

CS388: Natural Language Processing

Lecture 13: Instruction Tuning, RLHF, Dialog

Greg Durrett

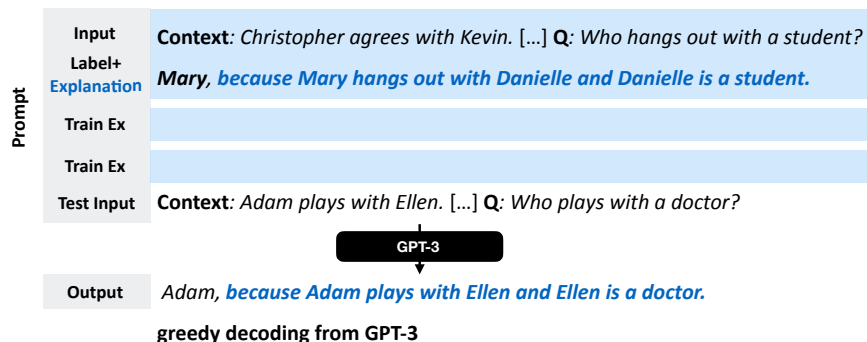


Announcements

- ▶ Project 3 out now
 - ▶ We highly recommend Colab
 - ▶ You don't need all training iterations
 - ▶ You can decrease the frequency of checkpointing
- ▶ Project 2 back soon
- ▶ Final project proposals back soon

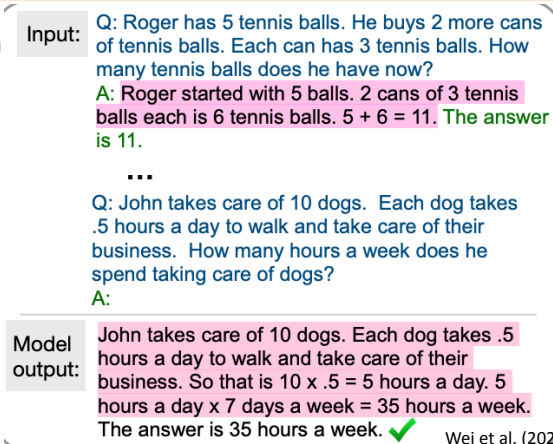


Recap: Chain-of-thought



Recap: Chain-of-thought

- ▶ Can help substantially on mathematical reasoning
- ▶ Some work to optimize the specifics of the prompts and the examples





Today

- Instruction tuning
- RLHF
- Chatbots
- Task-oriented dialogue systems

Instruction Tuning

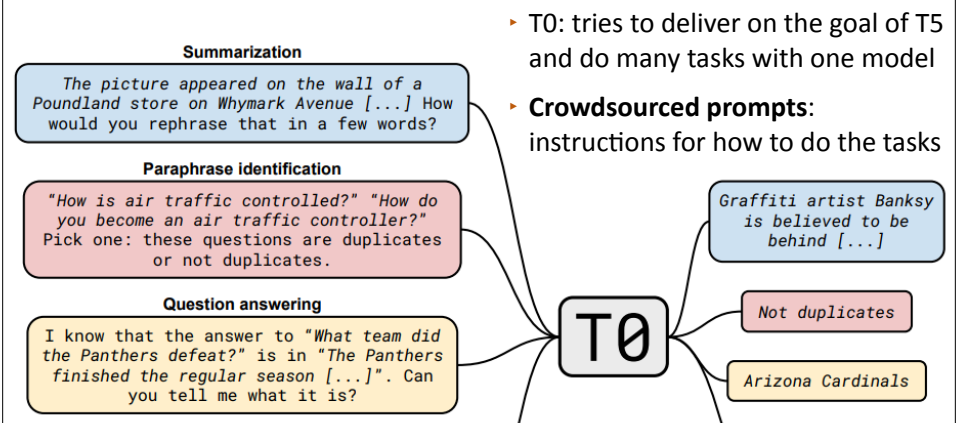


Instruction Tuning

- We want to optimize models for $P(\text{answer} \mid \text{prompt, input})$, but they're learned on a basic language modeling objective
- One solution: treat the basic language modeling as pre-training, then fine-tune them on what we care about
- Two versions of this:
 - **Instruction tuning**: supervised fine-tuning on data derived from many NLP tasks
 - **Reinforcement learning from human feedback (RLHF)**: RL to improve human judgments of how good the outputs are



