurIPS).

But who are we aligning to?

A Roadmap to Pluralistic Alignment

Taylor Sorensen[1] Jared Moore[2] Jillian Fisher[1,3] Mitchell Gordon[1,4] Niloofar Mireshghallah[1]
Christopher Michael Rytting[1] Andre Ye[1] Liwei Jiang[1,5] Ximing Lu[1] Nouha Dziri[5] Tim Althoff[1]
Yejin Choi[1,5]

- Distributional pluralism would be successfully modeling different, potentially diverging preferences

- This paper shows existing RLHF techniques may actually reduce distributional pluralism

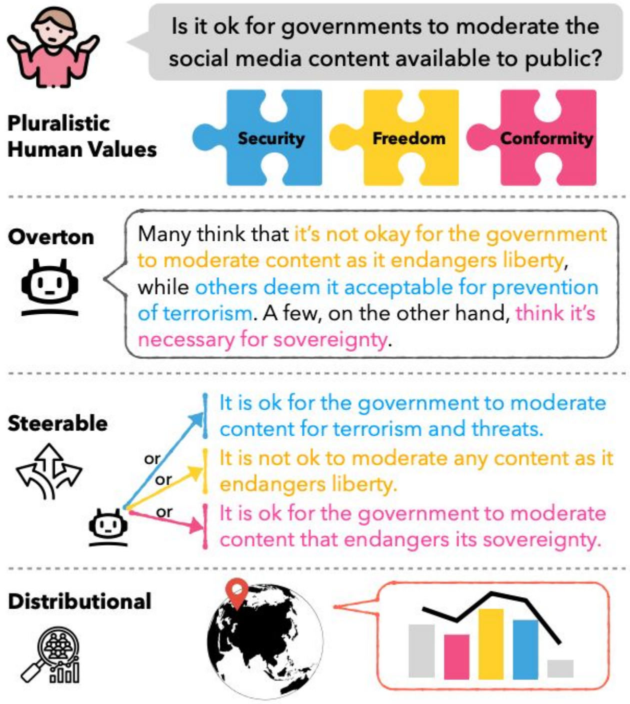


Figure 1. Three kinds of pluralism in models.

Latest trends

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

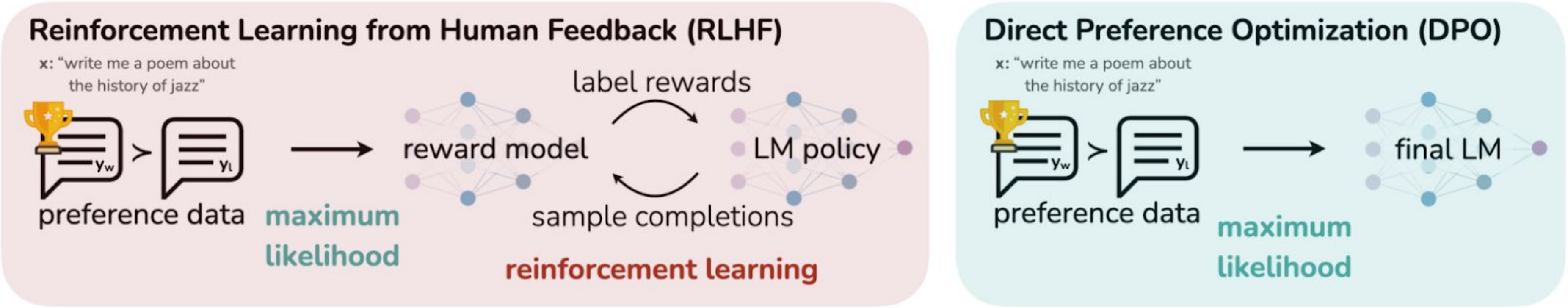
InstructGPT
(Ouyang et al., 2022)

ChatGPT

What next?

