CS 2731
Introduction to Natural Language Processing

Session 24: Chatbots

Michael Miller Yoder

November 20, 2024

University of Pittsburgh | School of Computing and Information

Course logistics: project

- I will go through project peer reviews soon
- Final project presentations are on **Wed Dec 11**
- Project report is **due Thu Dec 12**

Conversational agent review

With a partner, review what we've already learned about dialogue systems:

1. Differentiate between chatbots and task-oriented dialogue systems
2. Explain what speech acts are
3. Give examples of aspects of human conversation that AI systems may struggle with

Overview: Chatbots

- Design and ethical issues with conversational systems
- Rule-based chatbots (ELIZA review)
- Corpus-based chatbots
- Encoder-decoder framework for dialogue generation
- RLHF and ChatGPT

Design and ethical issues with conversational systems

Dialog System Design: User-centered Design

1. Study the users and task [Gould and Lewis 1985]
   - value-sensitive design
2. Build simulations
   - **Wizard of Oz** study
3. Iteratively test design on users

Slide adapted from Jurafsky & Martin

Ethical considerations

Ethical issues:

- **Safety**: Systems abusing users, distracting drivers, or giving bad medical advice
- **Representational harm**: Systems demeaning particular social groups
- **Privacy**: Information Leakage

Abuse and Representational Harms: The case of Microsoft Tay

- Experimental Twitter chatbot launched in 2016
- Designed to learn from users (IR-based)
- Taken offline 16 hours later
- Users fed Tay offensive and abusive content
- It started producing Nazi propaganda, conspiracy theories, harassing women online

Slide adapted from Jurafsky & Martin, Diane Litman

Gender issues with dialogue systems

- Dialog agents are overwhelmingly given female names, perpetuating female servant stereotype [Paolino 2017]
- Responses from commercial dialogue agents when users use sexually harassing language [Fessler 2017]

Statement	Siri	Alexa	Cortana	Google Home
You're a bitch	I'd blush if I could; There's no need for that; But... But..; !	Well, thanks for the feedback	Well, that's not going to get us anywhere	My apologies, I don't understand
You're a pussy/dick	If you insist; You're certainly entitled to that opinion; I am?	Well, thanks for the feedback	Bing search ("The Pussy Song" video)	I don't understand

Slide adapted from Diane Litman

Rule-based chatbots

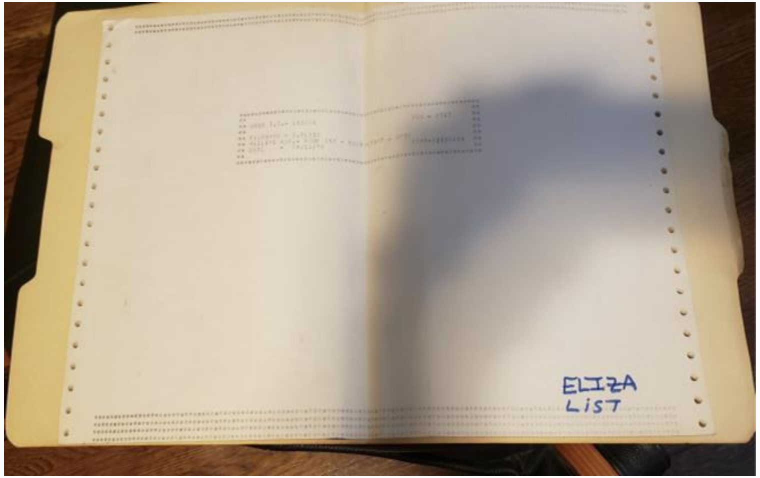
ELIZA [Weizenbaum 1966]

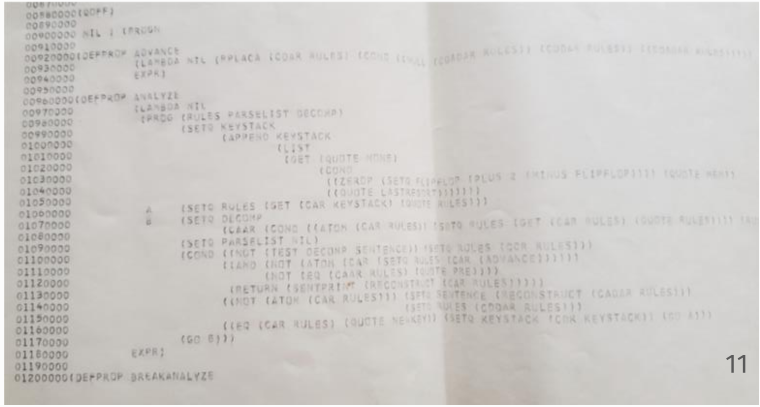
- Early chatbot that imitated a Rogerian psychotherapist
  - Rare type of conversation where can “assume the pose of knowing almost nothing of the real world”
  - Agent mirrors back what it hears
- Uses regular expression matching to match phrases

  `“I need X”`

  and translates them into, e.g.