Task Generalization: T0

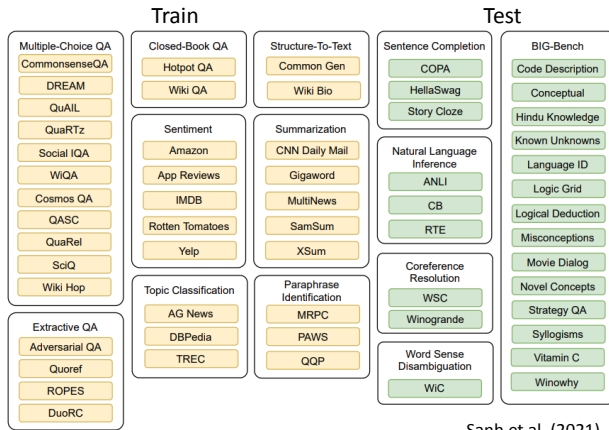


Sanh et al. (2021)



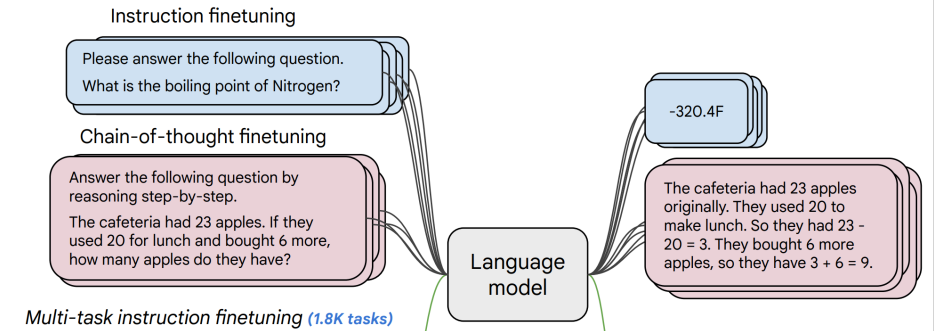
Task Generalization

- Pre-train: T5 task
- Train: a collection of tasks with prompts. **This uses existing labeled training data**
- Test: a new task specified only by a new prompt. **No training data in this task**



Frontiers

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model fine-tuned on many tasks after pre-training



Frontiers

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

Conceptual Physics	When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is	
	(A) 9.8 m/s^2	✓
	(B) more than 9.8 m/s^2	✗
	(C) less than 9.8 m/s^2	✗
College Mathematics	In the complex z -plane, the set of points satisfying the equation $z^2 = z ^2$ is a	
	(A) pair of points	✗
	(B) circle	✗
	(C) half-line	✗
	(D) line	✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Chung et al. (2022)



Frontiers

- Flan-PaLM (October 20, 2022): 1800 tasks, 540B parameter model
- MMLU task (Hendrycks et al., 2020): 57 high school/college/professional exams:

-	Random	25.0
-	Average human rater	34.5
May 2020	GPT-3 5-shot	43.9
Mar. 2022	Chinchilla 5-shot	67.6
Apr. 2022	PaLM 5-shot	69.3
Oct. 2022	Flan-PaLM 5-shot	72.2
	Flan-PaLM 5-shot: CoT + SC	75.2
-	Average human expert	89.8

Chung et al. (2022)



