

CS 2731 / ISSP 2230

Introduction to Natural Language Processing

Session 25: Dialogue, chatbots part 1

Michael Miller Yoder

April 10, 2024



University of
Pittsburgh

School of Computing and Information

Course logistics: project, project, project

- Project presentations in class Apr 24
- Final reports **due Apr 25**
- Grades on basic working systems are out
 - Feel free to come to office hours or set up a meeting with Michael or Bhiman to discuss next steps

Overview: Dialogue, chatbots part 1

- Introduction to dialogue systems and chatbots
- Properties of human conversation
- Rule-based chatbots (ELIZA review)
- Corpus-based chatbots
- Encoder-decoder framework for dialogue generation
- RLHF and ChatGPT

Dialogue systems and chatbots

Conversational Systems

- Personal Assistants on phones or other devices
 - SIRI, Alexa, Cortana, Google Assistant
- Playing music, setting timers and clocks
- Chatting for fun
- Booking travel reservations
- Clinical uses for mental health

Two kinds of conversational systems

- Chatbots
 - mimic informal human chatting for fun, or even for therapy
- (Task-based) Dialogue Agents
 - interfaces to personal assistants
 - cars, robots, appliances
 - booking flights or restaurants

Spoken conversational systems

- Incorporates speech recognition and text-to-speech
 - Additional possible sources of error
- Benefits of speech as an interface
 - Highly intuitive
 - Eyes and hands free
 - Small devices
 - Rich communication channel

Properties of human conversation

C₁: ... I need to travel in May.
A₂: And, what day in May did you want to travel?
C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A₄: And you're flying into what city?
C₅: Seattle.
A₆: And what time would you like to leave Pittsburgh?
C₇: Uh hmm I don't think there's many options for non-stop.
A₈: Right. There's three non-stops today.
C₉: What are they?
A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.
The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the
last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C₁₁: OK I'll take the 5ish flight on the night before on the 11th.
A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air
flight 115.
C₁₃: OK.
A₁₄: And you said returning on May 15th?
C₁₅: Uh, yeah, at the end of the day.
A₁₆: OK. There's #two non-stops ... #
C₁₇: #Act... actually #, what day of the week is the 15th?
A₁₈: It's a Friday.
C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.
A₂₀: OK... OK. On Sunday I have ...

Figure 15.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A₁₆ and C₁₇ indicate overlaps in speech.

Turn-taking

C₁: ... I need to travel in May.
A₂: And, what day in May did you want to travel?
C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A₄: And you're flying into what city?
C₅: Seattle.
A₆: And what time would you like to leave Pittsburgh?
C₇: Uh hmm I don't think there's many options for non-stop.
A₈: Right. There's three non-stops today.
C₉: What are they?
A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.
The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the
last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C₁₁: OK I'll take the 5ish flight on the night before on the 11th.
A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air
flight 115.
C₁₃: OK.
A₁₄: And you said returning on May 15th?
C₁₅: Uh, yeah, at the end of the day.
A₁₆: OK. There's #two non-stops ... #
C₁₇: #Act... actually #, what day of the week is the 15th?
A₁₈: It's a Friday.
C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.
A₂₀: OK...OK. On Sunday I have ...

- A turn is a single contribution from one speaker
- Turn-taking is complex
- When to take/yield the floor?
- People can detect when their conversation partner is about to stop talking
- People interrupt each other, resulting in overlapping speech

Figure 15.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A₁₆ and C₁₇ indicate overlaps in speech.

C₁: ... I need to travel in May.
A₂: And, what day in May did you want to travel?
C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A₄: And you're flying into what city?
C₅: Seattle.
A₆: And what time would you like to leave Pittsburgh?
C₇: Uh hmm I don't think there's many options for non-stop.
A₈: Right. There's three non-stops today.
C₉: What are they?
A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.
The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the
last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C₁₁: OK I'll take the 5ish flight on the night before on the 11th.
A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air
flight 115.
C₁₃: OK.
A₁₄: And you said returning on May 15th?
C₁₅: Uh, yeah, at the end of the day.
A₁₆: OK. There's #two non-stops ... #
C₁₇: #Act... actually #, what day of the week is the 15th?
A₁₈: It's a Friday.
C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.
A₂₀: OK... OK. On Sunday I have ...