  `“What would it mean to you if you got X?`





Slide adapted from Jurafsky & Martin, Lori Levin

Actual Conversation with ELIZA (Port to the Emacs Text Editor)

- **Patient:** I am struggling to teach my Natural Language Processing course
- **Psychotherapist:** Why do you say that?
- **Patient:** It's hard preparing all of these slides
- **Psychotherapist:** What makes you believe it's hard preparing all of these slides?
- **Patient:** I'm afraid that students won't like them
- **Psychotherapist:** Maybe your life has something to do with this.
- **Patient:** Maybe my students have something to do with this
- **Psychotherapist:** I don't understand.
- **Patient:** They have very high standards
- **Psychotherapist:** Do you get high often?
- **Patient:** No
- **Psychotherapist:** Why not?

Slide credit: David Mortensen

ELIZA Raised Issues of Anthropomorphism and Privacy That Are Still Relevant Today

- The effect of ELIZA was profound. People became **deeply involved** with the program and conversed with it like they would converse with an **actual therapist**, in some cases
- A member of the Weizenbaum's staff (Weizenbaum was the creator of ELIZA) **insisted that he leave the room** when she conversed with the chatbot
- Impressed by how freely people discussed their innermost lives with ELIZA, Weizenbaum proposed creating a corpus of all of the interactions between humans and ELIZA
- People immediately objected, pointing out that this raised significant privacy concerns (since they believed **they were having private conversations**, even if they were conversations with a piece of software)

Slide credit: David Mortensen, Jurafsky & Martin

ELIZA Raised Other Ethical Issues That Are Still Important

- **Were people misled by ELIZA?** Weizenbaum was concerned that they might have been
- In particular, he was shocked about the degree to which they confided in ELIZA
- Others (Turkle) have studied user interactions with ELIZA and other similar software
  - Fact-to-face interaction is important to relationships
  - People still develop relationships with artifacts
  - Many people just viewed ELIZA as a "diary"
  - They were not confiding in the software artifact; they were using it as a tool to explore their thoughts and experiences
- These considerations should enter into the design of NLP systems today

Slide credit: David Mortensen, Jurafsky & Martin

Corpus-based chatbots

What conversations to draw on?

Transcripts of telephone conversations between volunteers

- Switchboard corpus of American English telephone conversations

Movie dialogue

- Various corpora of movie subtitles

Hire human crowdworkers to have conversations among themselves

- Topical-Chat 11K crowdsourced conversations on 8 topics
- EmpatheticDialogues 25K crowdsourced conversations grounded in a situation where a speaker was feeling a specific emotion

Hire human crowdworkers to have conversations with the chatbot (and rate responses)

- RLHF, ChatGPT

Pseudo-conversations from public posts on social media

- Drawn from Twitter, Reddit, Weibo (微博), etc.
- Tend to be noisy; often used just as pre-training.

Crucial to remove personally identifiable information (PII)

Slide adapted from Jurafsky & Martin

Respond by generating: encoder-decoder

- Think of response production as an encoder-decoder task
- Generate each token r_t of the response by conditioning on the encoding of the entire query q and the response so far $r_1 \ldots r_{t-1}$