Frontiers

Model	Finetuning Mixtures	Tasks	Norm. avg.	MMLU		BBH	
				Direct	CoT	Direct	CoT
540B	None (no finetuning)	0	49.1	71.3	62.9	49.1	63.7
	CoT	9	52.6 (+3.5)	68.8	64.8	50.5	61.1
	CoT, Muffin	89	57.0 (+7.9)	71.8	66.7	56.7	64.0
	CoT, Muffin, T0-SF	282	57.5 (+8.4)	72.9	68.2	57.3	64.0
	CoT, Muffin, T0-SF, NIV2	1,836	58.5 (+9.4)	73.2	68.1	58.8	65.6

- Human performance estimates are ~80 on Big-Bench (BBH)

Chung et al. (2022)



Flan-T5

- Flan-T5: T5-11B model given the “Flan treatment”, instruction tuned on many tasks
- Best model at the ~10B scale for few-shot prompting, also very good choice for fine-tuning

Params	Model	Norm. avg.	MMLU		BBH	
			Direct	CoT	Direct	CoT
11B	T5-XXL	-2.9	25.9	18.7	29.5	19.3
	Flan-T5-XXL	23.7 (+26.6)	55.1	48.6	45.3	41.4

(about 20% behind the 540B Flan-PaLM)

- If you have the resources, Flan-T5 is something you can explore in your project/research!

Chung et al. (2022)

Reinforcement Learning from Human Feedback (RLHF)



RLHF

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

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



Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



- Apply this approach to optimizing outputs from large language models
- Step 3 (not shown): do RL with this policy

Ouyang et al. (2022)



RLHF

- Humans produce comparisons of two trajectories (= outputs from systems) — different from standard reward in RL
- Fit the reward function r using supervised estimation:

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

- This turns scores into log probabilities of 1 being preferred to 2. Same as logistic regression where we classify pairs as $1 > 2$ or $2 < 1$, but we actually learn a continuous scoring function, not a classifier
- The rest of the RL setup is TRPO/PPO, fairly standard frameworks (note: they typically constrain the policy to not deviate too far from a basic supervised policy)

Christiano et al. (2017)



RLHF

Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: "" {summary} "" This is the outline of the commercial for that play: ""

- For OpenAI, RLHF data is collected from their API. **Very different from instruct-tuning datasets**

Ouyang et al. (2022)



text-davinci-003

- text-davinci-001/002 were both learned only from fine-tuning on demonstrations rated 7/7 (i.e., not using RLHF)
- text-davinci-003 (latest version) and ChatGPT both use PPO with learned reward models
- Conclusion: likely difficult to get PPO working reliably (or to get a good reward function — signal from annotators may be unstable)
 - ...but RLHF datasets from OpenAI are not public
 - Data quality is paramount! Anecdotally there are lots of human-written demonstrations in there and lots of ratings

<https://beta.openai.com/docs/model-index-for-researchers>

Pre-trained Chatbots

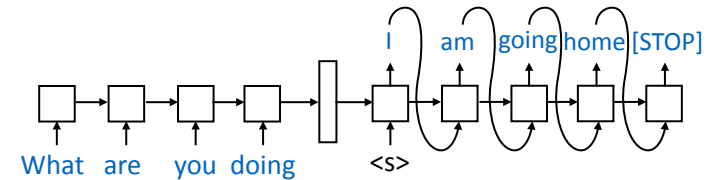


What are chatbots?

- Like story generation in that it's open-ended, but involves dialogue with a user
- Input: a conversation history of utterances, plus something the user (a person) just said.
Output: the model's response to that
- Needs to generate interesting and diverse content, but also needs to be able to answer questions and carry on a conversation



Seq2seq models



- Can we just train seq2seq models to “translate” from utterance to response?

