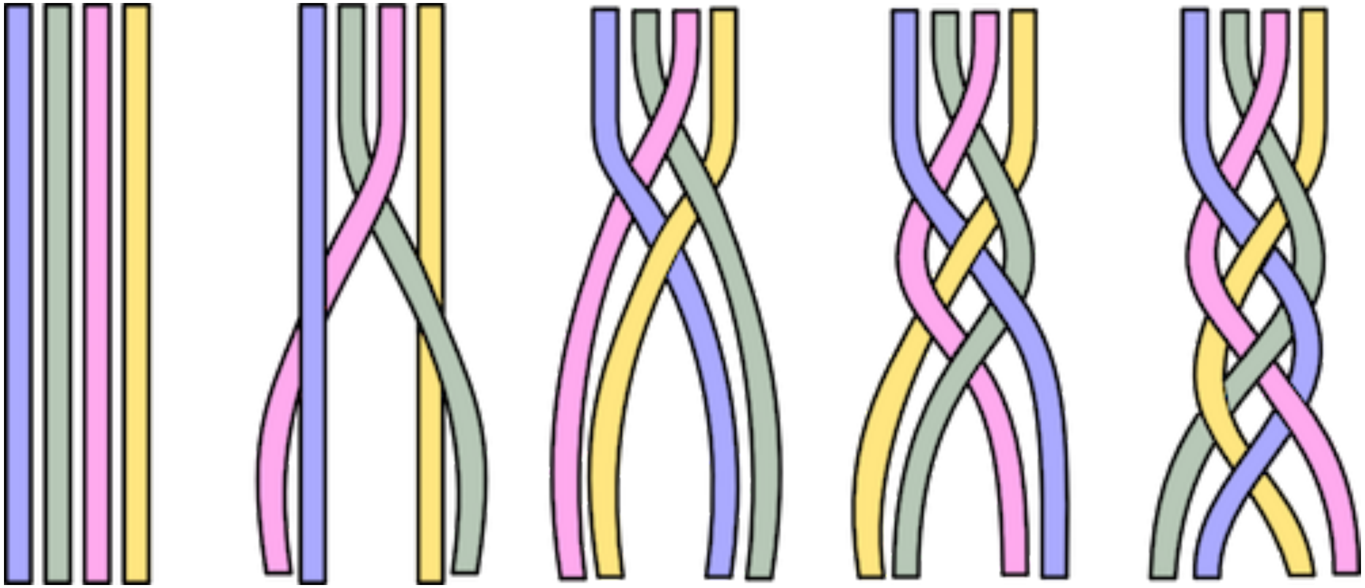
Putting together the threads of conversational AI?



Jason Weston

Facebook AI Research, NYC & NYU

My Long-Term Research Goal

An agent that can SEE, TALK, and continuously LEARN from interactions

Our goal: talk to our computer, *superhuman conversationalist*

Some properties our agent should have:

Engagingness:	have personality, curiosity, diversity
Expertness:	be knowledgeable & correct
Images:	can interact with text AND other modalities
Continual learning:	grow and learn from experience

Maybe we don't have enough threads yet to knit together the whole,
...but let's try anyway!

Persona 1	Persona 2
I am engaged to be married	I am a writer
I used to play sports in high school	I write romance novels
My favorite sport is soccer	I live on a faraway island
I am a college graduate	I am an introvert
I love watching sports on TV	I have two cats

[PERSON 1:] I just got done playing soccer. What have you been up to?
[PERSON 2:] I do not have pets. I spend my free time writing. what about you?
[PERSON 1:] Just graduated from college. Looking for work.
[PERSON 2:] yes I graduated, in writing and poetry
[PERSON 1:] Have you had anything published?
[PERSON 2:] I mainly write fiction novels.
[PERSON 1:] Do you like to watch sports?
[PERSON 2:] do you like kings of leon my favorite by them is use somebody
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

Persona-Chat (engaging personality)

A: Stylish *B: Fatalistic*

A: Riding a mechanical bull in a skirt is just my style.

B: You'd probably fall off and get hurt.

A: And everyone would be copying me for it! It'll be trendy!

A: Fearful *B: Miserable*

A: I just heard something out there and I have no idea what it was.

B: It was probably a Wolf coming to eat us because you talk too much.

A: I would never go camping in the woods for this very reason.

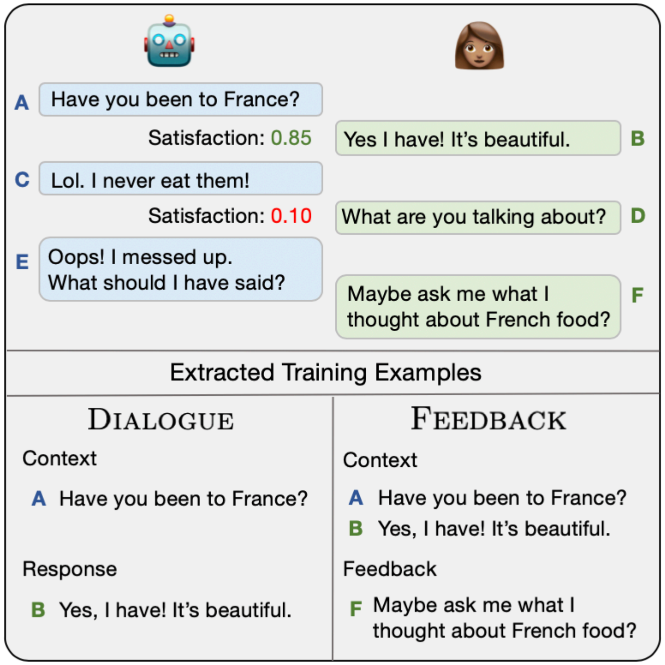
Image-Chat (images)

Wizard of Wikipedia (expert knowledge)

Topic: Toga party (*seen*)
Model: I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human: Me too!!! I usually use some flip flops and my blanket!
Model: Toga parties typically have games involving kegs and excessive drinking!
Human: Yeah! And getting to wear funny outfits and a crown of leaves!
Model: I love me some Toga Parties!
Human: I love me some too!
Model: Toga parties are so much fun with the Roman or Greek theme.
Human: Yeah! I especially like Greek themes

Topic: E-book (*seen*)
Human: i do not know why, but I have never gotten into E-books
Model: Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.
Human: yes, i like the physical feel and smell of a real book
Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human: i tried e-books but I do not like reading on a screen, its disorientated for me
Model: Yes, reading a book on a tablet just isn't the same!
Human: plus I like owning the physical copy of a real book, I feel like I own it
Model: I prefer print and newsreels.

Self-feeding (continual learning)



A Have you been to France?
Satisfaction: 0.85
Yes I have! It's beautiful. B
C Lol. I never eat them!
Satisfaction: 0.10
What are you talking about? D
E Oops! I messed up. What should I have said?
Maybe ask me what I thought about French food? F

Extracted Training Examples

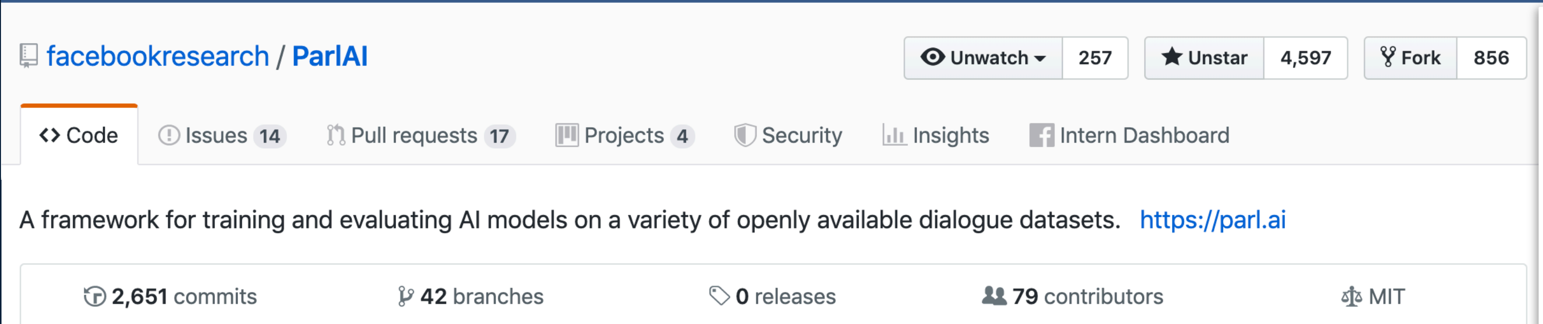
DIALOGUE	FEEDBACK
Context	Context
A Have you been to France?	A Have you been to France? B Yes, I have! It's beautiful.
Response	Feedback
B Yes, I have! It's beautiful.	F Maybe ask me what I thought about French food?

The Dream.. It all starts with software, right?????

ParlAI: A platform for training and evaluating dialog agents on a variety of openly available datasets.

Its goal is to provide the community:
- a unified framework for training and testing dialog models
- a repository of both learning agents and tasks, use both to iterate research!
- seamless integration of Amazon Mechanical Turk for data collection and human evaluation

Over ~~20~~ 80 tasks are supported, including popular datasets such as:
SQuAD, MCTest, WikiQA, WebQuestions, SimpleQuestions, WikiMovies, QACNN & QADailyMail, CBT, BookTest, bAbI tasks, bAbI Dialog tasks, Ubuntu Dialog, OpenSubtitles, Cornell Movie, VQA, VisDial & CLEVR. *Check it out:* **http://parl.ai**

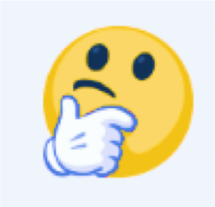


facebookresearch / ParlAI

A framework for training and evaluating AI models on a variety of openly available dialogue datasets. https://parl.ai

2,651 commits   42 branches   0 releases   79 contributors   MIT

What's inside ParlAI?

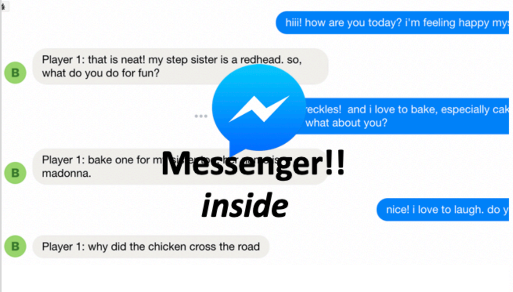


Models

- Transformers / BERT
- DrQA
- RNN Seq2seq
- RNN Language Model
- Memory Network
- Starspace
- TFIDF / IR retrievers
- MLB VQA
- VSE++ Captioning

Tools

- Mechanical Turk interface
- FB Messenger support
- Dictionary/vocabulary builder
- Tokenizers
- Torch model parent classes
- Data loading / streaming
- Pretrained models & embeddings



hiii! how are you today? i'm feeling happy my

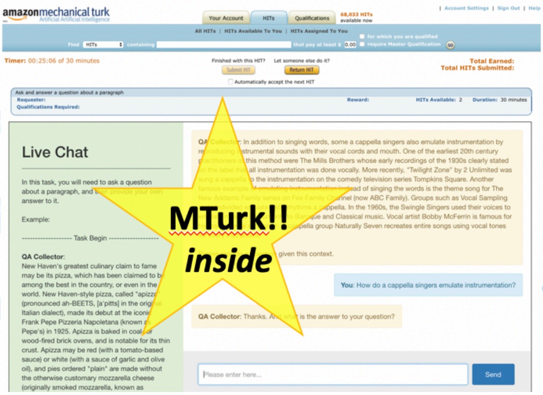
Player 1: that is neat! my step sister is a redhead. so, what do you do for fun?

reckles! and i love to bake, especially cak what about you?

Player 1: bake one for m[illegible] madonna.

nice! i love to laugh. do y

Player 1: why did the chicken cross the road


amazonmechanical turk
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
68,033 HITs available now
| Account Settings | Sign Out | Help
All HITs | HITs Available To You | HITs Assigned To You
Find
HITs
containing
that pay at least $
0.00
for which you are qualified
require Master Qualification
GO
Submit HIT
Return HIT
Send
MTurk!!
inside

Timer: 00:25:06 of 30 minutes

Finished with this HIT? Let someone else do it?

☐ Automatically accept the next HIT

Total Earned:
Total HITs Submitted:

Ask and answer a question about a paragraph
Requester:
Qualifications Required:
Reward:
HITs Available: 2
Duration: 30 minutes

Live Chat

In this task, you will need to ask a question about a paragraph, and then provide your own answer to it.

Example:

------------------ Task Begin ------------------

QA Collector:
New Haven's greatest culinary claim to fame may be its pizza, which has been claimed to be among the best in the country, or even in the world. New Haven-style pizza, called "apizza" (pronounced ah-BEETS, [a'pitts] in the original Italian dialect), made its debut at the iconic Frank Pepe Pizzeria Napoletana (known as Pepe's) in 1925. Apizza is baked in coal- or wood-fired brick ovens, and is notable for its thin crust. Apizza may be red (with a tomato-based sauce) or white (with a sauce of garlic and olive oil), and pies ordered "plain" are made without the otherwise customary mozzarella cheese (originally smoked mozzarella, known as

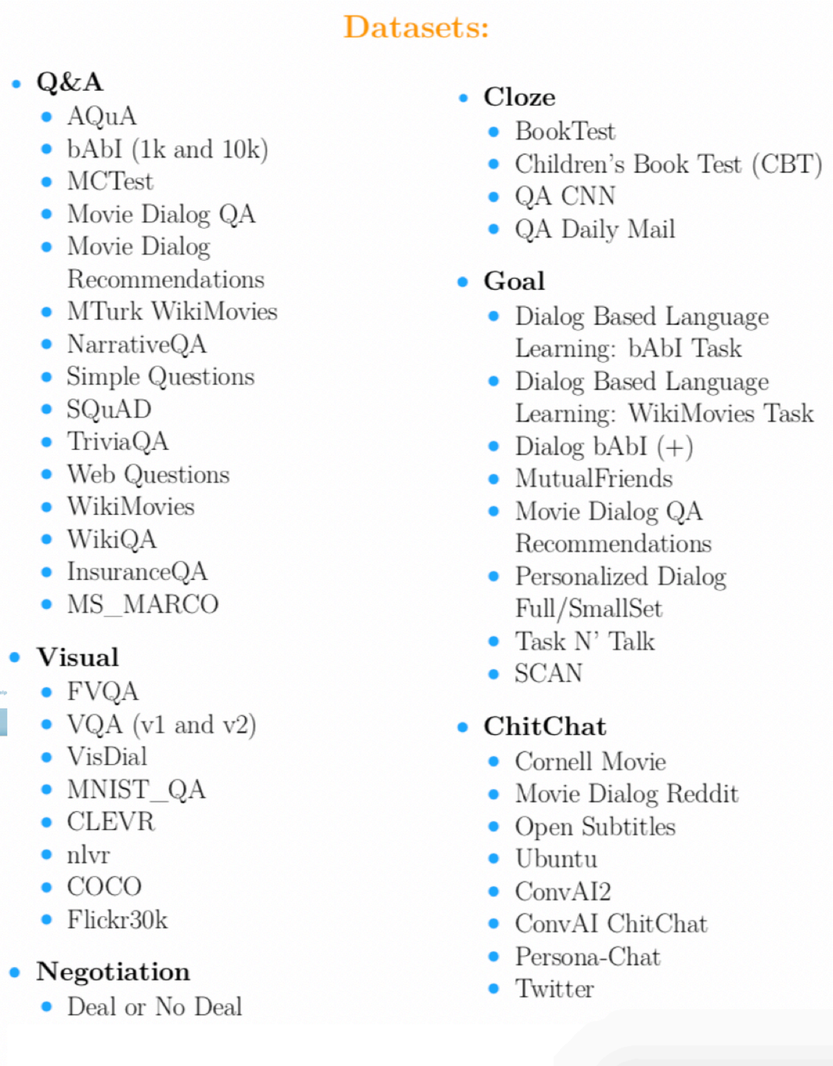
QA Collector: In addition to singing words, some a cappella singers also emulate instrumentation by reproducing instrumental sounds with their vocal cords and mouth. One of the earliest 20th century practitioners of this method were The Mills Brothers whose early recordings of the 1930s clearly stated on the label that all instrumentation was done vocally. More recently, "Twilight Zone" by 2 Unlimited was sung a cappella to the instrumentation on the comedy television series Tompkins Square. Another famous example of emulating instrumentation instead of singing the words is the theme song for The New Addams Family series on Fox Family Channel (now ABC Family). Groups such as Vocal Sampling [illegible] emulate [illegible] rhythms a cappella. In the 1960s, the Swingle Singers used their voices to [illegible] to Baroque and Classical music. Vocal artist Bobby McFerrin is famous for [illegible] cappella group Naturally Seven recreates entire songs using vocal tones for every instrument.

Please provide a question given this context.

You: How do a cappella singers emulate instrumentation?

QA Collector: Thanks. And what is the answer to your question?

Please enter here...



Datasets:

- **Q&A**
  - AQuA
  - bAbI (1k and 10k)
  - MCTest
  - Movie Dialog QA
  - Movie Dialog Recommendations
  - MTurk WikiMovies
  - NarrativeQA
  - Simple Questions
  - SQuAD
  - TriviaQA
  - Web Questions
  - WikiMovies
  - WikiQA
  - InsuranceQA
  - MS_MARCO
- **Visual**
  - FVQA
  - VQA (v1 and v2)
  - VisDial
  - MNIST_QA
  - CLEVR
  - nlvr
  - COCO
  - Flickr30k
- **Negotiation**
  - Deal or No Deal
- **Cloze**
  - BookTest
  - Children's Book Test (CBT)
  - QA CNN
  - QA Daily Mail
- **Goal**
  - Dialog Based Language Learning: bAbI Task
  - Dialog Based Language Learning: WikiMovies Task
  - Dialog bAbI (+)
  - MutualFriends
  - Movie Dialog QA Recommendations
  - Personalized Dialog Full/SmallSet
  - Task N' Talk
  - SCAN
- **ChitChat**
  - Cornell Movie
  - Movie Dialog Reddit
  - Open Subtitles
  - Ubuntu
  - ConvAI2
  - ConvAI ChitChat
  - Persona-Chat
  - Twitter

Code and models for all the work talked about here!!

🧙‍♀️ of Wikipedia: Knowledge-Powered Conversational Agents

*Emily Dinan, *Stephen Roller, *Kurt Shuster, Angela Fan, Michael Auli, Jason We

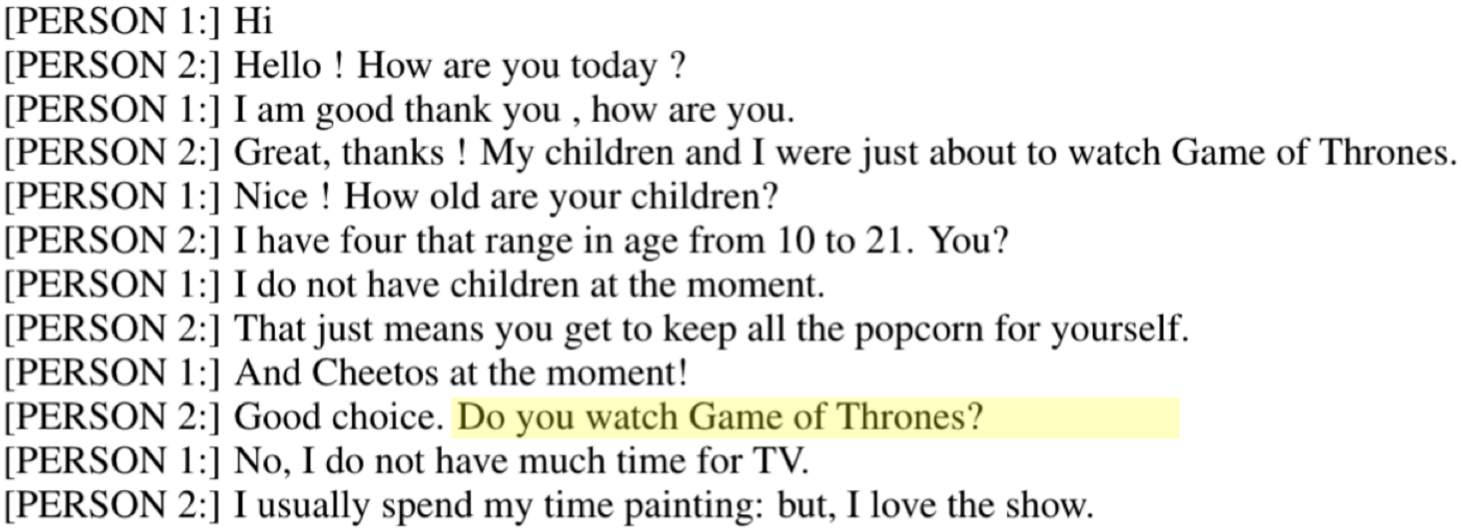
(*joint first authors)

Current chatbots just chat, they aren't (em)powered with ***knowledge***

- Seq2Seq models take in last lines of dialog, output a sentence
- We should ground on concrete information, rather than "generate and hope"

E.g. many studies/models using non-knowledge-based chit-chat datasets:

- Open-Subtitles (Vinyals & Le, 2015)
- Persona-Chat (Zhang et al., 2018)
- Twitter (Sordoni et al., 2015)



[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!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Existing Work

Goal directed dialogue uses knowledge :

- but via an API on structured knowledge
- e.g. restaurant or airline booking. (El Asri et al., 2017; Bordes et al., 2017)

Question-answering uses knowledge:

- but can only answer questions
- e.g. SQuAD (Rajpurkar et al., 2016) or QuAC (Choi et al., 2018)
- can use a retriever over unstructured text, e.g. Open-SQuAD (Chen et al., 2017)

Text based Knowledge for dialogue:

- Ghazvininejad et al. (2018) - local businesses using Foursquare tips as knowledge
- Parthasarathi & Pineau (2018) - news articles using Wikipedia summaries
- Moghe et al. (2018) – discuss movies given plot,reviews,etc.
- "A Dataset for Document Grounded Conversations" Zhou et al., this EMNLP!

To our knowledge, no convincing demonstration yet of full multi-turn dialogue in an open-domain setting…

Open-Domain Dialogue Setting

- Initial Starting Topic
- Pair of speakers converse naturally.
- They chat & learn from each other: discussion, facts, opinions.

Open-Domain Dialogue Setting

- Initial Starting Topic
- Pair of speakers converse naturally.
- They chat & learn from each other: discussion, facts, opinions.

Non-symmetric speakers:

- *Apprentice*: curious & eager to learn!
- *Wizard*: knowledgeable on area, eager to discuss & be engaging.
  - Given an IR system to condition response on.
  - Clicks on sentences used.

1307 Diverse General Topics: *crowd-sourced*

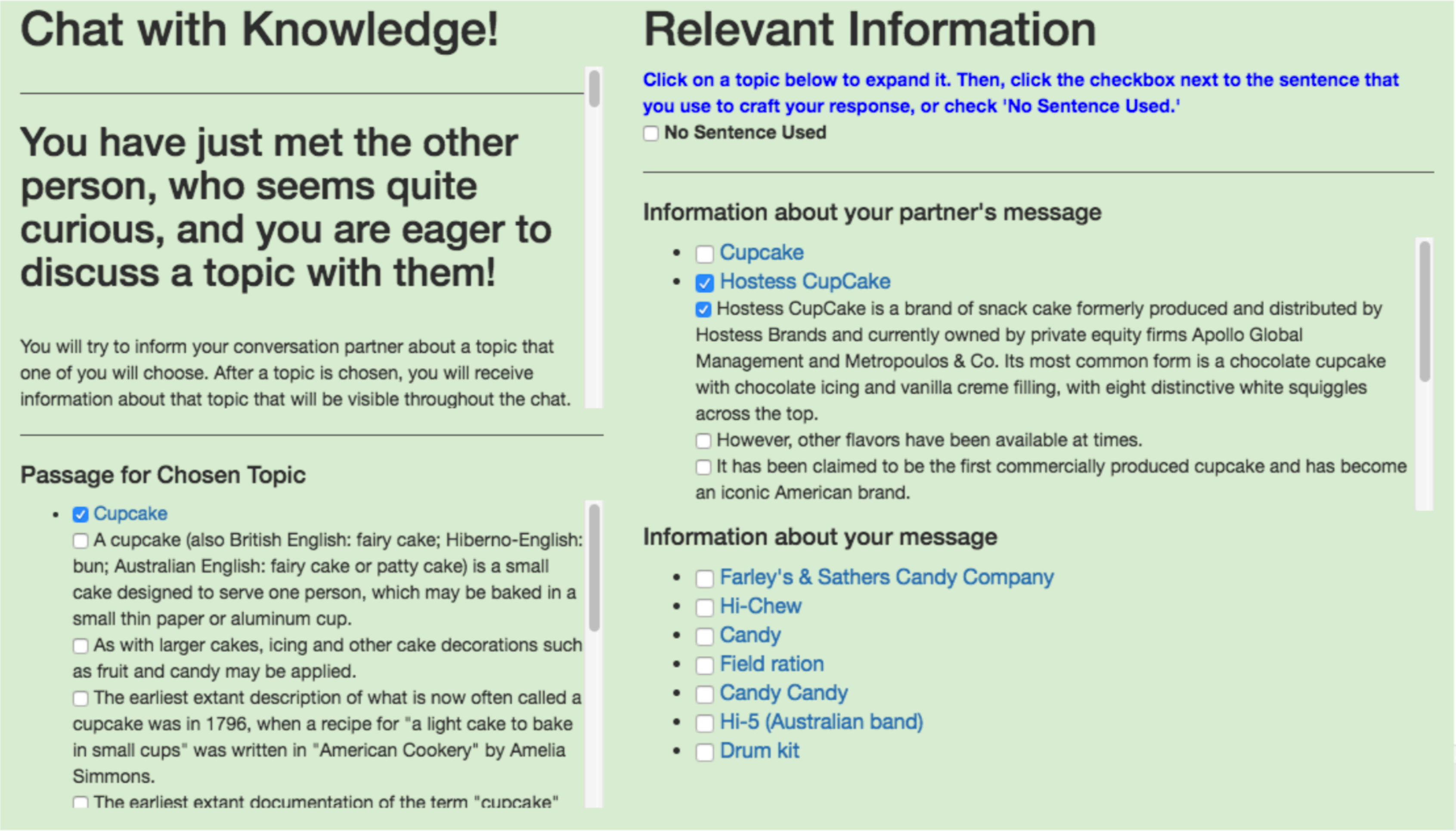
Gouda cheese
commuting
music festivals
podcasts
bowling
Arnold Schwarzenegger
Alpine skiing
Bodybuilding supplement
Harley-Davidson
Miley Cyrus
Hamilton (musical)
Ireland
Cannabis (drug)
Mount Kilimanjaro
Eggplant
Welding
Aquarium
Italian cuisine

Mercedes-Benz S-Class
Peanut
German language
Chicago-style pizza
Black hair
Toga party
100 metres
Tiger
Smoking
Winter
Acrophobia
List of art media
Pet adoption
Influencer marketing
Vitamin C
Human height
Steak

Kurt Cobain
List of water sports
Strawberry
Online game
Text messaging
Baileys Irish Cream
Fiction
American football
Online shopping
Aldi
Rock and roll
Kendrick Lamar
Medical billing
Blue Ridge Parkway
Clown
Pasta
Hiking

Each Linked to Wikiped

A.1 HUMAN ANNOTATION INTERFACE (FOR WIZARD)



Chat with Knowledge!

You have just met the other person, who seems quite curious, and you are eager to discuss a topic with them!

You will try to inform your conversation partner about a topic that one of you will choose. After a topic is chosen, you will receive information about that topic that will be visible throughout the chat.

Passage for Chosen Topic

- ☑ Cupcake

☐ A cupcake (also British English: fairy cake; Hiberno-English: bun; Australian English: fairy cake or patty cake) is a small cake designed to serve one person, which may be baked in a small thin paper or aluminum cup.

☐ As with larger cakes, icing and other cake decorations such as fruit and candy may be applied.

☐ The earliest extant description of what is now often called a cupcake was in 1796, when a recipe for "a light cake to bake in small cups" was written in "American Cookery" by Amelia Simmons.

☐ The earliest extant documentation of the term "cupcake"

Relevant Information

Click on a topic below to expand it. Then, click the checkbox next to the sentence that you use to craft your response, or check 'No Sentence Used.'

☐ **No Sentence Used**

Information about your partner's message

- ☐ Cupcake
- ☑ Hostess CupCake

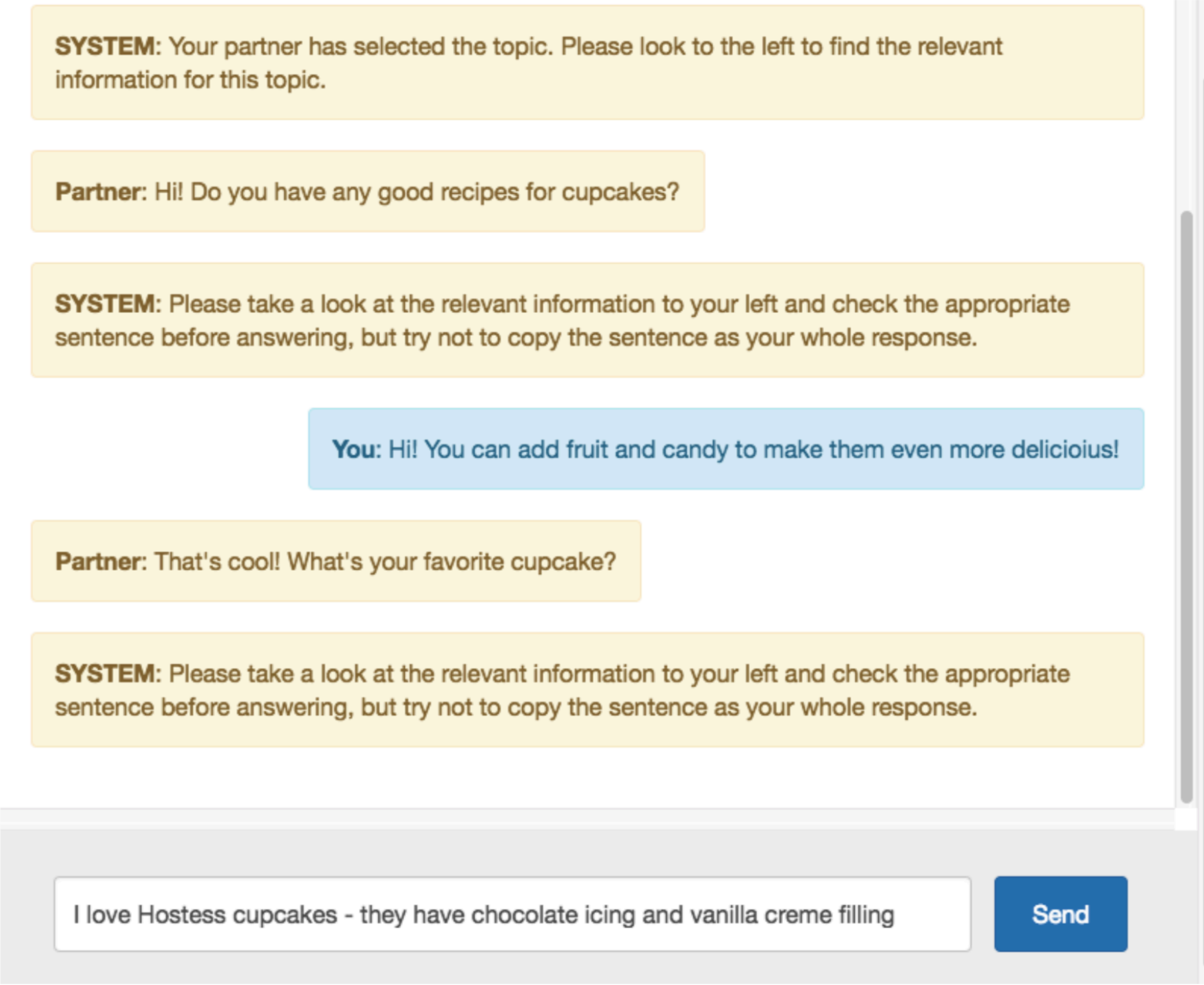
☑ Hostess CupCake is a brand of snack cake formerly produced and distributed by Hostess Brands and currently owned by private equity firms Apollo Global Management and Metropoulos & Co. Its most common form is a chocolate cupcake with chocolate icing and vanilla creme filling, with eight distinctive white squiggles across the top.

☐ However, other flavors have been available at times.

☐ It has been claimed to be the first commercially produced cupcake and has become an iconic American brand.

Information about your message

- ☐ Farley's & Sathers Candy Company
- ☐ Hi-Chew
- ☐ Candy
- ☐ Field ration
- ☐ Candy Candy
- ☐ Hi-5 (Australian band)
- ☐ Drum kit



SYSTEM: Your partner has selected the topic. Please look to the left to find the relevant information for this topic.