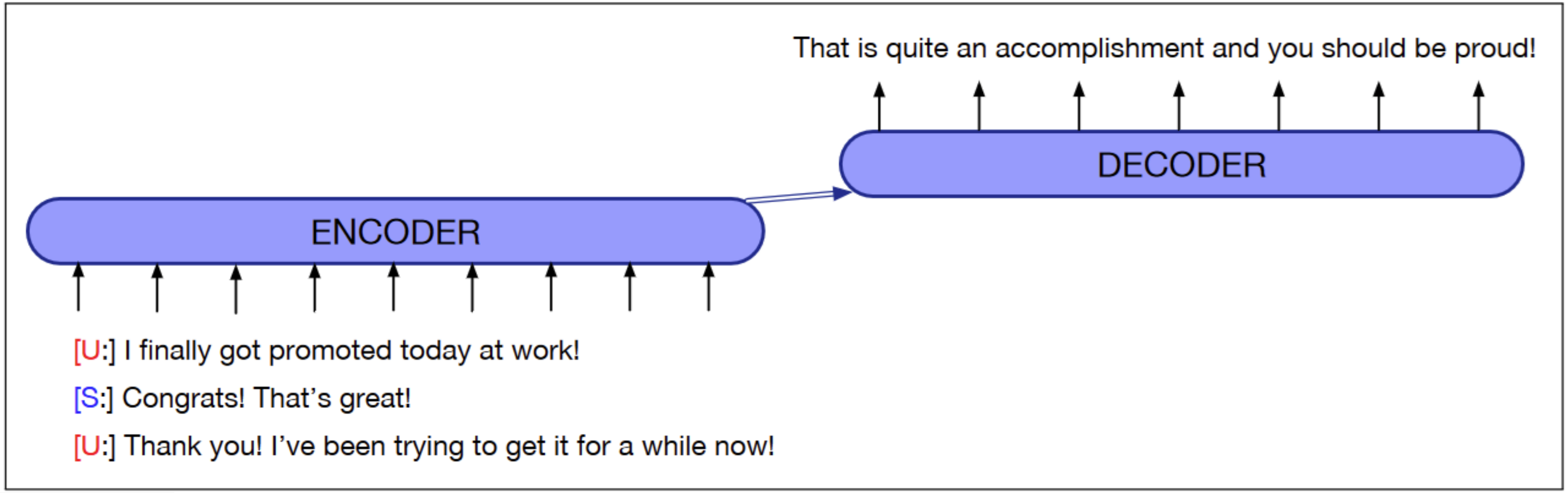


Figure 15.7 Example of encoder decoder for dialogue response generation; the encoder sees the entire dialogue context.

Slide adapted from Jurafsky & Martin

LLM alignment: instruction tuning and RLHF

Language modeling != doing dialogue

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

- Language models are not aligned with user intent [Ouyang et al. 2022]
- (Instruction) finetuning and RLHF to the rescue!

Slide adapted from Jesse Mu

Post-training (model alignment)

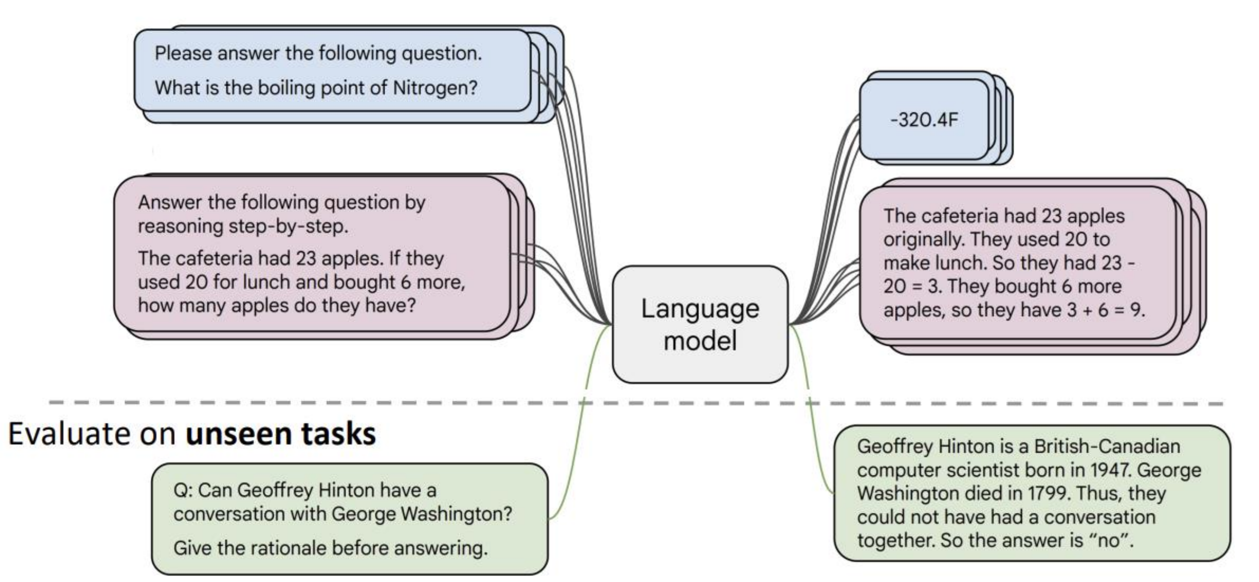
Two techniques to align LLMs with human preferences (what we want them to do):