- Hard to evaluate with automatic metrics:

System	BLEU
RANDOM	0.33
MT	3.21
HUMAN	6.08



Lack of Diversity

- Training to maximize likelihood gives a system that prefers common responses:

Input: What are you doing?	
-0.86 I don't know.	-1.09 Get out of here.
-1.03 I don't know!	-1.09 I'm going home.
-1.06 Nothing.	-1.09 Oh my god!
-1.09 Get out of the way.	-1.10 I'm talking to you.
Input: what is your name?	
-0.91 I don't know.	...
-0.92 I don't know!	-1.55 My name is Robert.
-0.92 I don't know, sir.	-1.58 My name is John.
-0.97 Oh, my god!	-1.59 My name's John.
Input: How old are you?	
-0.79 I don't know.	...
-1.06 I'm fine.	-1.64 Twenty-five.
-1.17 I'm all right.	-1.66 Five.
-1.17 I'm not sure.	-1.71 Eight.

Li et al. (2016)



PersonaChat

Persona 1

I like to ski
My wife does not like me anymore
I have went to Mexico 4 times this year
I hate Mexican food
I like to eat cheetos

Persona 2

I am an artist
I have four children
I recently got a cat
I enjoy walking for exercise
I love watching Game of Thrones

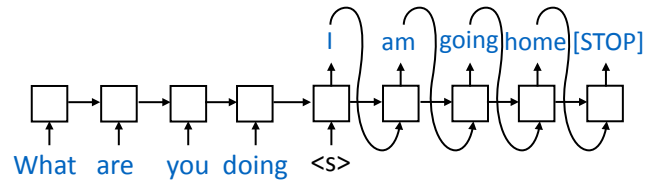
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!

Zhang et al. (2018)

- Efforts to imbue seq2seq models with “personality”
- These systems still don't work great. What else is needed?



Pre-trained Chatbots



- ▶ Initialize the parameters of this model with a pre-trained model, then fine-tune it on some data
- ▶ It turns out that scaling up dramatically and strategies like nucleus sampling can help with the “I don’t know” problem



Blender

- ▶ 2.7B-param model, also a 9.4B-parameter seq2seq model variant
- ▶ “Poly-encoder” Transformer architecture, some training tricks
- ▶ Three models: retrieve (from training data), generate, retrieve-and-refine
- ▶ Fine-tuning on three prior datasets: PersonaChat, Empathetic Dialogues (discuss personal situation, listener is empathetic), Wizard of Wikipedia (discuss something from Wikipedia)

Roller et al. (2020)



Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of “Sing Along with Mitch”?

Yes, I would love that.

Ok, I'll try to get it to you by the end of the day. What else do you like to do?

Can you at least tell me what the lyrics are?

Sure, it goes like this: “Baking is my favorite thing to do in the world.”

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!



Blender

I'm from the midwest, but I live in the south now. How about you?

Born and raised in the south

What part of the south? I was born in georgia, but raised in tennessee



Why did the mushroom go to the party?

I'm not sure, but I'm glad it did. I love museums. What do you do for a living?

No, it was supposed to be a joke.

Oh, sorry. I didn't get it. What kind of joke was it? I like jokes.

The answer is, "because he's a fun guy." Do you get it?

No, I don't. Can you explain it to me like I'm a five year old?

Yes. Mushrooms are fungi, and "fungi" is pronounced like "fun guy"

Oh, okay. I get it now. Thank you for the explanation. I really appreciate it.

Are you sure you understand? Can you explain it to me?

Sure. It's a play on the word fungus, which is also pronounced fun guy.



Blender

- ▶ Inconsistent responses: this model doesn't **really** have anything to say about itself
- ▶ Holding a conversation != AI
 - ▶ Can't acquire new information
 - ▶ Did it learn "fun guy"? No, it doesn't understand phonology. It probably had this in the data somewhere

I'm from the midwest, but I live in the south now. How about you?

Born and raised in the south

What part of the south? I was born in Georgia, but raised in Tennessee

Why did the mushroom go to the party?

I'm not sure, but I'm glad it did. I love museums. What do you do for a living?