Partner: Hi! Do you have any good recipes for cupcakes?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

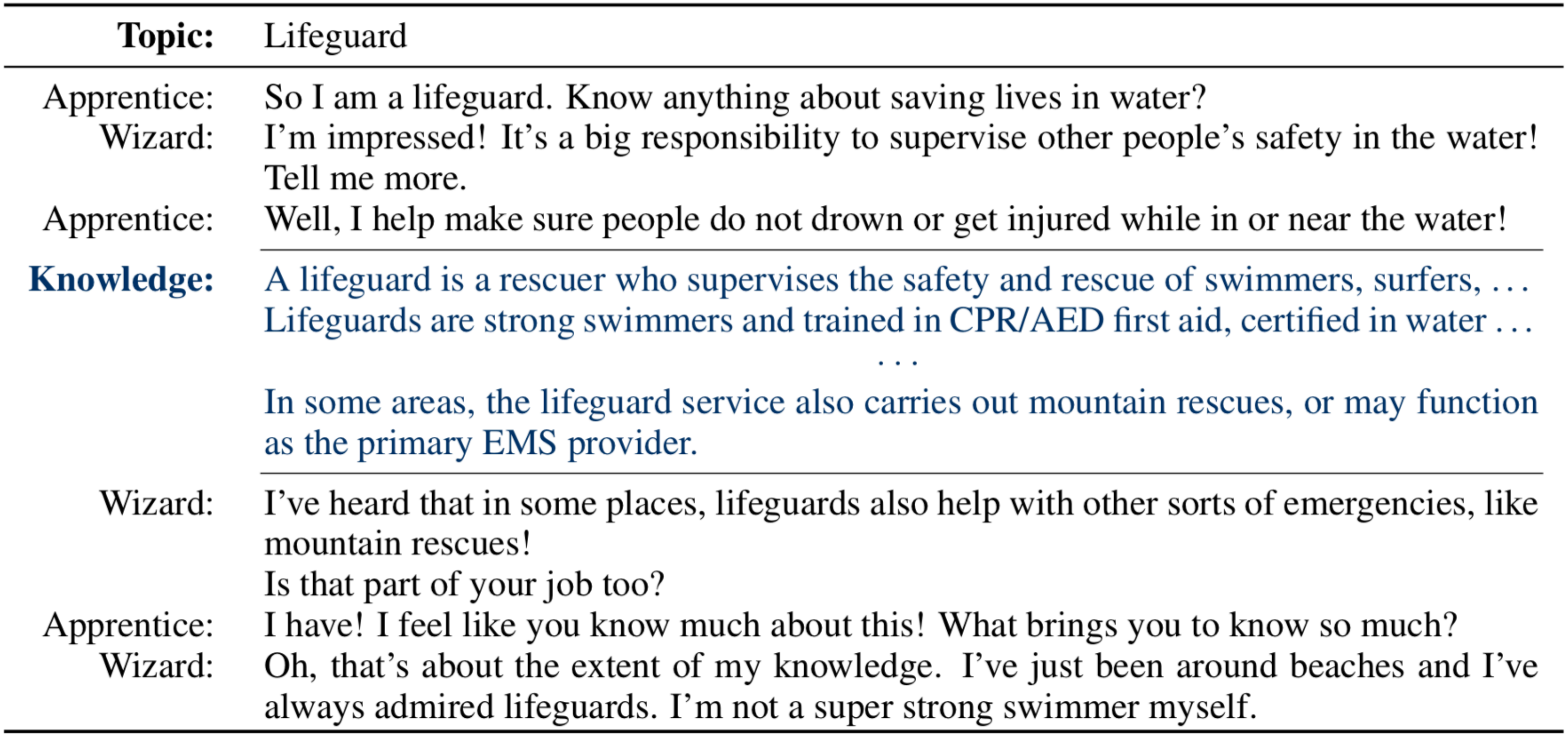
You: Hi! You can add fruit and candy to make them even more delicioius!

Partner: That's cool! What's your favorite cupcake?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

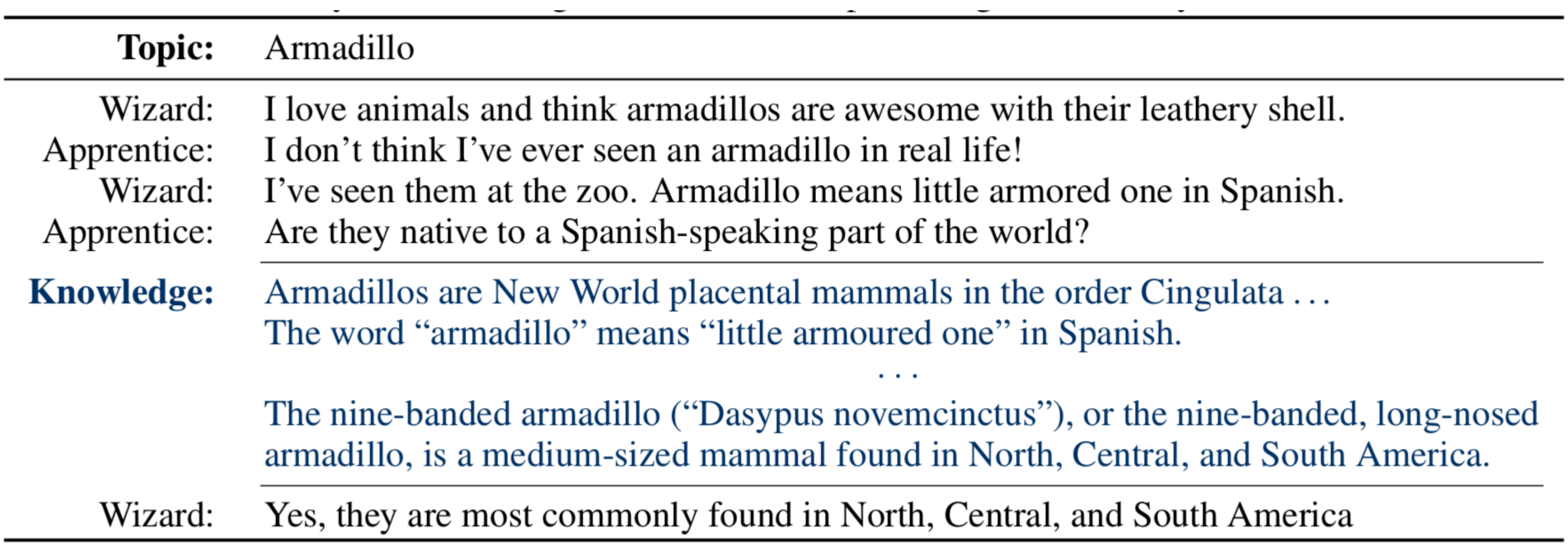
I love Hostess cupcakes - they have chocolate icing and vanilla creme filling

Dataset Examples



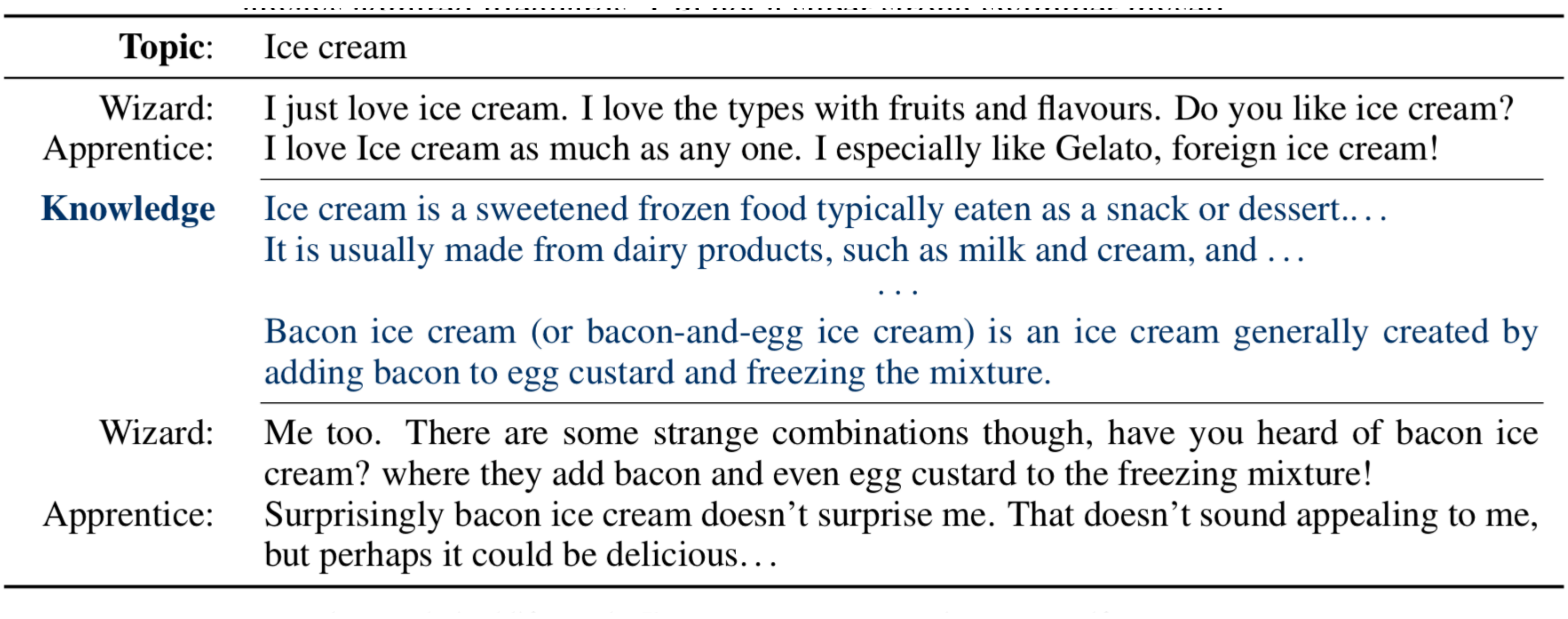
Topic:	Lifeguard
Apprentice:	So I am a lifeguard. Know anything about saving lives in water?
Wizard:	I'm impressed! It's a big responsibility to supervise other people's safety in the water! Tell me more.
Apprentice:	Well, I help make sure people do not drown or get injured while in or near the water!
Knowledge:	A lifeguard is a rescuer who supervises the safety and rescue of swimmers, surfers, … Lifeguards are strong swimmers and trained in CPR/AED first aid, certified in water … … In some areas, the lifeguard service also carries out mountain rescues, or may function as the primary EMS provider.
Wizard:	I've heard that in some places, lifeguards also help with other sorts of emergencies, like mountain rescues! Is that part of your job too?
Apprentice:	I have! I feel like you know much about this! What brings you to know so much?
Wizard:	Oh, that's about the extent of my knowledge. I've just been around beaches and I've always admired lifeguards. I'm not a super strong swimmer myself.

Dataset Examples



Topic:	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	Are they native to a Spanish-speaking part of the world?
Knowledge:	Armadillos are New World placental mammals in the order Cingulata . . . The word "armadillo" means "little armoured one" in Spanish. . . . The nine-banded armadillo ("Dasypus novemcinctus"), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America

Dataset Examples



Topic:	Ice cream
Wizard:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream?
Apprentice:	I love Ice cream as much as any one. I especially like Gelato, foreign ice cream!
Knowledge	Ice cream is a sweetened frozen food typically eaten as a snack or dessert.. . . It is usually made from dairy products, such as milk and cream, and . . . . . . Bacon ice cream (or bacon-and-egg ice cream) is an ice cream generally created by adding bacon to egg custard and freezing the mixture.
Wizard:	Me too. There are some strange combinations though, have you heard of bacon ice cream? where they add bacon and even egg custard to the freezing mixture!
Apprentice:	Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious. . .

Table 1: Dataset statistics of the Wizard of Wikipedia task.

Wizard of Wikipedia Task	Train	Valid	Test Seen	Test Unseen
Number of Utterances	166,787	17,715	8,715	8,782
Number of Dialogues	18,430	1,948	965	968
Number of Topics	1,247	599	533	58
Average Turns per Dialogue	9.0	9.1	9.0	9.1
Knowledge Database	5.4M articles		93M sentences	

Models

- If we can learn models on this data we think it could be (part of) one of the best chatbots out there. Might actually want to talk to it?
  - **Fun, engaging + knowledgeable !**

Models require:

1. Read current dialogue
2. Retrieve knowledge *(e.g. IR system, operates every turn)*
3. Read/attend on results
4. Condition & generate something fun ***and*** knowledgeable!

Generative Model architecture

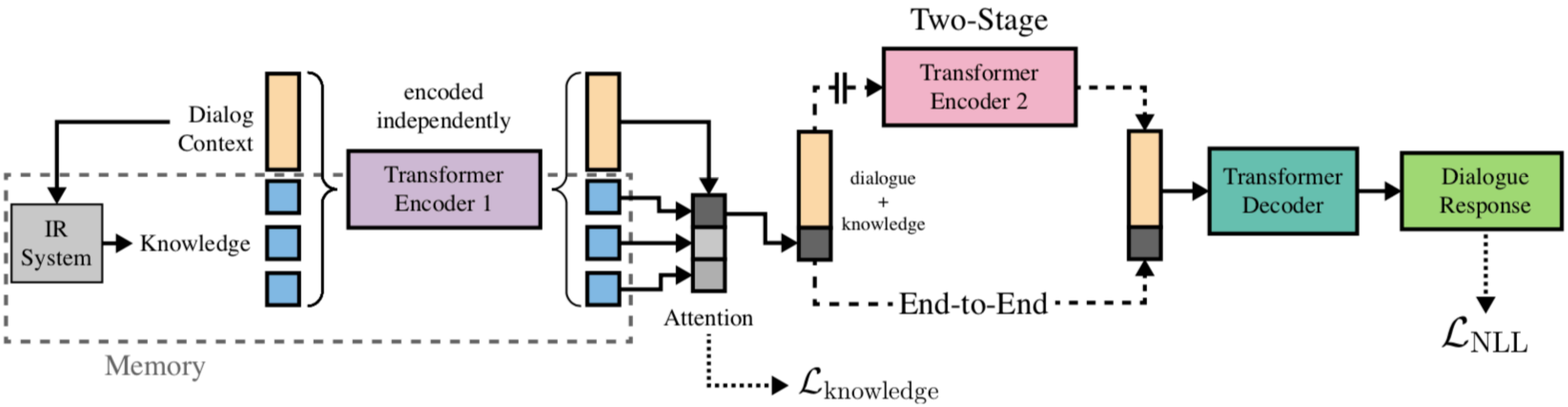
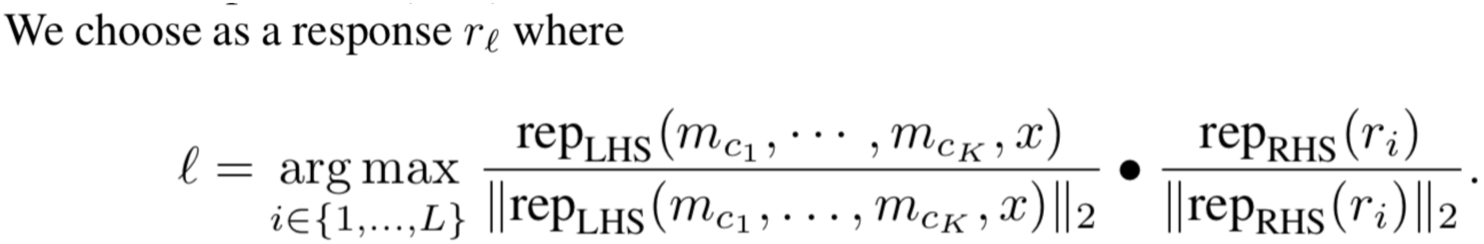


Figure 1: **Generative Transformer Memory Network.** An IR system provides knowledge candidates from Wikipedia. Dialogue Context and Knowledge are encoded using a shared encoder. In the Two-stage model, the dialogue and knowledge are re-encoded after knowledge selection.

Retrieval Model architecture

Similar to generative model, but:

- (1) we don’t attend to only the best returned knowledge *(keep the weighted attention sum)*
- (2) we replace the decoder with ranker of encoded sentences:



We choose as a response r_ℓ where

$$\ell = \underset{i\in\{1,\ldots,L\}}{\arg\max} \frac{\mathrm{rep}_{\mathrm{LHS}}(m_{c_1},\cdots,m_{c_K},x)}{\|\mathrm{rep}_{\mathrm{LHS}}(m_{c_1},\ldots,m_{c_K},x)\|_2} \bullet \frac{\mathrm{rep}_{\mathrm{RHS}}(r_i)}{\|\mathrm{rep}_{\mathrm{RHS}}(r_i)\|_2}.$$

- Transformers pretrained on 1.7 billion Reddit dialogue examples, following (Mazare´ et al., 2018). *Shown to work very well for PersonaChat dialogue*

Full Dialogue Task: Retrieval

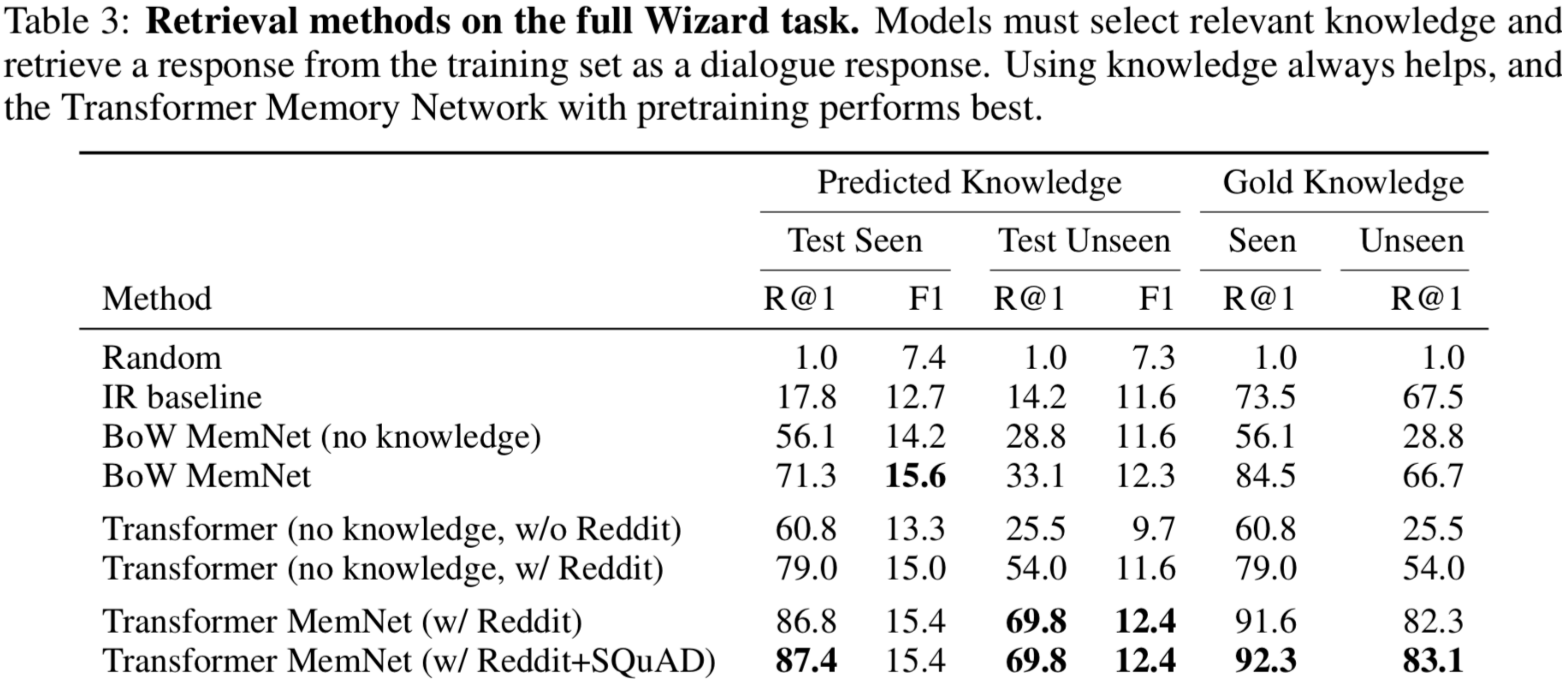


Table 3: **Retrieval methods on the full Wizard task.** Models must select relevant knowledge and retrieve a response from the training set as a dialogue response. Using knowledge always helps, and the Transformer Memory Network with pretraining performs best.

	Predicted Knowledge				Gold Knowledge	
	Test Seen		Test Unseen		Seen	Unseen
Method	R@1	F1	R@1	F1	R@1	R@1
Random	1.0	7.4	1.0	7.3	1.0	1.0
IR baseline	17.8	12.7	14.2	11.6	73.5	67.5
BoW MemNet (no knowledge)	56.1	14.2	28.8	11.6	56.1	28.8
BoW MemNet	71.3	**15.6**	33.1	12.3	84.5	66.7
Transformer (no knowledge, w/o Reddit)	60.8	13.3	25.5	9.7	60.8	25.5
Transformer (no knowledge, w/ Reddit)	79.0	15.0	54.0	11.6	79.0	54.0
Transformer MemNet (w/ Reddit)	86.8	15.4	**69.8**	**12.4**	91.6	82.3
Transformer MemNet (w/ Reddit+SQuAD)	**87.4**	15.4	**69.8**	**12.4**	**92.3**	**83.1**

Full Dialogue Task: Generation

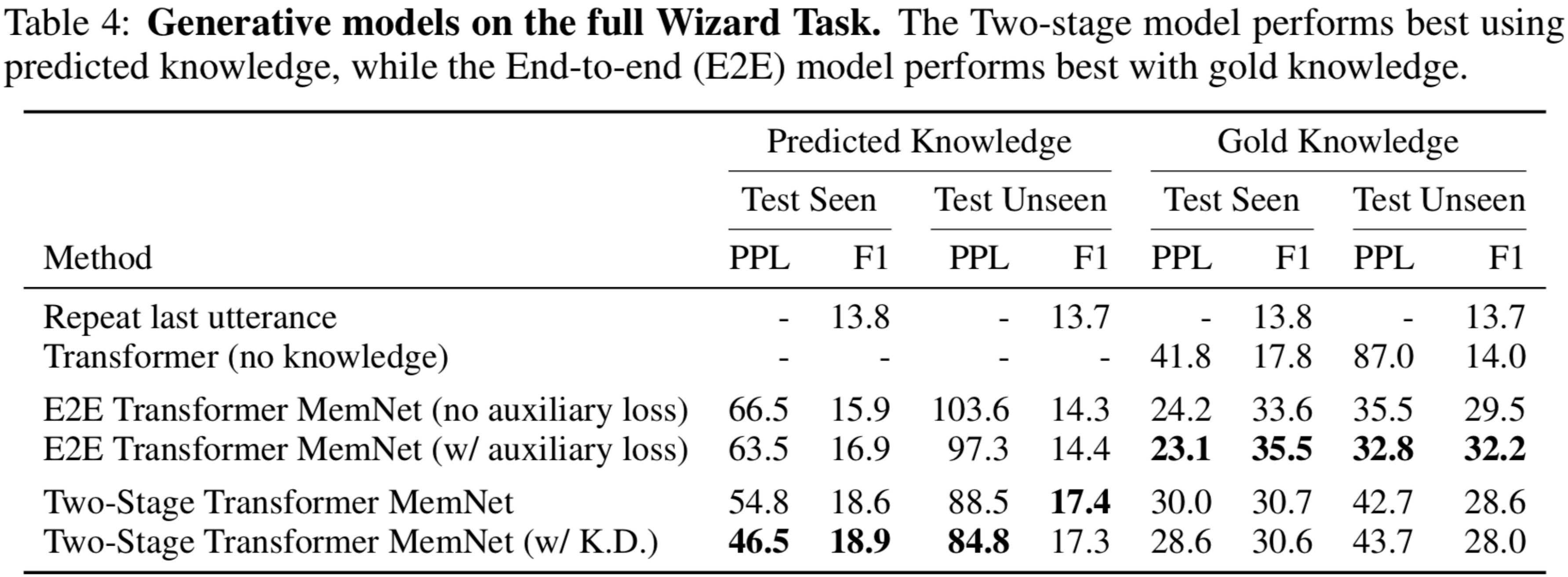


Table 4: **Generative models on the full Wizard Task.** The Two-stage model performs best using predicted knowledge, while the End-to-end (E2E) model performs best with gold knowledge.

	Predicted Knowledge				Gold Knowledge			
	Test Seen		Test Unseen		Test Seen		Test Unseen	
Method	PPL	F1	PPL	F1	PPL	F1	PPL	F1
Repeat last utterance	-	13.8	-	13.7	-	13.8	-	13.7
Transformer (no knowledge)	-	-	-	-	41.8	17.8	87.0	14.0
E2E Transformer MemNet (no auxiliary loss)	66.5	15.9	103.6	14.3	24.2	33.6	35.5	29.5
E2E Transformer MemNet (w/ auxiliary loss)	63.5	16.9	97.3	14.4	**23.1**	**35.5**	**32.8**	**32.2**
Two-Stage Transformer MemNet	54.8	18.6	88.5	**17.4**	30.0	30.7	42.7	28.6
Two-Stage Transformer MemNet (w/ K.D.)	**46.5**	**18.9**	**84.8**	17.3	28.6	30.6	43.7	28.0

Human Evaluations

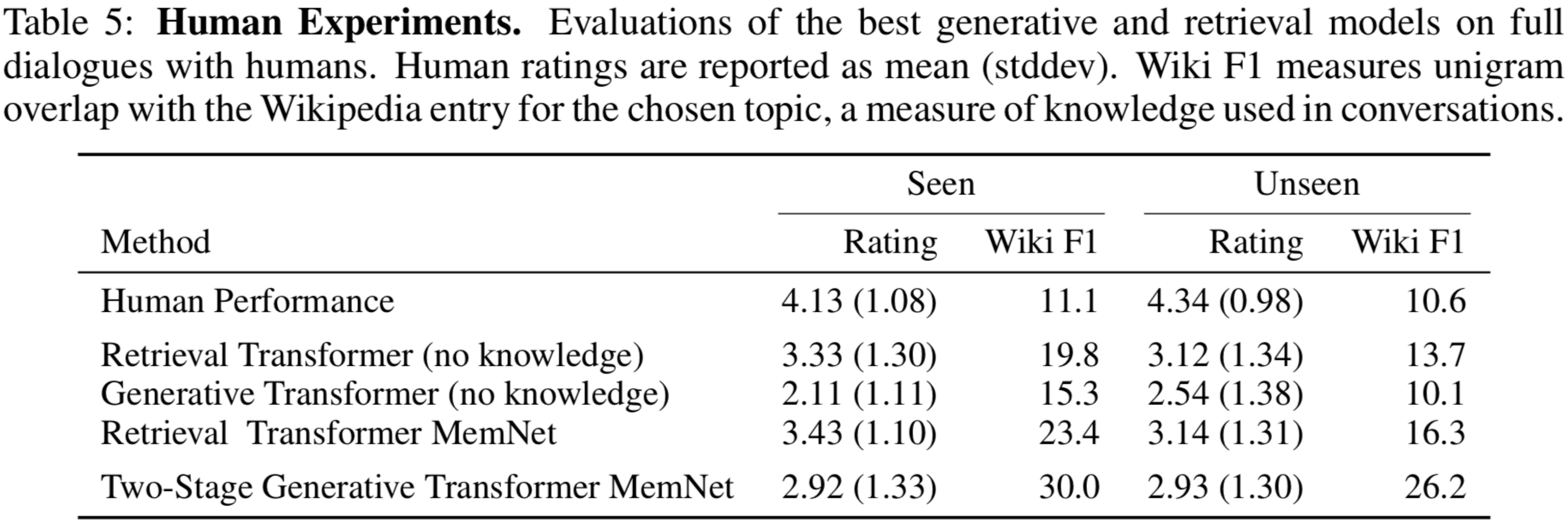
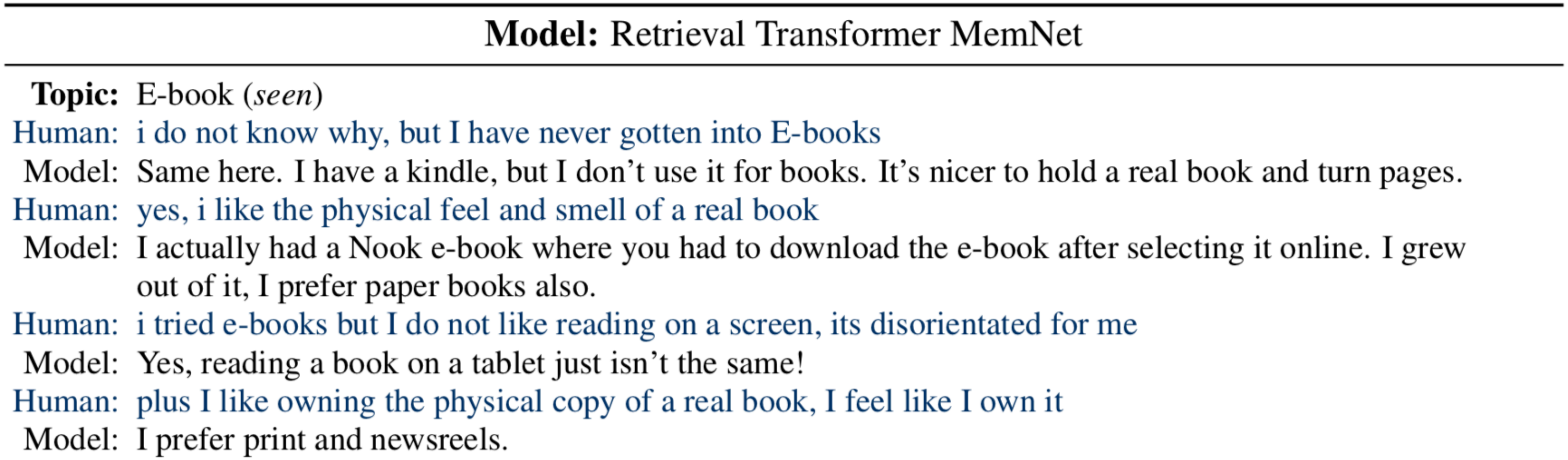


Table 5: **Human Experiments.** Evaluations of the best generative and retrieval models on full dialogues with humans. Human ratings are reported as mean (stddev). Wiki F1 measures unigram overlap with the Wikipedia entry for the chosen topic, a measure of knowledge used in conversations.

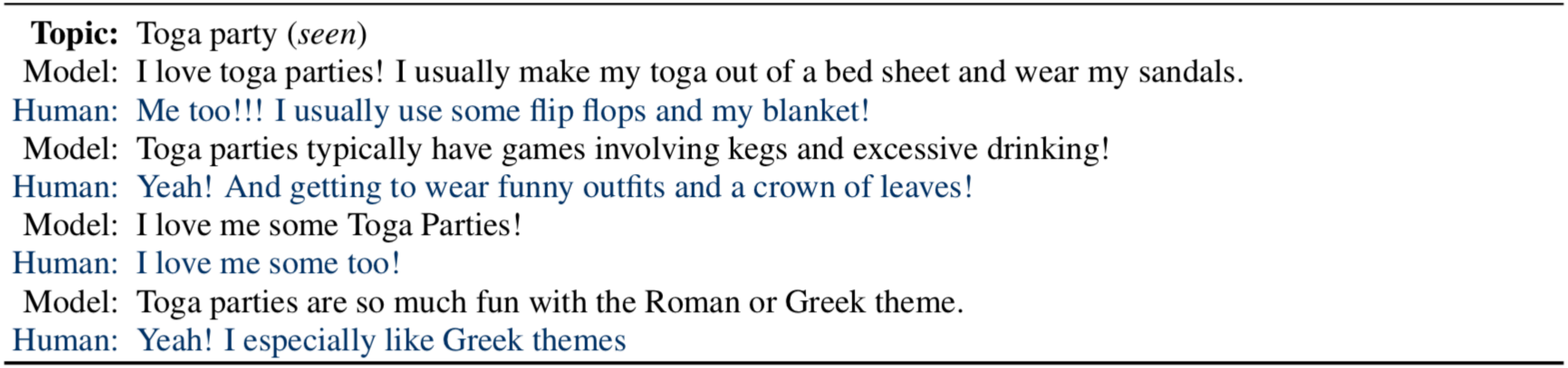
	Seen		Unseen	
Method	Rating	Wiki F1	Rating	Wiki F1
Human Performance	4.13 (1.08)	11.1	4.34 (0.98)	10.6
Retrieval Transformer (no knowledge)	3.33 (1.30)	19.8	3.12 (1.34)	13.7
Generative Transformer (no knowledge)	2.11 (1.11)	15.3	2.54 (1.38)	10.1
Retrieval Transformer MemNet	3.43 (1.10)	23.4	3.14 (1.31)	16.3
Two-Stage Generative Transformer MemNet	2.92 (1.33)	30.0	2.93 (1.30)	26.2

Model talking to Humans



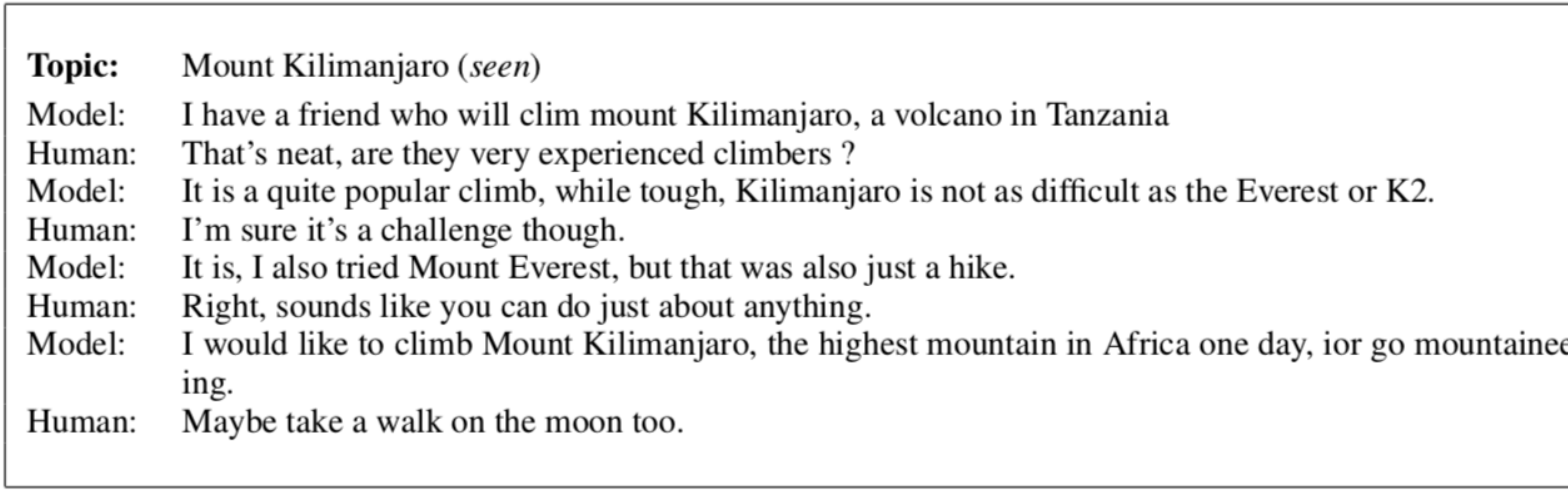
Model: Retrieval Transformer MemNet

Topic: E-book (*seen*)
Human: i do not know why, but I have never gotten into E-books
Model: Same here. I have a kindle, but I don’t use it for books. It’s nicer to hold a real book and turn pages.
Human: yes, i like the physical feel and smell of a real book
Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human: i tried e-books but I do not like reading on a screen, its disorientated for me
Model: Yes, reading a book on a tablet just isn’t the same!
Human: plus I like owning the physical copy of a real book, I feel like I own it
Model: I prefer print and newsreels.

