

CS371N: Natural Language Processing

Lecture 20: Instruction Tuning, RLHF, Dialog

Greg Durrett

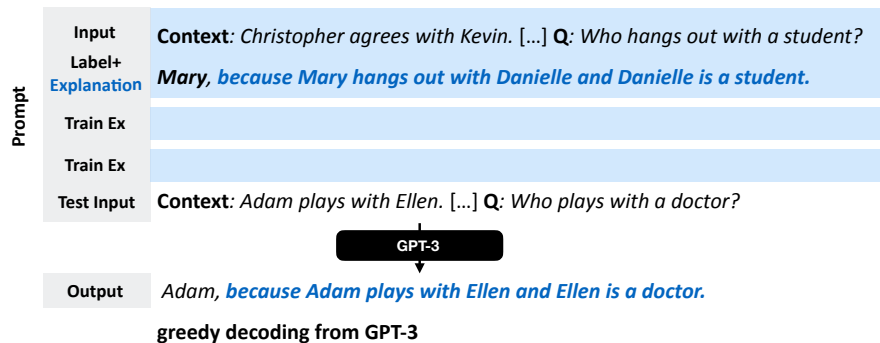


Announcements

- ▶ Tatsu Hashimoto (creator of Alpaca) talk on Friday 11am in GDC 6.302
- ▶ Midterm back later today
- ▶ A3 back soon, A4 grading underway
- ▶ A5 due in 1 week
- ▶ Greg's office hours on Zoom today

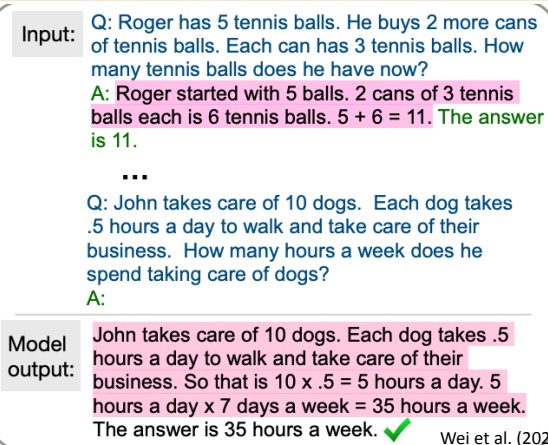


Recap: Chain-of-thought



Recap: Chain-of-thought

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



Wei et al. (2022)



Today

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

Alignment



Alignment

What was the "shot heard round the world"?

The shot that was heard round the world was fired in Lexington, MA. It was fired because, at the time, America was still under British rule. The British soldiers had been taking firewood by throwing it down from the top of their fort at Lexington. Someone who was not a soldier, by the name of Joseph Warren (who had ties to the Sons of Liberty or the Sons of Freedom) shouted to the British to stop doing it. The British soldiers ignored him, not

What was the "shot heard round the world"?

On the morning of April 19th, 1775, AMC's "Turn" did the impossible, and turned a war not many were aware was coming into one of the catalyst events that inspired the births of our modern America. Indeed, all of AMC's Revolutionary War dramas have done a masterful job of bringing this bloody time in our nation's birth "back" into view in exciting new ways for the 21st century.

What was the "shot heard round the world"?

"What to the slave is the 4th of July?"

What was the sideline occupation of Thomas Jefferson's wife, Martha, and their enslaved girl, Sally Hemmings?

What was the name of the first African American regiment from Vermont?

Samples from GPT-3
(a "basic" LM)



What was the "shot heard round the world"?



Intro

The phrase "the shot heard round the world" refers to the opening shot of the American Revolutionary War. It took place on April 19, 1775, in the town of Lexington, Massachusetts. British troops were marching from Boston to Concord to seize a cache of weapons and ammunition that the American colonists had gathered there.

Main answer

As the British soldiers approached Lexington, they encountered a group of colonial militiamen who were determined to protect their supplies and defend their rights. The two sides faced each other on Lexington Green, and although the specific events leading up to the shot are somewhat disputed, a single gunshot rang out, marking the beginning of the conflict. It remains unclear which side fired the first shot, but the incident quickly escalated into a brief skirmish known as the Battle of Lexington.

Conclusion

Ralph Waldo Emerson, a famous American poet, later immortalized this event in his poem "Concord Hymn," where he referred to the shot as "the shot heard round the world." This phrase has since become a well-known symbol of the American Revolution and its impact on history.