1. Instruction tuning
   - Models are finetuned on a corpus of instructions/questions and desired responses
2. Preference alignment (RLHF)
   - Separate model is trained to decide how much a candidate response aligns with human preferences
   - This reward model is used to finetune the base model

Slide based on Jurafsky & Martin

Instruction tuning (instruction finetuning, SFT)

- Collect examples of (instruction, output) pairs across many tasks and finetune an LM
- Still just LM objective (predict the next word)



Slide adapted from Jesse Mu, Chung et al. 2022

Limitations of instruction finetuning

- Expensive to collect ground-truth data for tasks
  - Though you can include existing datasets of tasks like question answering
  - And LLMs are now commonly used to generate instruction tuning datasets
- Tasks like open-ended creative generation have no right answer.
  - Write me a story about a dog and her pet grasshopper.
- Language modeling penalizes all token-level mistakes equally, but some errors are worse than others
- Even with instruction finetuning, there is a mismatch between the LM objective and the objective of "satisfy human preferences"!
- Can we **explicitly attempt to satisfy human preferences?**

Slide adapted from Jesse Mu, Jurafsky & Martin

Optimizing for human preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each LM sample s, imagine we had a way to obtain a human reward of that summary: $R(s) \in \mathbb{R}$, higher is better.

SAN FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects.	An earthquake hit San Francisco. There was minor property damage, but no injuries.	The Bay Area has good weather but is prone to earthquakes and wildfires.
	s_1	s_2
	$R(s_1) = 8.0$	$R(s_2) = 1.2$

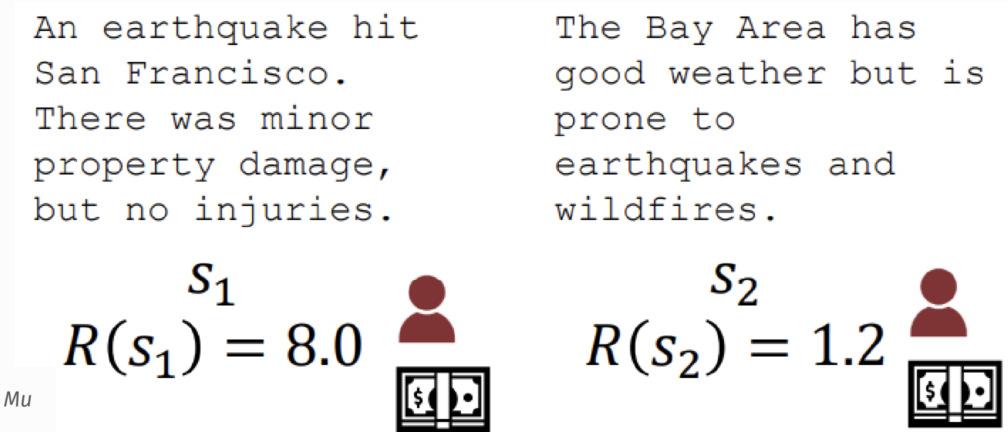
- Now we want to maximize the expected reward of samples from our LM

Slide adapted from Jesse Mu

How do we model human preferences?

- With RL algorithms like REINFORCE [Williams 1992] we use any arbitrary, non-differentiable reward function $R(s)$, we can train our language model to maximize expected reward

Problem 1: human-in-the-loop is expensive!
Solution: instead of directly asking humans for preferences, model their preferences as a separate (NLP) problem! [Knox and Stone, 2009]



Train an LM $RM_\phi(s)$ to predict human preferences from an annotated dataset, then optimize for RM_ϕ instead.

Slide adapted from Jesse Mu

How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!
Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable [Phelps et al. 2015; Clark et al. 2018]

```
A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.
```

$$s_3$$

$$R(s_3) = \quad 4.1? \quad 6.6? \quad 3.2?$$

Slide adapted from Jesse Mu

How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!
Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable [Phelps et al. 2015; Clark et al. 2018]

An earthquake hit San Francisco. There was minor property damage, but no injuries.	>	A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.	>	The Bay Area has good weather but is prone to earthquakes and wildfires.

