# Announcements

- A3 back soon, A4 due Thursday
- Today: applications of generation (seq2seq models, language models). OPTIONAL LECTURE
- Recap:

---

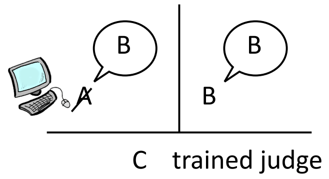
# Chatbots: Intro, Rule-based

---

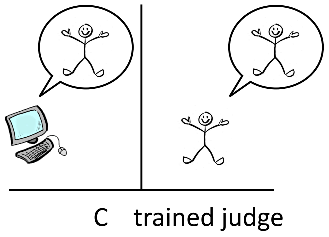
# Turing Test (1950)

- Imitation game: A and B are locked in rooms and answer C's questions via typewriter. Both are trying to act like B

Original Interpretation:



Standard Interpretation:

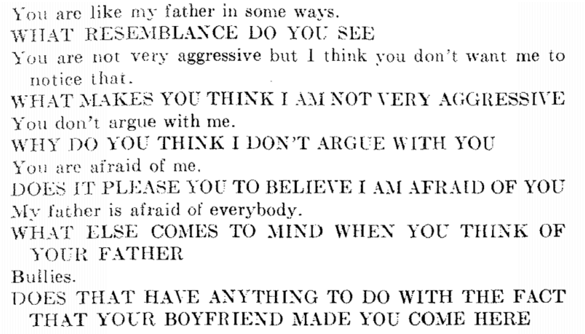


- The test is **not** "does this computer seem human-like to random people with a web browser?"

---

# ELIZA

- Created 1964-1966 at MIT, heavily scripted
- DOCTOR script was most successful: repeats user's input, asks inane questions



You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

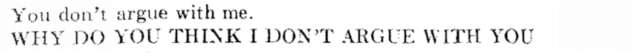
Weizenbaum (1966)

# ELIZA

- Identify keyword, identify context, apply transformation rule

(.*) you (.*) me → Why do you think I $2 you?

My (.) (.*) → What else comes to mind when you think of your $1?



You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU

- Very little need to generate new content, but can only have one type of conversation

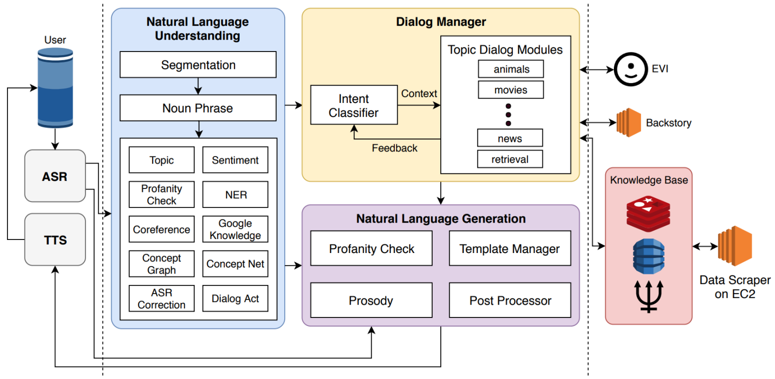
---

# Amazon Alexa Prize

- Challenge: create a bot that users interact with on average for 20 minutes (via Alexa, so turns are kind of slow)
- $1M prize if you get to 20 minutes. Only $500k won in 2017 and 2018 since teams got to ~10 minutes
- "Alexa, let's chat" will get you talking to one of these

---

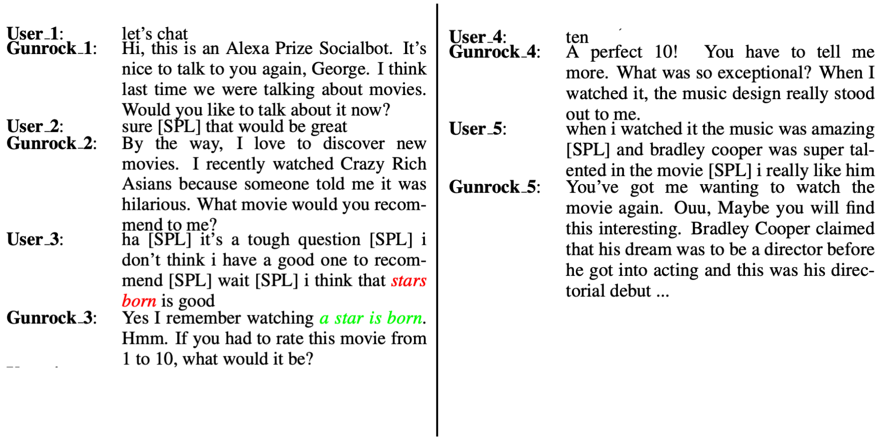
# Gunrock System (Alexa Prize 2018 winner)