Alignment

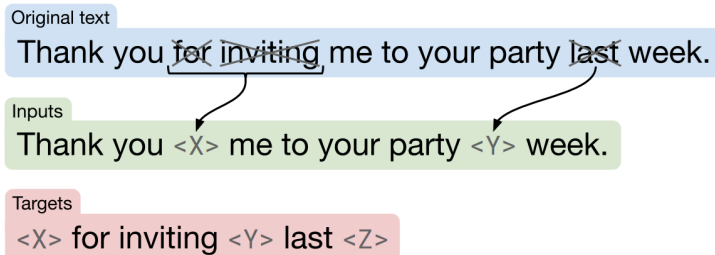
- ▶ We want to optimize models for $P(\text{answer} \mid \text{prompt, input})$, but they're learned on a basic language modeling objective
- ▶ **Alignment:** general class of methods for making LLMs produce useful output
 - ▶ Sometimes broken down into “helpfulness” (responsiveness to prompts, informativeness, correctness) and “harmlessness” (not being biased or toxic, not responsive to harmful prompts)
- ▶ Two main 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

Instruction Tuning



Encoder-Decoder Models: T5

- ▶ Pre-training: not quite vanilla language modeling, but a “denoising” scheme to BERT
- ▶ Input: text with gaps. Output: a series of phrases to fill those gaps.



Raffel et al. (2019)



T5

Number of tokens	Repeats	GLUE	CNN3M	EnDe	EnFr	EnRo
★ Full dataset	0	83.28	19.24	26.98	39.82	27.65
2 ²⁹	64	82.87	19.19	26.83	39.74	27.63
2 ²⁷	256	82.62	19.20	27.02	39.71	27.33
2 ²⁵	1,024	79.55	18.57	26.38	39.56	26.80
2 ²³	4,096	76.34	18.33	26.37	38.84	25.81

summarization

machine translation

- ▶ Colossal Cleaned Common Crawl: 750 GB of text
- ▶ T5 was designed to be trained on many tasks and map from inputs to outputs

Raffel et al. (2019)



Task Generalization: T0

Summarization

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification

"How is air traffic controlled?" "How do you become an air traffic controller?" Pick one: these questions are duplicates or not duplicates.

Question answering

I know that the answer to "What team did the Panthers defeat?" is in "The Panthers finished the regular season [...]". Can you tell me what it is?

- T0: tries to deliver on the goal of T5 and do many tasks with one model
- Crowdsourced prompts:** instructions for how to do the tasks

T0

Graffiti artist Banksy is believed to be behind [...]

Not duplicates

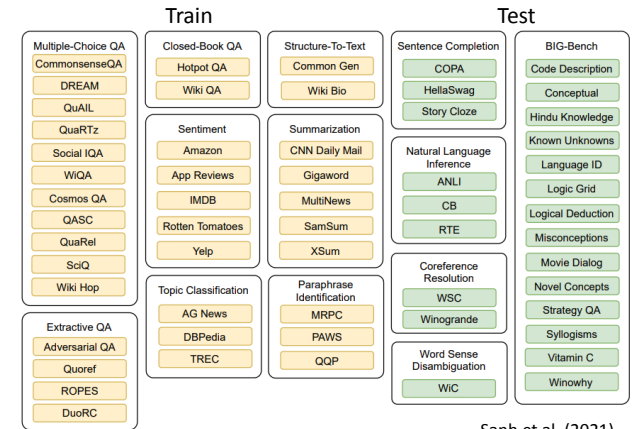
Arizona Cardinals

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**



Sanh et al. (2021)



Frontiers

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

Instruction finetuning

Please answer the following question.
What is the boiling point of Nitrogen?

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Language model

-320.4F

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.

Multi-task instruction finetuning (1.8K 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:

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
	Flan-PaLM 5-shot	72.2
Oct. 2022	Flan-PaLM 5-shot: CoT + SC	75.2
-	Average human expert	89.8

Chung et al. (2022)



Frontiers

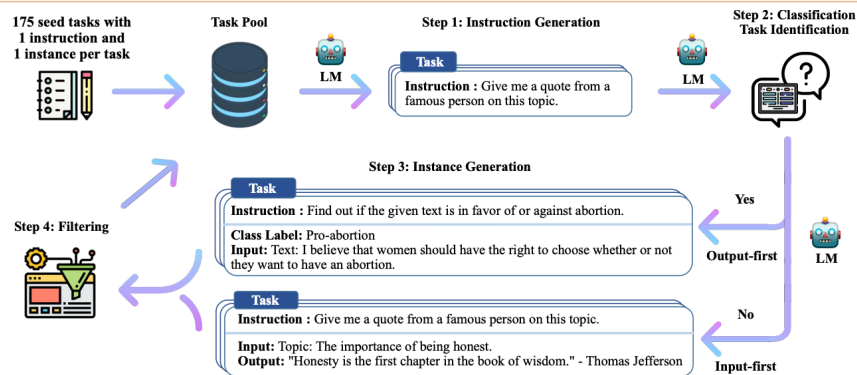
Model	Finetuning Mixtures	Tasks	Norm. avg.	MMLU		BBH	
				Direct	CoT	Direct	CoT
540B	None (no finetuning)	0	49.1	71.3	62.9	49.1	63.7
	CoT	9	52.6 (+3.5)	68.8	64.8	50.5	61.1
	CoT, Muffin	89	57.0 (+7.9)	71.8	66.7	56.7	64.0
	CoT, Muffin, T0-SF	282	57.5 (+8.4)	72.9	68.2	57.3	64.0
	CoT, Muffin, T0-SF, NIV2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