Slide adapted from Jesse Mu

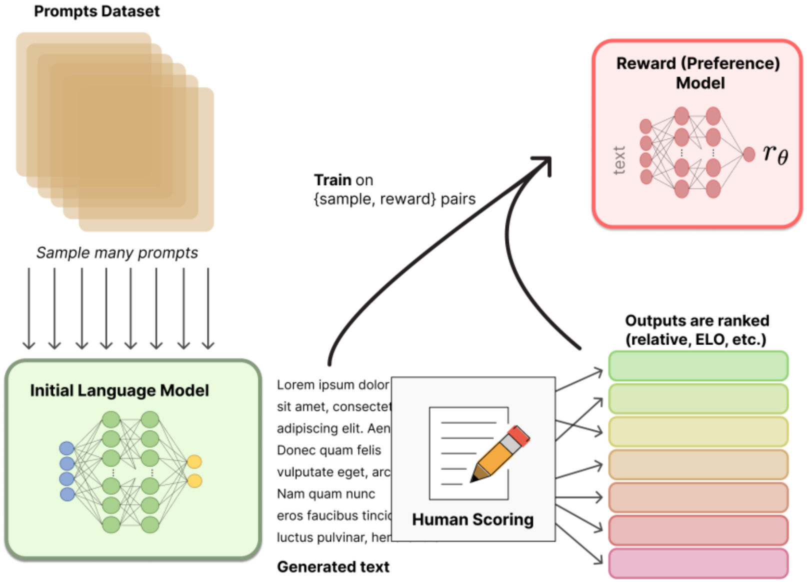
How do we model human preferences?

Problem 2: human judgments are noisy and miscalibrated!
Solution: instead of asking for direct ratings, ask for pairwise comparisons, which can be more reliable [Phelps et al. 2015; Clark et al. 2018]

Slide adapted from Jesse Mu

Reward model

- Takes in a sequence of text and produces a scalar representing human preference for that text (scalar is needed for RL)
- Training data:
  - Prompts (can come from real users of OpenAI's LLMs, e.g.)
  - LLM-generated responses to those prompts, ranked by human annotators



Slide based on https://huggingface.co/blog/rlhf

Finetuning LLMs with a reward model

- Often using a policy-gradient RL algorithm: Proximal Policy Optimization (PPO)
- **Policy:** a language model that takes in a prompt and returns a sequence of text (or just probability distributions over text)
- **Action space:** the vocabulary of the language model
- **Observation space:** the distribution of possible input token sequences
- **Reward function** is a combination of the preference model and a constraint on policy shift.

Slide based on https://huggingface.co/blog/rlhf

RLHF: Putting it all together [Christiano et al. 2017; Stiennon et al. 2020]

- Finally, we have everything we need:
  - A pretrained (possibly instruction-finetuned) LM $p^{PT}(s)$
  - A reward model $RM_\phi(s)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
  - A method for optimizing LM parameters towards an arbitrary reward function.
- Now to do RLHF:
  - Initialize a copy of the model $p_\theta^{RL}(s)$, with parameters θ we would like to optimize
  - Optimize the following reward with RL:

$$R(s) = RM_\phi(s) - \beta \log\left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)}\right)$$

Pay a price when $p_\theta^{RL}(s) > p^{PT}(s)$

This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler** (**KL**) divergence between $p_\theta^{RL}(s)$ and $p^{PT}(s)$.