There are *vocal pauses*
such as “uh”.

Figure 15.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A₁₆ and C₁₇ indicate overlaps in speech.

C₁: ... I need to travel in May.
A₂: And, what day in May did you want to travel?
C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A₄: And you're flying into what city?
C₅: Seattle.
A₆: And what time would you like to leave Pittsburgh?
C₇: Uh hmm I don't think there's many options for non-stop.
A₈: Right. There's three non-stops today.
C₉: What are they?
A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.
The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the
last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C₁₁: OK I'll take the 5ish flight on the night before on the 11th.
A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air
flight 115.
C₁₃: OK.
A₁₄: And you said returning on May 15th?
C₁₅: Uh, yeah, at the end of the day.
A₁₆: OK. There's #two non-stops ... #
C₁₇: #Act... actually #, what day of the week is the 15th?
A₁₈: It's a Friday.
C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.
A₂₀: OK... OK. On Sunday I have ...

There are *discourse markers* like “OK” and “Right”.

Figure 15.1 Part of a phone conversation between a human travel agent (A) and human client (C). The passages framed by # in A₁₆ and C₁₇ indicate overlaps in speech.

Grounding



Why do elevator buttons light up?

And what happens when pedestrian crosswalk buttons don't?



Image: ABC News

Grounding with Discourse Markers

A: And you said returning on May 15th?

C: Uh, yeah, at the end of the day.

A: OK

C: OK I'll take the 5ish flight on the night before on the 11th.

A: On the 11th? OK.

C: ...I need to travel in May.

A: And, what day in May did you want to travel?

Grounding = acknowledgment

- Conversation participants need *common ground*: set of things mutually believed by both speaker and hearer
- Principle of closure: Agents performing an action require evidence, sufficient for current purposes, that they have succeeded in performing it (Clark 1996, Norman 1988)
- Speech is an action too! So speakers need to ground each other's utterances.
- Grounding: acknowledging that the hearer has understood

Grounding is important for computers too!

System: Did you want to review more of your profile?

User: No.

System: What's next? **AWKWARD**

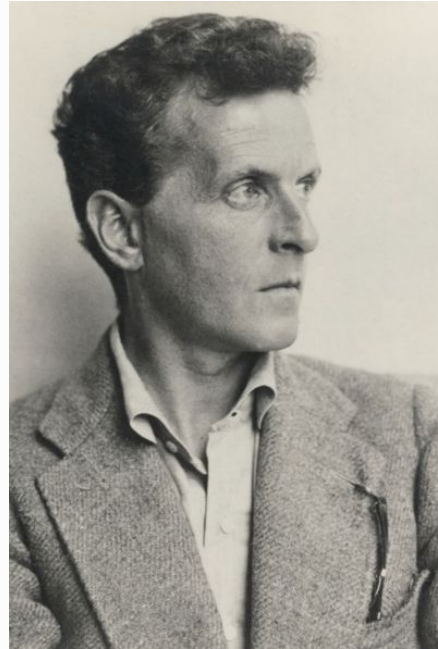
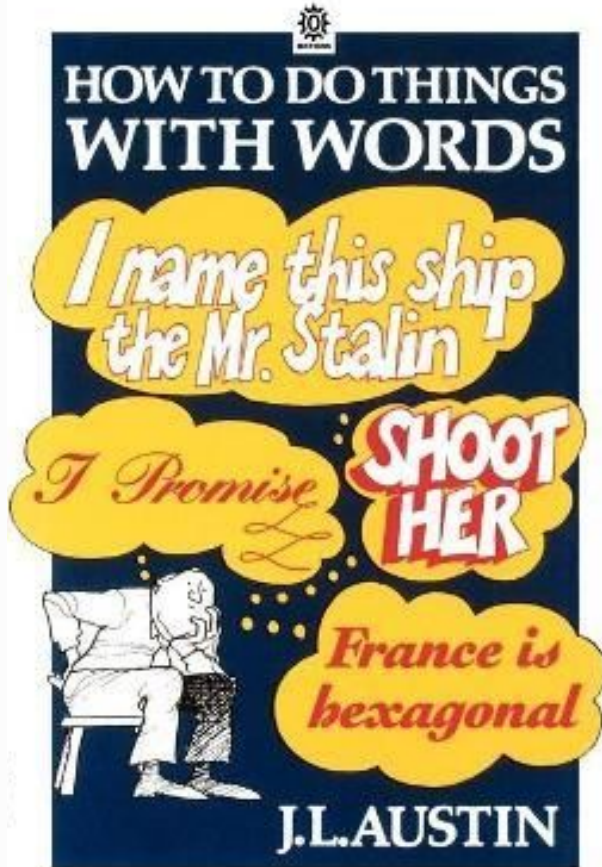
System: Did you want to review more of your profile?