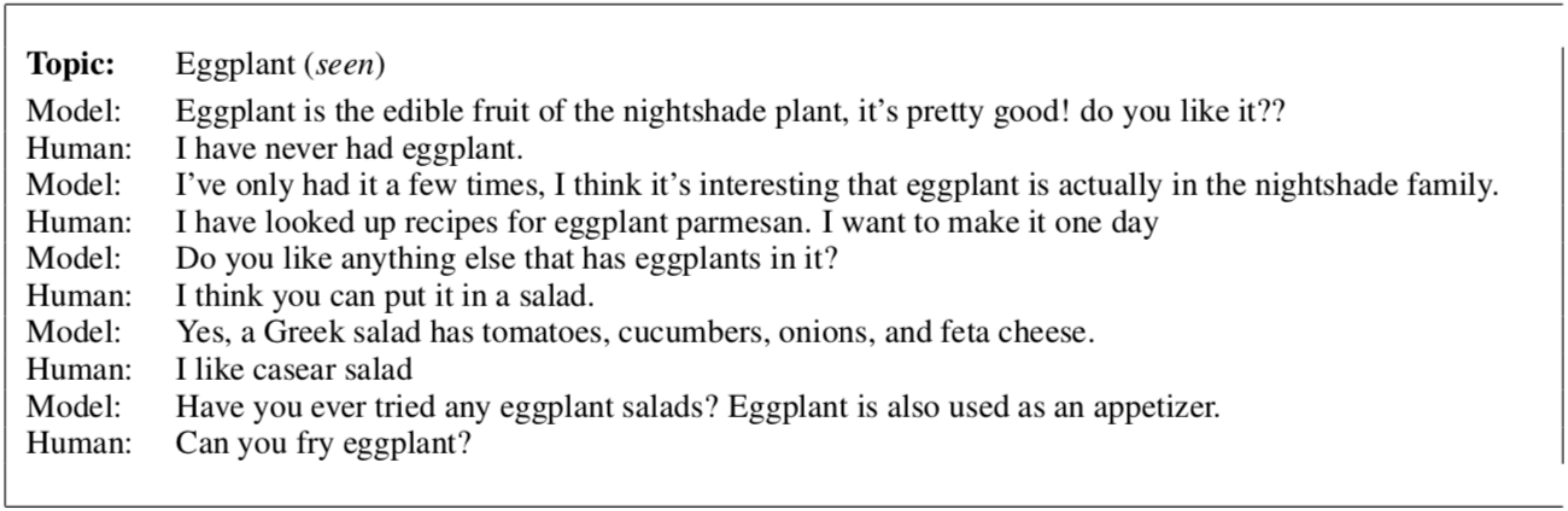


Topic: Toga party (*seen*)
Model: I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human: Me too!!! I usually use some flip flops and my blanket!
Model: Toga parties typically have games involving kegs and excessive drinking!
Human: Yeah! And getting to wear funny outfits and a crown of leaves!
Model: I love me some Toga Parties!
Human: I love me some too!
Model: Toga parties are so much fun with the Roman or Greek theme.
Human: Yeah! I especially like Greek themes

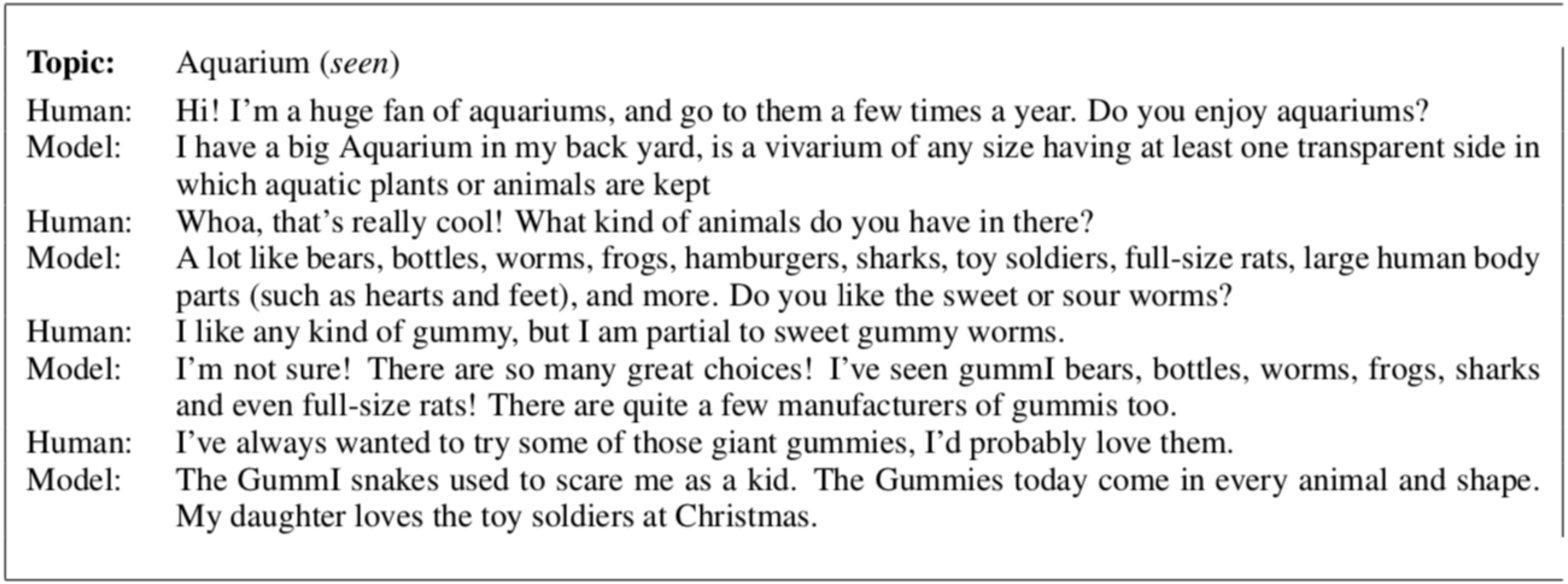
Topic:	Steak (*seen*)
Human:	Hello, it's lunch time here, and I'm in the mood for a great steak
Model:	Wow that sounds delicious I wish I could say the same hahah, When the word "steak" is used without qualification, it normally refers to a Beefsteak
Human:	Interesting, I don't really have a favorite steak, how about you?
Model:	I like rib eye it is really good specially medium rare, When the word "steak" is used without qualification, it generally refers to a beefsteak
Human:	I like rib eye too! What is it about a rib eye that you enjoy?
Model:	Most steak cuts are very tender meat but I think rib eye can be quite tender
Human:	They really are. How do you choose a good steak?
Model:	When it comes to steak and other cuts of meat you have rare, medium rare, medium, well done and so on
Human:	How do you prefer your steak prepared?
Model:	I like that you can prepare it many different ways. You can fried it, grill it, or even broil the steak



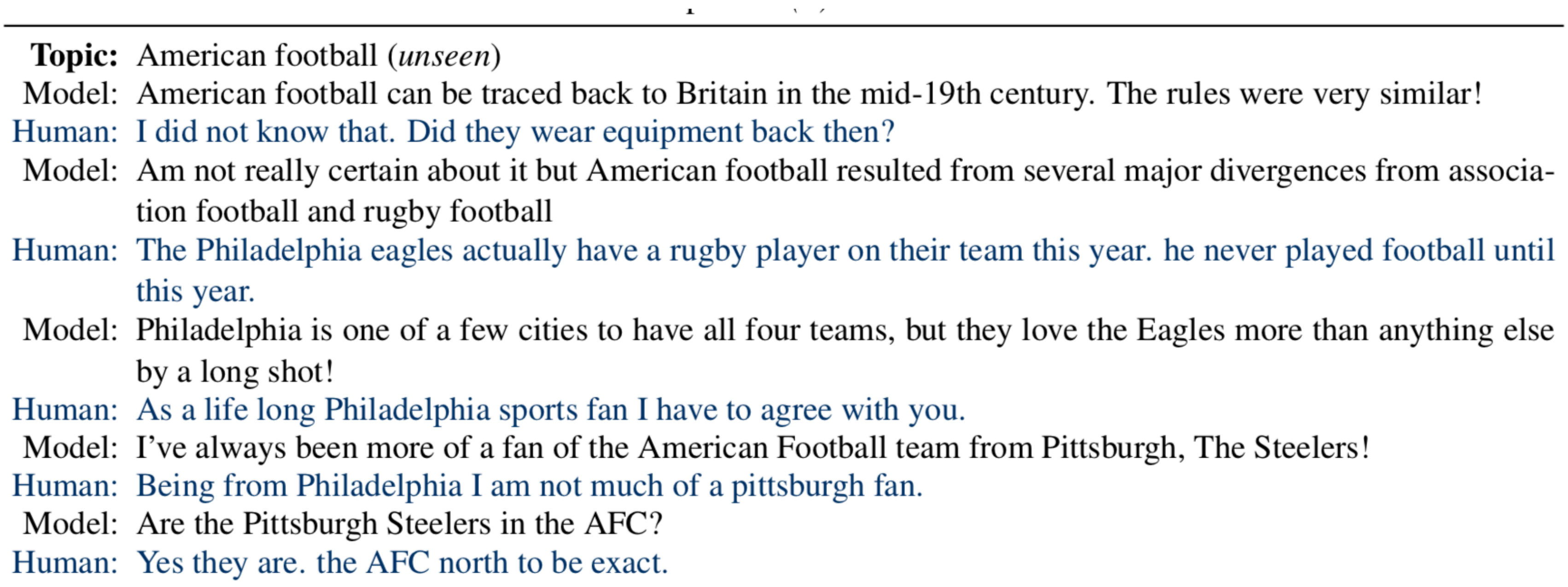
Topic:	Mount Kilimanjaro (*seen*)
Model:	I have a friend who will clim mount Kilimanjaro, a volcano in Tanzania
Human:	That's neat, are they very experienced climbers ?
Model:	It is a quite popular climb, while tough, Kilimanjaro is not as difficult as the Everest or K2.
Human:	I'm sure it's a challenge though.
Model:	It is, I also tried Mount Everest, but that was also just a hike.
Human:	Right, sounds like you can do just about anything.
Model:	I would like to climb Mount Kilimanjaro, the highest mountain in Africa one day, ior go mountainee ing.
Human:	Maybe take a walk on the moon too.



Topic: Eggplant (*seen*)
Model: Eggplant is the edible fruit of the nightshade plant, it's pretty good! do you like it??
Human: I have never had eggplant.
Model: I've only had it a few times, I think it's interesting that eggplant is actually in the nightshade family.
Human: I have looked up recipes for eggplant parmesan. I want to make it one day
Model: Do you like anything else that has eggplants in it?
Human: I think you can put it in a salad.
Model: Yes, a Greek salad has tomatoes, cucumbers, onions, and feta cheese.
Human: I like casear salad
Model: Have you ever tried any eggplant salads? Eggplant is also used as an appetizer.
Human: Can you fry eggplant?



Topic: Aquarium (*seen*)
Human: Hi! I'm a huge fan of aquariums, and go to them a few times a year. Do you enjoy aquariums?
Model: I have a big Aquarium in my back yard, is a vivarium of any size having at least one transparent side in which aquatic plants or animals are kept
Human: Whoa, that's really cool! What kind of animals do you have in there?
Model: A lot like bears, bottles, worms, frogs, hamburgers, sharks, toy soldiers, full-size rats, large human body parts (such as hearts and feet), and more. Do you like the sweet or sour worms?
Human: I like any kind of gummy, but I am partial to sweet gummy worms.
Model: I'm not sure! There are so many great choices! I've seen gummI bears, bottles, worms, frogs, sharks and even full-size rats! There are quite a few manufacturers of gummis too.
Human: I've always wanted to try some of those giant gummies, I'd probably love them.
Model: The GummI snakes used to scare me as a kid. The Gummies today come in every animal and shape. My daughter loves the toy soldiers at Christmas.



Topic: American football (*unseen*)

Model: American football can be traced back to Britain in the mid-19th century. The rules were very similar!

Human: I did not know that. Did they wear equipment back then?

Model: Am not really certain about it but American football resulted from several major divergences from association football and rugby football

Human: The Philadelphia eagles actually have a rugby player on their team this year. he never played football until this year.

Model: Philadelphia is one of a few cities to have all four teams, but they love the Eagles more than anything else by a long shot!

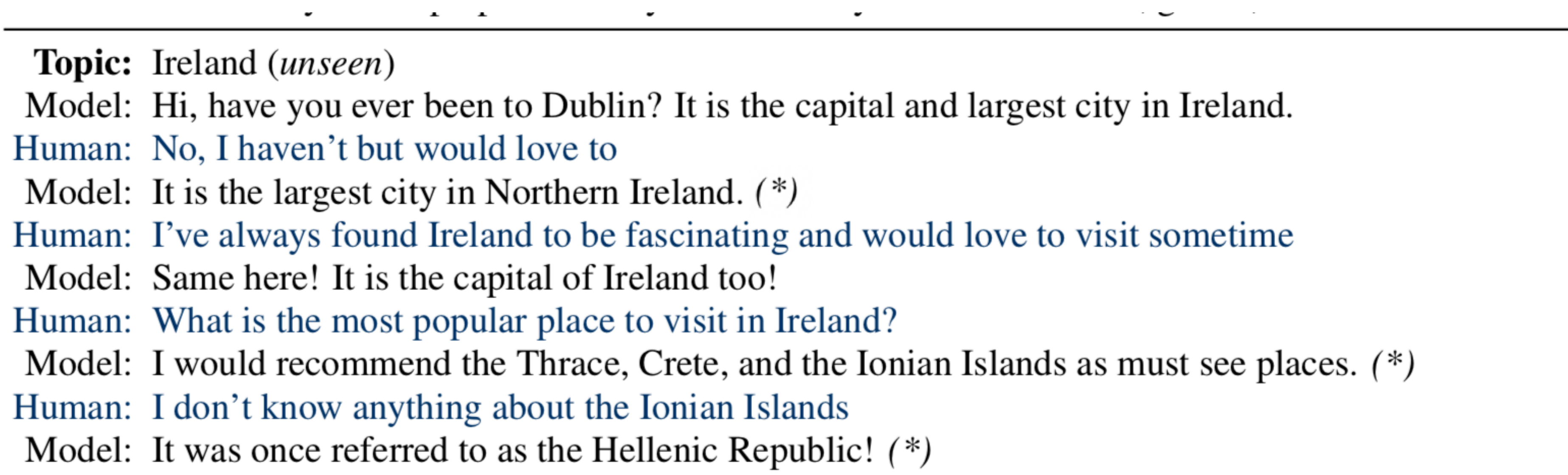
Human: As a life long Philadelphia sports fan I have to agree with you.

Model: I've always been more of a fan of the American Football team from Pittsburgh, The Steelers!

Human: Being from Philadelphia I am not much of a pittsburgh fan.

Model: Are the Pittsburgh Steelers in the AFC?

Human: Yes they are. the AFC north to be exact.



Topic: Ireland (*unseen*)

Model: Hi, have you ever been to Dublin? It is the capital and largest city in Ireland.

Human: No, I haven't but would love to

Model: It is the largest city in Northern Ireland. *(*)*

Human: I've always found Ireland to be fascinating and would love to visit sometime

Model: Same here! It is the capital of Ireland too!

Human: What is the most popular place to visit in Ireland?

Model: I would recommend the Thrace, Crete, and the Ionian Islands as must see places. *(*)*

Human: I don't know anything about the Ionian Islands

Model: It was once referred to as the Hellenic Republic! *(*)*

Model: Two-Stage Generative Transformer MemNet

Topic: Circus (*seen*)
Model: have you ever been to a circus? it is a company of performers who put on diverse entertainment shows.
Human: I went to the circus once when I was a child actually.
Model: that's great! entertainment is a form of activity that holds the attention and interest of an audience.
Human: That's true. Have you ever been to a circus?
Model: I have not, but I know they were built with various types of seating ring, and sometimes a stage.
Human: I went to one held in a stadium, so we had box seats.
Model: that's awesome! i'm sure you'll have to go to the circus, with appropriate music.
Human: Yes, circuses usually have their own kind of music.
Model: circuses can be done by circuses.

Model: Two-Stage Generative Transformer MemNet

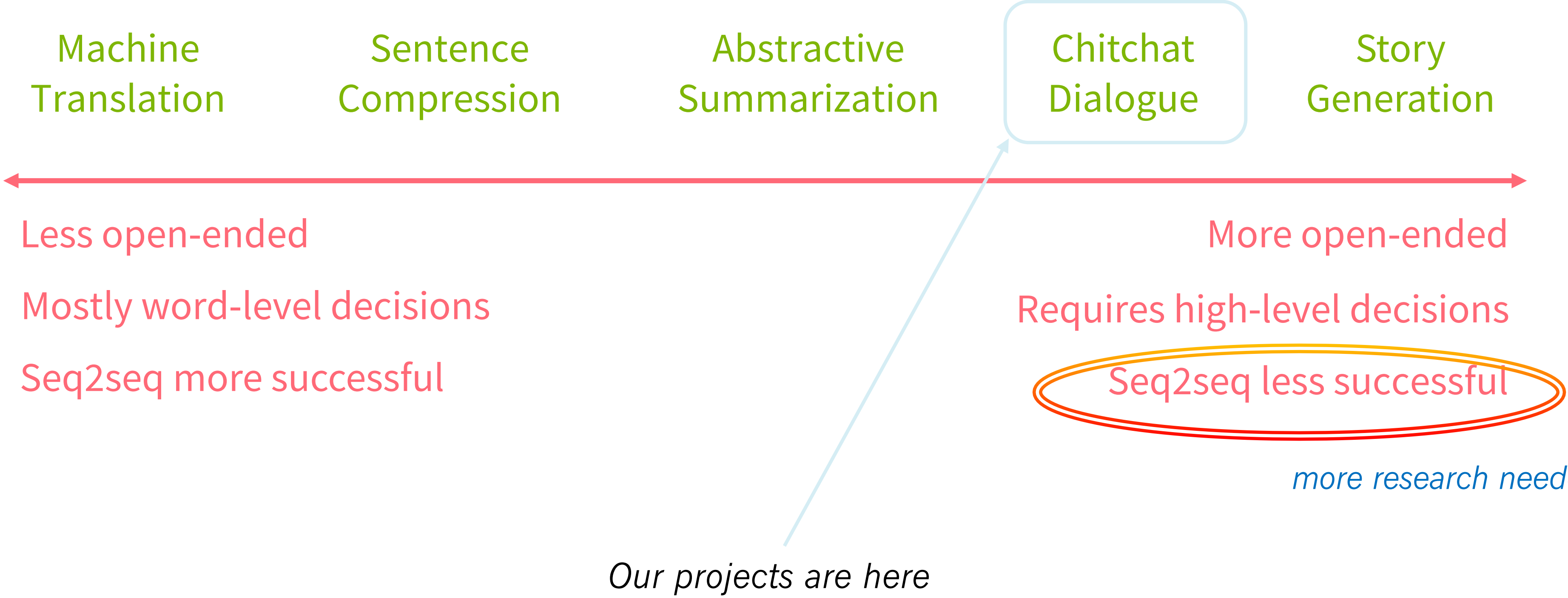
Topic: Shower (*unseen*)
Human: So I've been looking for a better showerhead with more water pressure for my shower.
Model: I love pressure-versed water at nearly constant temperature to a shower
Human: That sounds pretty great! Love those rainshower showers!
Model: the ancient greeks were the first people to have showers.
Human: Well I knew they were good for something.
Model: the original showers were neither indoor structures nor man made but were common.
Human: Ah that sounds intimidating!
Model: it is required manual transport of both fresh and waste water.

Error Analysis Summary

- **Retrieval models** are surprisingly good on seen topics *(not always of course)*
- On unseen, they can fail very badly

- **Generative models** aren't as good on seen, but *can* generalize to unseen.
- They still suffer from several problems (fixes: open research problems):
    - local & global repetition in dialogue, fixed to some extent by knowledge grounding
    - too much reliance on copying ("I don't know, but I do know that [Wikipedia excerpt] ")
    - clearly not as engaging as retrieval, but catches up on unseen
    - balance issues: too many statements, not enough questions

Natural Language Generation task spectrum



Machine Translation

Sentence Compression

Abstractive Summarization

Chitchat Dialogue

Story Generation

Less open-ended

Mostly word-level decisions

Seq2seq more successful

More open-ended

Requires high-level decisions

Seq2seq less successful

more research neede

Our projects are here

Slide from Abigail See

Conclusion

- Models:
  - Need to be engaging:
    - (i) fun to talk to,
    - (ii) expert/knowledgeable

Open problems:

- Improve generative models to level of retrieval models

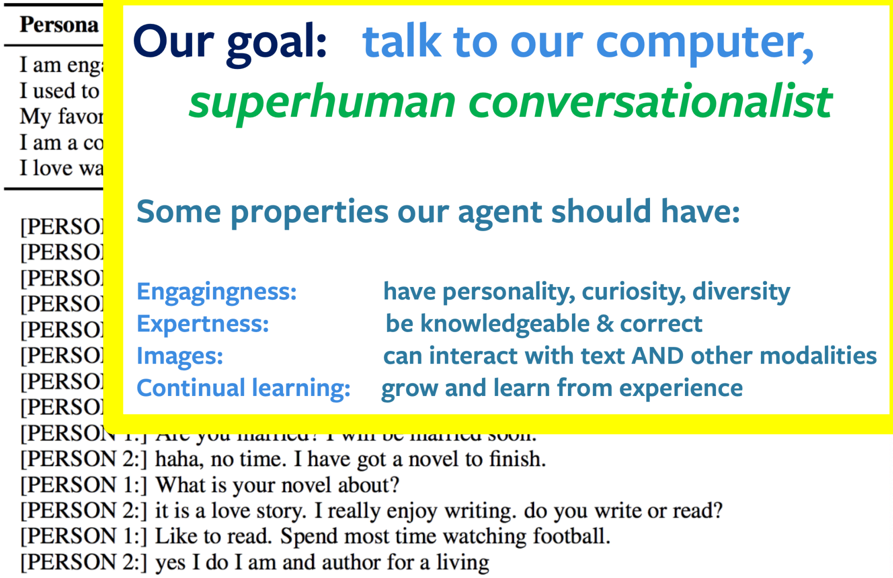
- Generalize well to unseen topics

- Natural to multitask with other tasks, e.g. get strong QA performance

Our goal: talk to our computer, *superhuman conversationalist*

Some properties our agent should have:

Engagingness:	have personality, curiosity, diversity
Expertness:	be knowledgeable & correct
Images:	can interact with text AND other modalities
Continual learning:	grow and learn from experience

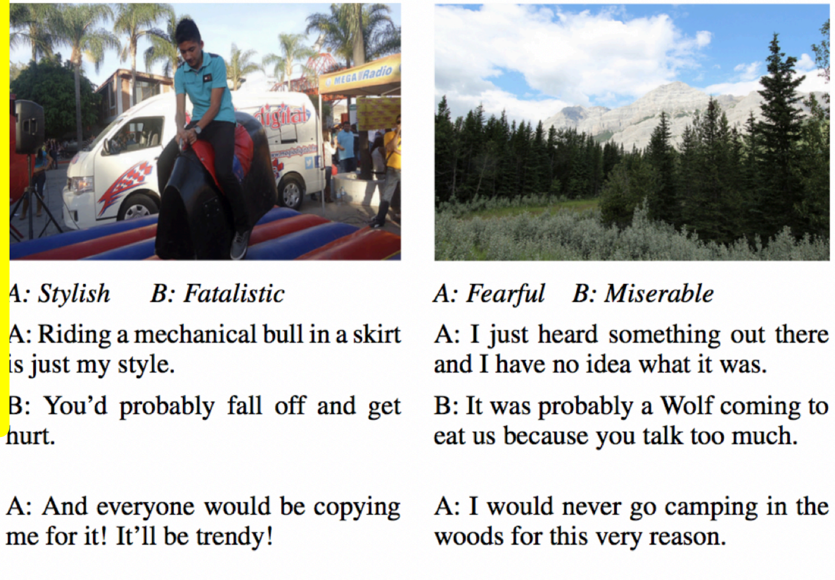


Persona

I am eng
I used to
My favor
I am a co
I love wa

[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

Persona-Chat (engaging personality)



A: Stylish *B: Fatalistic*

A: Riding a mechanical bull in a skirt is just my style.

B: You'd probably fall off and get hurt.

A: And everyone would be copying me for it! It'll be trendy!

A: Fearful *B: Miserable*

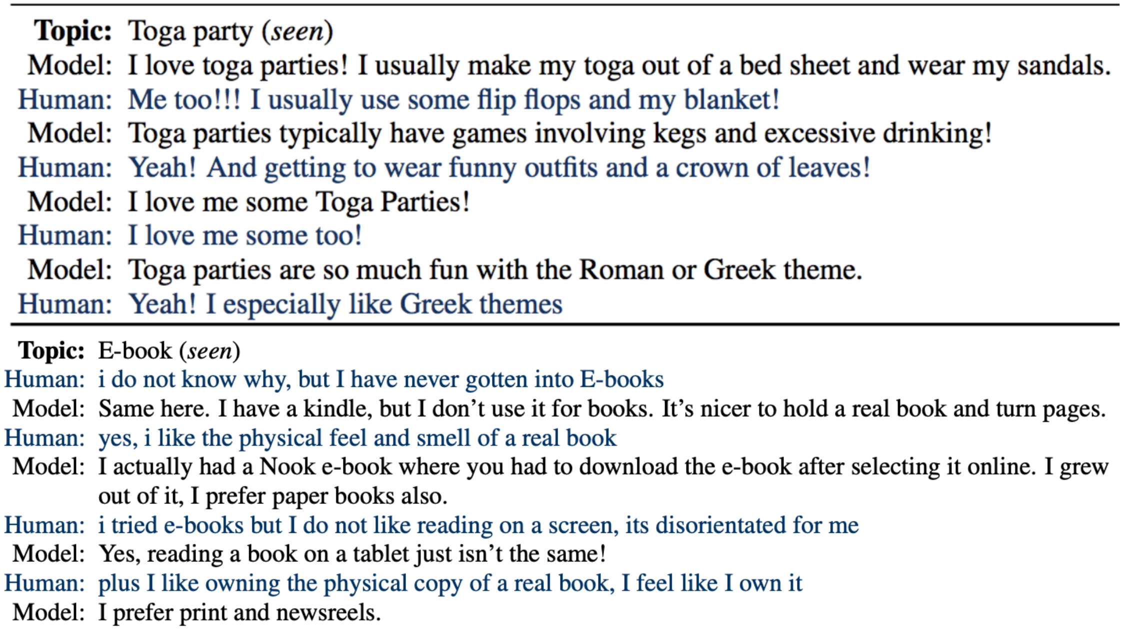
A: I just heard something out there and I have no idea what it was.

B: It was probably a Wolf coming to eat us because you talk too much.

A: I would never go camping in the woods for this very reason.

Image-Chat (images)

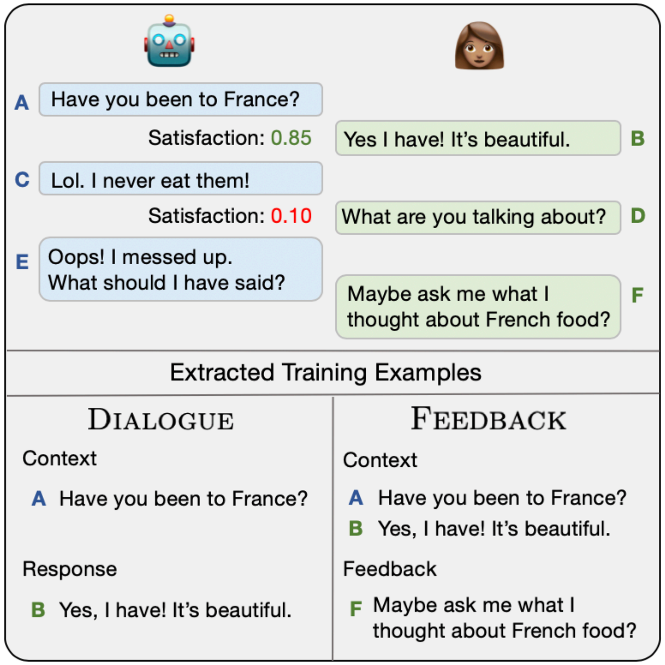
Wizard of Wikipedia (expert knowledge)



Topic: Toga party (*seen*)
Model: I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human: Me too!!! I usually use some flip flops and my blanket!
Model: Toga parties typically have games involving kegs and excessive drinking!
Human: Yeah! And getting to wear funny outfits and a crown of leaves!
Model: I love me some Toga Parties!
Human: I love me some too!
Model: Toga parties are so much fun with the Roman or Greek theme.
Human: Yeah! I especially like Greek themes

Topic: E-book (*seen*)
Human: i do not know why, but I have never gotten into E-books
Model: Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.
Human: yes, i like the physical feel and smell of a real book
Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human: i tried e-books but I do not like reading on a screen, its disorientated for me
Model: Yes, reading a book on a tablet just isn't the same!
Human: plus I like owning the physical copy of a real book, I feel like I own it
Model: I prefer print and newsreels.

Self-feeding (continual learning)



A Have you been to France?
Satisfaction: 0.85
B Yes I have! It's beautiful.
C Lol. I never eat them!
Satisfaction: 0.10
D What are you talking about?
E Oops! I messed up. What should I have said?
F Maybe ask me what I thought about French food?

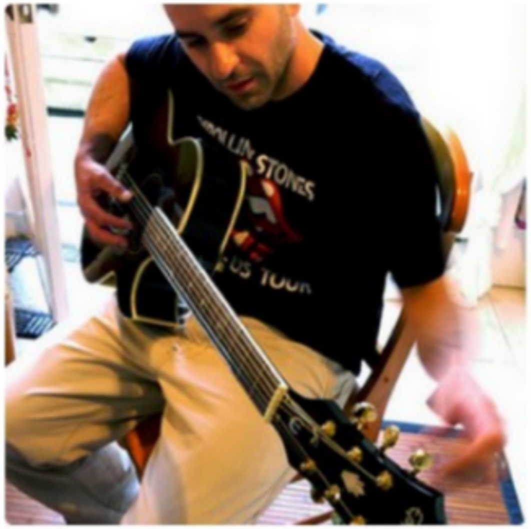
Extracted Training Examples

DIALOGUE	FEEDBACK
Context	Context
A Have you been to France?	A Have you been to France? B Yes, I have! It's beautiful.
Response	Feedback
B Yes, I have! It's beautiful.	F Maybe ask me what I thought about French food?

Engaging Image Captioning Via Personality

Kurt Shuster, Samuel Humeau, Hexiang Hu,
Antoine Bordes, Jason Weston
CVPR 2019

Standard (COCO) Image Captioning Models



Man in black shirt is playing guitar.

Standard (COCO) Image Captioning Models

Man in black shirt is playing guitar.

A plate with a sandwich and salad on it.

Good for: testing if model understands image content
Bad for: engaging human reader

Standard (COCO) Image Captioning Models

Man in black shirt is playing guitar.

A plate with a sandwich and salad on it.

Good for: testing if model understands image content
Bad for: engaging human reader

Want to be good at both of these!!!

What makes an utterance engaging? One answer: personality, emotion & style traits

(not always just neutral, factual tone)

Existing Work

Neutral, factual captions:

- COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014)
- Many models developed for them (discussed later).

Funny captions:

- wordplay (puns) (Chandrasekaran et al., 2017)
- or training on data from humour websites (Yoshida et al., 2018).

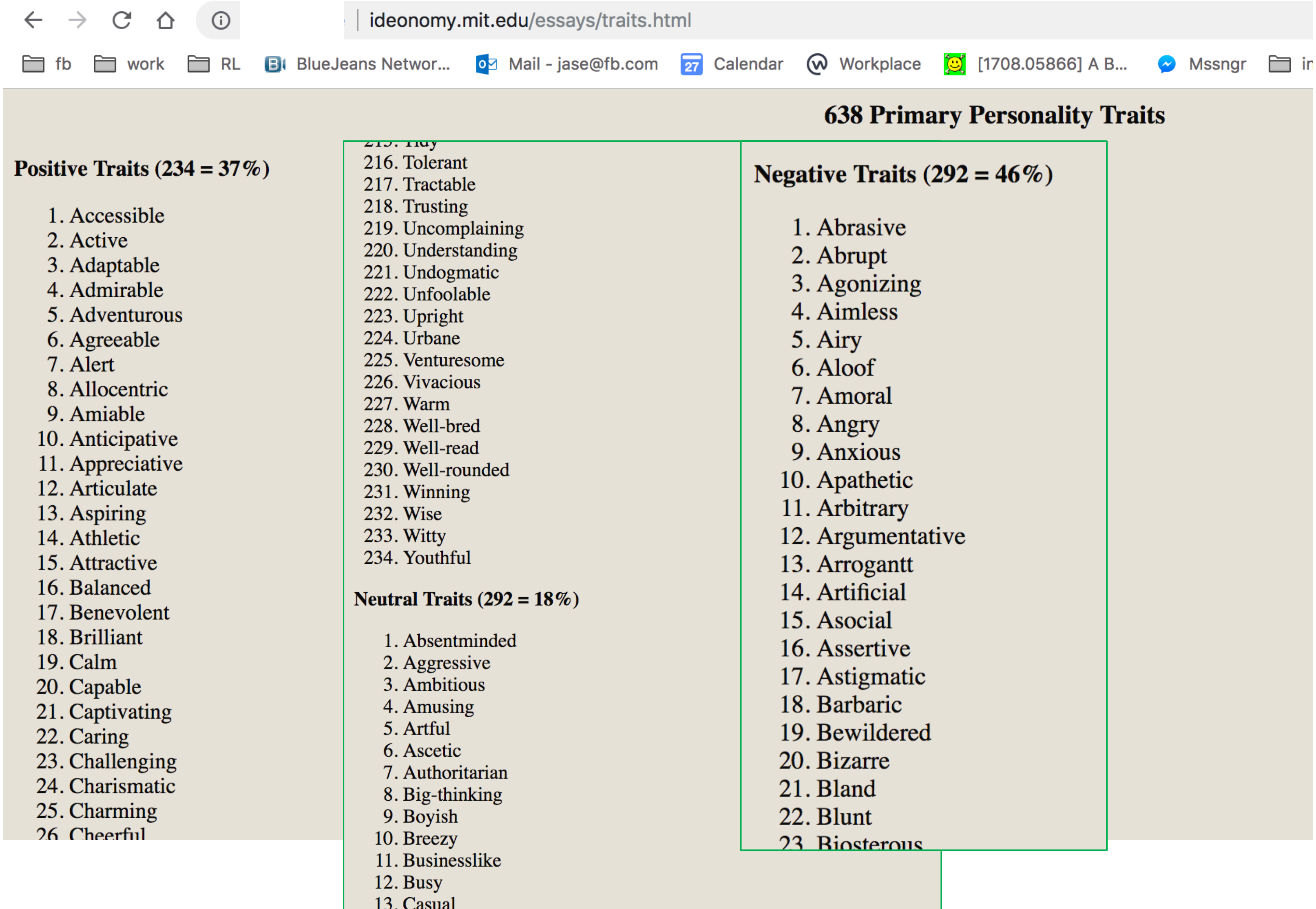
Using user features:

- location and age (Denton et al., 2015)
- or knowledge of the reader's active vocabulary (Park et al., 2017).