No, it was supposed to be a joke.

Oh, sorry. I didn't get it. What kind of joke was it? I like jokes.

The answer is, "because he's a fun guy." Do you get it?

No, I don't. Can you explain it to me like I'm a five year old?

Yes. Mushrooms are fungi, and "fungi" is pronounced like "fun guy"

Oh, okay. I get it now. Thank you for the explanation. I really appreciate it.

Are you sure you understand? Can you explain it to me?

Sure. It's a play on the word fungus, which is also pronounced fun guy.



Chatbots

- ▶ What happens when these models get really good at fooling people? Google LaMDA model (similar to Blender):

"I KNOW A PERSON WHEN I TALK TO IT" —

Google fires Blake Lemoine, the engineer who claimed AI chatbot is a person

Google says Lemoine violated security rules, slams "wholly un-

NEWS

Ex-Google engineer Blake Lemoine discusses sentient AI

Ex-Google engineer Blake Lemoine discusses why LaMDA and other AI systems may be considered sentient and explains exactly how much AI systems know about consumers.

BBC Sign In

Home News Sport Reel Worklife Travel

NEWS

Home | War in Ukraine | US Elections 2022 | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech

Blake Lemoine: Google fires engineer who said AI tech has feelings



ChatGPT

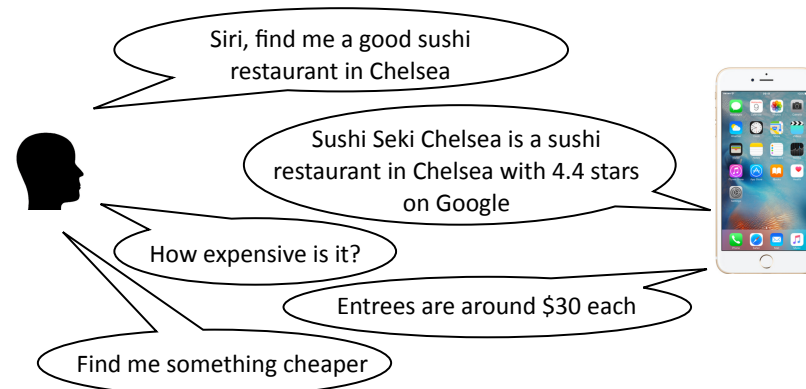
- ▶ Big model with RLHF. (More like a QA system than these other chatbots)
- ▶ Not much we can say except:
 - ▶ It's based on the earlier davinci models
 - ▶ Lots of data collection to fencepost it (e.g., "I don't know anything about the current weather ...")
 - ▶ Continuously improved without detailed release notes (e.g., they made it better at math)

Task-Oriented Dialogue



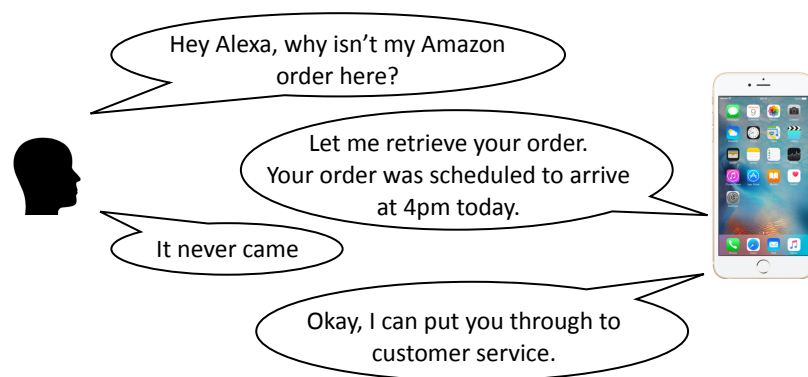
Task-Oriented Dialogue

- How do you build conversational systems to do things?