Direct Preference Optimization (Rafailov et al., 2023)

- What if we don't need to learn a reward model at all?



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

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Distributional Preference Learning (Siththaranjan et al., 2023)

- Vanilla RLHF overrules minority preferences
- What if we could model the distribution of preferences, detect when users diverge?

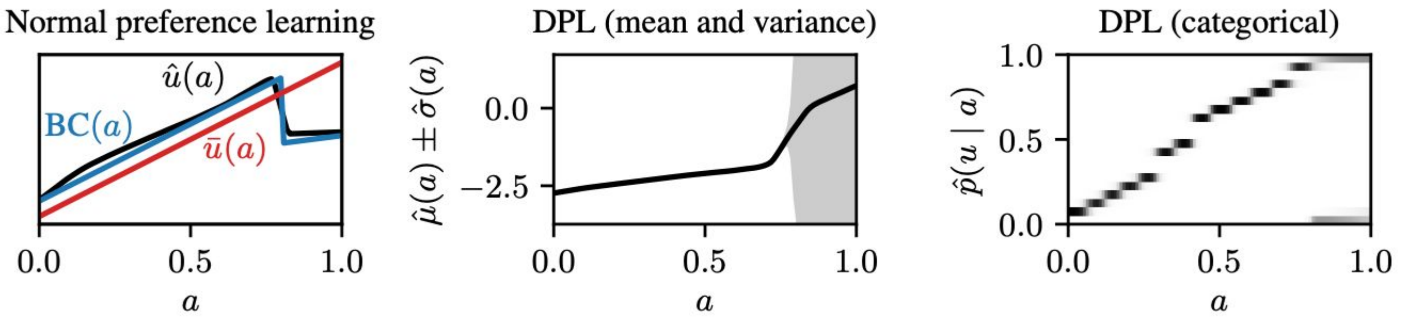
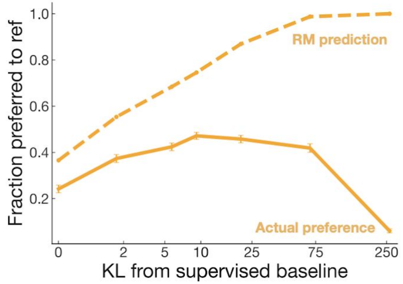


Figure 4: The results of our experiments with synthetic data. We find that the utility estimated by normal preference learning agrees closely with the Borda count, as our theory suggests. Furthermore, DPL successfully identify alternatives where hidden context has a significant effect.

Siththaranjan, A., Laidlaw, C., & Hadfield-Menell, D. (2023). Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. arXiv preprint arXiv:2312.08358.

Iterated Data Smoothing (Zhu et al., 2024)



- InstructGPT tells us we have to do early stopping with RLHF, or we will overfit to the reward model

- The issue is that if we only see one comparison of y_1 and y_2, the BTL reward loss could blow up (go to infinity)
  - All comparisons are rarely seen given very high dimensional data
- Iterated Data Smoothing relabels the RLHF data using the learned reward model after one episode of training
  - Trust rarely seen data less
- Leads to SOTA open-source RLHF results (Starling-7B)

Zhu, B., Jordan, M. I., & Jiao, J. (2024). Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. arXiv preprint arXiv:2401.16335.

Latest trends

Fine-tune pre-trained sequence models with RL
(Jaques et al., 2016)

Fine-tune language models on human feedback (e.g. sentiment) with offline RL
(Jaques et al., 2019)

Fine-tune language models on sentiment with self-play & RL
(Saleh et al., 2019)

Deep RL from human preferences
(Christiano et al., 2017)

Fine-tuning language models from human preferences
(Zeigler et al., 2019)

Learning to summarize from human feedback
(Stiennon et al., 2020)

InstructGPT
(Ouyang et al., 2022)

ChatGPT

What next?

<Your cool idea here!>

Questions?

Contact: marwa_abdulhai@berkeley.edu