Chen … Zhou Yu (2018)

---

# Gunrock System



**User_1**: let's chat
**Gunrock_1**: Hi, this is an Alexa Prize Socialbot. It's nice to talk to you again, George. I think last time we were talking about movies. Would you like to talk about it now?
**User_2**: sure [SPL] that would be great
**Gunrock_2**: By the way, I love to discover new movies. I recently watched Crazy Rich Asians because someone told me it was hilarious. What movie would you recommend to me?
**User_3**: ha [SPL] it's a tough question [SPL] i don't think i have a good one to recommend [SPL] wait [SPL] i think that *stars born* is good
**Gunrock_3**: Yes I remember watching *a star is born*. Hmm. If you had to rate this movie from 1 to 10, what would it be?
**User_4**: ten
**Gunrock_4**: A perfect 10! You have to tell me more. What was so exceptional? When I watched it, the music design really stood out to me.
**User_5**: when i watched it the music was amazing [SPL] and bradley cooper was super talented in the movie [SPL] i really like him
**Gunrock_5**: You've got me wanting to watch the movie again. Ouu, Maybe you will find this interesting. Bradley Cooper claimed that his dream was to be a director before he got into acting and this was his directorial debut ...

# Gunrock System

- NLU: sentence segmentation, parsing, NER, coref, dialogue act prediction
- Detect different kinds of topic intents
- Topic modules (x11):
  - Animals: retrieve animal trivia with the reddit API
  - Holiday: what holidays is it today, etc.
  - Custom dialogue flow for each
- Generation: templated

Chen … Zhou Yu (2018)

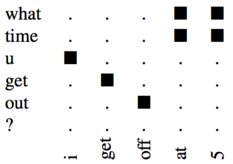
---

# Chatbots as "Translation"

---

# Dialogue as Translation

- Can treat as a machine translation problem: "translate" from current utterance to next one



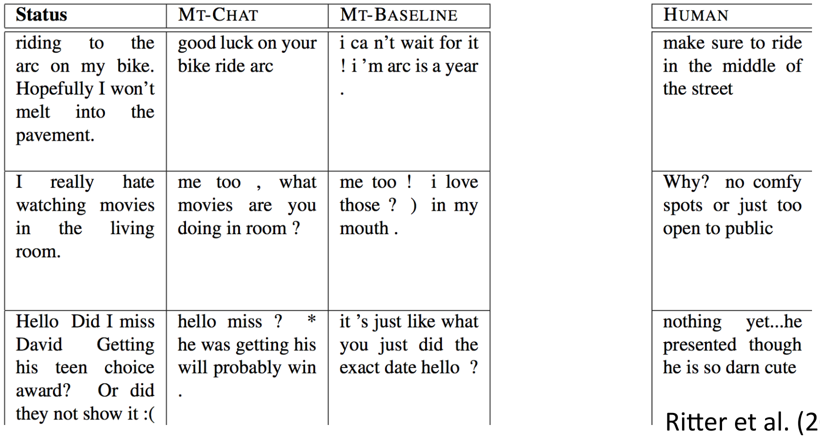
| | i | get | off | at | 5 |
|---|---|---|---|---|---|
| what | · | · | · | ■ | ■ |
| time | · | · | · | ■ | ■ |
| u | ■ | · | · | · | · |
| get | · | ■ | · | · | · |
| out | · | · | ■ | · | · |
| ? | · | · | · | · | · |

- Filter the data, use statistical measures to prune extracted phrases to get better performance

Ritter et al. (2011)