Slide adapted from Jesse Mu

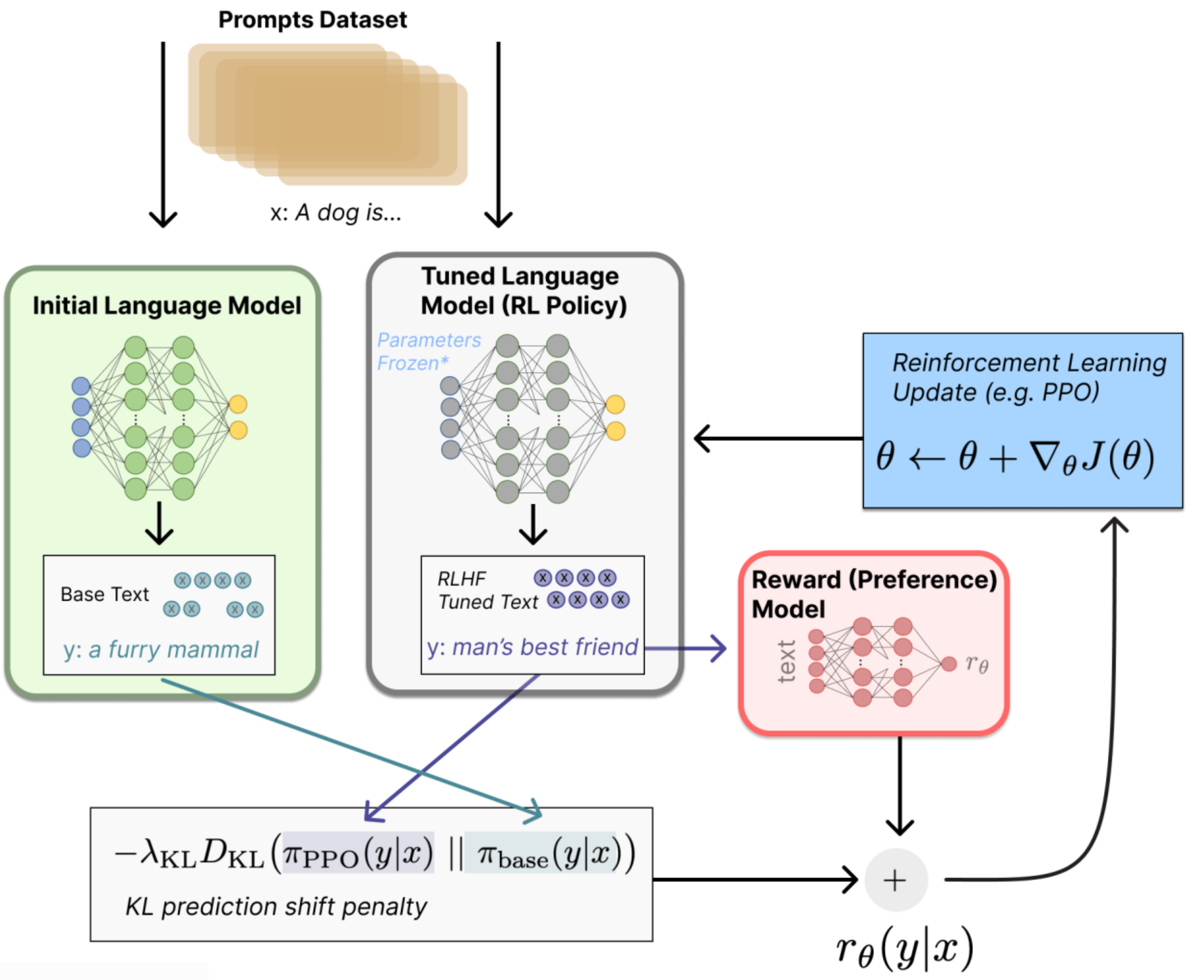
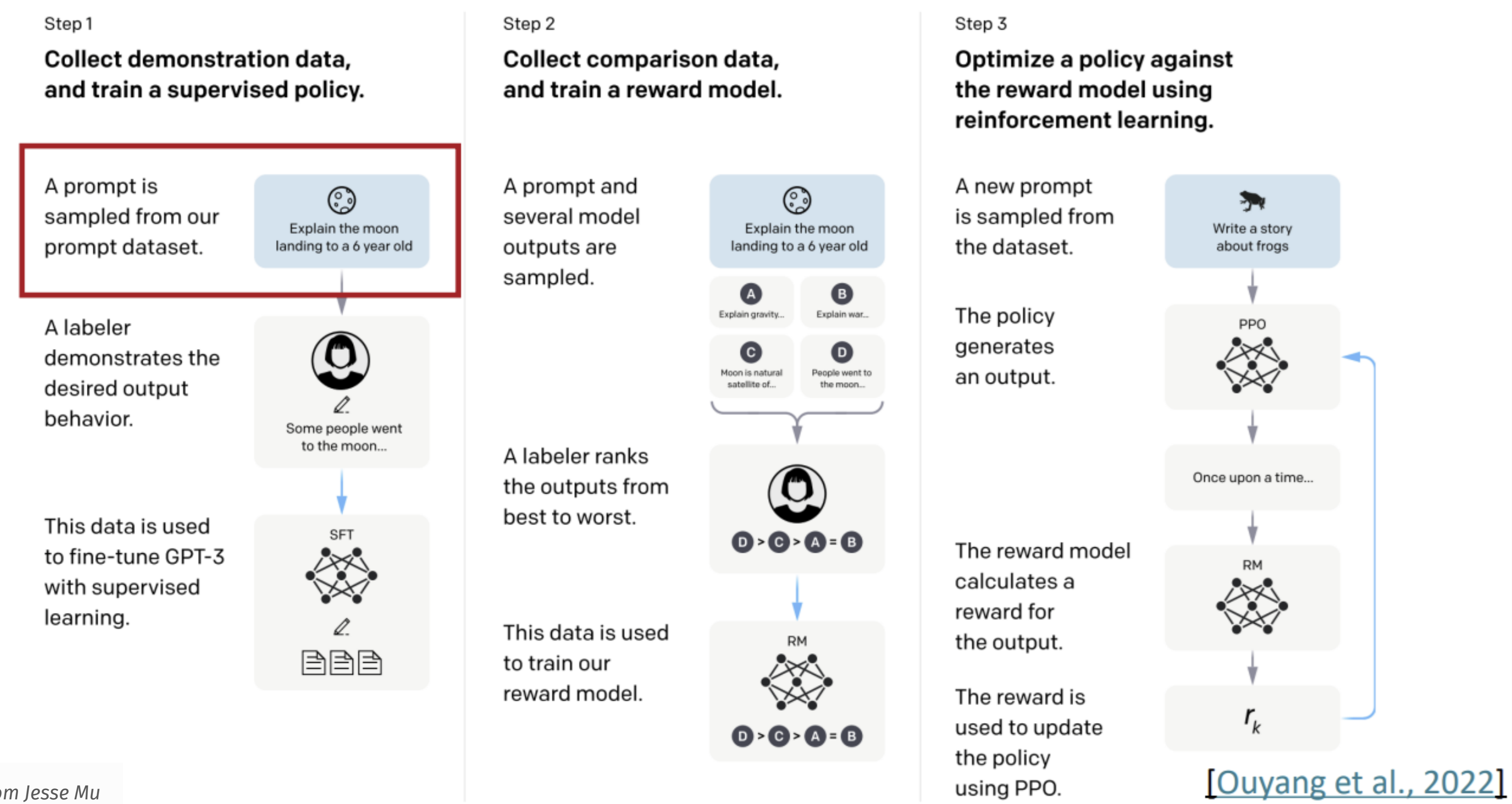