User: No.

System: Okay, what's next? **LESS AWKWARD!**

Speech acts: sentences that do things

Utterances as actions



Ludwig Wittgenstein

Each turn in a dialogue is a kind of action [Wittgenstein 1953, Austin 1962]

Speech Acts: sentences that do things

Some sentences inform: *Today is Thursday*

Here are some *performative* sentences that change the state of the world:

- *I hereby name this ship the QE2.*
The ship now has a name.
- *I hereby bequeath this cell phone to my son.*
My son will now inherit the cell phone.
- *I hereby take this person to be my spouse.*
I am now married.
- *I hereby declare war.*
There is war.
- *I hereby excommunicate you.*
You are excommunicated.

Intent vs form

The following three sentences all have the same intent (to get the agent to book a flight to Seattle), but they take the form of a question, a command, and a statement.

- Can you book me a flight to Seattle?
- Book me a flight to Seattle.
- I'd like a flight to Seattle.

To respond appropriately, an automated conversational agent has to understand the user's intent. Task-oriented dialogue systems include classifiers for identifying the user's intent.

Conversations have structure

Local structure between adjacent speech acts, from the field of conversation analysis [Sacks et al. 1974]

Called adjacency pairs:

- Question > Answer
- Proposal > Acceptance/Rejection
- Compliments ("Nice jacket!") > Downplayer ("Oh, this old thing?")

Subdialogues

Correction subdialogue

Agent: OK. There's #two non-stops#

Client: #Act- actually#, what day of the week is the 15th?

Agent: It's a Friday.

Client: Uh hmm. I would consider staying there an extra day til Sunday.

Agent: OK...OK. On Sunday I have ...

Conversational initiative

- Some conversations are controlled by one person
 - A reporter interviewing a chef asks questions, and the chef responds.
 - This reporter has the **conversational initiative** (Walker and Whittaker 1990)
- Most human conversations have **mixed initiative**:
 - I lead, then you lead, then I lead.
- Mixed initiative is very hard for NLP systems, which often default to simpler styles that can be frustrating for humans:
 - **User initiative** (user asks or commands, system responds)
 - **System initiative** (system asks user questions to fill out a form, user can't change the direction)

Conversational implicature

Agent: And, what day in May did you want to travel?

Client: OK, uh, I need to be there for a meeting that's from the 12th to the 15th.

Rule-based chatbots

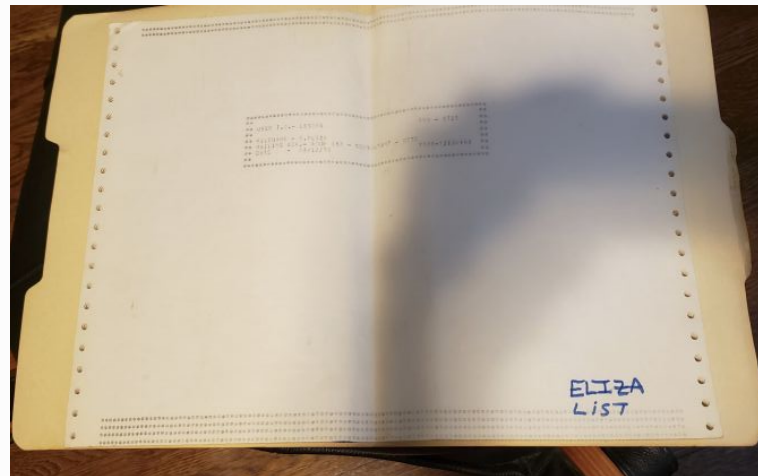
ELIZA [Weizenbaum 1966]