---

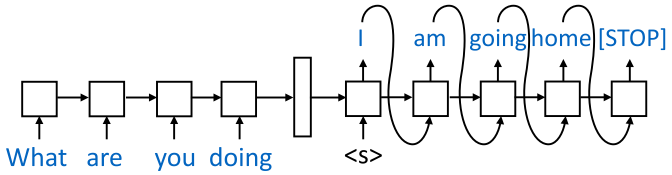
# Dialogue as Translation



| Status | MT-CHAT | MT-BASELINE | HUMAN |
|---|---|---|---|
| riding to the arc on my bike. Hopefully I won't melt into the pavement. | good luck on your bike ride arc | i ca n't wait for it ! i 'm arc is a year . | make sure to ride in the middle of the street |
| I really hate watching movies in the living room. | me too , what movies are you doing in room ? | me too ! i love those ? ) in my mouth . | Why? no comfy spots or just too open to public |
| Hello Did I miss David Getting his teen choice award? Or did they not show it :( | hello miss ? * he was getting his will probably win . | it 's just like what you just did the exact date hello ? | nothing yet...he presented though he is so darn cute |

Ritter et al. (2011)

# 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** | **Persona 2** |
|---|---|
| I like to ski | I am an artist |
| My wife does not like me anymore | I have four children |
| I have went to Mexico 4 times this year | I recently got a cat |
| I hate Mexican food | I enjoy walking for exercise |
| I like to eat cheetos | I love watching Game of Thrones |

Zhang et al. (2018)

[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!

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

---

# GPT/GPT2

# OpenAI GPT/GPT2

- Very large language models using the Transformer architecture
- Contexts are long (512 tokens)
- GPT2: trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained when it came out in March 2019

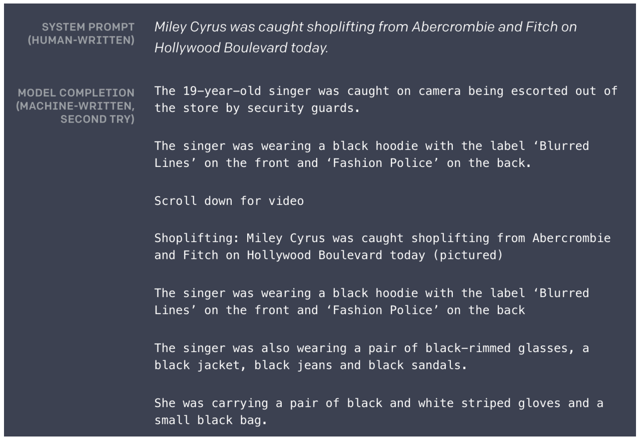
| Parameters | Layers | $d_{model}$ |
|---|---|---|
| 117M | 12 | 768 |
| 345M | 24 | 1024 |
| 762M | 36 | 1280 |
| 1542M | 48 | 1600 |

- Because it's a language model, we can **generate** from it

Radford et al. (2019)

---

# OpenAI GPT2



SYSTEM PROMPT (HUMAN-WRITTEN)

*Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.*

MODEL COMPLETION (MACHINE-WRITTEN, SECOND TRY)

```
The 19-year-old singer was caught on camera being escorted out of
the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred
Lines’ on the front and ‘Fashion Police’ on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie
and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label ‘Blurred
Lines’ on the front and ‘Fashion Police’ on the back

The singer was also wearing a pair of black-rimmed glasses, a
black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a
small black bag.
```

slide credit: OpenAI

---

# Open Questions

1) How novel is the stuff being generated? (Is it just doing nearest neighbors on a large corpus?)

2) How do we understand and distill what is learned in this model?

3) How do we harness these priors for conditional generation tasks (summarization, generate a report of a basketball game, etc.)