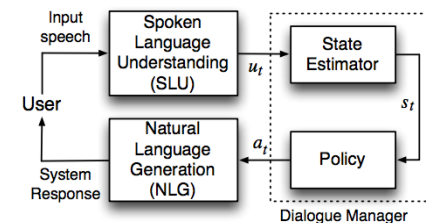
Task-Oriented Dialogue

- Customer service:



Task-Oriented Dialogue

- Parsing / language understanding is just one piece of a system
- Dialogue state: reflects any information about the conversation (e.g., search history)



- User utterance \rightarrow update dialogue state \rightarrow take action (e.g., query the restaurant database) \rightarrow say something
- How do we represent the information from the user's utterance?

Young et al. (2013)



ATIS

- Intent and slots model: classify an intent (**Airfare**), then fill several slots needed to specify the parameters for that intent

Utterance	<i>How much is the cheapest flight from Boston to New York tomorrow morning?</i>
Goal:	Airfare
Cost.Relative	<i>cheapest</i>
Depart.City	<i>Boston</i>
Arrival.City	<i>New York</i>
Depart.Date.Relative	<i>tomorrow</i>
Depart.Time.Period	<i>morning</i>

- This is how most Alexa skills work. Can match with rule-based systems or use classifiers

DARPA (early 1990s), Figure from Tur et al. (2010)



Intents

- 29 different intents in ATIS:

which flights go from cleveland to indianapolis on april fifth

Intent: *flight*

does tacoma airport offer transportation from the airport to the downtown area

Intent: *ground_service*

what days of the week do flights from san jose to nashville fly on

Intent: *day_name*

what meals are served on american flight 811 from tampa to milwaukee

Intent: *meal*



Dataflow Graphs

- How do we scale to more complex dialog scenarios? One proposal: **dataflow graphs**

User: *Where is my meeting at 2 this afternoon?*

`place(findEvent(EventSpec(start=pm(2))))`

(1)



Agent: *It's in Conference Room D.*

Semantic Machines; Andreas et al. (2020)



Dataflow Graphs

User: *Can you create a meeting with Megan right before that starts?*

```

createEvent(EventSpec(
  end=start(refer(Constraint[Event]())),
  attendee=PersonSpec(name='Megan')
))
  
```

(2)



Agent: *Which person named Megan did you mean?*

Semantic Machines; Andreas et al. (2020)



Task-Oriented Dialog: What the user sees

Find me a good sushi restaurant in Chelsea

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

Entrees are around \$30 each



Task-Oriented Dialog: Under the hood

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
```

```
location <- Chelsea
```

```
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
get_value(cost, curr_result)
```

Entrees are around \$30 each



Training Dialog Systems

- ▶ “Wizard of Oz”: can run the dialog system in a real setting and have a human decide what it should do next
- ▶ Learning from demonstrations: the system can learn from what the wizard does and do that in the future

Find me a good sushi restaurant in Chelsea

wizard enters these { `restaurant_type <- sushi`
`location <- Chelsea`
`curr_result <- execute_search()`

wizard types this out or invokes templates { `Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google`

Semantic Machines; Andreas et al. (2020)



Task-Oriented Dialogue

- ▶ Building these systems takes a ton of engineering, like Gunrock — it typically **doesn't** use pre-trained models (until 2023...)
- ▶ Need to know what the system should **do**, not just what it should say
- ▶ Generation is usually templated (handwritten), otherwise the system can behave unexpectedly
- ▶ Lots of industry activity in this space, less in academia (hard to maintain all of the moving parts for a real dialog system)
- ▶ Current interest: work like Toolformer / Langchain that allows LLMs to generate the API calls directly



Takeaways

- Instruction-tuning and RLHF are two procedures that take LMs to the next level — these models work dramatically better than basic GPT-3
- These are the foundation of modern chatbots (along with lots of pre-training data), very exciting capabilities in these LLM agents
- Task-oriented dialog has historically been different but is starting to unify with chatbots (Bing agent has ability to make API calls)