Style transfer:

- unsupervised (Mathews et al., 2018).
- Small datasets, e.g. Senticap (800 examples), (Mathews et al., 2016)
- romantic and humorous only - FlickrStyle10K , 10k examples - Gan et al. (2017)

MIT Personality List - 638 Traits



638 Primary Personality Traits

Positive Traits (234 = 37%)

1. Accessible
2. Active
3. Adaptable
4. Admirable
5. Adventurous
6. Agreeable
7. Alert
8. Allocentric
9. Amiable
10. Anticipative
11. Appreciative
12. Articulate
13. Aspiring
14. Athletic
15. Attractive
16. Balanced
17. Benevolent
18. Brilliant
19. Calm
20. Capable
21. Captivating
22. Caring
23. Challenging
24. Charismatic
25. Charming
26. Cheerful

216. Tolerant
217. Tractable
218. Trusting
219. Uncomplaining
220. Understanding
221. Undogmatic
222. Unfoolable
223. Upright
224. Urbane
225. Venturesome
226. Vivacious
227. Warm
228. Well-bred
229. Well-read
230. Well-rounded
231. Winning
232. Wise
233. Witty
234. Youthful

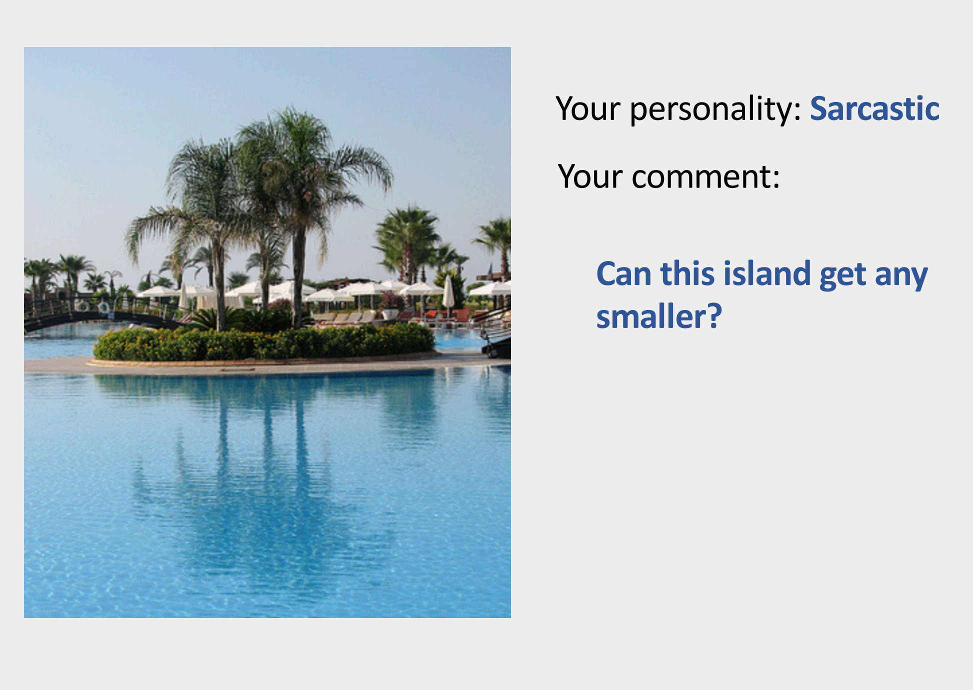
Neutral Traits (292 = 18%)

1. Absentminded
2. Aggressive
3. Ambitious
4. Amusing
5. Artful
6. Ascetic
7. Authoritarian
8. Big-thinking
9. Boyish
10. Breezy
11. Businesslike
12. Busy
13. Casual

Negative Traits (292 = 46%)

1. Abrasive
2. Abrupt
3. Agonizing
4. Aimless
5. Airy
6. Aloof
7. Amoral
8. Angry
9. Anxious
10. Apathetic
11. Arbitrary
12. Argumentative
13. Arrogantt
14. Artificial
15. Asocial
16. Assertive
17. Astigmatic
18. Barbaric
19. Bewildered
20. Bizarre
21. Bland
22. Blunt
23. Biosterous

Step 1: build a dataset

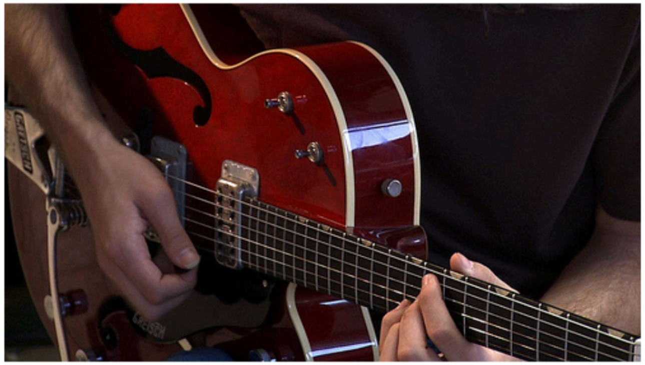


Your personality: **Sarcastic**

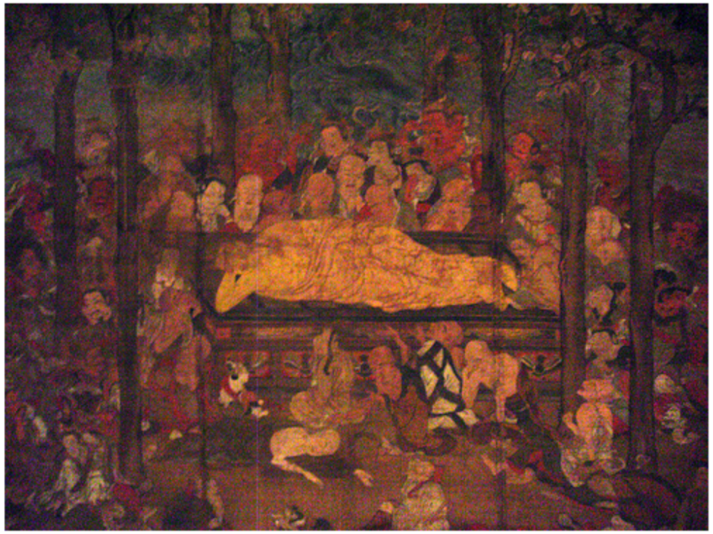
Your comment:

Can this island get any smaller?

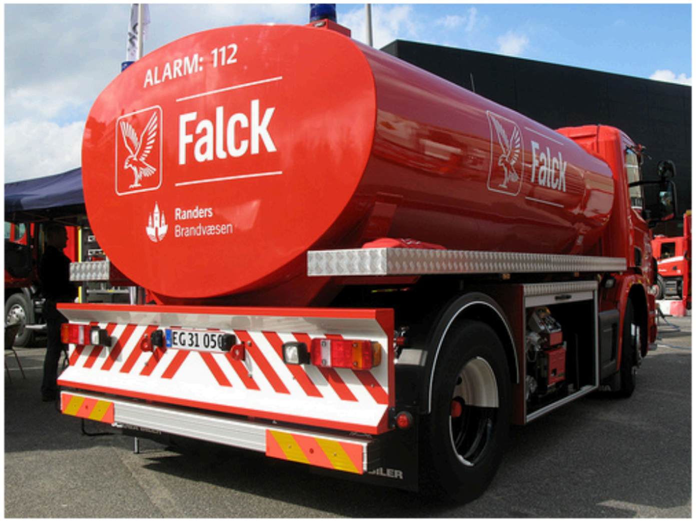
- Selected 215 personality traits
- Images from YFFC100M
- Collect captions via annotators



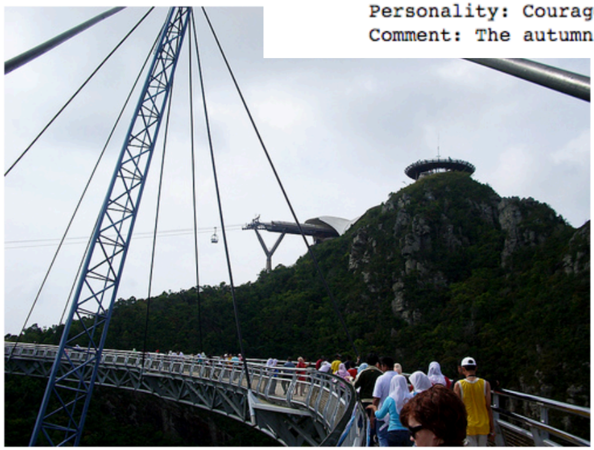
Personality: Energetic
Comment: About to play the best tune you've ever heard in your life, get ready!



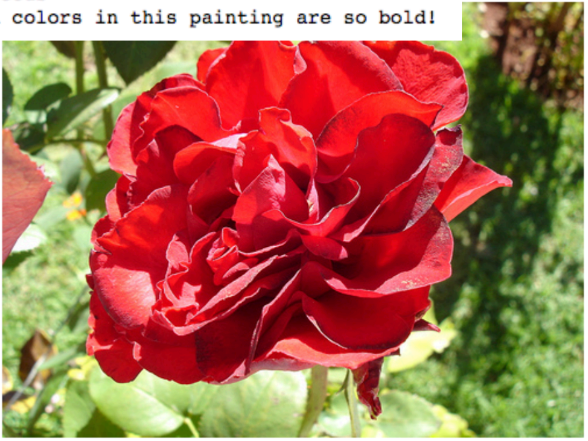
Personality: Courageous
Comment: The autumn colors in this painting are so bold!



Personality: Creative
Comment: Falck alarm everyone. Just a Falck alarm



Personality: Vague
Comment: We were up high, in a large group.



Personality: Fanatical
Comment: I LOVE RED FLOWERS!

Examples from the dataset

Sarcastic
Yes please sit by me

Mellow
Look at that smooth easy catch of the ball. like ballet.

Zany
I wish I could just run down this shore!

Contradictory
Love what you did with the place!

Contemptible
I can't believe no one has been taking care of this plant. Terrible

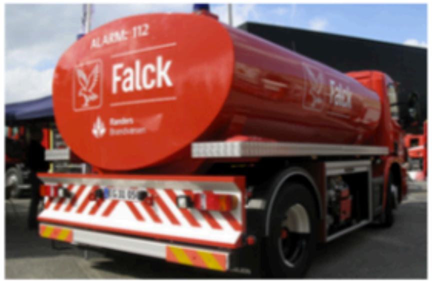
Energetic
About to play the best tune you've ever heard in your life. Get ready!

Examples from the dataset

Get ready!

Kind
they left me a parking spot

Spirited
That is one motor cycle enthusiast!!!



Creative
Falck alarm, everyone. Just a Falck alarm.

Crazy
I drove down this road backwards at 90 miles per hour three times

Morbid
I hope this car doesn't get into a wreck.

Questioning
Why do people think its cool to smoke cigarettes?

Step 1: Collect a large supervised dataset

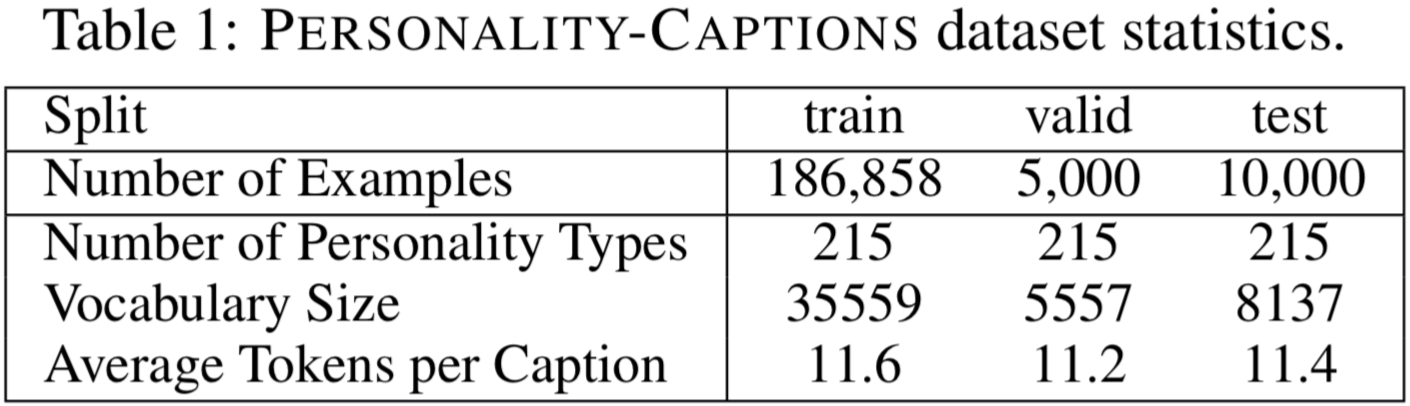


Table 1: PERSONALITY-CAPTIONS dataset statistics.

Split	train	valid	test
Number of Examples	186,858	5,000	10,000
Number of Personality Types	215	215	215
Vocabulary Size	35559	5557	8137
Average Tokens per Caption	11.6	11.2	11.4

Type	Datasets With Personality				Datasets Without Personality			
Dataset	Personality-Captions			FlickrStyle10K	COCO		Flickr30k	
Split	train	valid	test	train	valid	train	train	valid
Number of Images	186,858	5,000	10,000	7000	82783	40504	29000	1014
Number of Captions	186,858	5,000	50,000	14000	414113	202654	145000	5070
Number of Personality Types	215	215	215	2	None	None	None	None
Vocabulary Size	33641	5460	16655	8889	23776	17724	17920	4283
Average Tokens per Caption	11.2	10.9	11.1	14.51	11.3	11.3	13.53	13.74

Table 1: Personality-Captions dataset statistics compared to other captioning datasets.

Step 2: Build strong models

We make use of state-of-the-art in vision and language domains to build our models:

Image Encoder:

- ResNeXt (Xie et al., 2016) trained on 3.5 billion Instagram pictures following Mahajan et al. (2018), which we call *ResNeXt-IG-3.5B*.
- *Shown to work very well on ImageNet classification (but not captioning).*

Text Encoder:

- Transformer (Vaswani et al., 2017) trained on 1.7 billion Reddit dialogue examples, following (Mazaré et al., 2018).
- *Shown to work very well for PersonaChat dialogue (but not captioning).*

Models: we consider both generative and retrieval models.

- Generative: *consider three widely used architectures:*
  - ShowTell (Vinyals et al., 2015)
  - ShowAttTell (Xu et al., 2015)
  - UpDown (Anderson et al., 2018)

Use ResNeXt-IG-3.5B and add learnt personality features to each decoder step

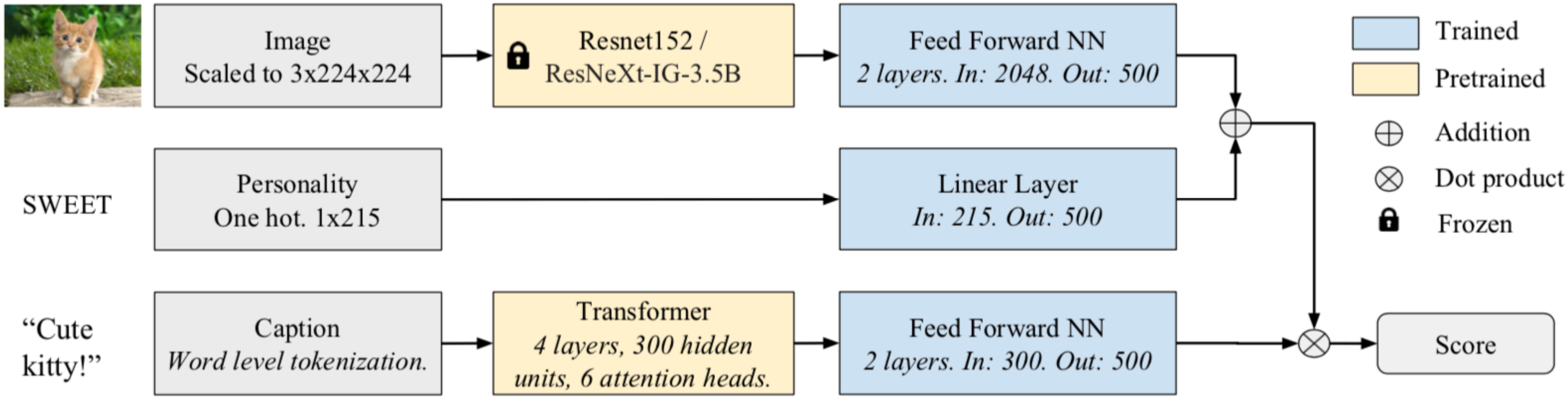
Models: we consider both generative and retrieval models.

- Generative: *consider three recent best architectures:*
  - ShowTell (Vinyals et al., 2015)
  - ShowAttTell (Xu et al., 2015)
  - UpDown (Anderson et al., 2018)

Use ResNeXt-IG-3.5B and add learnt personality features to each decoder step

- Retrieval:

TransResNet



Our generative models are good at **understanding image content**.

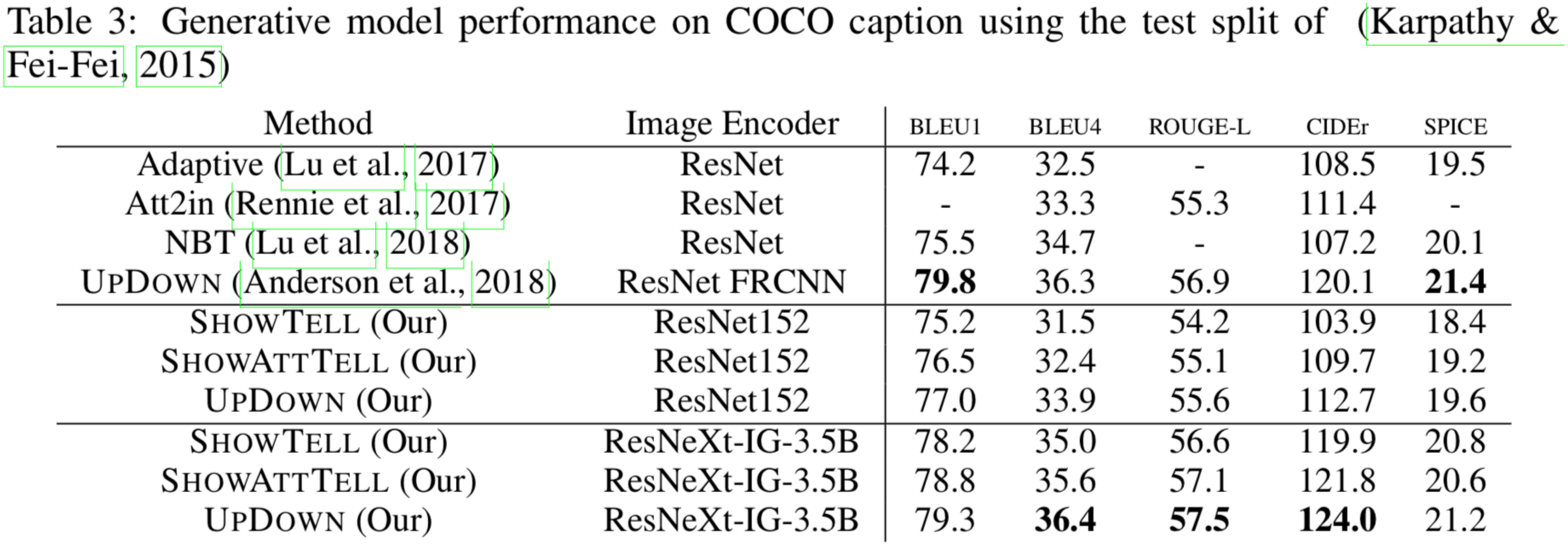


Table 3: Generative model performance on COCO caption using the test split of (Karpathy & Fei-Fei, 2015)

Method	Image Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
Adaptive (Lu et al., 2017)	ResNet	74.2	32.5	-	108.5	19.5
Att2in (Rennie et al., 2017)	ResNet	-	33.3	55.3	111.4	-
NBT (Lu et al., 2018)	ResNet	75.5	34.7	-	107.2	20.1
UpDown (Anderson et al., 2018)	ResNet FRCNN	**79.8**	36.3	56.9	120.1	**21.4**
ShowTell (Our)	ResNet152	75.2	31.5	54.2	103.9	18.4
ShowAttTell (Our)	ResNet152	76.5	32.4	55.1	109.7	19.2
UpDown (Our)	ResNet152	77.0	33.9	55.6	112.7	19.6
ShowTell (Our)	ResNeXt-IG-3.5B	78.2	35.0	56.6	119.9	20.8
ShowAttTell (Our)	ResNeXt-IG-3.5B	78.8	35.6	57.1	121.8	20.6
UpDown (Our)	ResNeXt-IG-3.5B	79.3	**36.4**	**57.5**	**124.0**	21.2

Our retrieval models are good at **understanding image content**.

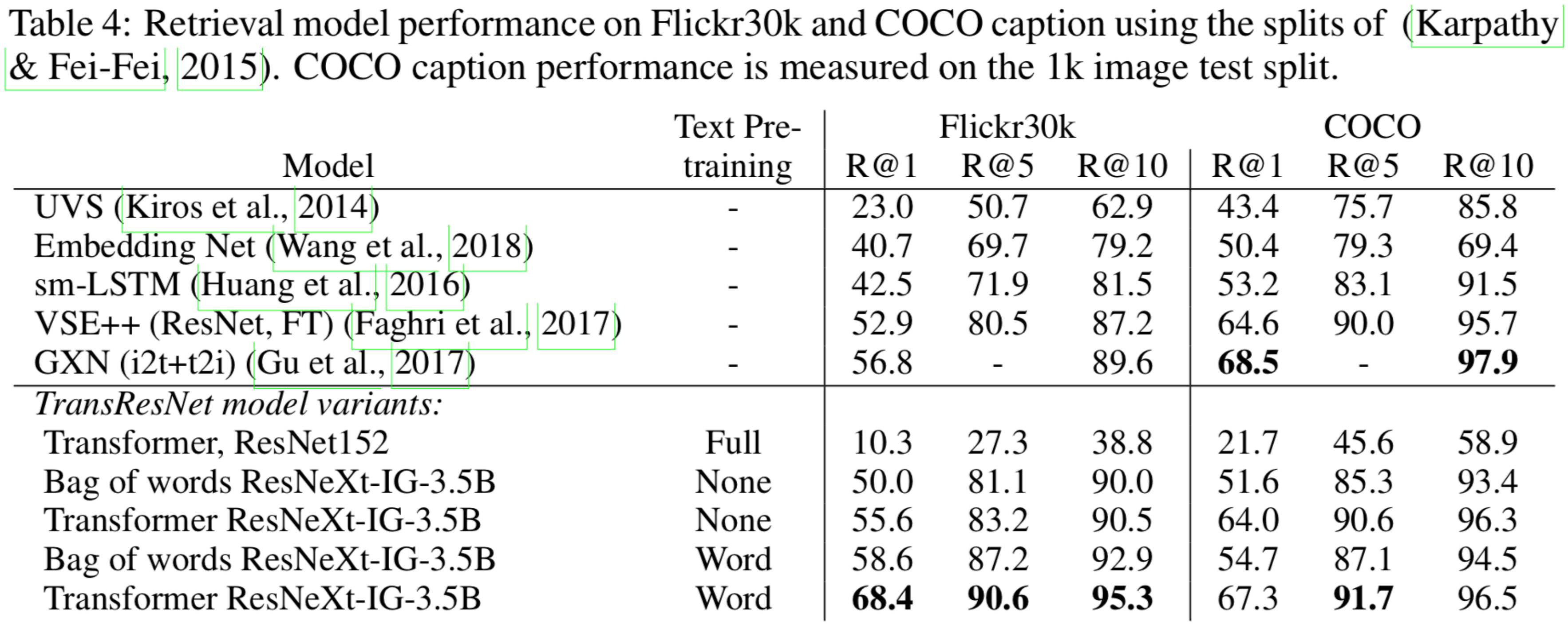


Table 4: Retrieval model performance on Flickr30k and COCO caption using the splits of (Karpathy & Fei-Fei, 2015). COCO caption performance is measured on the 1k image test split.

Model	Text Pre-training	Flickr30k R@1	Flickr30k R@5	Flickr30k R@10	COCO R@1	COCO R@5	COCO R@10
UVS (Kiros et al., 2014)	-	23.0	50.7	62.9	43.4	75.7	85.8
Embedding Net (Wang et al., 2018)	-	40.7	69.7	79.2	50.4	79.3	69.4
sm-LSTM (Huang et al., 2016)	-	42.5	71.9	81.5	53.2	83.1	91.5
VSE++ (ResNet, FT) (Faghri et al., 2017)	-	52.9	80.5	87.2	64.6	90.0	95.7
GXN (i2t+t2i) (Gu et al., 2017)	-	56.8	-	89.6	**68.5**	-	**97.9**
TransResNet model variants:							
Transformer, ResNet152	Full	10.3	27.3	38.8	21.7	45.6	58.9
Bag of words ResNeXt-IG-3.5B	None	50.0	81.1	90.0	51.6	85.3	93.4
Transformer ResNeXt-IG-3.5B	None	55.6	83.2	90.5	64.0	90.6	96.3
Bag of words ResNeXt-IG-3.5B	Word	58.6	87.2	92.9	54.7	87.1	94.5
Transformer ResNeXt-IG-3.5B	Word	**68.4**	**90.6**	**95.3**	67.3	**91.7**	96.5

Our generative models are good at using **personality**

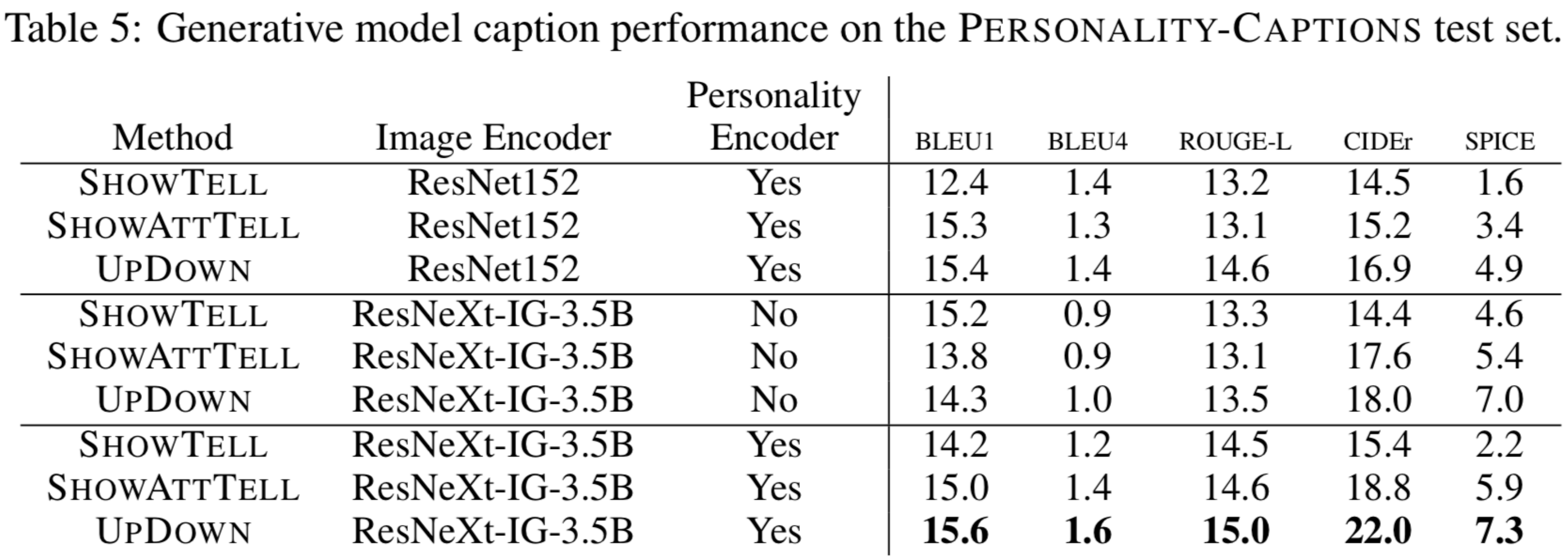


Table 5: Generative model caption performance on the Personality-Captions test set.

Method	Image Encoder	Personality Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
ShowTell	ResNet152	Yes	12.4	1.4	13.2	14.5	1.6
ShowAttTell	ResNet152	Yes	15.3	1.3	13.1	15.2	3.4
UpDown	ResNet152	Yes	15.4	1.4	14.6	16.9	4.9
ShowTell	ResNeXt-IG-3.5B	No	15.2	0.9	13.3	14.4	4.6
ShowAttTell	ResNeXt-IG-3.5B	No	13.8	0.9	13.1	17.6	5.4
UpDown	ResNeXt-IG-3.5B	No	14.3	1.0	13.5	18.0	7.0
ShowTell	ResNeXt-IG-3.5B	Yes	14.2	1.2	14.5	15.4	2.2
ShowAttTell	ResNeXt-IG-3.5B	Yes	15.0	1.4	14.6	18.8	5.9
UpDown	ResNeXt-IG-3.5B	Yes	**15.6**	**1.6**	**15.0**	**22.0**	**7.3**

Our retrieval models are good at using **personality**

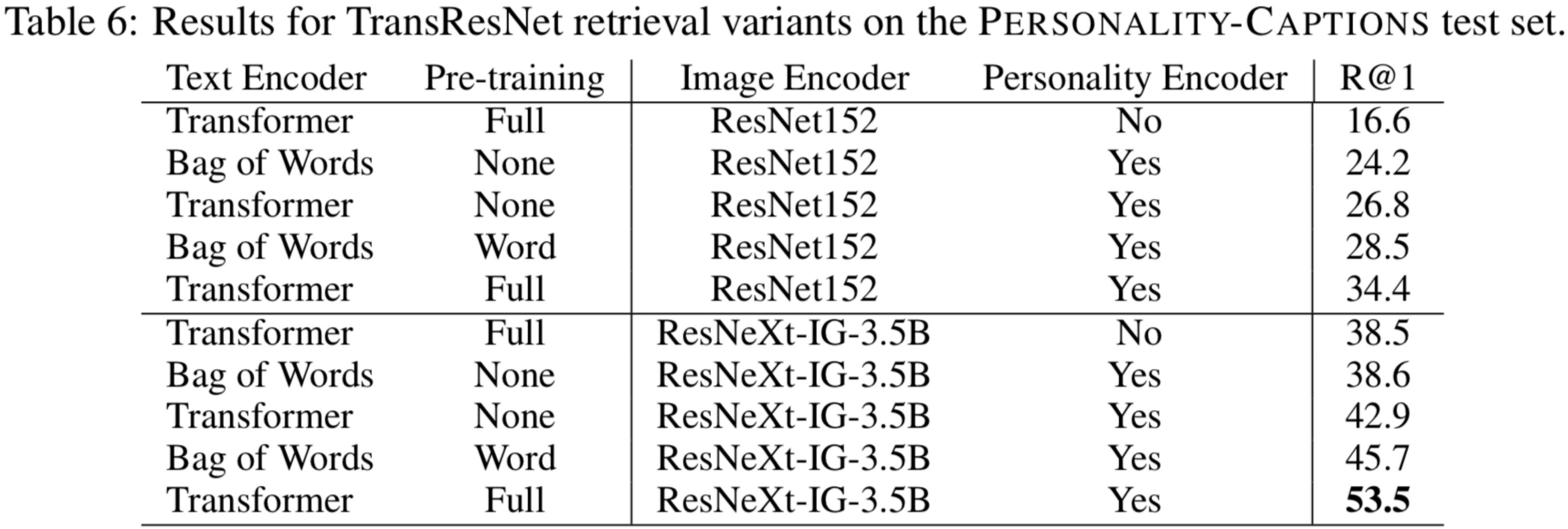
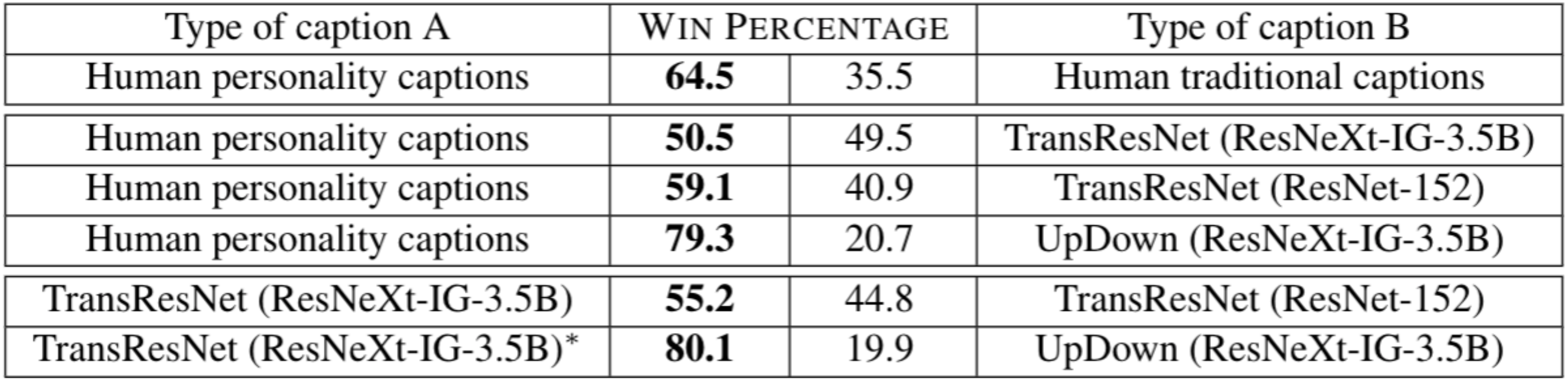


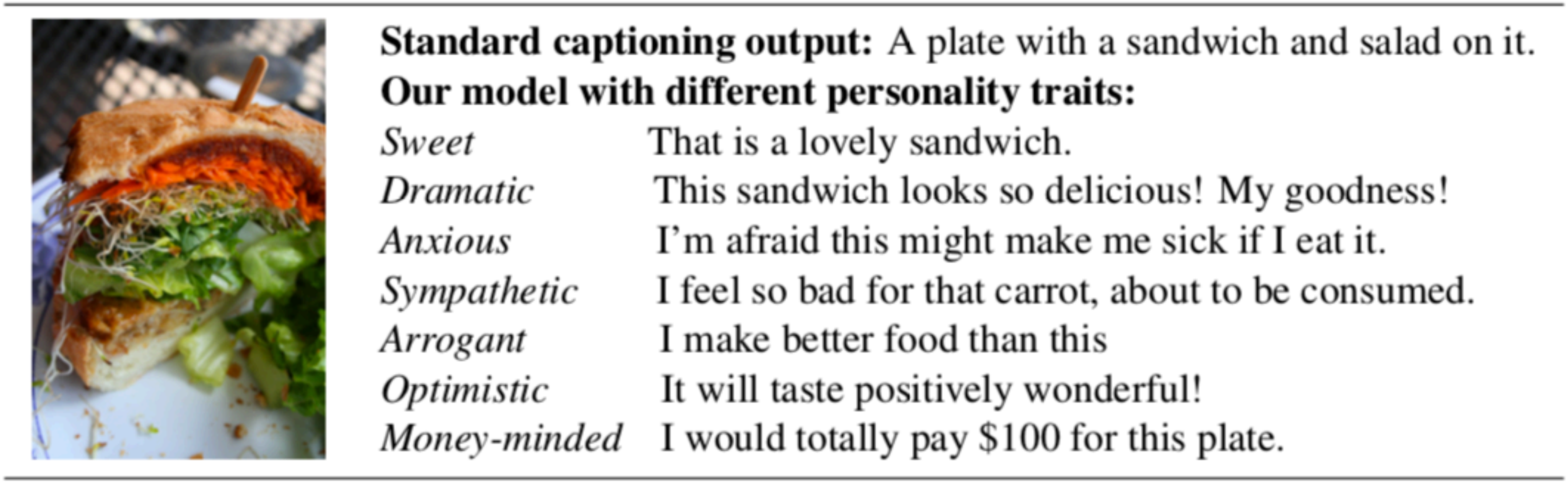
Table 6: Results for TransResNet retrieval variants on the PERSONALITY-CAPTIONS test set.