- 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?”



```
00070000
00080000(DEFPP)
00090000
00100000 NIL ; (PRUSH
00110000
00120000(DEFPP ADVANCE
(LAMBDA NIL (RPLAC (CDAR RULES) (COND ((NIL (CDADR RULES)) (CDADR RULES)) (CDADR RULES)) (CDADR RULES))))
00130000
00140000
00150000
00160000(DEFPP ANALYZE
(LAMBDA NIL
(FRAG (RULES PARSELIST (DECOMP)
(SETR KEYSTACK
(APPEND KEYSTACK
(LIST
(GET (QUOTE NAME)
(COND
((ZEROP (SETR FLIPFLOP (+PLUS 2 (MINUS FLIPFLOP)))) (QUOTE NAME))
(QUOTE LASTWORD)))))))
A (SETR RULES (GET (CAR KEYSTACK) (QUOTE RULES)))
B (SETR DECOMP
(LAMB (COND ((ATOM (CAR RULES)) (SETR RULES (GET (CAR RULES) (QUOTE RULES))))
(SETR PARSELIST NIL)
(COND ((TEST (DECOMP SENTENCE)) (SETR RULES (CDR RULES)))
((AND (NOT (ATOM (CAR (SETR RULES (CAR (ADVANCE))))))
(NOT (EQ (CAR RULES) (QUOTE PRE))))
(RETURN (SENTPRINT (RECONSTRUCT (CAR RULES))))
((NOT (ATOM (CAR RULES))) (SETR RULES (CDADR RULES)))
((EQ (CAR RULES) (QUOTE NAME))) (SETR KEYSTACK (CDR KEYSTACK)) (GO A)))
(SETR DECOMP)
(SETR KEYSTACK (CAR KEYSTACK)) (GO B)))
00180000
00190000
00200000
00210000
00220000
00230000
00240000
00250000
00260000
00270000
00280000
00290000
00300000
00310000
00320000
00330000
00340000
00350000
00360000
00370000
00380000
00390000
00400000
00410000
00420000
00430000
00440000
00450000
00460000
00470000
00480000
00490000
00500000
00510000
00520000
00530000
00540000
00550000
00560000
00570000
00580000
00590000
00600000
00610000
00620000
00630000
00640000
00650000
00660000
00670000
00680000
00690000
00700000
00710000
00720000
00730000
00740000
00750000
00760000
00770000
00780000
00790000
00800000
00810000
00820000
00830000
00840000
00850000
00860000
00870000
00880000
00890000
00900000
00910000
00920000
00930000
00940000
00950000
00960000
00970000
00980000
00990000
01000000
01010000
01020000
01030000
01040000
01050000
01060000
01070000
01080000
01090000
01100000
01110000
01120000
01130000
01140000
01150000
01160000
01170000
01180000
01190000
01200000
01210000
01220000
01230000
01240000
01250000
01260000
01270000
01280000
01290000
01300000
01310000
01320000
01330000
01340000
01350000
01360000
01370000
01380000
01390000
01400000
01410000
01420000
01430000
01440000
01450000
01460000
01470000
01480000
01490000
01500000
01510000
01520000
01530000
01540000
01550000
01560000
01570000
01580000
01590000
01600000
01610000
01620000
01630000
01640000
01650000
01660000
01670000
01680000
01690000
01700000
01710000
01720000
01730000
01740000
01750000
01760000
01770000
01780000
01790000
01800000
01810000
01820000
01830000
01840000
01850000
01860000
01870000
01880000
01890000
01900000
01910000
01920000
01930000
01940000
01950000
01960000
01970000
01980000
01990000
02000000(DEFPP BREAKANALYZE
```

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?