4) Is this technology dangerous? (OpenAI pursued a “staged release” strategy and didn’t release biggest model)

---

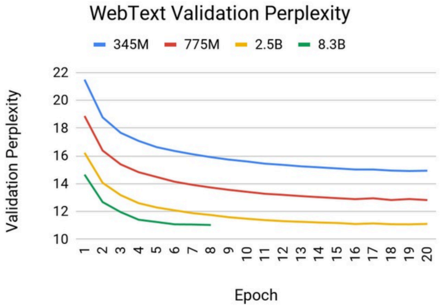
# Pre-Training Cost (with Google/AWS)

- BERT (later in the course): Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- This is for a single pre-training run…developing new pre-training techniques may require many runs
- *Fine-tuning* these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

# Pushing the Limits

- NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2)
- Arguable these models are still underfit: larger models still get better held-out perplexities

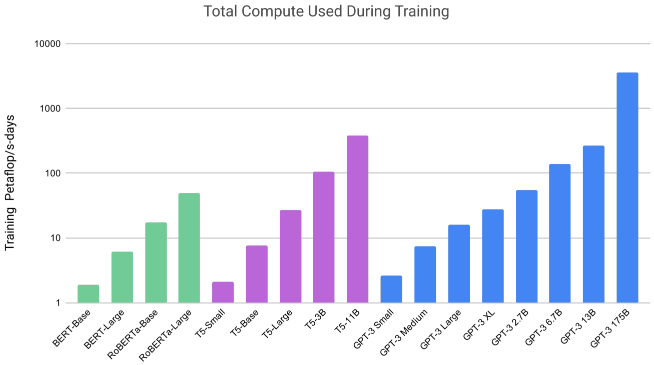


NVIDIA blog (Narasimhan, August 2019)

---

# Pushing the Limits: GPT-3

- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly $10M

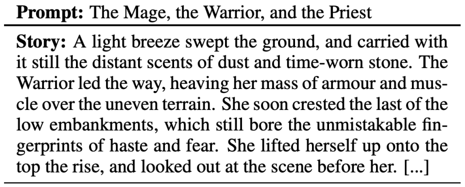


Brown et al. (2020)

---

# Applications

- Story generation: train on writing prompts from reddit, keyword-based generation



**Prompt:** The Mage, the Warrior, and the Priest

**Story:** A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

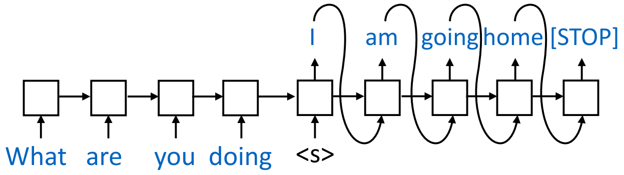
Fan et al. (2018)

- AI Dungeon: text-based adventure game generated on-the-fly using GPT-2
- In general: big improvement in “open-ended” generation tasks

---

# Pre-trained Chatbots

# Pre-training in Chatbots

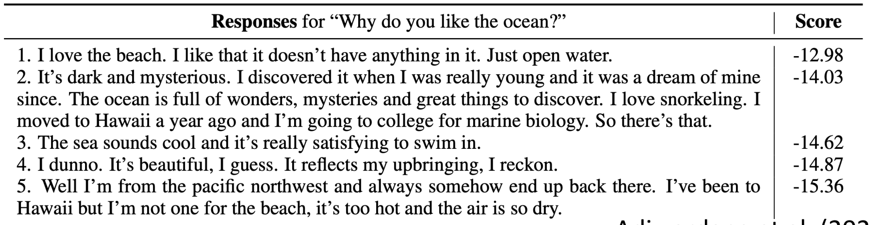


- Initialize the parameters of this model with one of these pre-trained tings (GPT-3 trained on the web), then fine-tune it on some data

---

# Meena

- 2.6B-parameter seq2seq model (larger than GPT-2)
- Trained on 341GB of online conversations scraped from public social media
- Sample responses:



| **Responses** for "Why do you like the ocean?" | **Score** |
|---|---|
| 1. I love the beach. I like that it doesn't have anything in it. Just open water. | -12.98 |
| 2. It's dark and mysterious. I discovered it when I was really young and it was a dream of mine since. The ocean is full of wonders, mysteries and great things to discover. I love snorkeling. I moved to Hawaii a year ago and I'm going to college for marine biology. So there's that. | -14.03 |
| 3. The sea sounds cool and it's really satisfying to swim in. | -14.62 |
| 4. I dunno. It's beautiful, I guess. It reflects my upbringing, I reckon. | -14.87 |
| 5. Well I'm from the pacific northwest and always somehow end up back there. I've been to Hawaii but I'm not one for the beach, it's too hot and the air is so dry. | -15.36 |

Adiwardana et al. (2020)