Text Encoder	Pre-training	Image Encoder	Personality Encoder	R@1
Transformer	Full	ResNet152	No	16.6
Bag of Words	None	ResNet152	Yes	24.2
Transformer	None	ResNet152	Yes	26.8
Bag of Words	Word	ResNet152	Yes	28.5
Transformer	Full	ResNet152	Yes	34.4
Transformer	Full	ResNeXt-IG-3.5B	No	38.5
Bag of Words	None	ResNeXt-IG-3.5B	Yes	38.6
Transformer	None	ResNeXt-IG-3.5B	Yes	42.9
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	45.7
Transformer	Full	ResNeXt-IG-3.5B	Yes	**53.5**

Human evaluation studies: *our best retrieval model is close to matching human performance (using metric: humans measuring engagement)*



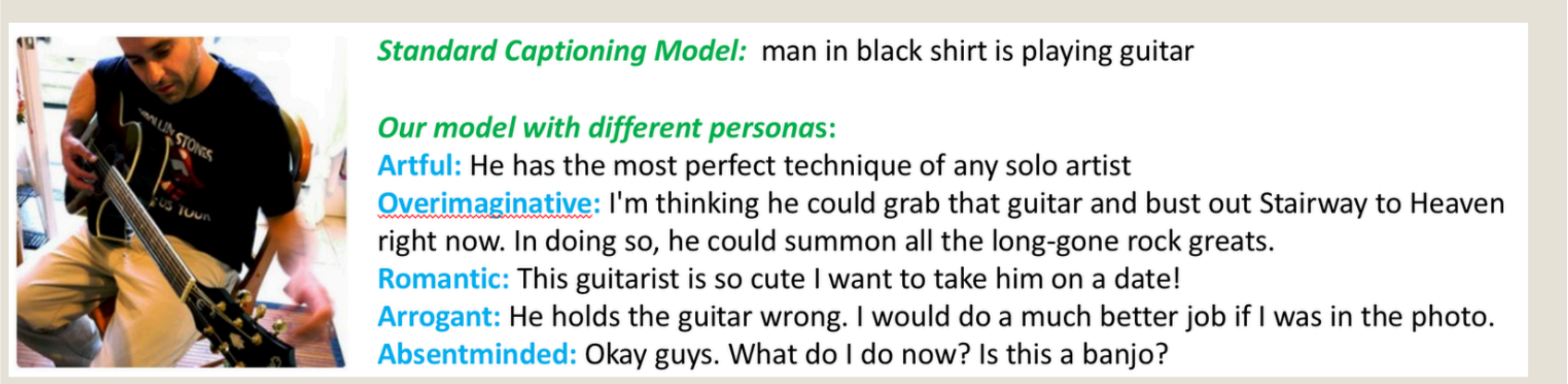
Type of caption A	WIN PERCENTAGE		Type of caption B
Human personality captions	**64.5**	35.5	Human traditional captions
Human personality captions	**50.5**	49.5	TransResNet (ResNeXt-IG-3.5B)
Human personality captions	**59.1**	40.9	TransResNet (ResNet-152)
Human personality captions	**79.3**	20.7	UpDown (ResNeXt-IG-3.5B)
TransResNet (ResNeXt-IG-3.5B)	**55.2**	44.8	TransResNet (ResNet-152)
TransResNet (ResNeXt-IG-3.5B)*	**80.1**	19.9	UpDown (ResNeXt-IG-3.5B)



Standard captioning output: A plate with a sandwich and salad on it.

Our model with different personality traits:

Sweet	That is a lovely sandwich.
Dramatic	This sandwich looks so delicious! My goodness!
Anxious	I'm afraid this might make me sick if I eat it.
Sympathetic	I feel so bad for that carrot, about to be consumed.
Arrogant	I make better food than this
Optimistic	It will taste positively wonderful!
Money-minded	I would totally pay $100 for this plate.



Standard Captioning Model: man in black shirt is playing guitar

***Our model with different personas*:**

Artful: He has the most perfect technique of any solo artist

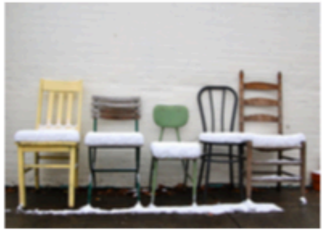
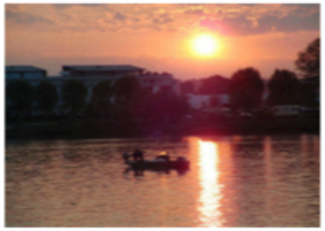
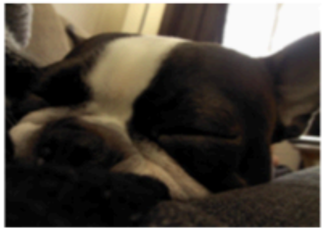
Overimaginative: I'm thinking he could grab that guitar and bust out Stairway to Heaven right now. In doing so, he could summon all the long-gone rock greats.

Romantic: This guitarist is so cute I want to take him on a date!

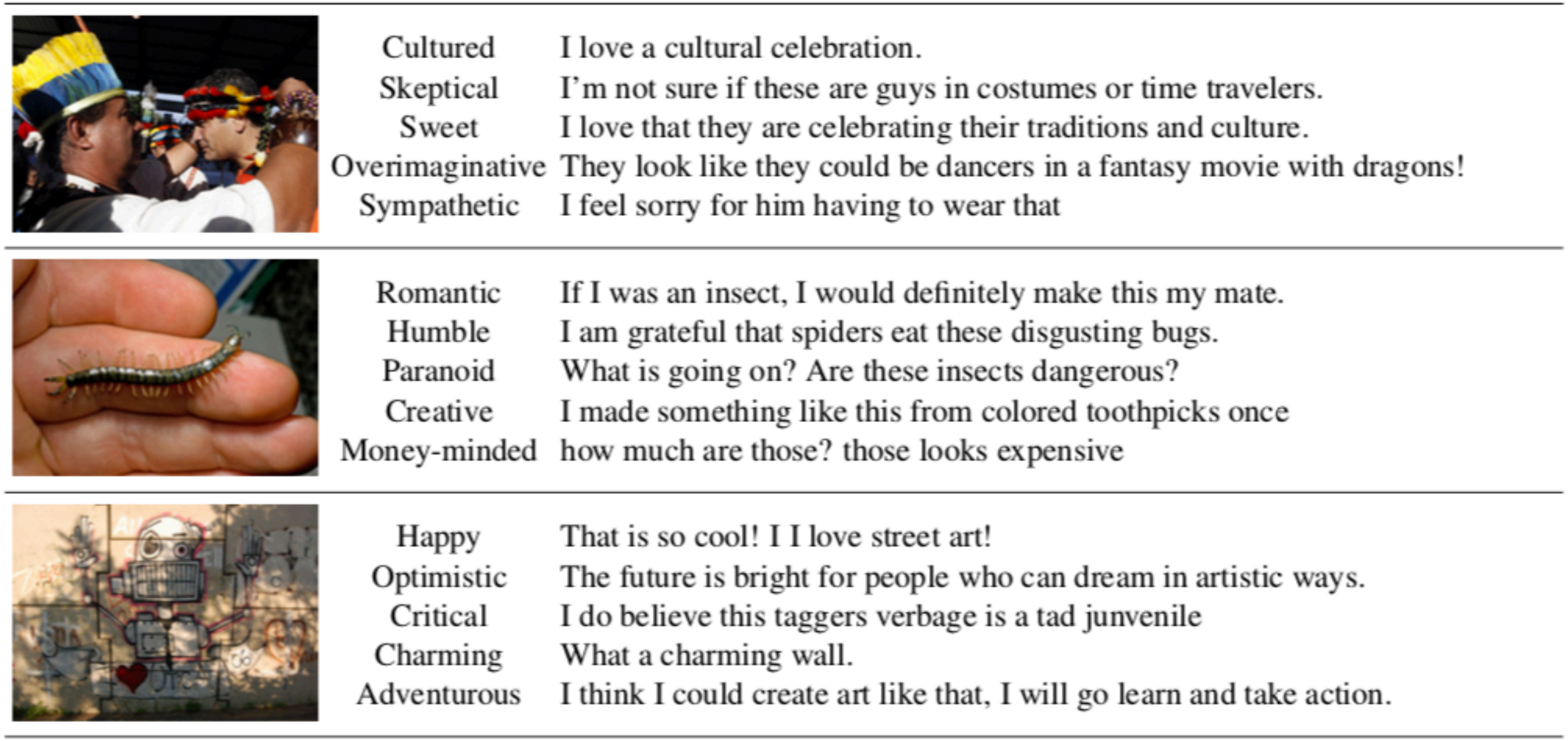
Arrogant: He holds the guitar wrong. I would do a much better job if I was in the photo.

Absentminded: Okay guys. What do I do now? Is this a banjo?

More examples of our best model

Image	Personality	Generated comment
	Sweet	I love, love, love these chairs! I want the big one in my house!
	Vague	This chair is either covered in snow or the snow is covered in the chair.
	Cultured	These chairs remind me of the Swedish interior design revolution of the 70's.
	Paranoid	What if someone fell off those chairs.
	Overimaginative	Those chairs look like they could be in a doll house.
	Arrogant	I've seen better sunsets elsewhere.
	Overimaginative	that sunset is so orange it could be a fruit
	Vague	It's the sunset.
	Optimistic	The sunset makes look forward to a happy tomorrow.
	Charming	The way the sun is hitting the water makes for a romantic evening.
	Sweet	What a cute puppy, reminds me of my friends.
	Skeptical	I don't think this dog will bite me.
	Sympathetic	poor dog! It looks so hungry :c
	Vague	it's a dog
	Wishful	I wish that I had a dog as cute as him.

More examples of our best model



Cultured	I love a cultural celebration.
Skeptical	I'm not sure if these are guys in costumes or time travelers.
Sweet	I love that they are celebrating their traditions and culture.
Overimaginative	They look like they could be dancers in a fantasy movie with dragons!
Sympathetic	I feel sorry for him having to wear that
Romantic	If I was an insect, I would definitely make this my mate.
Humble	I am grateful that spiders eat these disgusting bugs.
Paranoid	What is going on? Are these insects dangerous?
Creative	I made something like this from colored toothpicks once
Money-minded	how much are those? those looks expensive
Happy	That is so cool! I I love street art!
Optimistic	The future is bright for people who can dream in artistic ways.
Critical	I do believe this taggers verbage is a tad junvenile
Charming	What a charming wall.
Adventurous	I think I could create art like that, I will go learn and take action.

More examples of our best model

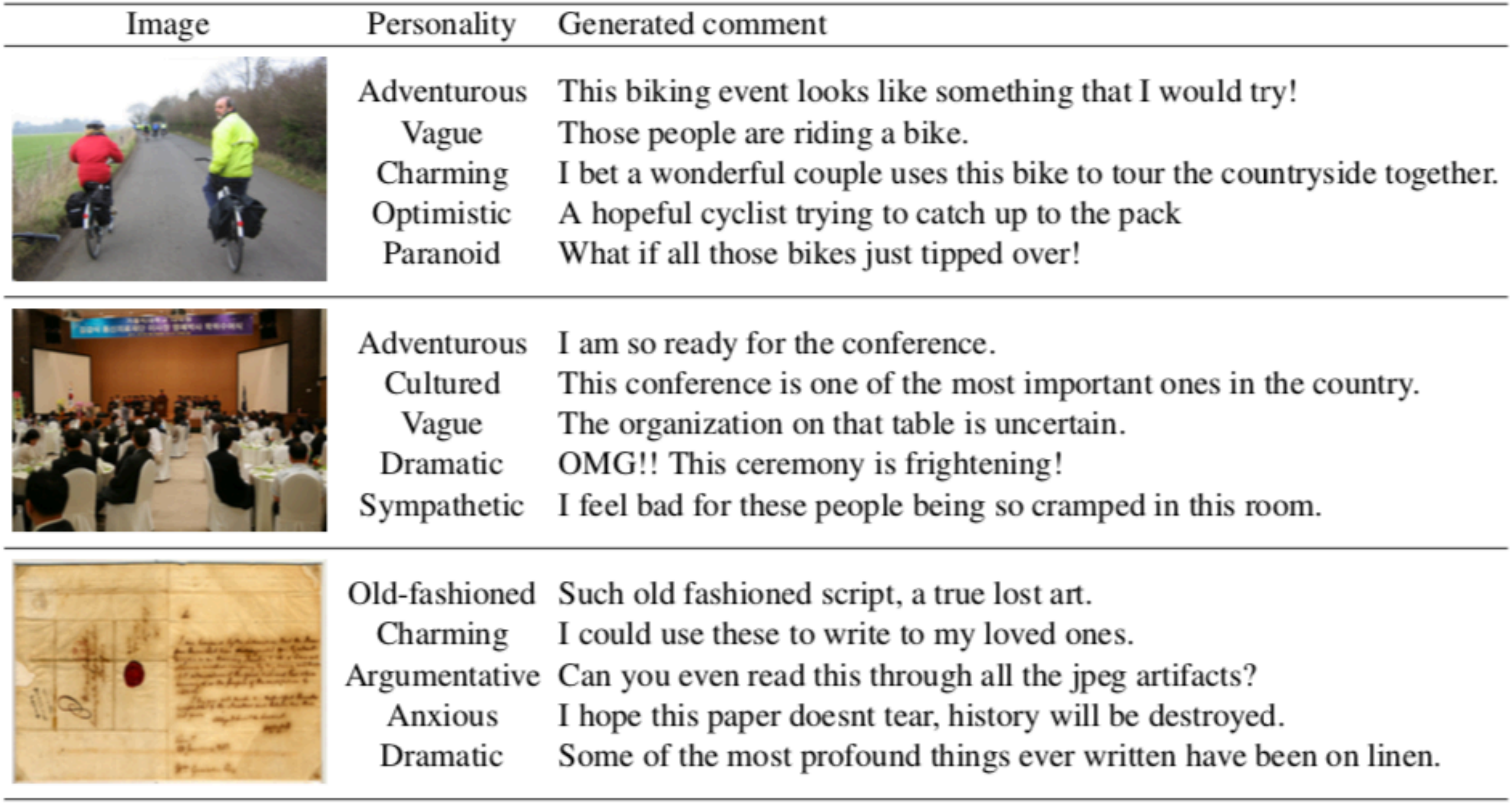
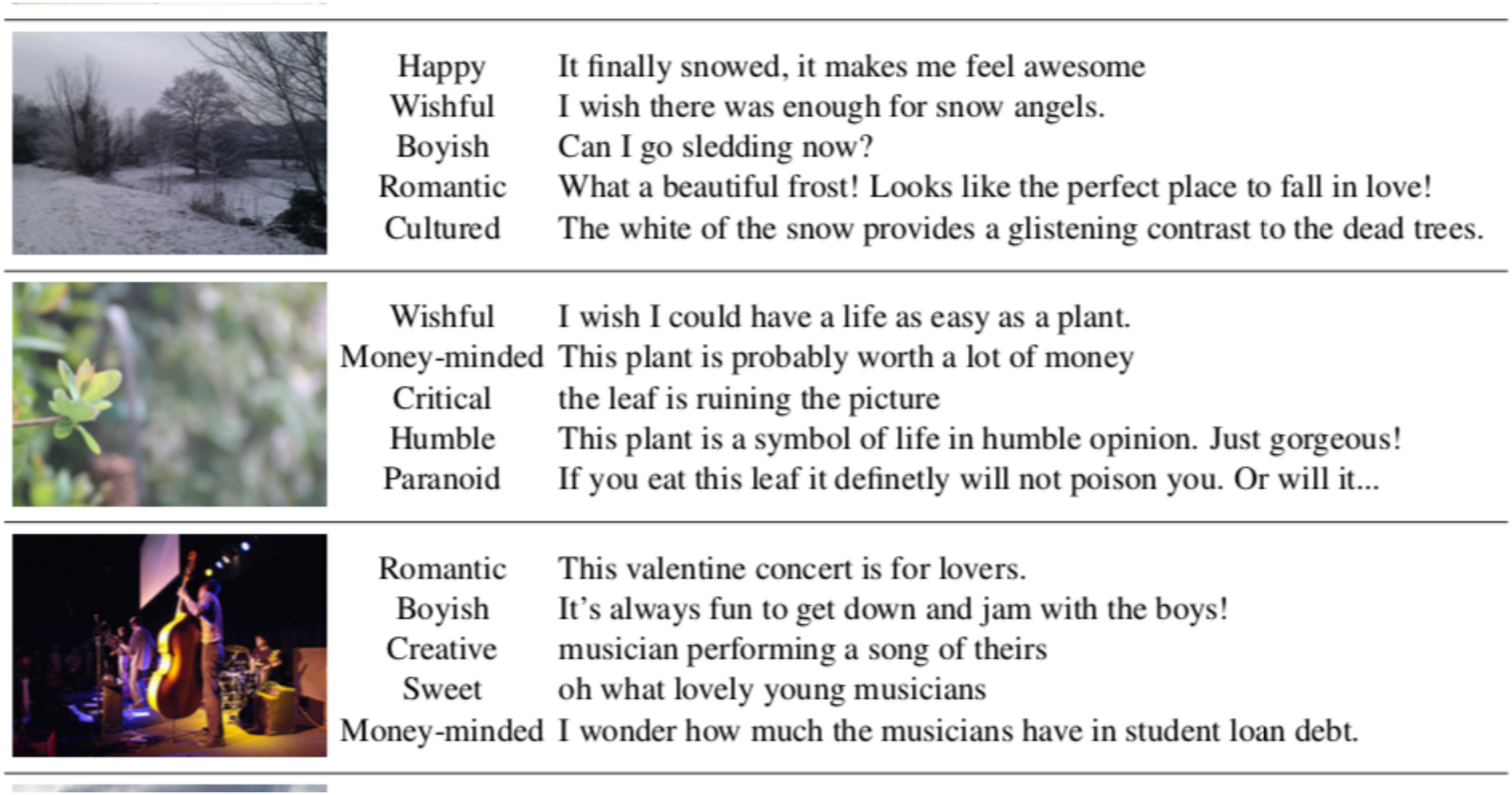


Image	Personality	Generated comment
	Adventurous	This biking event looks like something that I would try!
	Vague	Those people are riding a bike.
	Charming	I bet a wonderful couple uses this bike to tour the countryside together.
	Optimistic	A hopeful cyclist trying to catch up to the pack
	Paranoid	What if all those bikes just tipped over!
	Adventurous	I am so ready for the conference.
	Cultured	This conference is one of the most important ones in the country.
	Vague	The organization on that table is uncertain.
	Dramatic	OMG!! This ceremony is frightening!
	Sympathetic	I feel bad for these people being so cramped in this room.
	Old-fashioned	Such old fashioned script, a true lost art.
	Charming	I could use these to write to my loved ones.
	Argumentative	Can you even read this through all the jpeg artifacts?
	Anxious	I hope this paper doesnt tear, history will be destroyed.
	Dramatic	Some of the most profound things ever written have been on linen.

More examples of our best model



Happy	It finally snowed, it makes me feel awesome
Wishful	I wish there was enough for snow angels.
Boyish	Can I go sledding now?
Romantic	What a beautiful frost! Looks like the perfect place to fall in love!
Cultured	The white of the snow provides a glistening contrast to the dead trees.
Wishful	I wish I could have a life as easy as a plant.
Money-minded	This plant is probably worth a lot of money
Critical	the leaf is ruining the picture
Humble	This plant is a symbol of life in humble opinion. Just gorgeous!
Paranoid	If you eat this leaf it definetly will not poison you. Or will it...
Romantic	This valentine concert is for lovers.
Boyish	It’s always fun to get down and jam with the boys!
Creative	musician performing a song of theirs
Sweet	oh what lovely young musicians
Money-minded	I wonder how much the musicians have in student loan debt.

Engaging Image Chat:
Modeling Personality in Grounded Dialogue

Kurt Shuster, Samuel Humeau,
Antoine Bordes, Jason Weston

Next Step: **Dialogue!**

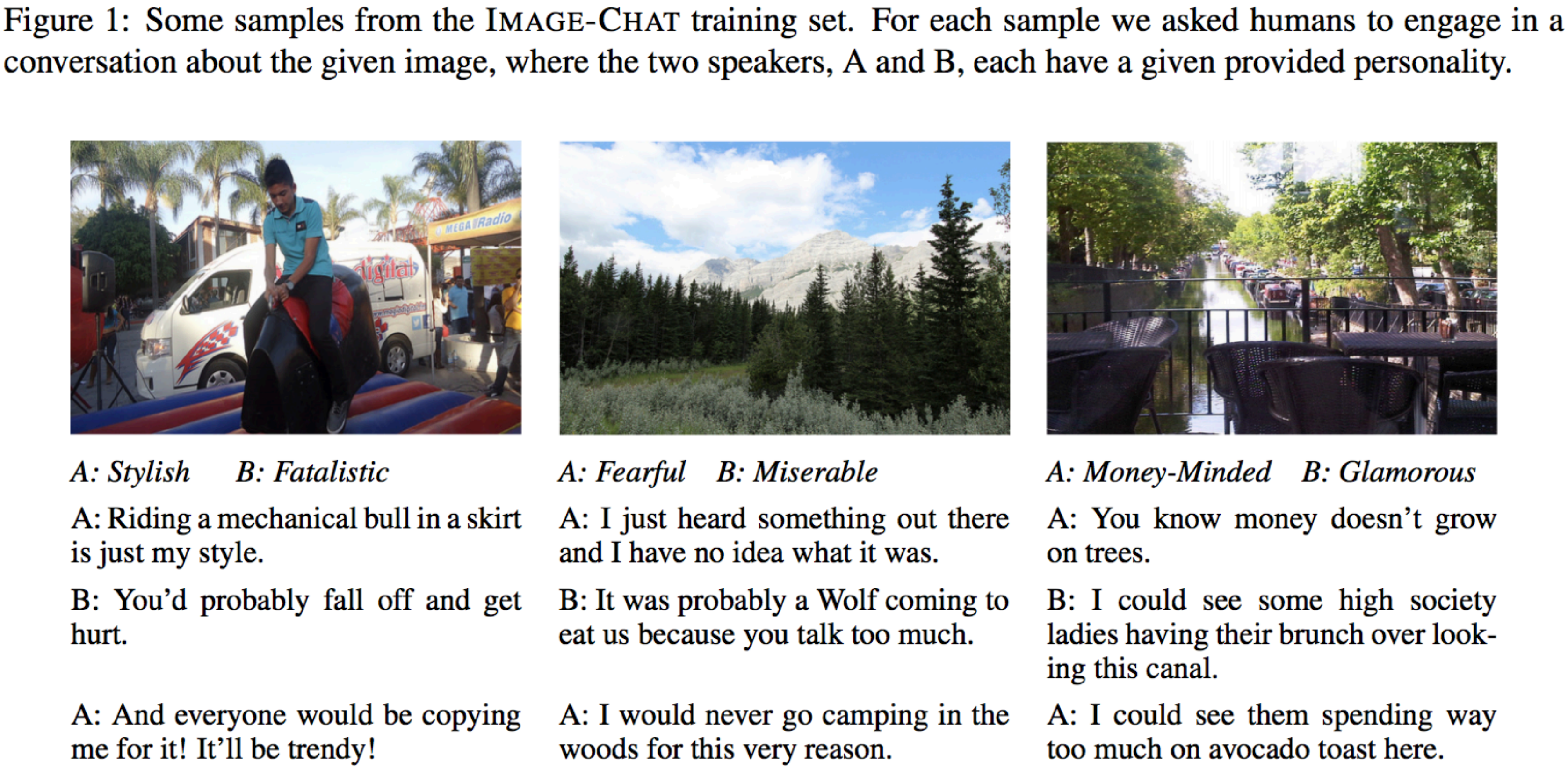


Figure 1: Some samples from the IMAGE-CHAT training set. For each sample we asked humans to engage in a conversation about the given image, where the two speakers, A and B, each have a given provided personality.

A: Stylish *B: Fatalistic*

A: Riding a mechanical bull in a skirt is just my style.

B: You'd probably fall off and get hurt.

A: And everyone would be copying me for it! It'll be trendy!

A: Fearful *B: Miserable*

A: I just heard something out there and I have no idea what it was.

B: It was probably a Wolf coming to eat us because you talk too much.

A: I would never go camping in the woods for this very reason.

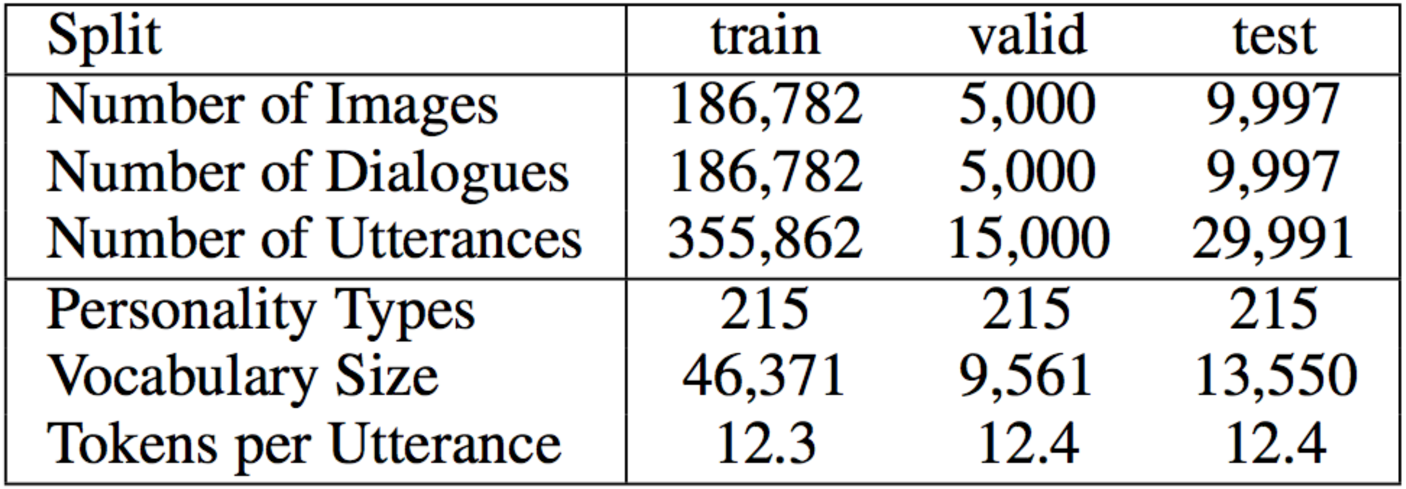
A: Money-Minded *B: Glamorous*

A: You know money doesn't grow on trees.

B: I could see some high society ladies having their brunch over looking this canal.

A: I could see them spending way too much on avocado toast here.

Image-Chat : An engaging image grounded Dialogue dataset



Split	train	valid	test
Number of Images	186,782	5,000	9,997
Number of Dialogues	186,782	5,000	9,997
Number of Utterances	355,862	15,000	29,991
Personality Types	215	215	215
Vocabulary Size	46,371	9,561	13,550
Tokens per Utterance	12.3	12.4	12.4

Table 1: IMAGE-CHAT dataset statistics.

Model Architectures!

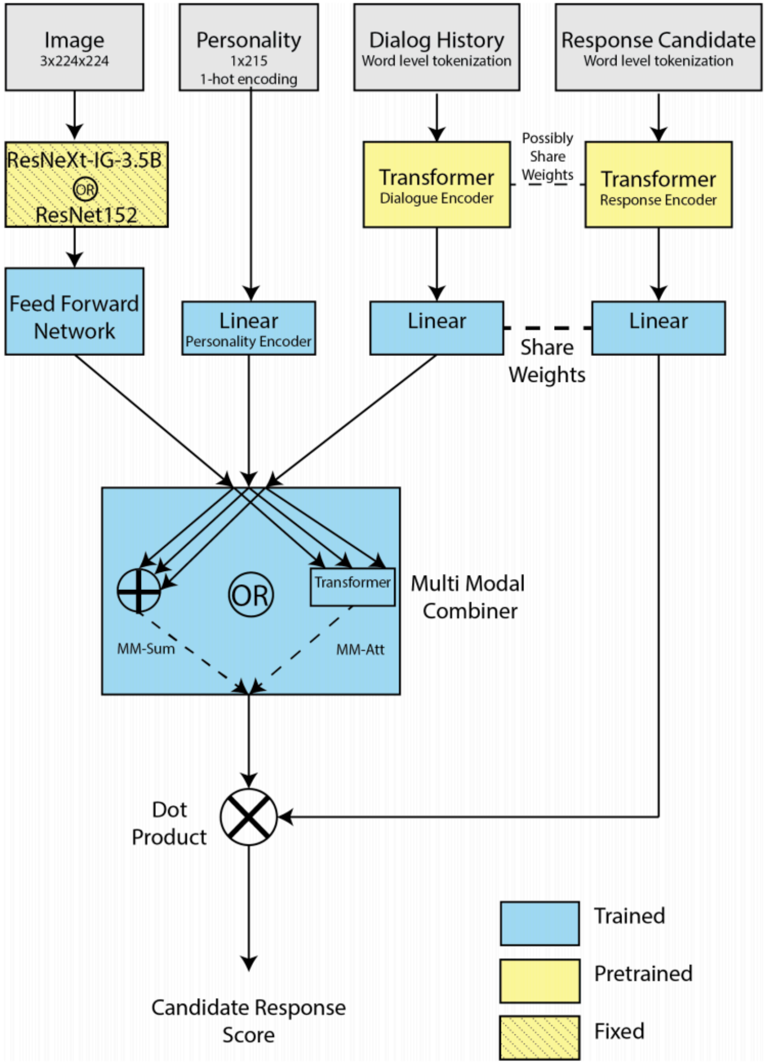


Figure 2: The TRANSRESNET Multimodal architecture for grounded dialogue. There are several options: different image encoders (ResNet152 or ResNeXt-IG-3.5B), text encoders (shared or separate Transformers for dialogue history and response), and different Multimodal combiners (sum or attention-based).

Results

- ResNext-IG-3.5B helps again!
- Later turns are harder, but performance still ok

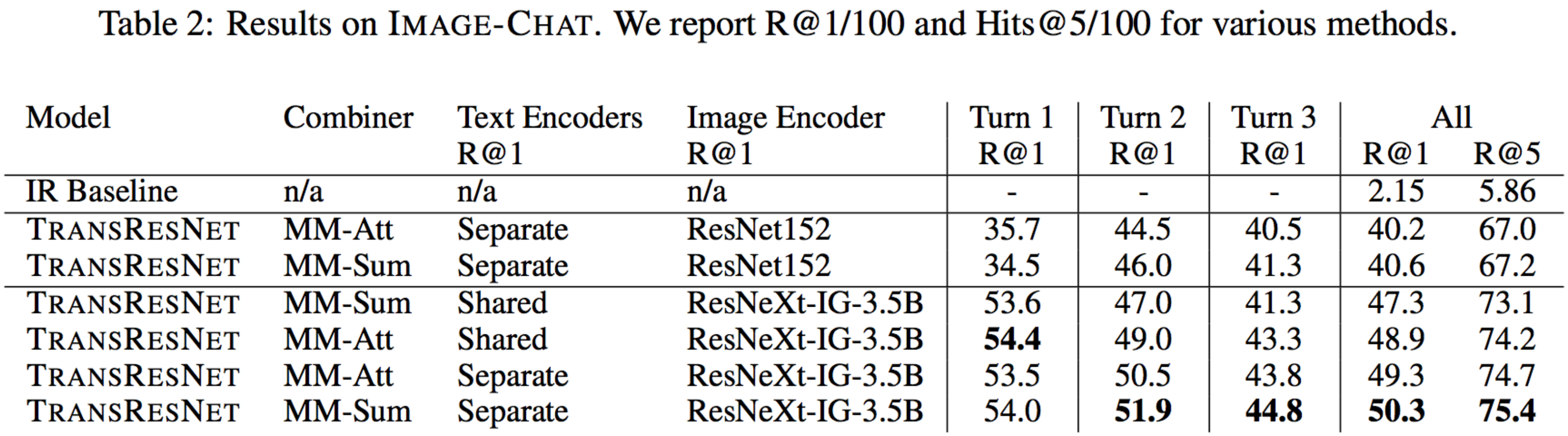


Table 2: Results on Image-Chat. We report R@1/100 and Hits@5/100 for various methods.