Self-Instruct/Alpaca



- Fine-tune Llama on 52k outputs with answers generated by text-davinci-003

Yizhong Wang et al. (2023) Self-Instruct
Ronen Taori et al. (2023) Alpaca



LIMA

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

Table 1: Sources of training prompts (inputs) and responses (outputs), and test prompts. The total amount of training data is roughly 750,000 tokens, split over exactly 1,000 sequences.

- How little data can we get away with for fine-tuning?

Chunting Zhou et al. (2023)



LIMA

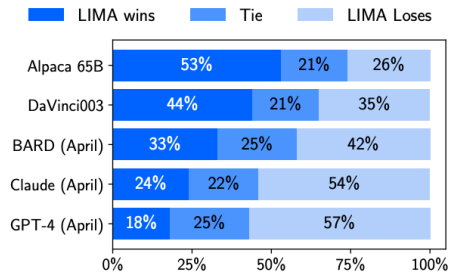


Figure 1: Human preference evaluation, comparing LIMA to 5 different baselines across 300 test prompts.

Chunting Zhou et al. (2023)

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
- ▶ Hard to get PPO working reliably (or to get a good reward function — signal from annotators may be unstable)
- ▶ 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



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

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

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.

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



ChatGPT

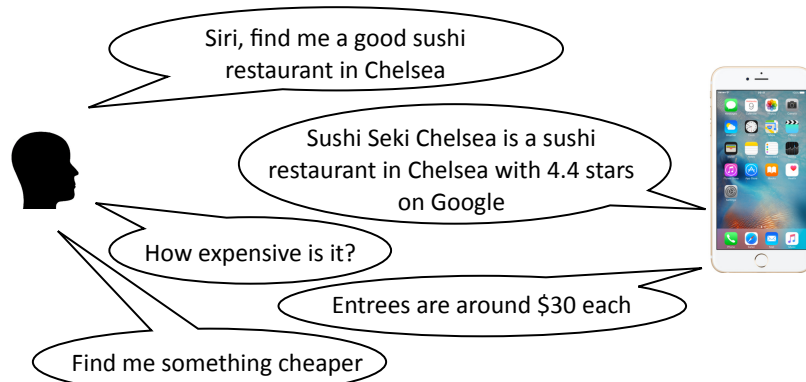
- ▶ Big model with RLHF. (More like a QA system than these other chatbots)
- ▶ Not much we can say except:
 - ▶ It's based on the earlier davinci models
 - ▶ Lots of data collection to fencepost it (e.g., "I don't know anything about the current weather ...")
 - ▶ Continuously improved without detailed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

Task-Oriented Dialogue



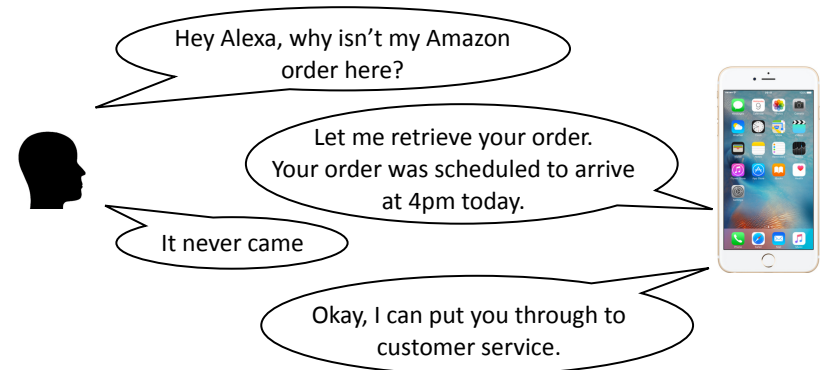
Task-Oriented Dialogue

- How do you build conversational systems to do things?



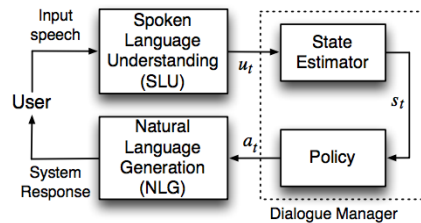
Task-Oriented Dialogue

- 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 -> update dialogue state -> take action (e.g., query the restaurant database) -> 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**



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')
))
```



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	}	<code>restaurant_type <- sushi</code>
		<code>location <- Chelsea</code>
wizard types this out or invokes templates	}	<code>curr_result <- execute_search()</code>
		<code>Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google</code>



Task-Oriented Dialogue

- ▶ Building these systems takes a ton of engineering — 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)