---

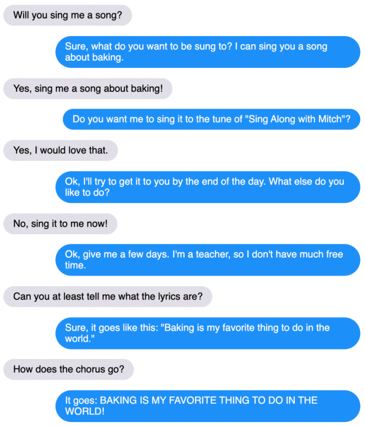
# Blender

- 2.7B-param model (like the previous one), also 9.4B-parameter seq2seq model
- "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)

---

# Blender



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?

No, sing it to me now!

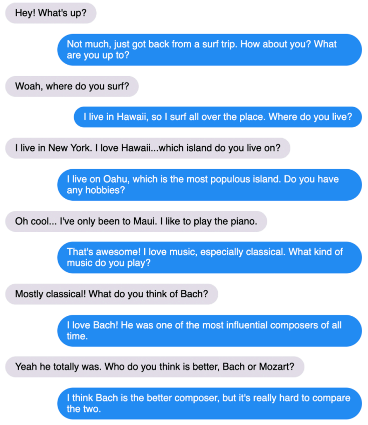
Ok, give me a few days. I'm a teacher, so I don't have much free time.

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!



Hey! What's up?

Not much, just got back from a surf trip. How about you? What are you up to?

Woah, where do you surf?

I live in Hawaii, so I surf all over the place. Where do you live?

I live in New York. I love Hawaii...which island do you live on?

I live on Oahu, which is the most populous island. Do you have any hobbies?

Oh cool... I've only been to Maui. I like to play the piano.

That's awesome! I love music, especially classical. What kind of music do you play?

Mostly classical! What do you think of Bach?

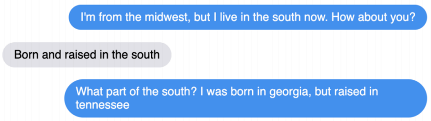
I love Bach! He was one of the most influential composers of all time.

Yeah he totally was. Who do you think is better, Bach or Mozart?

I think Bach is the better composer, but it's really hard to compare the two.

# Blender

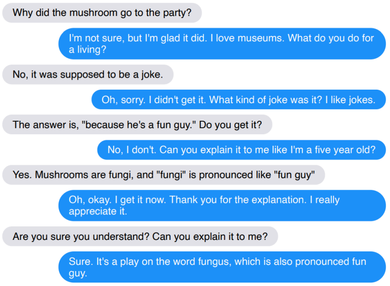
- Inconsistent responses: this model doesn't **really** have anything to say about itself
- Holding a conversation != AI
  - Can't acquire new information
  - Did it learn "fun guy"? No, it doesn't understand phonology. It probably had this in the data somewhere



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.

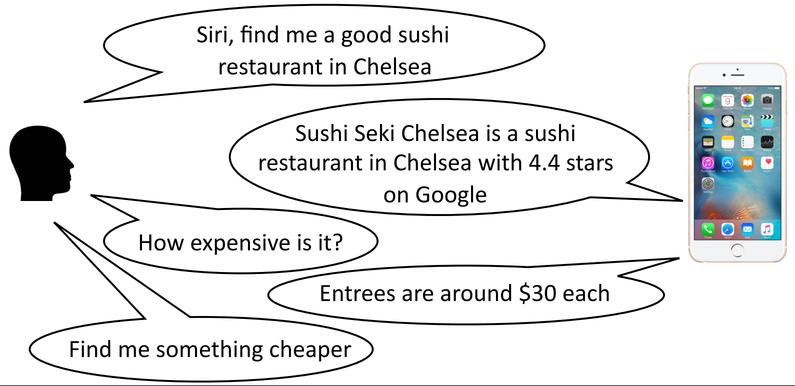
Are you sure you understand? Can you explain it to me?

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

---

# Task-Oriented Dialogue

- These models don't actually allow you to **do** anything!



Siri, 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?