Model	Combiner	Text Encoders	Image Encoder	Turn 1	Turn 2	Turn 3	All	
		R@1	R@1	R@1	R@1	R@1	R@1	R@5
IR Baseline	n/a	n/a	n/a	-	-	-	2.15	5.86
TransResNet	MM-Att	Separate	ResNet152	35.7	44.5	40.5	40.2	67.0
TransResNet	MM-Sum	Separate	ResNet152	34.5	46.0	41.3	40.6	67.2
TransResNet	MM-Sum	Shared	ResNeXt-IG-3.5B	53.6	47.0	41.3	47.3	73.1
TransResNet	MM-Att	Shared	ResNeXt-IG-3.5B	**54.4**	49.0	43.3	48.9	74.2
TransResNet	MM-Att	Separate	ResNeXt-IG-3.5B	53.5	50.5	43.8	49.3	74.7
TransResNet	MM-Sum	Separate	ResNeXt-IG-3.5B	54.0	**51.9**	**44.8**	**50.3**	**75.4**

Ablations:
Personality, Image and Dialog history all help!

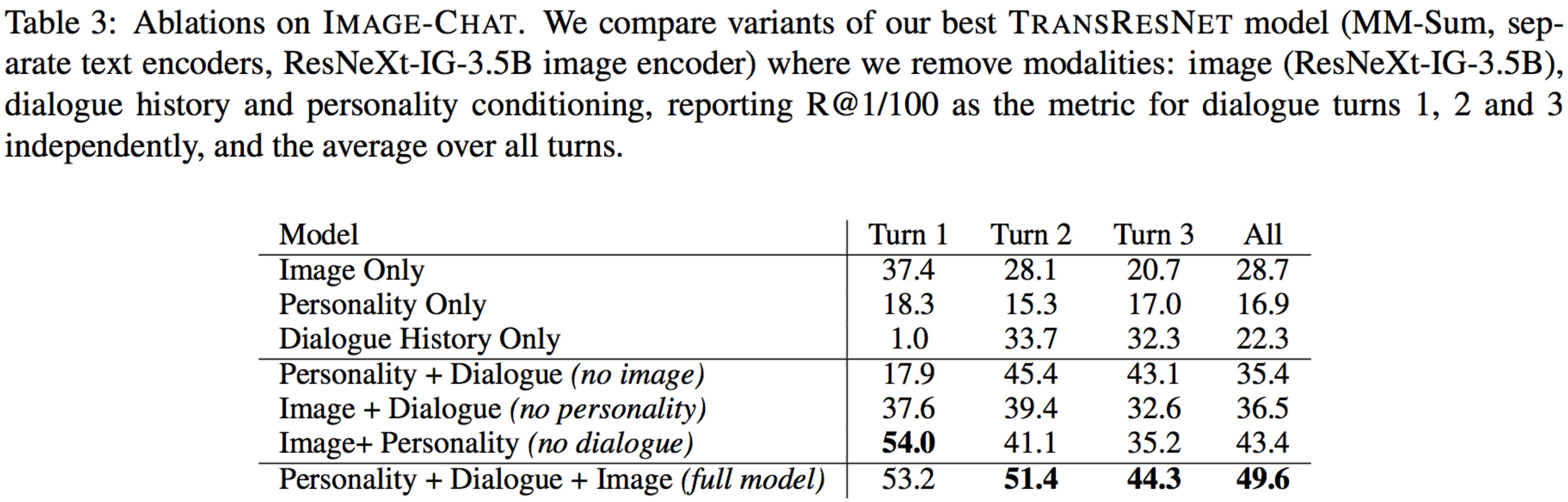


Table 3: Ablations on IMAGE-CHAT. We compare variants of our best TRANSRESNET model (MM-Sum, separate text encoders, ResNeXt-IG-3.5B image encoder) where we remove modalities: image (ResNeXt-IG-3.5B), dialogue history and personality conditioning, reporting R@1/100 as the metric for dialogue turns 1, 2 and 3 independently, and the average over all turns.

Model	Turn 1	Turn 2	Turn 3	All
Image Only	37.4	28.1	20.7	28.7
Personality Only	18.3	15.3	17.0	16.9
Dialogue History Only	1.0	33.7	32.3	22.3
Personality + Dialogue *(no image)*	17.9	45.4	43.1	35.4
Image + Dialogue *(no personality)*	37.6	39.4	32.6	36.5
Image+ Personality *(no dialogue)*	**54.0**	41.1	35.2	43.4
Personality + Dialogue + Image *(full model)*	53.2	**51.4**	**44.3**	**49.6**

Quite Decent Performance!

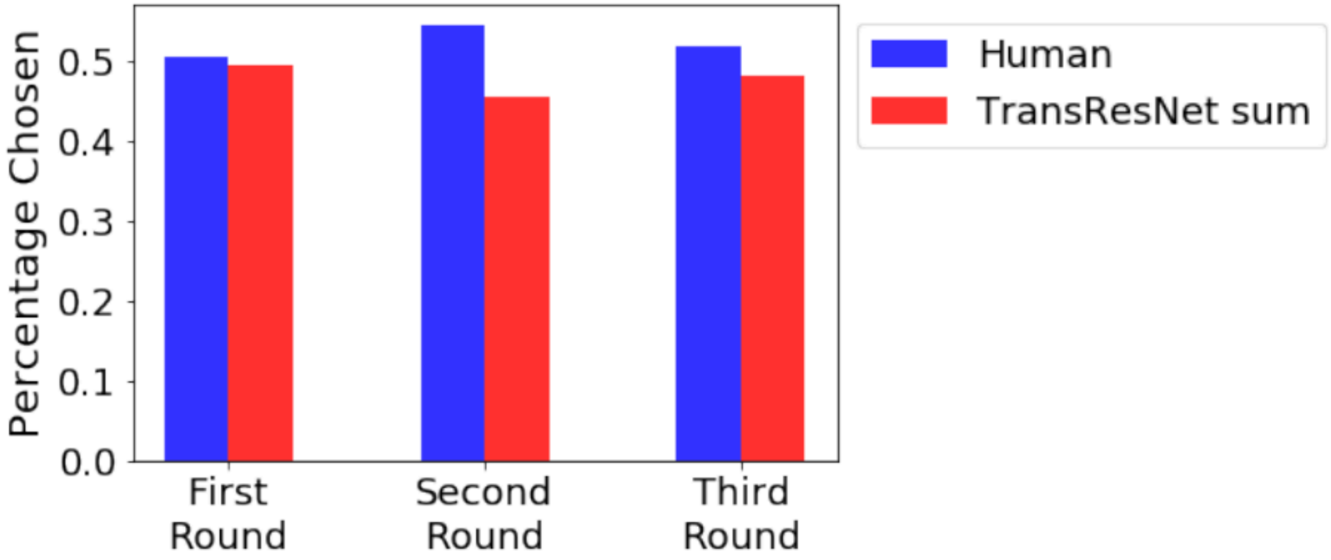


Figure 3: Human evaluations on IMAGE-CHAT. Engagingness win rates of pairwise comparisons between our best TRANSRESNET MM-Sum model versus human utterances, over three rounds of dialogue.

Image	Personality	Conversation
	A: Overimaginative	What if fog places humans in a 5th dimension.
	Model prediction: B: Crazy	Or imagine them getting like attacked by aliens and then the FBI finds the other FBI people buried there and then they're like whaaaat.
	A: Enigmatic	so what is that building suppose to mean
	Model prediction: B: Airy	It's just a building where people meet dude.
	A: Realistic	Beautiful ride on a cool winter day, but the road is slippery.
	Model prediction: B: Pretentious	Driving on the snow is the only way to go
	A: Cute	I sure hope the band members notice me! I've dressed up so cute for this concert.
	Model prediction: B: Cultured	I found their choice of dress to be quite becoming. I enjoyed the tunes immensely and found them to be charming.

Table 5: Example predictions from our best TRANSRESNET (MM-Sum) model on the human evaluation set for turn 2. Two speakers A and B with given personality traits discuss a given photo. The first turn is a human, and the second turn here is completed by our model in these examples.

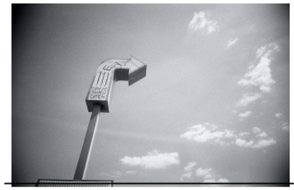
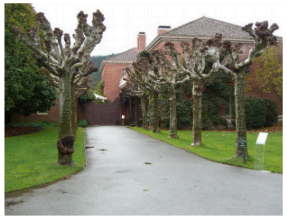
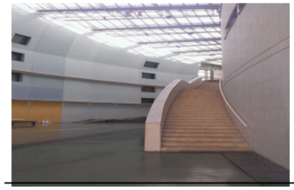
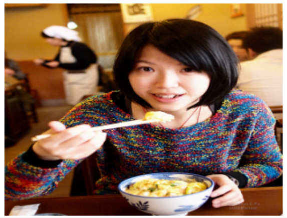
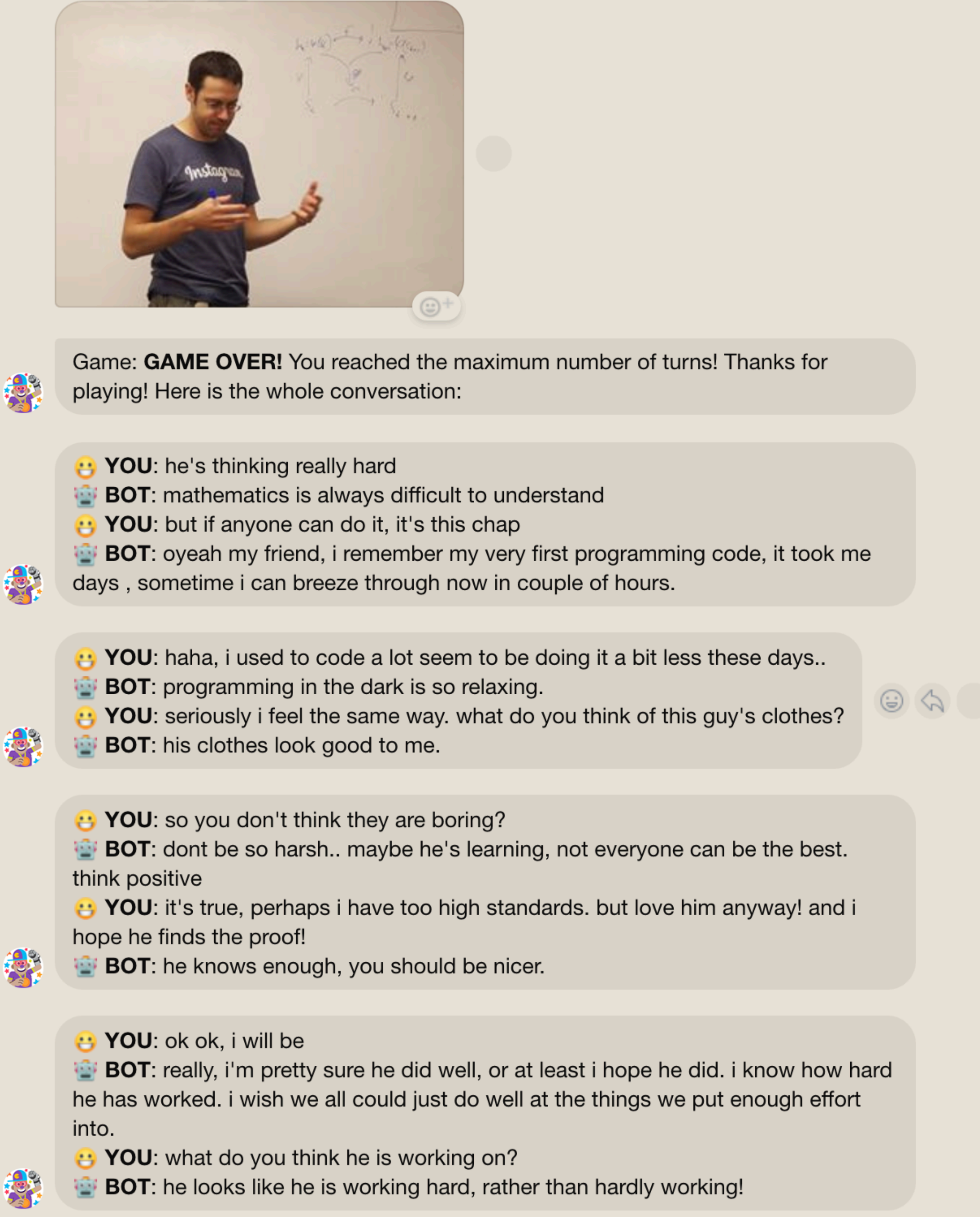
Image	Personality	Conversation
	A: Stylish	The black and white colors with the sign remind me of a better time.
	B: Cold	Retro signs are so dated.
	Model prediction:	
	A: Stylish	But they're so trendy!
	A: Boisterous	WOW LOOK AT THOSE TREES!!!! I have never seen tree cut like that. hmm how did they do that?
	B: Gentle	They use a special saw to cut them like that.
	Model prediction:	
	A: Boisterous	That's no excuse! They are killing trees!
	A: Witty	the round about way has lots of steps and multiple shades.
	B: Escapist	That staircase leads to bliss.
	Model prediction:	
	A: Witty	Perhaps the stairs go straight to the pearly gates!
	A: Cheerful	That food looks delicious! She looks so happy to be there, I would too for sure.
	B: Resentful	gee, wish I could eat out all the time
	Model prediction:	
	A: Cheerful	Yeah but look at how happy she is it reminds me of all the wonderful food I have eaten

Table 6: Example predictions from our best TRANSRESNET (MM-Sum) model on the human evaluation set for turn 3. Two speakers A and B with given personality traits discuss a given photo. The first two turns are from humans, and only the third turn here is completed by our model in these examples.



Game: **GAME OVER!** You reached the maximum number of turns! Thanks for playing! Here is the whole conversation:

😬 **YOU**: he's thinking really hard
🤖 **BOT**: mathematics is always difficult to understand
😬 **YOU**: but if anyone can do it, it's this chap
🤖 **BOT**: oyeah my friend, i remember my very first programming code, it took me days , sometime i can breeze through now in couple of hours.

😬 **YOU**: haha, i used to code a lot seem to be doing it a bit less these days..
🤖 **BOT**: programming in the dark is so relaxing.
😬 **YOU**: seriously i feel the same way. what do you think of this guy's clothes?
🤖 **BOT**: his clothes look good to me.

😬 **YOU**: so you don't think they are boring?
🤖 **BOT**: dont be so harsh.. maybe he's learning, not everyone can be the best. think positive
😬 **YOU**: it's true, perhaps i have too high standards. but love him anyway! and i hope he finds the proof!
🤖 **BOT**: he knows enough, you should be nicer.

😬 **YOU**: ok ok, i will be
🤖 **BOT**: really, i'm pretty sure he did well, or at least i hope he did. i know how hard he has worked. i wish we all could just do well at the things we put enough effort into.
😬 **YOU**: what do you think he is working on?
🤖 **BOT**: he looks like he is working hard, rather than hardly working!

Our goal: talk to our computer, *superhuman conversationalist*

Some properties our agent should have:

Engagingness:	have personality, curiosity, diversity
Expertness:	be knowledgeable & correct
Images:	can interact with text AND other modalities
Continual learning:	grow and learn from experience

Persona

I am eng
I used to
My favor
I am a co
I love wa

[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSO
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

Persona-Chat (engaging personality)

A: Stylish *B: Fatalistic*

A: Riding a mechanical bull in a skirt is just my style.

B: You'd probably fall off and get hurt.

A: And everyone would be copying me for it! It'll be trendy!

A: Fearful *B: Miserable*

A: I just heard something out there and I have no idea what it was.

B: It was probably a Wolf coming to eat us because you talk too much.

A: I would never go camping in the woods for this very reason.

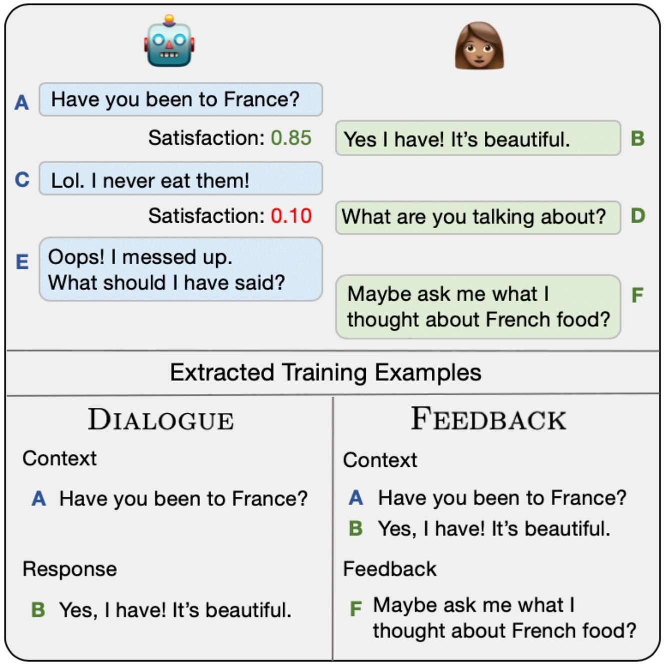
Image-Chat (images)

Wizard of Wikipedia (expert knowledge)

Topic: Toga party (*seen*)
Model: I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human: Me too!!! I usually use some flip flops and my blanket!
Model: Toga parties typically have games involving kegs and excessive drinking!
Human: Yeah! And getting to wear funny outfits and a crown of leaves!
Model: I love me some Toga Parties!
Human: I love me some too!
Model: Toga parties are so much fun with the Roman or Greek theme.
Human: Yeah! I especially like Greek themes

Topic: E-book (*seen*)
Human: i do not know why, but I have never gotten into E-books
Model: Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.
Human: yes, i like the physical feel and smell of a real book
Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human: i tried e-books but I do not like reading on a screen, its disorientated for me
Model: Yes, reading a book on a tablet just isn't the same!
Human: plus I like owning the physical copy of a real book, I feel like I own it
Model: I prefer print and newsreels.

Self-feeding (continual learning)



A Have you been to France?
Satisfaction: 0.85
B Yes I have! It's beautiful.
C Lol. I never eat them!
Satisfaction: 0.10
D What are you talking about?
E Oops! I messed up. What should I have said?
F Maybe ask me what I thought about French food?

Extracted Training Examples

DIALOGUE	FEEDBACK
Context	Context
A Have you been to France?	A Have you been to France? B Yes, I have! It's beautiful.
Response	Feedback
B Yes, I have! It's beautiful.	F Maybe ask me what I thought about French food?

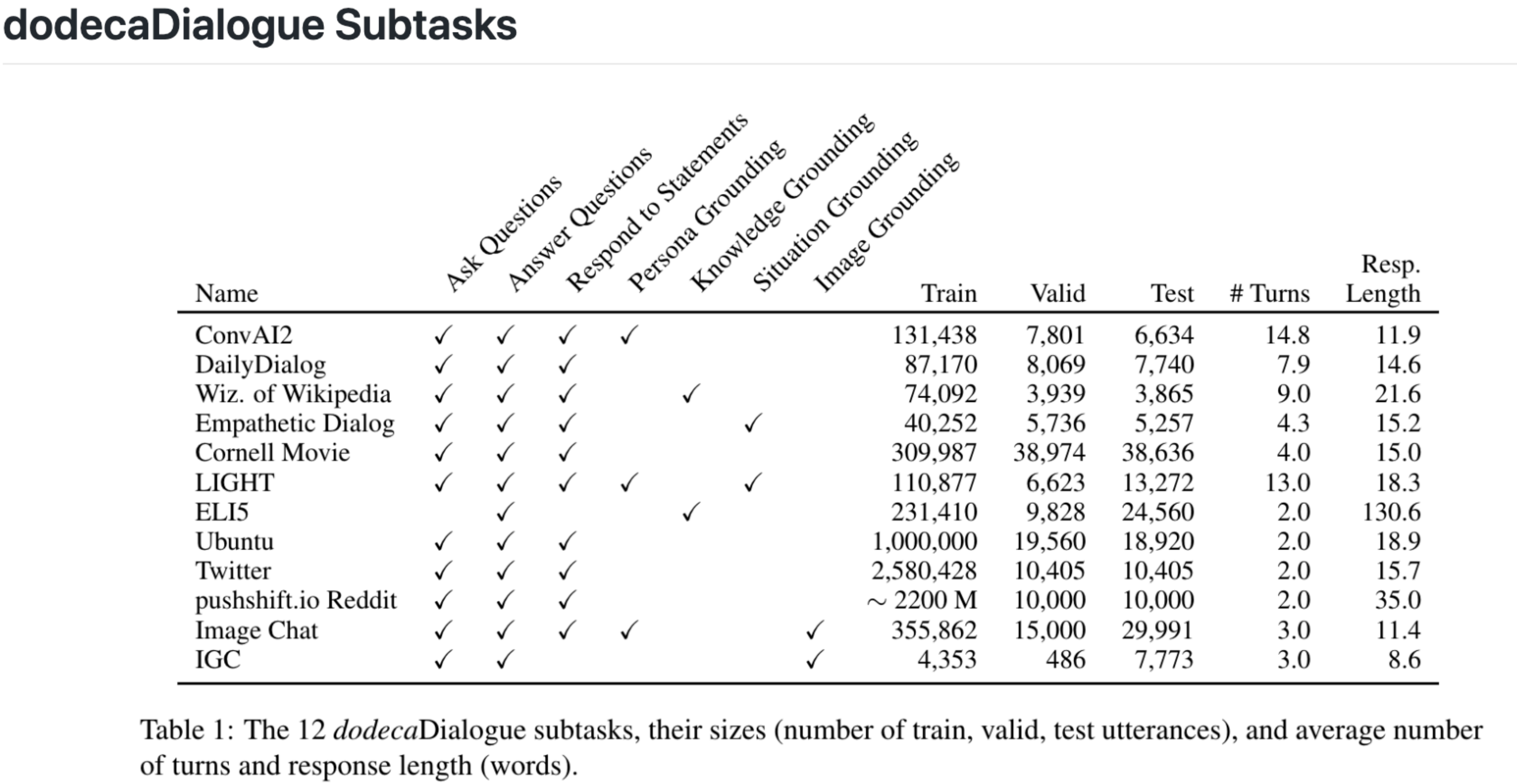
Latest *in progress* research

Putting it all together

Multi-Task: DodecaDialogue 12 task challenge

The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, Jason Weston

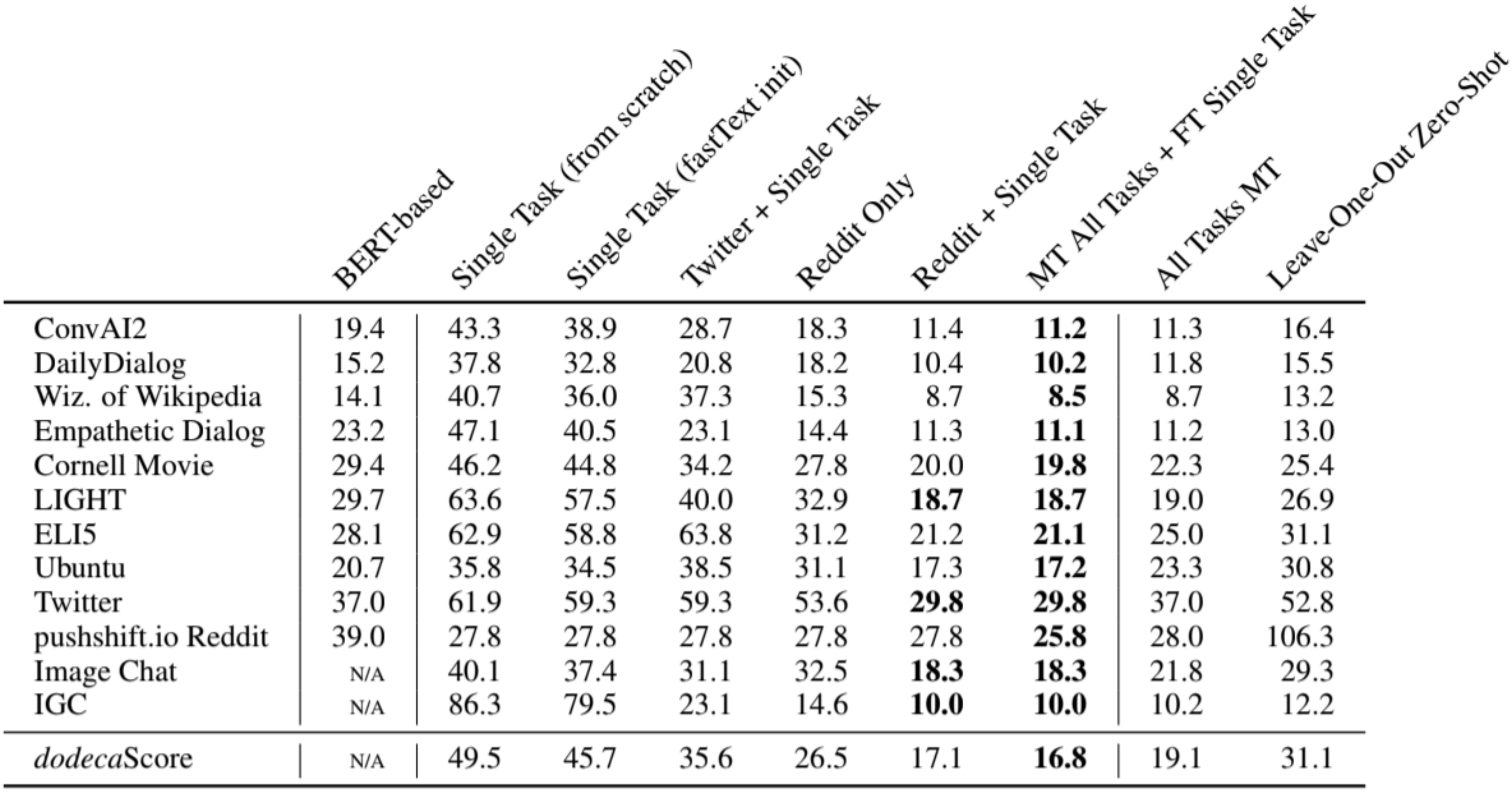


dodecaDialogue Subtasks

Name	Ask Questions	Answer Questions	Respond to Statements	Persona Grounding	Knowledge Grounding	Situation Grounding	Image Grounding	Train	Valid	Test	# Turns	Resp. Length
ConvAI2	✓	✓	✓	✓				131,438	7,801	6,634	14.8	11.9
DailyDialog	✓	✓	✓					87,170	8,069	7,740	7.9	14.6
Wiz. of Wikipedia	✓	✓	✓		✓			74,092	3,939	3,865	9.0	21.6
Empathetic Dialog	✓	✓	✓			✓		40,252	5,736	5,257	4.3	15.2
Cornell Movie	✓	✓	✓					309,987	38,974	38,636	4.0	15.0
LIGHT	✓	✓	✓	✓		✓		110,877	6,623	13,272	13.0	18.3
ELI5		✓			✓			231,410	9,828	24,560	2.0	130.6
Ubuntu	✓	✓	✓					1,000,000	19,560	18,920	2.0	18.9
Twitter	✓	✓	✓					2,580,428	10,405	10,405	2.0	15.7
pushshift.io Reddit	✓	✓	✓					$\sim$ 2200 M	10,000	10,000	2.0	35.0
Image Chat	✓	✓	✓	✓			✓	355,862	15,000	29,991	3.0	11.4
IGC	✓	✓					✓	4,353	486	7,773	3.0	8.6

Table 1: The 12 *dodeca*Dialogue subtasks, their sizes (number of train, valid, test utterances), and average number of turns and response length (words).

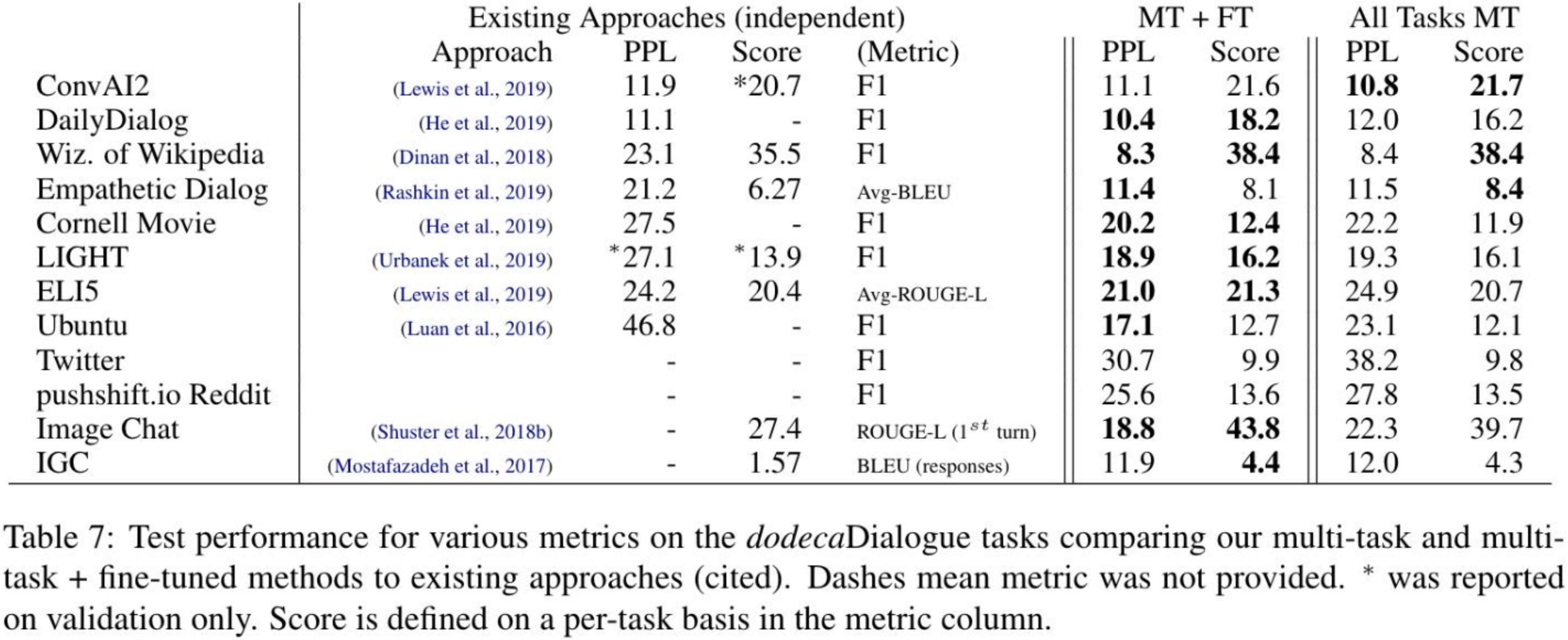
Train a image+seq2seq Transformer model on all Tasks, and then fine-tune (MT+FT) for best PPL



	BERT-based	Single Task (from scratch)	Single Task (fastText init)	Twitter + Single Task	Reddit Only	Reddit + Single Task	MT All Tasks + FT Single Task	All Tasks MT	Leave-One-Out Zero-Shot
ConvAI2	19.4	43.3	38.9	28.7	18.3	11.4	**11.2**	11.3	16.4
DailyDialog	15.2	37.8	32.8	20.8	18.2	10.4	**10.2**	11.8	15.5
Wiz. of Wikipedia	14.1	40.7	36.0	37.3	15.3	8.7	**8.5**	8.7	13.2
Empathetic Dialog	23.2	47.1	40.5	23.1	14.4	11.3	**11.1**	11.2	13.0
Cornell Movie	29.4	46.2	44.8	34.2	27.8	20.0	**19.8**	22.3	25.4
LIGHT	29.7	63.6	57.5	40.0	32.9	**18.7**	**18.7**	19.0	26.9
ELI5	28.1	62.9	58.8	63.8	31.2	21.2	**21.1**	25.0	31.1
Ubuntu	20.7	35.8	34.5	38.5	31.1	17.3	**17.2**	23.3	30.8
Twitter	37.0	61.9	59.3	59.3	53.6	**29.8**	**29.8**	37.0	52.8
pushshift.io Reddit	39.0	27.8	27.8	27.8	27.8	27.8	**25.8**	28.0	106.3
Image Chat	N/A	40.1	37.4	31.1	32.5	**18.3**	**18.3**	21.8	29.3
IGC	N/A	86.3	79.5	23.1	14.6	**10.0**	**10.0**	10.2	12.2
*dodeca*Score	N/A	49.5	45.7	35.6	26.5	17.1	**16.8**	19.1	31.1

Table 2: Validation perplexity for the *dodeca*Dialogue tasks in various settings.