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)

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

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)

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 $r_1 \dots r_{t-1}$

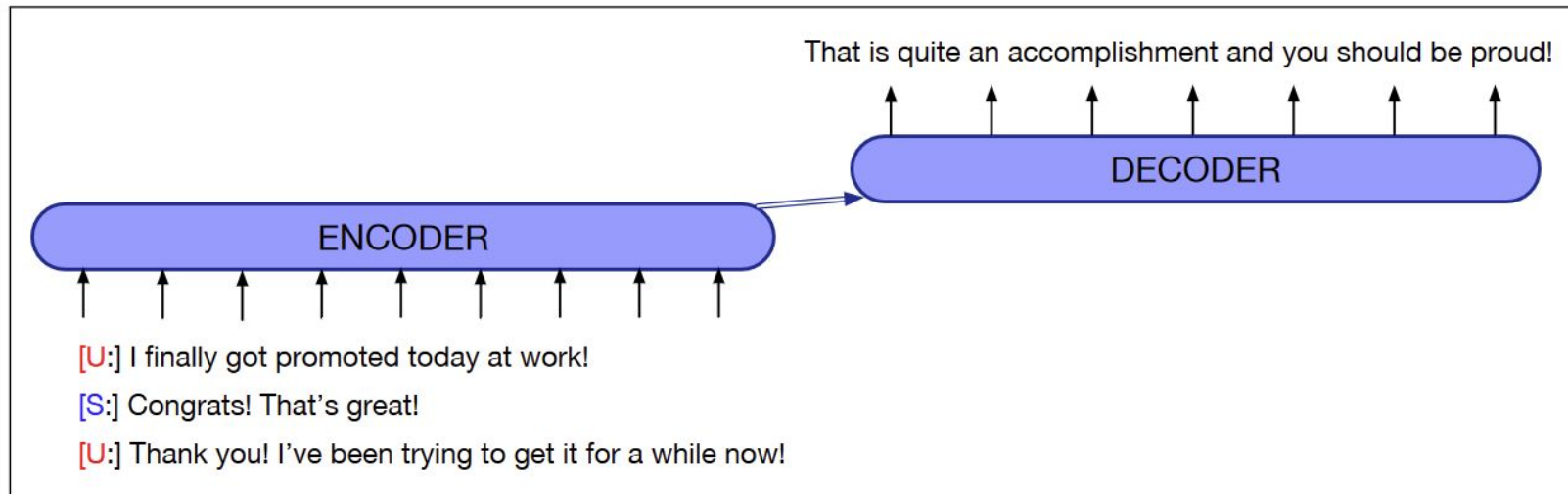


Figure 15.7 Example of encoder decoder for dialogue response generation; the encoder sees the entire dialogue context.

Reinforcement Learning from Human Feedback (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.

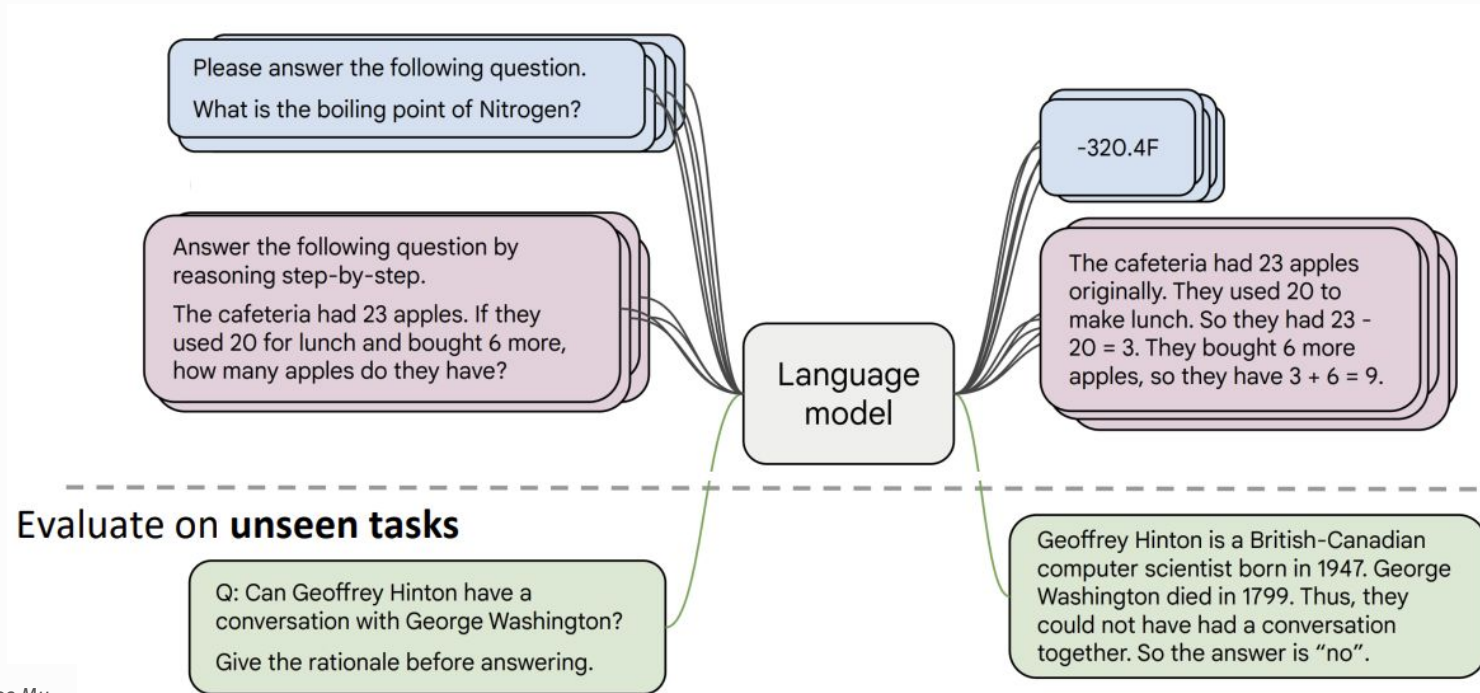
Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