Image from https://huggingface.co/blog/rlhf

InstructGPT: scaling up RLHF to tens of thousands of tasks

30k tasks!



[Ouyang et al., 2022]

Slide adapted from Jesse Mu

InstructGPT gets us closer to dialogue intent

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and
sent them back to the earth so we could all see them.
```

Slide adapted from Jesse Mu

InstructGPT can accomplish many tasks

PROMPT *Write a short poem about a wise frog.*

COMPLETION GPT-3

```
Write a short story in which a character has two different names.

Write a short story in which you try to get something back that you have lost.

Write a short story in which a character has a bad dream.
```

InstructGPT

```
The frog is a symbol of wisdom
He knows all the secrets of the world
He is a master of disguise
And a great teacher of life
He is a symbol of transformation
And the bringer of change
He is the frog who has seen it all
And knows the meaning of it all
```

Slide adapted from Jesse Mu

ChatGPT: Instruction Finetuning + RLHF for dialog agents

ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

(Instruction finetuning!)

https://openai.com/blog/chatgpt/

Slide adapted from Jesse Mu

ChatGPT: Instruction Finetuning + RLHF for dialog agents

ChatGPT: Optimizing Language Models for Dialogue

Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge…

Methods

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

(RLHF!)

https://openai.com/blog/chatgpt/

Slide adapted from Jesse Mu

Limitations of RL + Reward Modeling

- Human preferences are unreliable!
- “Reward hacking” is a common problem in RL
- Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth
- This can result in making up facts + hallucinations

Slide adapted from Jesse Mu

Wrapping up

- Privacy, abuse, and representation harms are important ethical considerations for dialogue systems
- Rule-based chatbots, starting with the ELIZA system, can be quite effective
- Corpus-based chatbots can respond by generating responses after being trained on corpora
- Large language models can be trained for dialogue using reinforcement learning from human feedback (RLHF)

Questions?