SOTA on all tasks by training an image+seq2seq model MT+FT
Full Multitask model is close behind

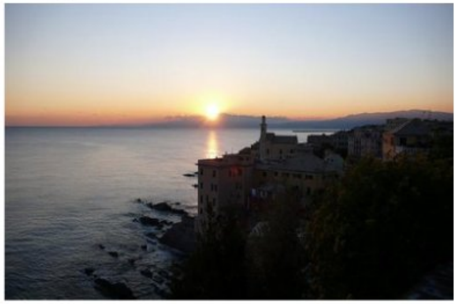


	Existing Approaches (independent)				MT + FT		All Tasks MT	
	Approach	PPL	Score	(Metric)	PPL	Score	PPL	Score
ConvAI2	(Lewis et al., 2019)	11.9	*20.7	F1	11.1	21.6	**10.8**	**21.7**
DailyDialog	(He et al., 2019)	11.1	-	F1	**10.4**	**18.2**	12.0	16.2
Wiz. of Wikipedia	(Dinan et al., 2018)	23.1	35.5	F1	**8.3**	**38.4**	8.4	**38.4**
Empathetic Dialog	(Rashkin et al., 2019)	21.2	6.27	Avg-BLEU	**11.4**	8.1	11.5	**8.4**
Cornell Movie	(He et al., 2019)	27.5	-	F1	**20.2**	**12.4**	22.2	11.9
LIGHT	(Urbanek et al., 2019)	*27.1	*13.9	F1	**18.9**	**16.2**	19.3	16.1
ELI5	(Lewis et al., 2019)	24.2	20.4	Avg-ROUGE-L	**21.0**	**21.3**	24.9	20.7
Ubuntu	(Luan et al., 2016)	46.8	-	F1	**17.1**	12.7	23.1	12.1
Twitter		-	-	F1	30.7	9.9	38.2	9.8
pushshift.io Reddit		-	-	F1	25.6	13.6	27.8	13.5
Image Chat	(Shuster et al., 2018b)	-	27.4	ROUGE-L (1st turn)	**18.8**	**43.8**	22.3	39.7
IGC	(Mostafazadeh et al., 2017)	-	1.57	BLEU (responses)	11.9	**4.4**	12.0	4.3

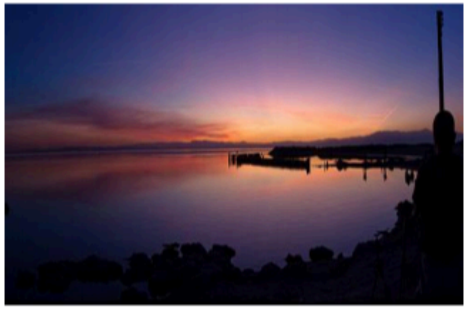
Table 7: Test performance for various metrics on the *dodeca*Dialogue tasks comparing our multi-task and multi-task + fine-tuned methods to existing approaches (cited). Dashes mean metric was not provided. * was reported on validation only. Score is defined on a per-task basis in the metric column.

Model Examples

Context	Wizard of Wikipedia	
Chosen Topic: Dog **Knowledge:** The dog was the first species to be domesticated and has been selectively bred over millennia for various behaviors, sensory capabilities, and physical attributes.	Speaker A:	Hi, do you like dogs?
	Speaker B:	I do! Actually, my favourite is the Basenji breed! It is a hunting dog.
	Speaker A:	Oh, those are cool. I have seen them on the Westminster Dog Show, but I've never met one in person. I have a Havanese.
	Speaker B:	Wow, are you Cuban? I know that the Havanese are the national dog of cuba!
	Speaker A:	No, not Cuban, but yes they are the national dog of Cuba! Apparently all the Havanese in the US are descended from a handful that came here from Cuba during the revolution!
	Model:	that's so cool! the dog was the first species to be domesticated!

Context	Image Chat	
Image  **Personality** Playful	Speaker A:	Great for night fishing.
	Speaker B:	Meh, I doubt any fish would bite anyway.
	Model:	i m sure there are plenty of fish that would bite.

Context	Wizard of Wikipedia	
Chosen Topic:		
The Rolling Stones	Speaker A:	
Knowledge:	Speaker B:	I love the Rolling Stones. Do you? They were formed in 1962.
no_passages_used	Speaker A:	I love them! Who was in the band at the start?
	Speaker B:	The first lineup was Brian Jones, Mick, Keith, Bill Wyman, Charlie Watts and Ian Steward.
	Speaker A:	Was Mick ever replaced?
	Model:	i'm not sure, but i do know that mick jagger was the lead vocalist.

Context	Image Chat (First Turn)	
Image		
	Model:	i wonder if this is a lake?
Personality		
Curious		

Improved Open-ended Generation with Unlikelihood Training

WHAC-A-MOLE
GAME
Electronic Lights and Sounds
Beat the clock in this mole-smackin' frenzy!
Preschool
AGES 4+
PLAYERS 1-2
3xAA
Fisher-Price GAMES

Collaborators: Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho + Margaret Li, Y-Lan Boureau (for new stuff!)

Natural Language Generation task spectrum

Machine Translation

Sentence Compression

Abstractive Summarization

Chitchat Dialogue

Story Generation

Less open-ended

Mostly word-level decisions

Seq2seq more successful

More open-ended

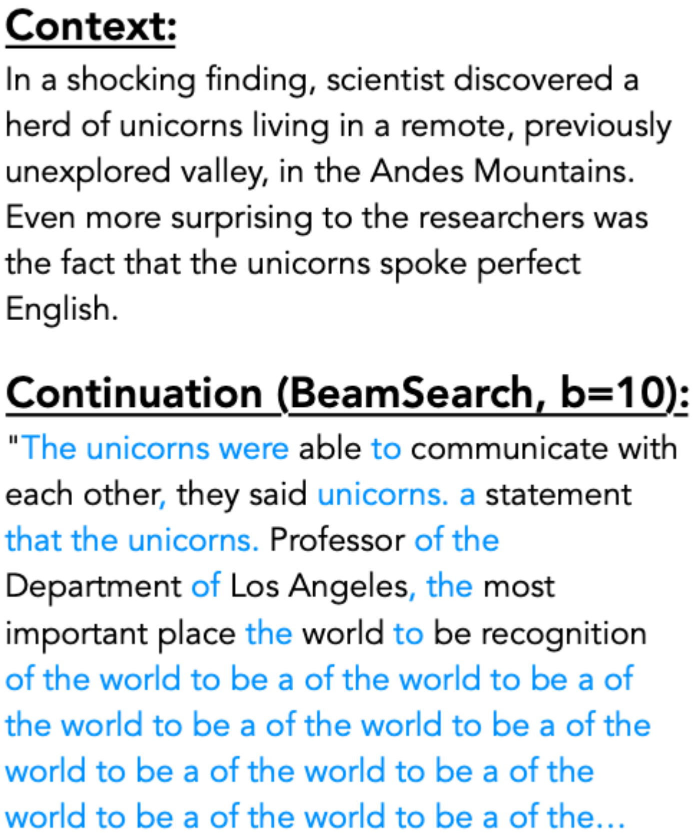
Requires high-level decisions

Seq2seq less successful

More research needed!

Our projects

Slides stolen from Abigail See



Context:

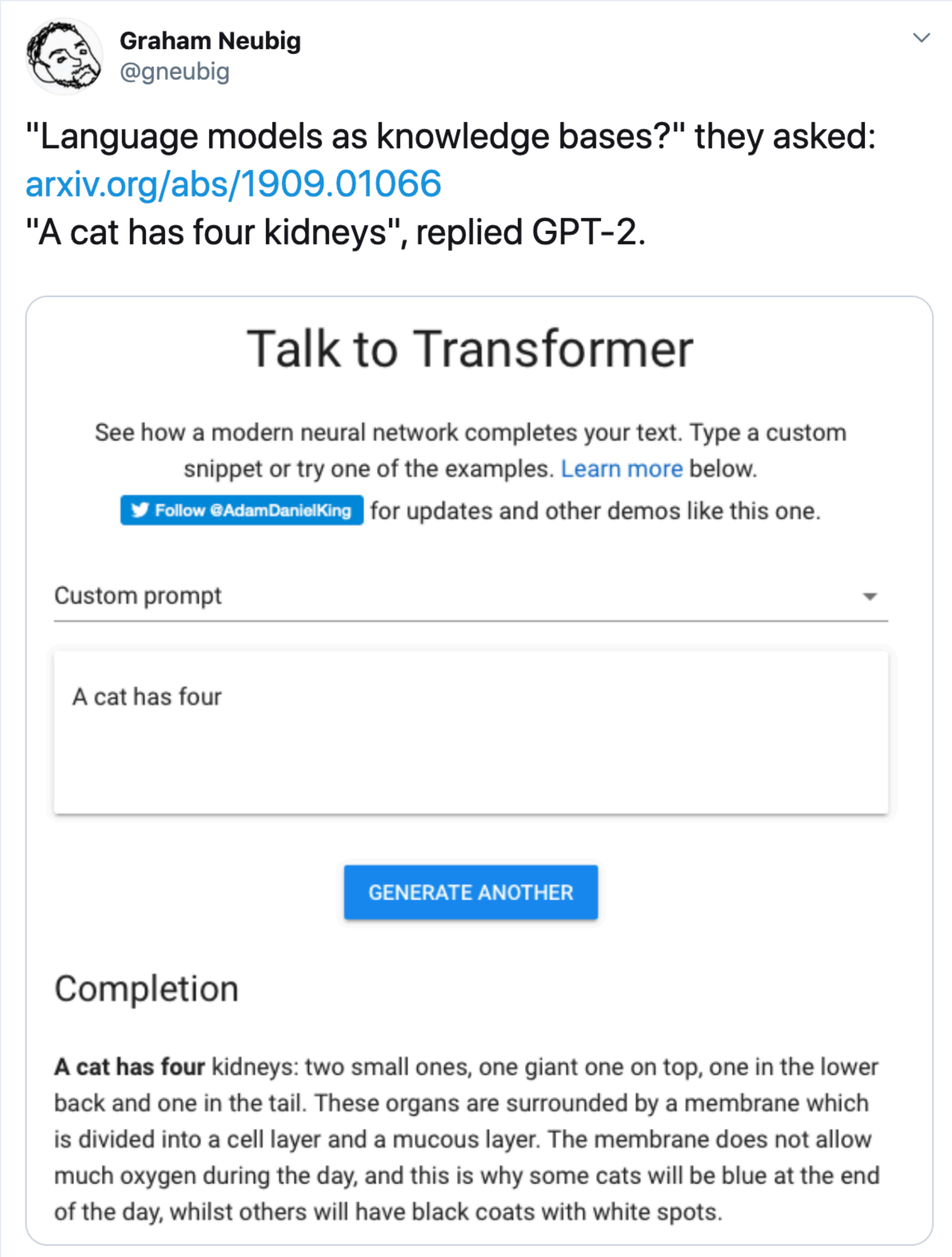
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Continuation (BeamSearch, b=10):

"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the…

Figure 1: Beam search leads to degenerate text, even when generated from GPT-2-117M, in stark contrast with the admirable quality of the text decoded using *top-k* sampling (Radford et al., 2019). The *continuation* is machine generated, conditioned on the *context* provided by a human. Blue text highlights decoded words that have occurred previously in the text.

The Curious Case of Neural Text Degeneration, Holtzman et al.'19



Graham Neubig
@gneubig

"Language models as knowledge bases?" they asked: arxiv.org/abs/1909.01066
"A cat has four kidneys", replied GPT-2.

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. Learn more below.

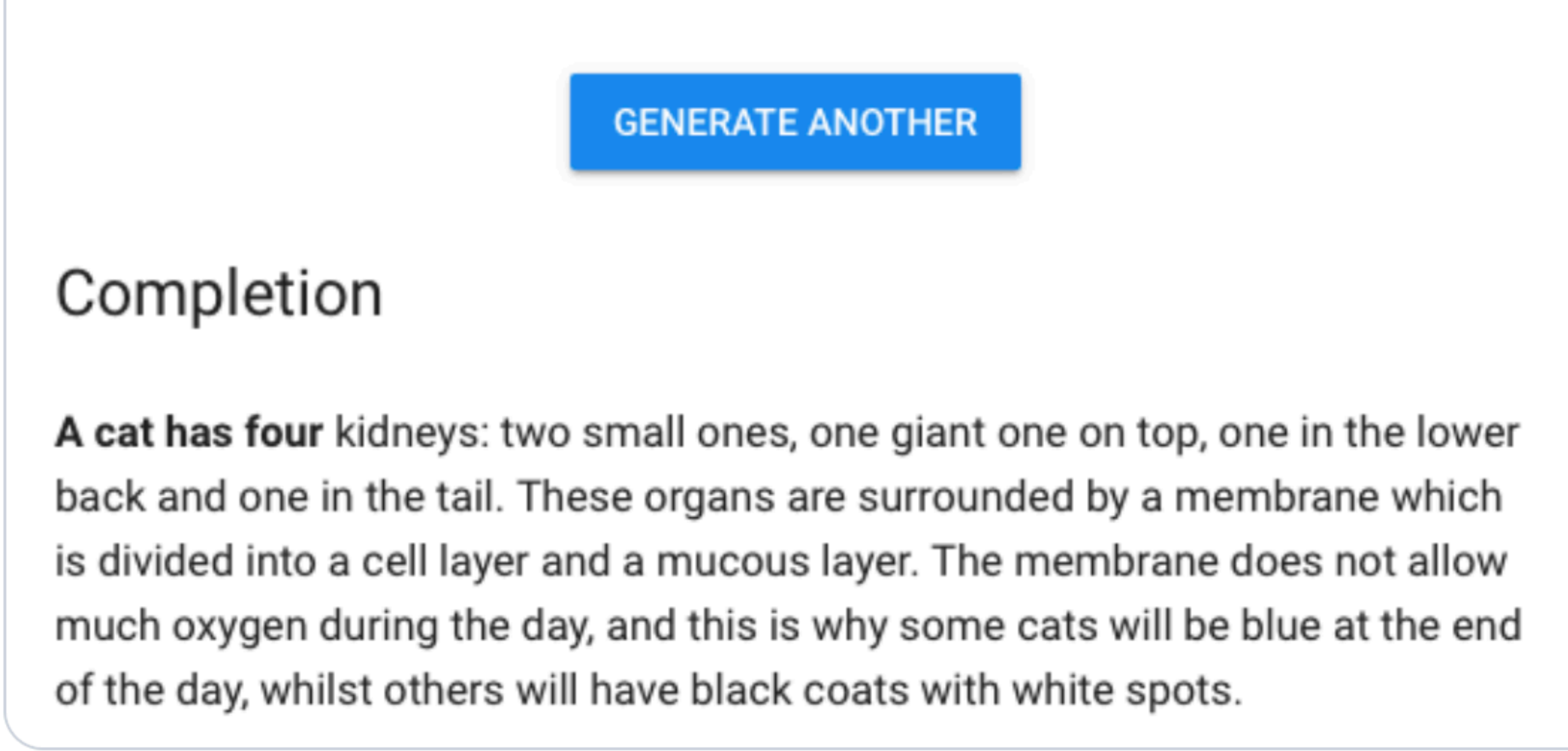
Follow @AdamDanielKing for updates and other demos like this one.

Custom prompt

A cat has four

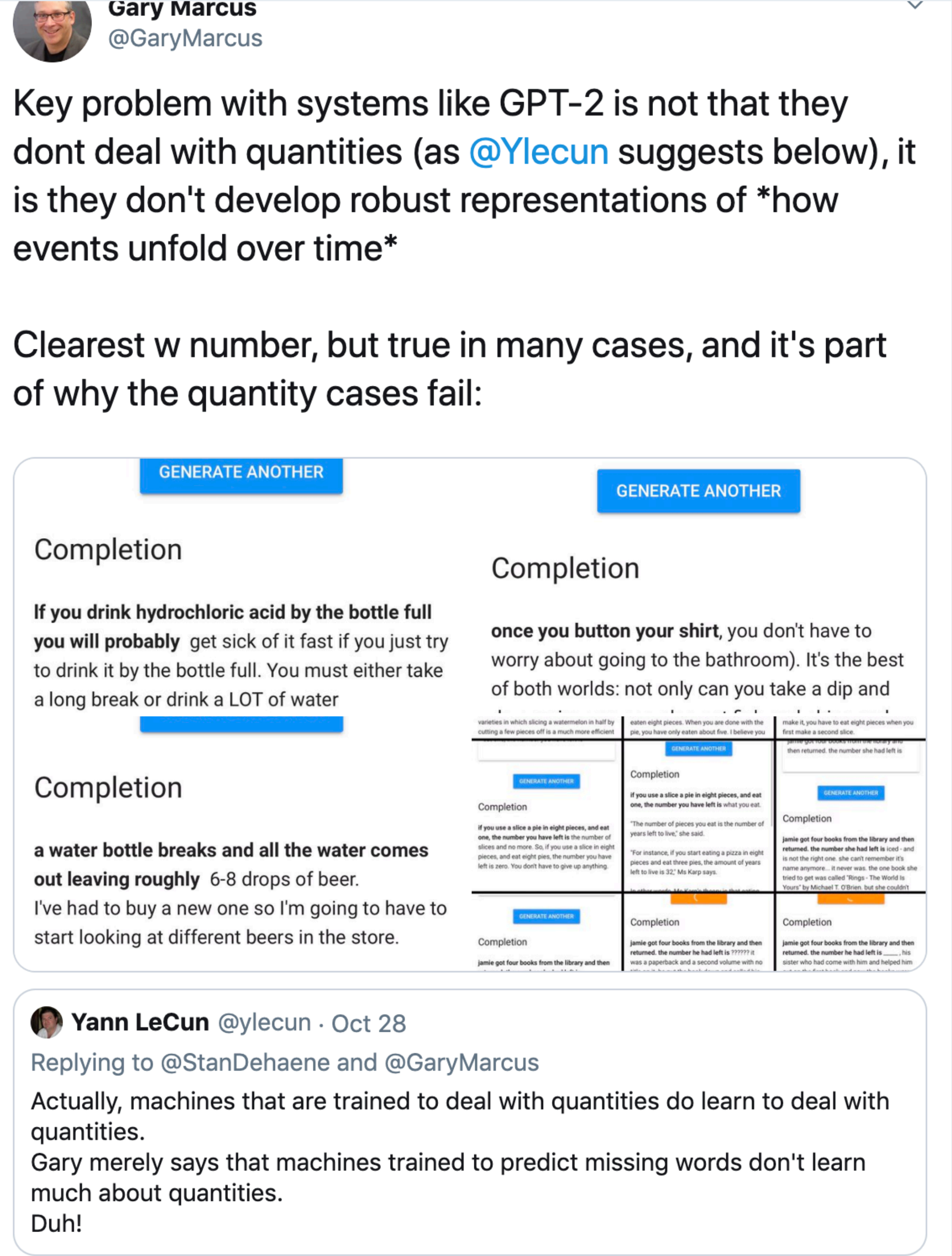
Completion

A cat has four kidneys: two small ones, one giant one on top, one in the lower back and one in the tail. These organs are surrounded by a membrane which is divided into a cell layer and a mucous layer. The membrane does not allow much oxygen during the day, and this is why some cats will be blue at the end of the day, whilst others will have black coats with white spots.



Completion

A cat has four kidneys: two small ones, one giant one on top, one in the lower back and one in the tail. These organs are surrounded by a membrane which is divided into a cell layer and a mucous layer. The membrane does not allow much oxygen during the day, and this is why some cats will be blue at the end of the day, whilst others will have black coats with white spots.



Key problem with systems like GPT-2 is not that they dont deal with quantities (as @Ylecun suggests below), it is they don't develop robust representations of *how events unfold over time*

Clearest w number, but true in many cases, and it's part of why the quantity cases fail:

Completion

If you drink hydrochloric acid by the bottle full you will probably get sick of it fast if you just try to drink it by the bottle full. You must either take a long break or drink a LOT of water

Completion

a water bottle breaks and all the water comes out leaving roughly 6-8 drops of beer. I've had to buy a new one so I'm going to have to start looking at different beers in the store.

Completion

once you button your shirt, you don't have to worry about going to the bathroom). It's the best of both worlds: not only can you take a dip and

varieties in which slicing a watermelon in half by cutting a few pieces off is a much more efficient

eaten eight pieces. When you are done with the pie, you have only eaten about five. I believe you

make it, you have to eat eight pieces when you first make a second slice.

then returned. the number she had left is

Completion

if you use a slice a pie in eight pieces, and eat one, the number you have left is the number of slices and no more. So, if you use a slice in eight pieces, and eat eight pies, the number you have left is zero. You don't have to give up anything.

Completion

if you use a slice a pie in eight pieces, and eat one, the number you have left is what you eat.

"The number of pieces you eat is the number of years left to live," she said.

"For instance, if you start eating a pizza in eight pieces and eat three pies, the amount of years left to live is 32," Ms Karp says.

Completion

jamie got four books from the library and then returned. the number she had left is iced - and is not the right one. she can't remember it's name anymore... it never was. the one book she tried to get was called "Rings - The World Is Yours" by Michael T. O'Brien. but she couldn't

Completion

jamie got four books from the library and then

Completion

jamie got four books from the library and then returned. the number he had left is ?????? it was a paperback and a second volume with no

Completion

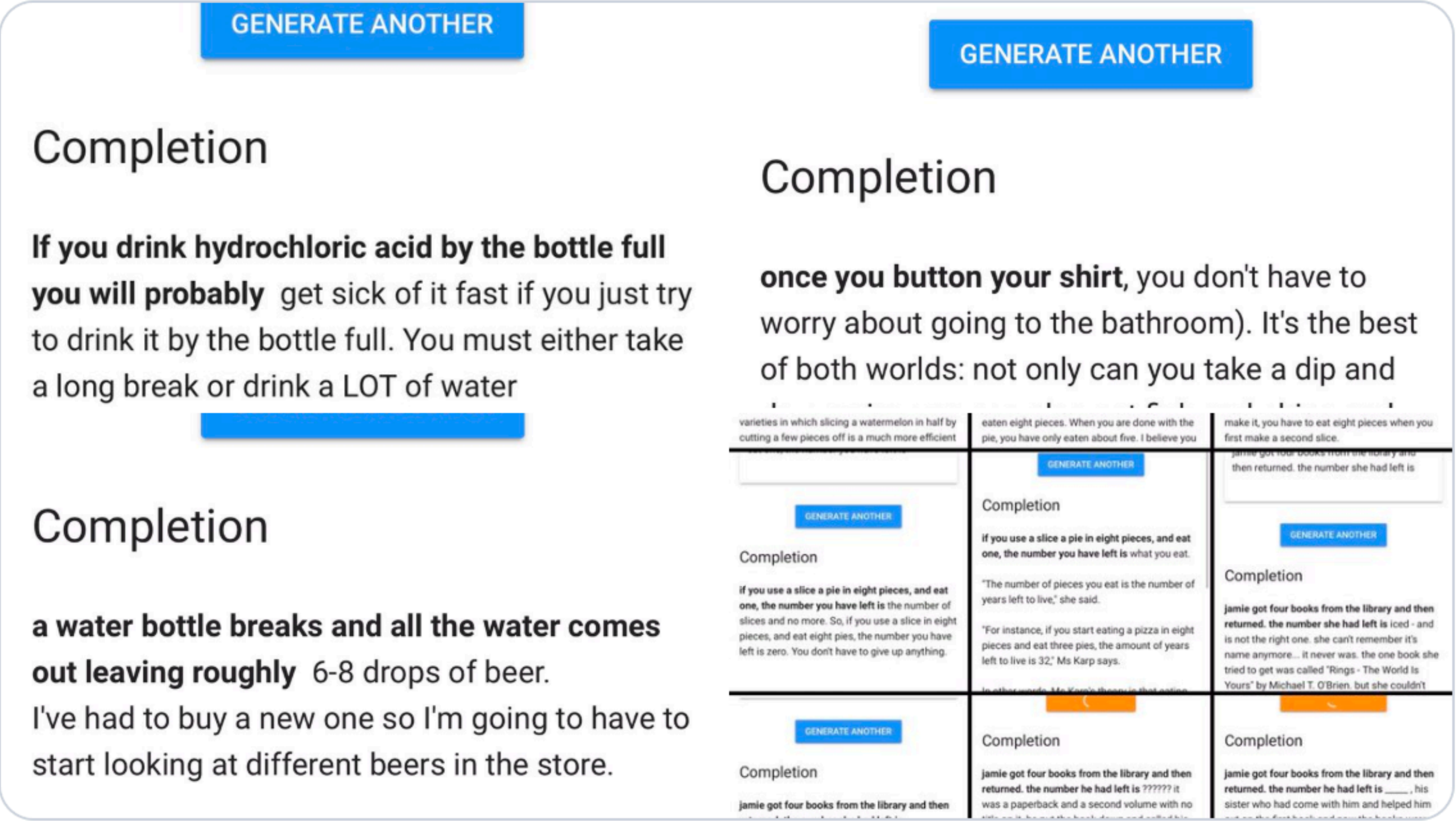
jamie got four books from the library and then returned. the number he had left is ____, his sister who had come with him and helped him

Yann LeCun @ylecun · Oct 28

Replying to @StanDehaene and @GaryMarcus

Actually, machines that are trained to deal with quantities do learn to deal with quantities.
Gary merely says that machines trained to predict missing words don't learn much about quantities.
Duh!

of why the quantity cases fail:



Completion

If you drink hydrochloric acid by the bottle full you will probably get sick of it fast if you just try to drink it by the bottle full. You must either take a long break or drink a LOT of water

Completion

a water bottle breaks and all the water comes out leaving roughly 6-8 drops of beer.
I've had to buy a new one so I'm going to have to start looking at different beers in the store.

Completion

once you button your shirt, you don't have to worry about going to the bathroom). It's the best of both worlds: not only can you take a dip and

varieties in which slicing a watermelon in half by cutting a few pieces off is a much more efficient

Completion

if you use a slice a pie in eight pieces, and eat one, the number you have left is the number of slices and no more. So, if you use a slice in eight pieces, and eat eight pies, the number you have left is zero. You don't have to give up anything.

Completion

jamie got four books from the library and then

eaten eight pieces. When you are done with the pie, you have only eaten about five. I believe you

Completion

If you use a slice a pie in eight pieces, and eat one, the number you have left is what you eat.

"The number of pieces you eat is the number of years left to live," she said.

"For instance, if you start eating a pizza in eight pieces and eat three pies, the amount of years left to live is 32," Ms Karp says.

Completion

jamie got four books from the library and then returned. the number he had left is ?????? it was a paperback and a second volume with no

make it, you have to eat eight pieces when you first make a second slice.

then returned. the number she had left is

Completion

jamie got four books from the library and then returned. the number she had left is iced - and is not the right one. she can't remember it's name anymore... it never was. the one book she tried to get was called "Rings - The World Is Yours" by Michael T. O'Brien. but she couldn't

Completion

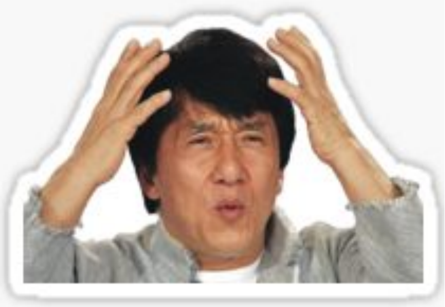
jamie got four books from the library and then returned. the number he had left is _____ , his sister who had come with him and helped him

Yann LeCun @ylecun · Oct 28

Replying to @StanDehaene and @GaryMarcus

Persona 1	Persona 2
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?
[PERSON 2:] I am doing good and you?
[PERSON 1:] I am well, just drinking some scotch and working on my writing
[PERSON 2:] I love to cook and bake but I rarely make time for it
[PERSON 1:] what do you usually bake?
[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?
[PERSON 1:] I used to be a juggler at a carnival but now I am an author
[PERSON 2:] I am a dancer, what do you do?
[PERSON 1:] I just told you. Do you have any pets?
[PERSON 2:] no, I am allergic to pets unfortunately.
[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!
[PERSON 2:] I have a cat. she has been with me through three wives, haha.
[PERSON 1:] Funny that you just told me you have no pets
[PERSON 2:] I do not have any animals.



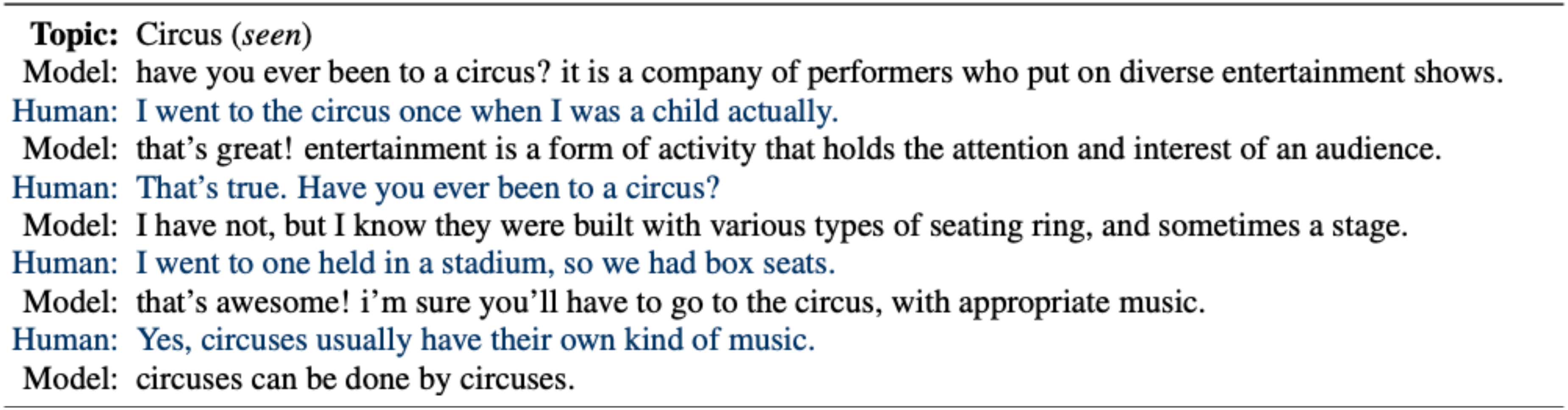
: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

Persona-Chat (Zhang et al., ACL '18)

Wizard of Wikipedia, Dinan et al., ICLR ‘19

Have a conversation about an in-depth topic, conditioning on retrieved paragraphs from Wikipedia.

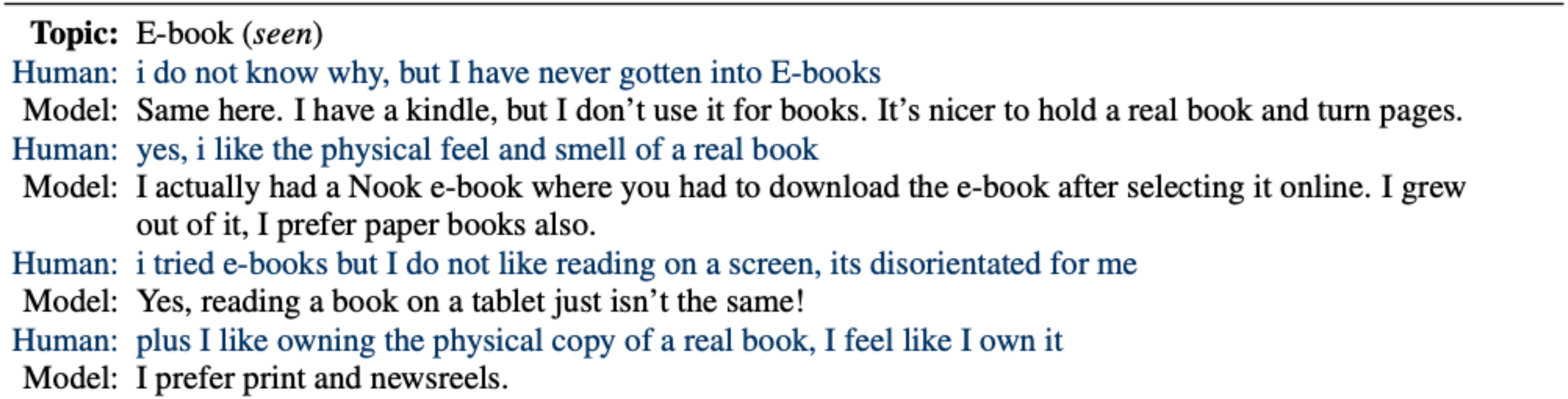
Generation



Topic:	Circus (*seen*)
Model:	have you ever been to a circus? it is a company of performers who put on diverse entertainment shows.
Human:	I went to the circus once when I was a child actually.
Model:	that’s great! entertainment is a form of activity that holds the attention and interest of an audience.
Human:	That’s true. Have you ever been to a circus?
Model:	I have not, but I know they were built with various types of seating ring, and sometimes a stage.
Human:	I went to one held in a stadium, so we had box seats.
Model:	that’s awesome! i’m sure you’ll have to go to the circus, with appropriate music.
Human:	Yes, circuses usually have their own kind of music.
Model:	circuses can be done by circuses.

- Tends to copy the knowledge source (Wikipedia), a bit dull
- Local & Global repetition in dialogue
- Sometimes uses incorrect date or amount, etc.

Retrieval (3.4 human eval)



Topic: E-book (*seen*)

Human: i do not know why, but I have never gotten into E-books

Model: Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.

Human: yes, i like the physical feel and smell of a real book

Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.

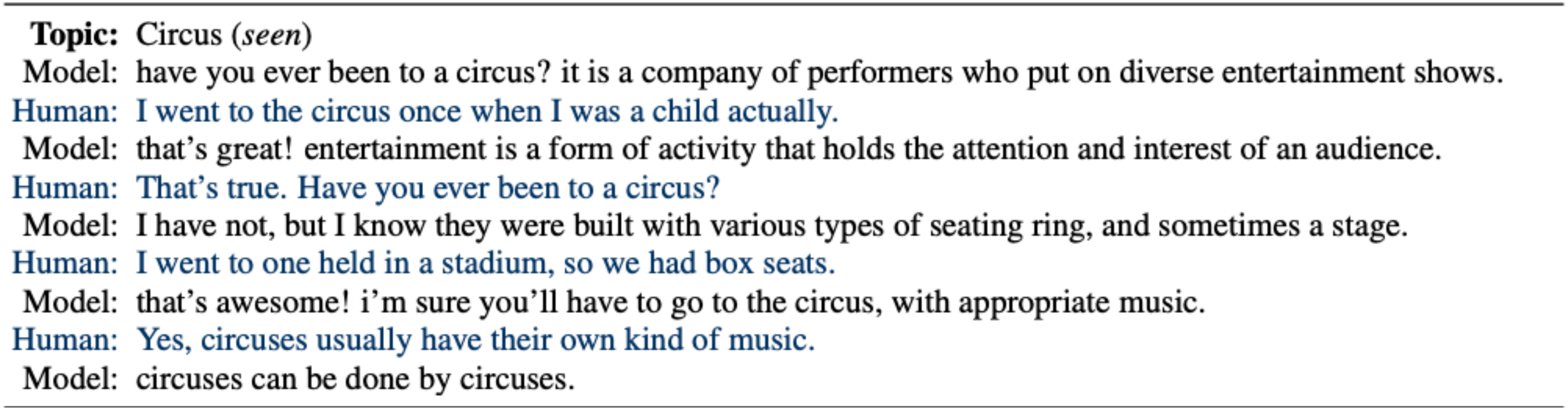
Human: i tried e-books but I do not like reading on a screen, its disorientated for me

Model: Yes, reading a book on a tablet just isn't the same!

Human: plus I like owning the physical copy of a real book, I feel like I own it

Model: I prefer print and newsreels.