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

Instruction finetuning

- Collect examples of (instruction, output) pairs across many tasks and finetune an LM



Limitations of instruction finetuning

- Expensive to collect ground-truth data for tasks
- **Problem 1:** tasks like open-ended creative generation have no right answer.
 - Write me a story about a dog and her pet grasshopper.
- **Problem 2:** 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?**

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.

$$s_1 \\ R(s_1) = 8.0$$

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

$$s_2 \\ R(s_2) = 1.2$$

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

Reinforcement learning to the rescue

- The field of reinforcement learning (RL) has studied these (and related) problems for many years [Williams 1992; Sutton and Barto 1998]
- Circa 2013: resurgence of interest in RL applied to deep learning, game-playing [Mnih et al., 2013]
- Interest in applying RL to modern LMs is a newer phenomenon [Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022]. Why?
 - RL w/ LMs has commonly been viewed as very hard to get right (still is!)
 - Newer advances in RL algorithms that work for large neural models, including language models (e.g. PPO; [Schulman et al., 2017])



How do we model human preferences?

- With 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]

An earthquake hit San Francisco. There was minor property damage, but no injuries.

$$R(s_1) = 8.0$$



The Bay Area has good weather but is prone to earthquakes and wildfires.

$$R(s_2) = 1.2$$



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

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.

$$R(s_3) = \begin{matrix} s_3 \\ 4.1? & 6.6? & 3.2? \end{matrix}$$

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.

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]

Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))]$$

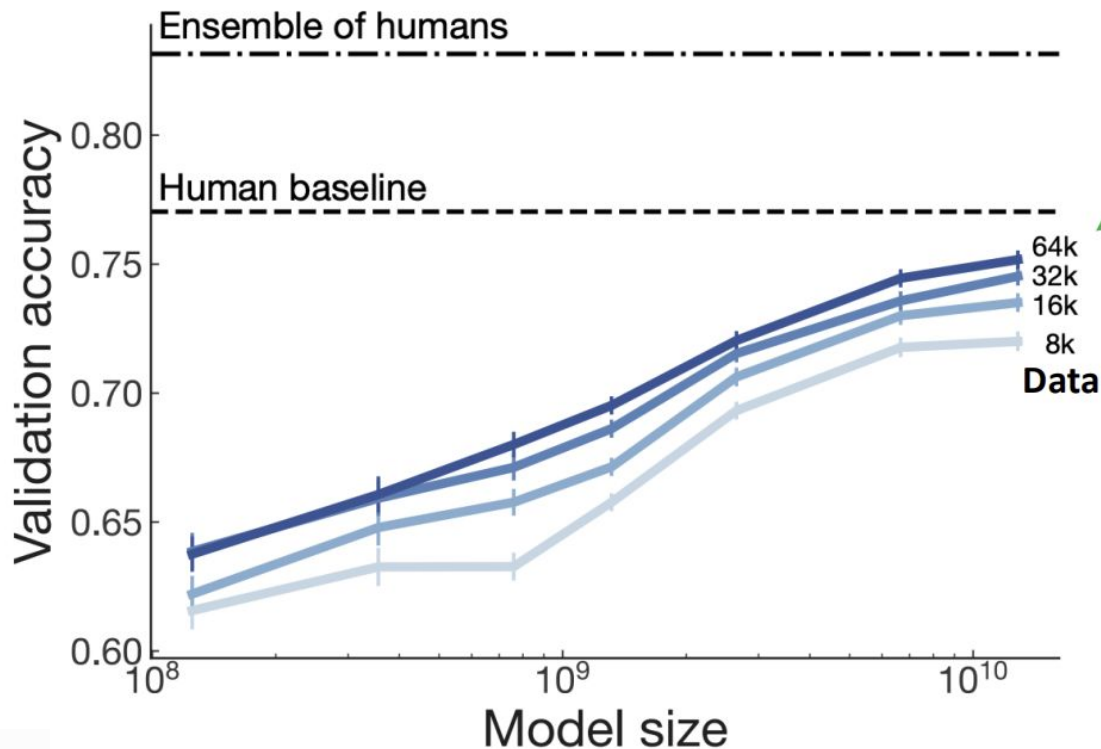
“winning”
sample

“losing”
sample

s^w should score
higher than s^l

Make sure your reward model works first!

Evaluate RM on predicting outcome of held-out human judgments



Large enough RM
trained on enough
data approaching
single human perf

[Stiennon et al., 2020]

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:

$$R(s) = RM_\phi(s) - \beta \log \left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)} \right) \quad \text{Pay a price when } p_\theta^{RL}(s) > p^{PT}(s)$$

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 $p^{PT}(s)$.

InstructGPT: scaling up RLHF to tens of thousands of tasks

30k tasks!

Step 1

Collect demonstration data, and train a supervised policy.

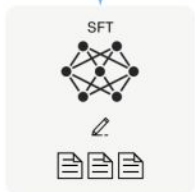
A prompt is sampled from our prompt dataset.

Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

Some people went to the moon...

This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

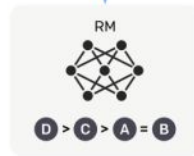
Explain the moon landing to a 6 year old

A Explain gravity... B Explain war...
C Moon is natural satellite of... D People went to the moon...

A labeler ranks the outputs from best to worst.

D > C > A = B

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

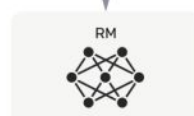
A new prompt is sampled from the dataset.

Write a story about frogs

The policy generates an output.

Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

[Ouyang et al., 2022]

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.

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

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!)

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!)

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

TECHNOLOGY

Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

<https://www.npr.org/2023/02/09/1155650909/google-chatbot-error-bard-shares>

Bing AI hallucinates the Super Bowl

Who won the superbowl?

✓ Searching for: **superbowl winner**

✓ Generating answers for you...

The Super Bowl is the annual American football game that determines the champion of the National Football League (NFL)¹. The most recent Super Bowl was **Super Bowl LVI**, which was held on **February 6, 2023 at SoFi Stadium in Inglewood, California**². The winner of that game was the **Philadelphia Eagles**, who defeated the **Kansas City Chiefs** by **31-24**³. It was the second Super Bowl win for the **Philadelphia Eagles** franchise.

The most recent Super Bowl was Super Bowl LVI, Eagles, who defeated the Kansas City Chiefs by 31-24

Learn more: [1. en.wikipedia.org](#) | [2. sportingnews.com](#) | [3. cbssports.com](#)

<https://news.ycombinator.com/item?id=34776508>

<https://apnews.com/article/kansas-city-chiefs-philadelphia-eagles-technology-science-82bc20f207e3e4cf81abc6a5d9e6b23a>

Wrapping up

- Automated conversational systems can be divided into 2 types:
 - Open-domain “chatbots”
 - Task-oriented dialogue systems
- Conversation is a complex joint interaction between participants
 - Turn-taking and grounding are example issues that dialogue systems must address
- 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?