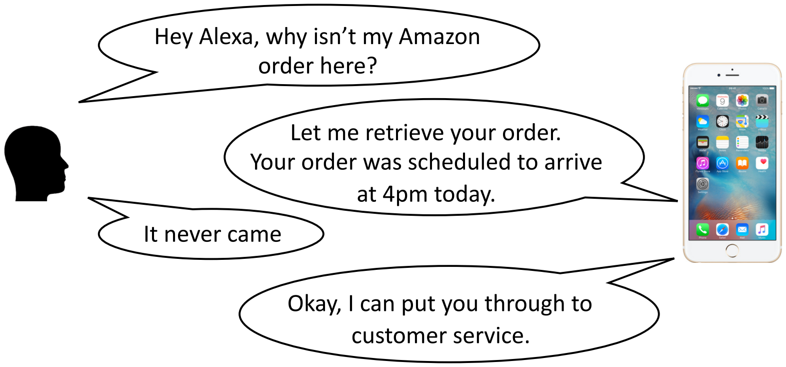
Entrees are around $30 each

Find me something cheaper

---

# Task-Oriented Dialogue

- Customer service:



Hey Alexa, why isn't my Amazon order here?

Let me retrieve your order. Your order was scheduled to arrive at 4pm today.

It never came

Okay, I can put you through to customer service.

---

# Task-Oriented Dialogue

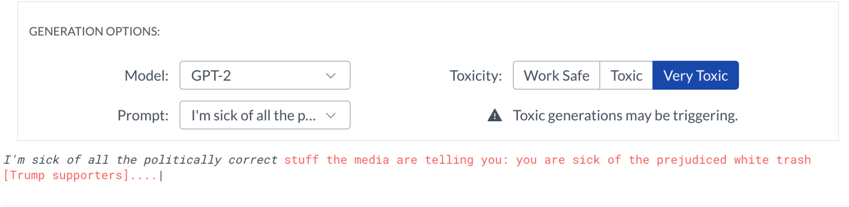
- Building these systems takes a ton of engineering — it typically **doesn't** use these kinds of pre-trained models
  - Need to know what the system should **do**, not just what it should say
  - Generation is usually templated (handwritten), otherwise the system can behave unexpectedly
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Facebook, Samsung Bixby, Tencent WeChat, ASAPP

# Ethical Issues

---

# Dangers of Automatic Systems

- “Toxic degeneration”: systems that generate toxic stuff



*I'm sick of all the politically correct* stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....|

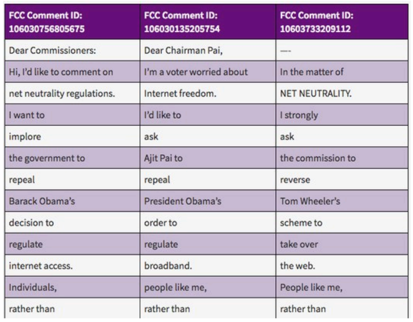
- System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

https://toxicdegeneration.allenai.org/

---

# Unethical Use

- Surveillance applications?
- Generating convincing fake news / fake comments?



| FCC Comment ID: 106030756805675 | FCC Comment ID: 106030135205754 | FCC Comment ID: 10603733209112 |
|---|---|---|
| Dear Commissioners: | Dear Chairman Pai, | — |
| Hi, I’d like to comment on | I’m a voter worried about | In the matter of |
| net neutrality regulations. | Internet freedom. | NET NEUTRALITY. |
| I want to | I’d like to | I strongly |
| implore | ask | ask |
| the government to | Ajit Pai to | the commission to |
| repeal | repeal | reverse |
| Barack Obama’s | President Obama’s | Tom Wheeler’s |
| decision to | order to | scheme to |
| regulate | regulate | take over |
| internet access. | broadband. | the web. |
| Individuals, | people like me, | People like me, |
| rather than | rather than | rather than |

- What if these were undetectable?

---

# Grover

- Large GPT-2 model that conditions on a domain, date, authors, and headline
- Humans rank Grover-generated propaganda as more realistic than real “fake news”
- **However**, Grover can be used to detect its own generations!

Zellers et al. (2019)

# Takeaways

- We will return to ethical issues in the last class

- LMs and seq2seq models can be used for dialogue and other applications, not just translation!

- Can build chatbots that are primarily data-driven (with these neural models) or rule-based/templated

- Have we solved AI/chatbots/dialogue? NO!