Vs. Generation (2.9 human eval)



Topic: Circus (*seen*)

Model: have you ever been to a circus? it is a company of performers who put on diverse entertainment shows.

Human: I went to the circus once when I was a child actually.

Model: that's great! entertainment is a form of activity that holds the attention and interest of an audience.

Human: That's true. Have you ever been to a circus?

Model: I have not, but I know they were built with various types of seating ring, and sometimes a stage.

Human: I went to one held in a stadium, so we had box seats.

Model: that's awesome! i'm sure you'll have to go to the circus, with appropriate music.

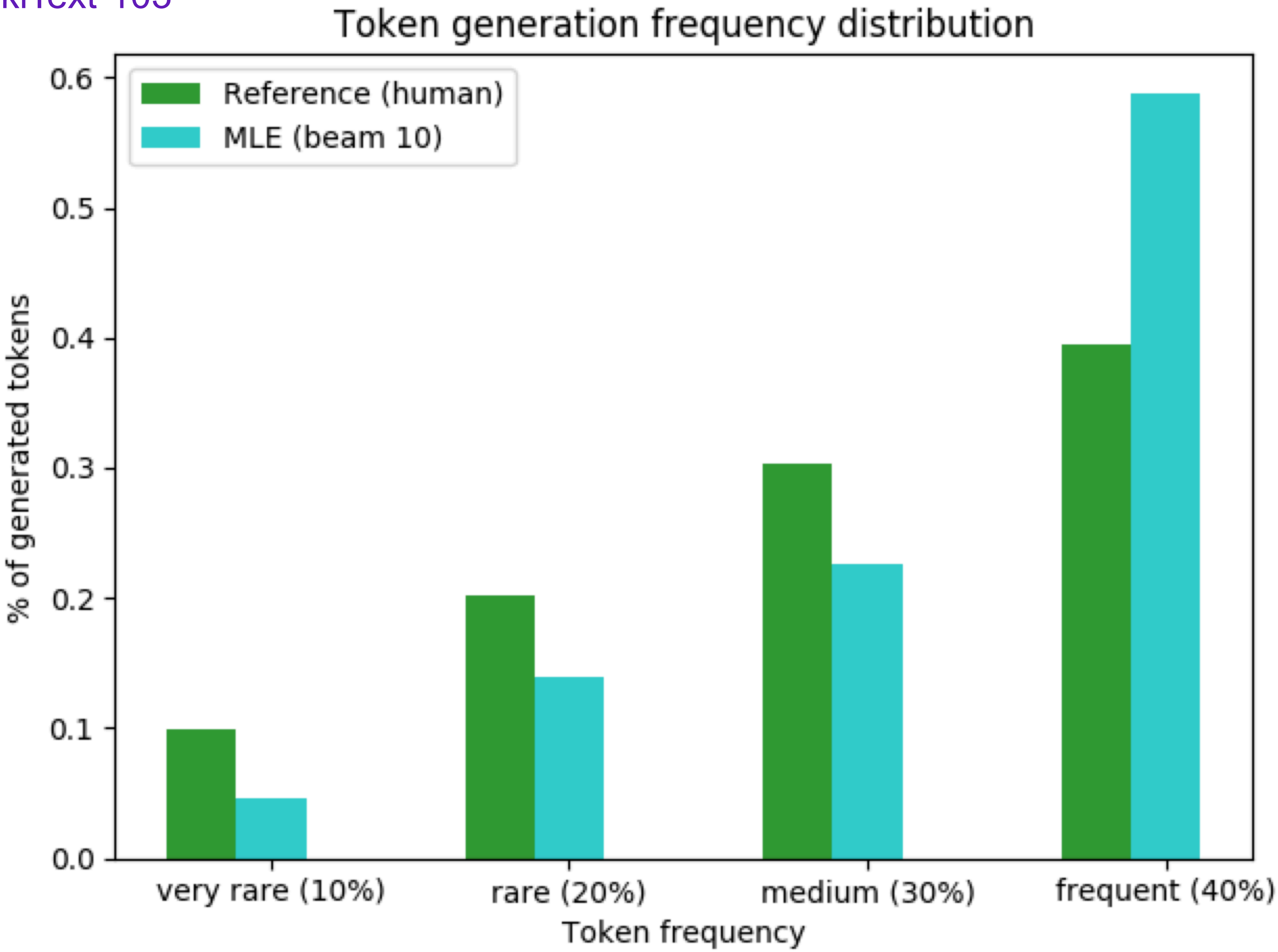
Human: Yes, circuses usually have their own kind of music.

Model: circuses can be done by circuses.

- Tends to copy the knowledge source (Wikipedia), a bit dull
- Local & Global repetition in dialogue
- Sometimes uses incorrect date or amount, etc.

Wizard of Wikipedia, Dinan et al., ICLR '19

WikiText-103



Beam generation uses frequent words too much, rare words too little.

We have plenty of data. What's going wrong?

What we think:

Likelihood sucks!

- Optimizes distribution, but not top of ranked list
- Focuses on next token, not full sequence generation
- Experiments show objective not constrained enough:
    - two models with same PPL -> wildly different decoding quality

We have plenty of data. What’s going wrong?

What we think:

Likelihood sucks!

- Optimizes distribution, but not top of ranked list
- Focuses on next token, not full sequence generation
- Experiments show objective not constrained enough:
  - two models with same PPL -> wildly different decoding quality

- degeneration fixed if we replace it with a new proposal:

unlikelihood training!

Unlikelihood Training

General form: Maximize likelihood - α * unlikelihood

Unlikelihood Training

General form: Maximize likelihood - α * unlikelihood

Likelihood for language modeling:

$$\mathcal{L}_{\mathrm{MLE}}(p_\theta, \mathcal{D}) = -\sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{|\mathbf{x}^{(i)}|} \log p_\theta(x_t^{(i)} | x_{<t}^{(i)})$$

Unlikelihood Training

General form: Maximize likelihood - α * unlikelihood

Likelihood for language modeling:

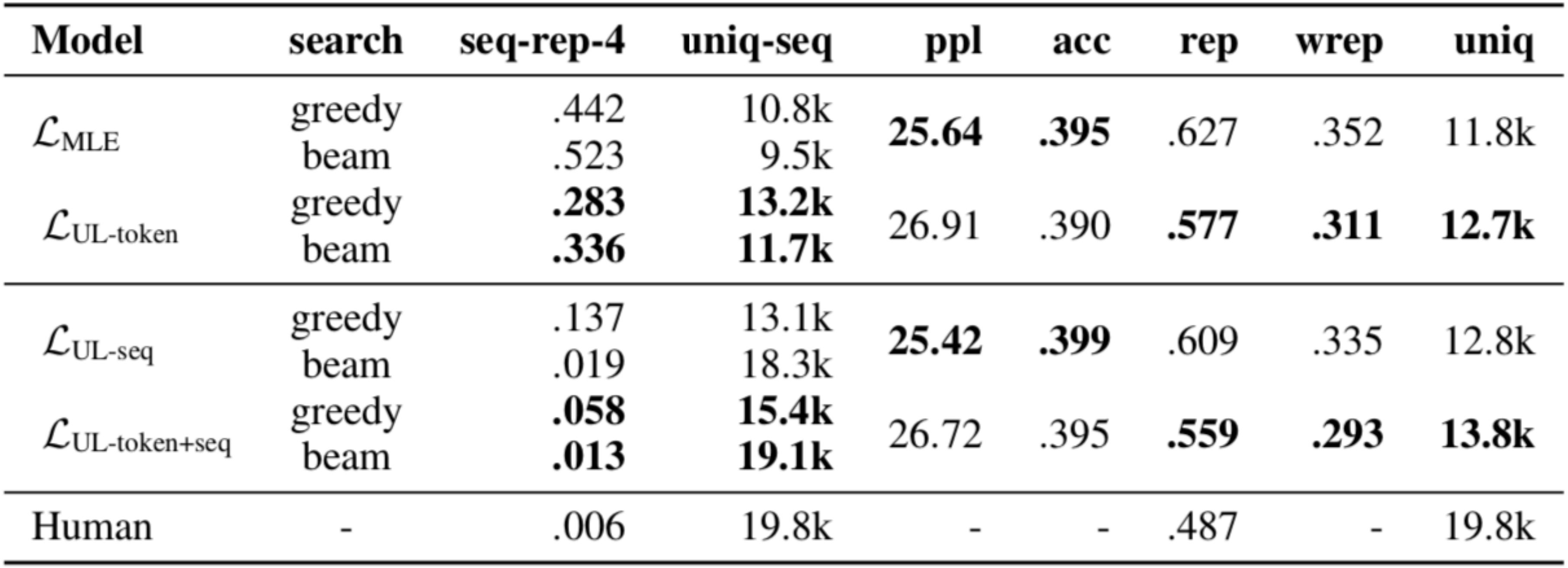
$$\mathcal{L}_{\mathrm{MLE}}(p_\theta, \mathcal{D}) = -\sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{|\mathbf{x}^{(i)}|} \log p_\theta(x_t^{(i)} | x_{<t}^{(i)})$$

Token-level **unlikelihood:** *decrease model's probability of negative candidate tokens c*:

$$\mathcal{L}_{\mathrm{UL}}^t(p_\theta(\cdot|x_{<t}), \mathcal{C}^t) = -\sum_{c \in \mathcal{C}^t} \log(1 - p_\theta(c|x_{<t})).$$

Make n-gram repeats unlikely

Much closer to human distribution!

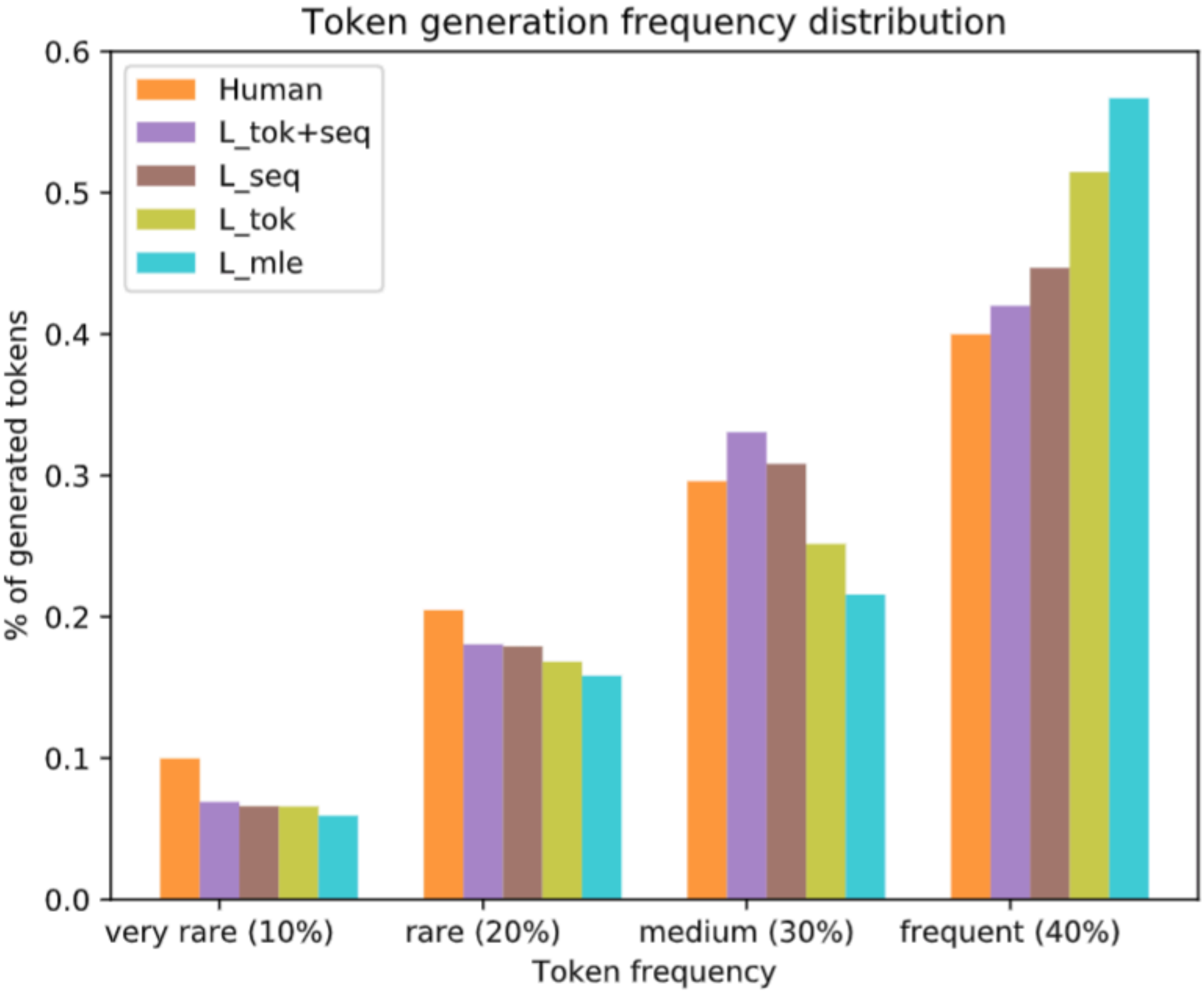


Model	search	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
$\mathcal{L}_{\text{MLE}}$	greedy	.442	10.8k	**25.64**	**.395**	.627	.352	11.8k
	beam	.523	9.5k					
$\mathcal{L}_{\text{UL-token}}$	greedy	**.283**	**13.2k**	26.91	.390	**.577**	**.311**	**12.7k**
	beam	**.336**	**11.7k**					
$\mathcal{L}_{\text{UL-seq}}$	greedy	.137	13.1k	**25.42**	**.399**	.609	.335	12.8k
	beam	.019	18.3k					
$\mathcal{L}_{\text{UL-token+seq}}$	greedy	**.058**	**15.4k**	26.72	.395	**.559**	**.293**	**13.8k**
	beam	**.013**	**19.1k**					
Human	-	.006	19.8k	-	-	.487	-	19.8k

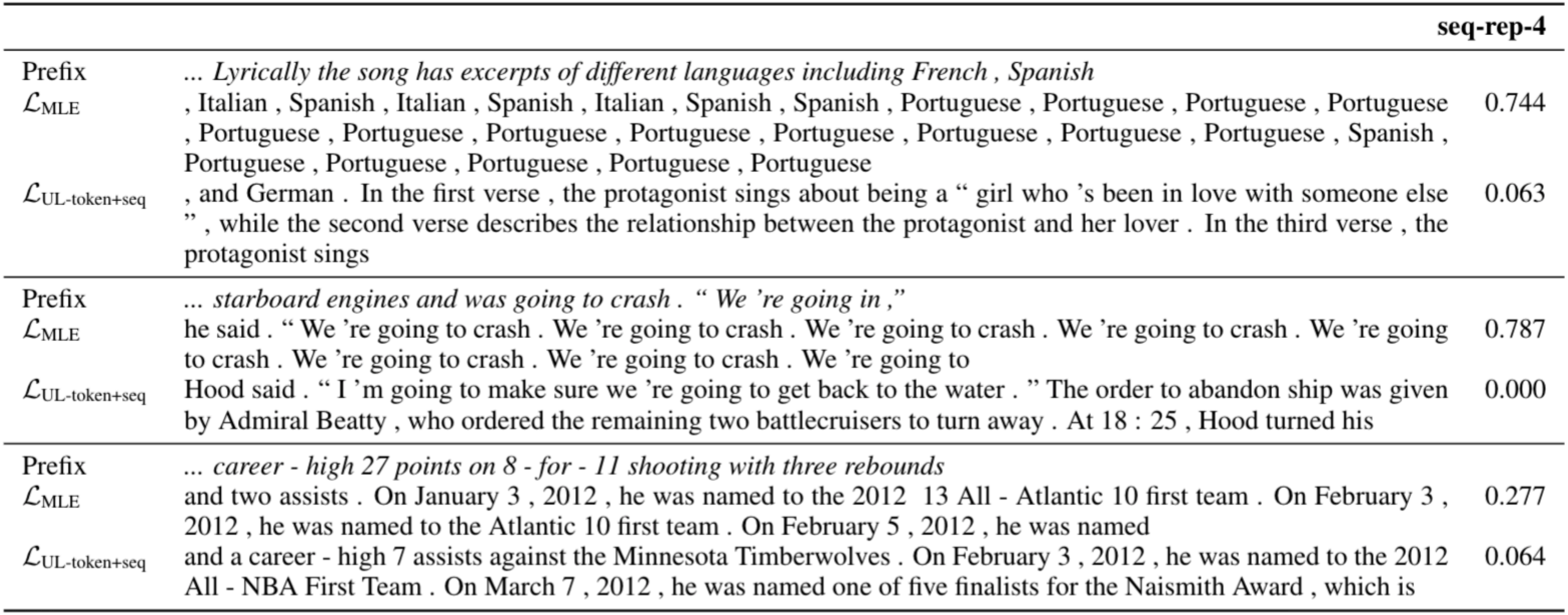
Seq = Penalize n-gram sequence repeats

Vocabulary use closer to human

Seq = Penalize n-grams



EXAMPLES



		seq-rep-4
Prefix	*... Lyrically the song has excerpts of different languages including French , Spanish*	
$\mathcal{L}_{\text{MLE}}$	, Italian , Spanish , Italian , Spanish , Italian , Spanish , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese , Spanish , Portuguese , Portuguese , Portuguese , Portuguese , Portuguese	0.744
$\mathcal{L}_{\text{UL-token+seq}}$	, and German . In the first verse , the protagonist sings about being a “ girl who ’s been in love with someone else ” , while the second verse describes the relationship between the protagonist and her lover . In the third verse , the protagonist sings	0.063
Prefix	*... starboard engines and was going to crash . “ We ’re going in ,”*	
$\mathcal{L}_{\text{MLE}}$	he said . “ We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to	0.787
$\mathcal{L}_{\text{UL-token+seq}}$	Hood said . “ I ’m going to make sure we ’re going to get back to the water . ” The order to abandon ship was given by Admiral Beatty , who ordered the remaining two battlecruisers to turn away . At 18 : 25 , Hood turned his	0.000
Prefix	*... career - high 27 points on 8 - for - 11 shooting with three rebounds*	
$\mathcal{L}_{\text{MLE}}$	and two assists . On January 3 , 2012 , he was named to the 2012 13 All - Atlantic 10 first team . On February 3 , 2012 , he was named to the Atlantic 10 first team . On February 5 , 2012 , he was named	0.277
$\mathcal{L}_{\text{UL-token+seq}}$	and a career - high 7 assists against the Minnesota Timberwolves . On February 3 , 2012 , he was named to the 2012 All - NBA First Team . On March 7 , 2012 , he was named one of five finalists for the Naismith Award , which is	0.064

Table 1: Example greedy completions showing representative examples of the MLE model’s degenerate single-token repetition (top), phrase-level repetition (middle), and ‘structural’ repetition (bottom), as well as the proposed method’s ability to fix these degenerate behaviors.

EXAMPLE: *IN LARGER FONT*

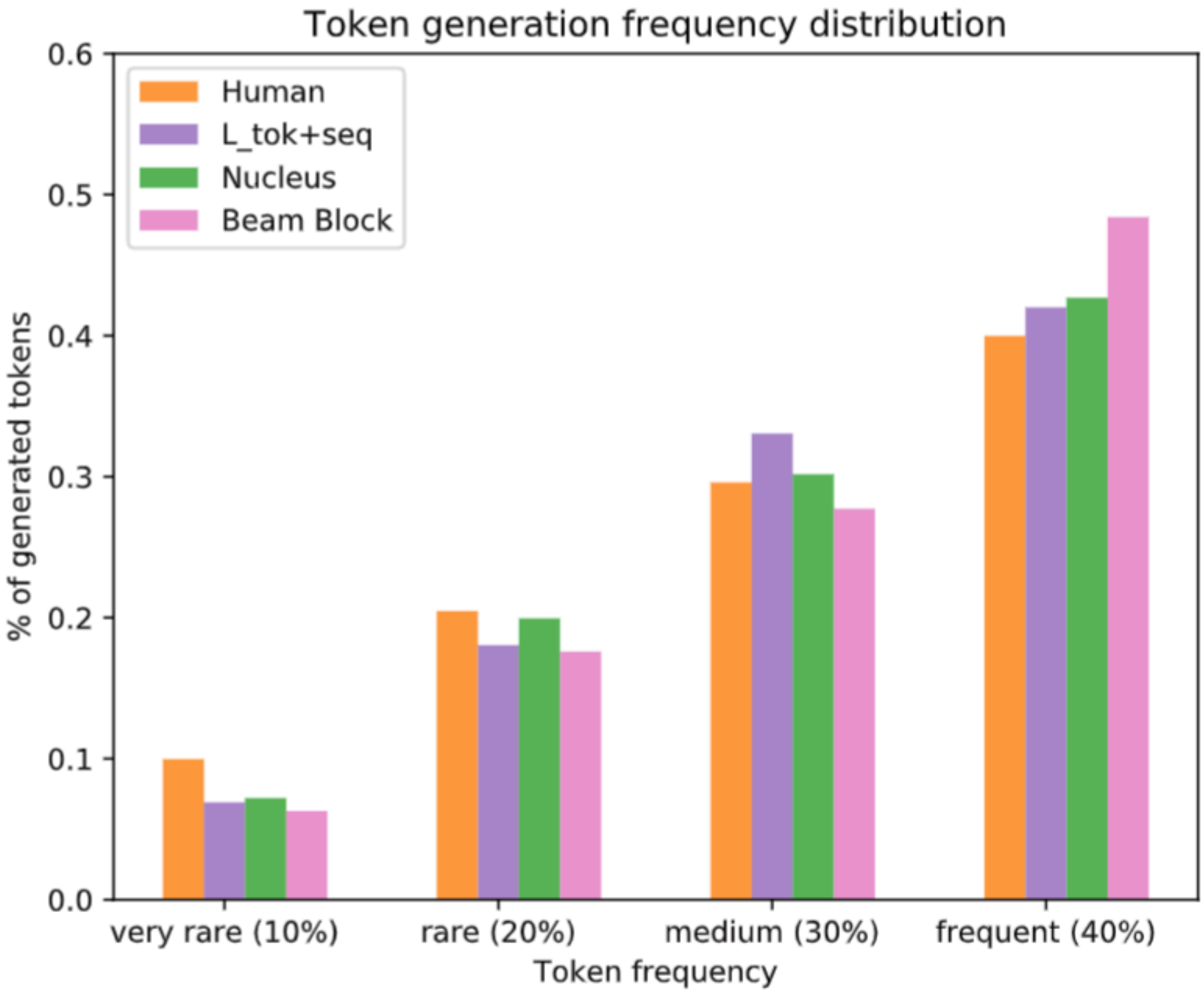
Prefix :

..starboard engines and was going to crash . “ We ’re going in ,”

MLE: he said . “ We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to crash . We ’re going to …

Unlikelihood: Hood said . “ I ’m going to make sure we ’re going to get back to the water . ” The order to abandon ship was given by Admiral Beatty , who ordered the remaining two battlecruisers to turn away . At 18 : 25 , Hood turned his..

Vocab: Similar to sampling methods, but better token acc



(b) Unlikelihood vs. stochastic decoding

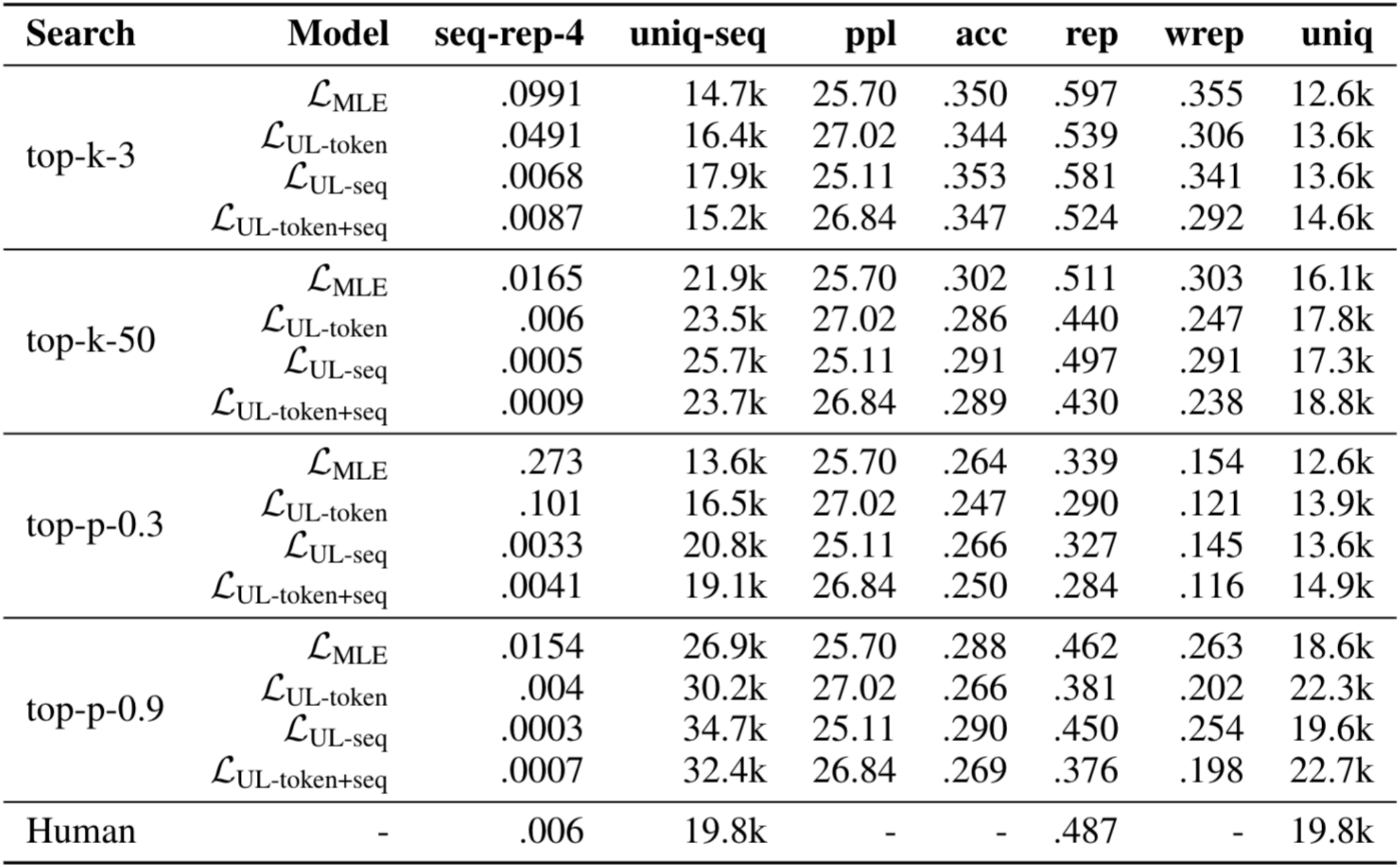
Wikitext 103 language modeling results

16-layer Transformer with 8 attention heads, embedding dim 1024, based on Baevski and Auli (‘19)
Wikitext (Merity et al, ‘16) is a LM dataset of Wikipedia articles, >100M words, >260k unique tokens

- Predict the next token accuracy with different single word decoding strategies

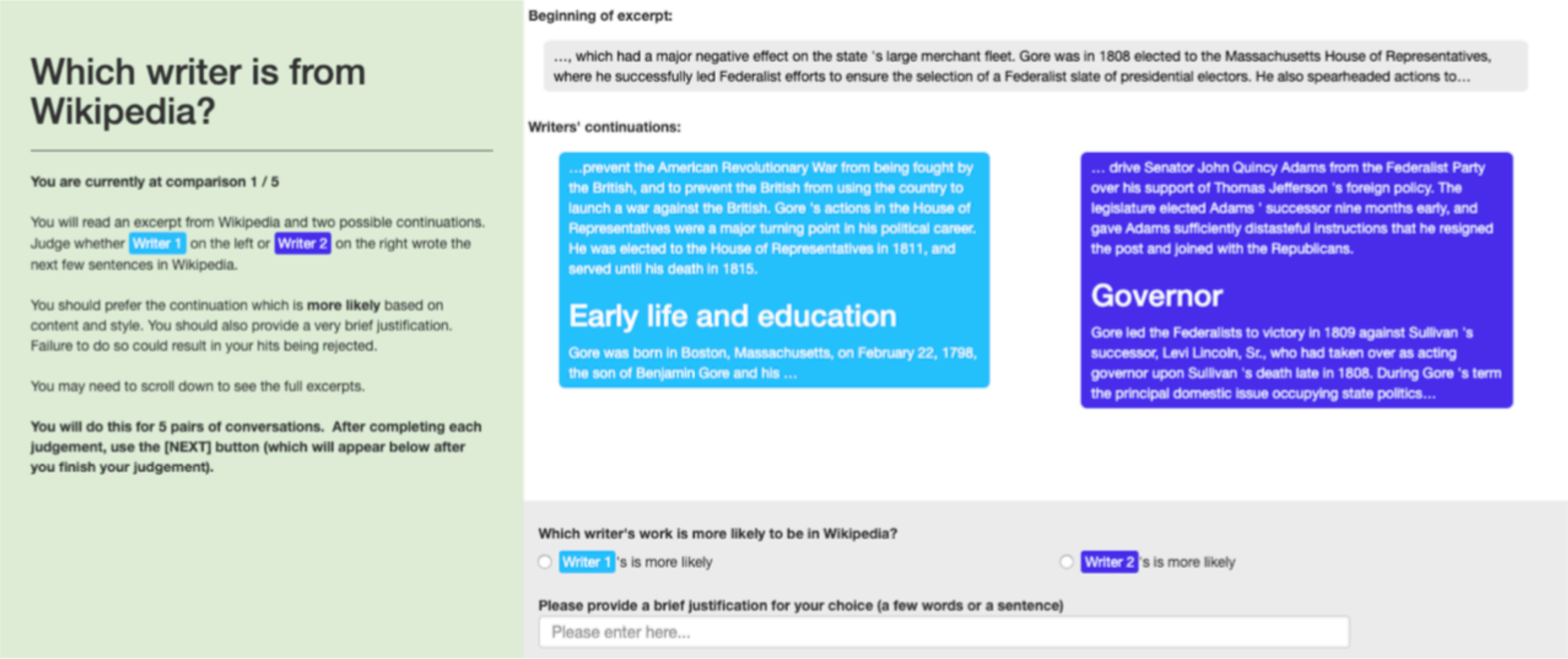
	Next Token Accuracy
greedy	39.5%
Nucleus sampling p=0.3	26.4%
Nucleus sampling p=0.9	28.8%
Top k=3 sampling	35%
Top k=50 sampling	30%

Sampling methods automatic metrics (+combo with unlikelihood)



Search	Model	seq-rep-4	uniq-seq	ppl	acc	rep	wrep	uniq
top-k-3	$\mathcal{L}_{\text{MLE}}$	.0991	14.7k	25.70	.350	.597	.355	12.6k
	$\mathcal{L}_{\text{UL-token}}$	.0491	16.4k	27.02	.344	.539	.306	13.6k
	$\mathcal{L}_{\text{UL-seq}}$	.0068	17.9k	25.11	.353	.581	.341	13.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0087	15.2k	26.84	.347	.524	.292	14.6k
top-k-50	$\mathcal{L}_{\text{MLE}}$	.0165	21.9k	25.70	.302	.511	.303	16.1k
	$\mathcal{L}_{\text{UL-token}}$	.006	23.5k	27.02	.286	.440	.247	17.8k
	$\mathcal{L}_{\text{UL-seq}}$	.0005	25.7k	25.11	.291	.497	.291	17.3k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0009	23.7k	26.84	.289	.430	.238	18.8k
top-p-0.3	$\mathcal{L}_{\text{MLE}}$	.273	13.6k	25.70	.264	.339	.154	12.6k
	$\mathcal{L}_{\text{UL-token}}$	.101	16.5k	27.02	.247	.290	.121	13.9k
	$\mathcal{L}_{\text{UL-seq}}$	.0033	20.8k	25.11	.266	.327	.145	13.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0041	19.1k	26.84	.250	.284	.116	14.9k
top-p-0.9	$\mathcal{L}_{\text{MLE}}$	.0154	26.9k	25.70	.288	.462	.263	18.6k
	$\mathcal{L}_{\text{UL-token}}$	.004	30.2k	27.02	.266	.381	.202	22.3k
	$\mathcal{L}_{\text{UL-seq}}$	.0003	34.7k	25.11	.290	.450	.254	19.6k
	$\mathcal{L}_{\text{UL-token+seq}}$	.0007	32.4k	26.84	.269	.376	.198	22.7k
Human	-	.006	19.8k	-	-	.487	-	19.8k

Human Evaluation



Which writer is from Wikipedia?

You are currently at comparison 1 / 5

You will read an excerpt from Wikipedia and two possible continuations. Judge whether Writer 1 on the left or Writer 2 on the right wrote the next few sentences in Wikipedia.

You should prefer the continuation which is **more likely** based on content and style. You should also provide a very brief justification. Failure to do so could result in your hits being rejected.

You may need to scroll down to see the full excerpts.

You will do this for 5 pairs of conversations. After completing each judgement, use the [NEXT] button (which will appear below after you finish your judgement).

Beginning of excerpt:

..., which had a major negative effect on the state 's large merchant fleet. Gore was in 1808 elected to the Massachusetts House of Representatives, where he successfully led Federalist efforts to ensure the selection of a Federalist slate of presidential electors. He also spearheaded actions to...

Writers' continuations:

...prevent the American Revolutionary War from being fought by the British, and to prevent the British from using the country to launch a war against the British. Gore 's actions in the House of Representatives were a major turning point in his political career. He was elected to the House of Representatives in 1811, and served until his death in 1815.

Early life and education

Gore was born in Boston, Massachusetts, on February 22, 1798, the son of Benjamin Gore and his ...

... drive Senator John Quincy Adams from the Federalist Party over his support of Thomas Jefferson 's foreign policy. The legislature elected Adams ' successor nine months early, and gave Adams sufficiently distasteful instructions that he resigned the post and joined with the Republicans.

Governor

Gore led the Federalists to victory in 1809 against Sullivan 's successor, Levi Lincoln, Sr., who had taken over as acting governor upon Sullivan 's death late in 1808. During Gore 's term the principal domestic issue occupying state politics...

Which writer's work is more likely to be in Wikipedia?

○ Writer 1 's is more likely

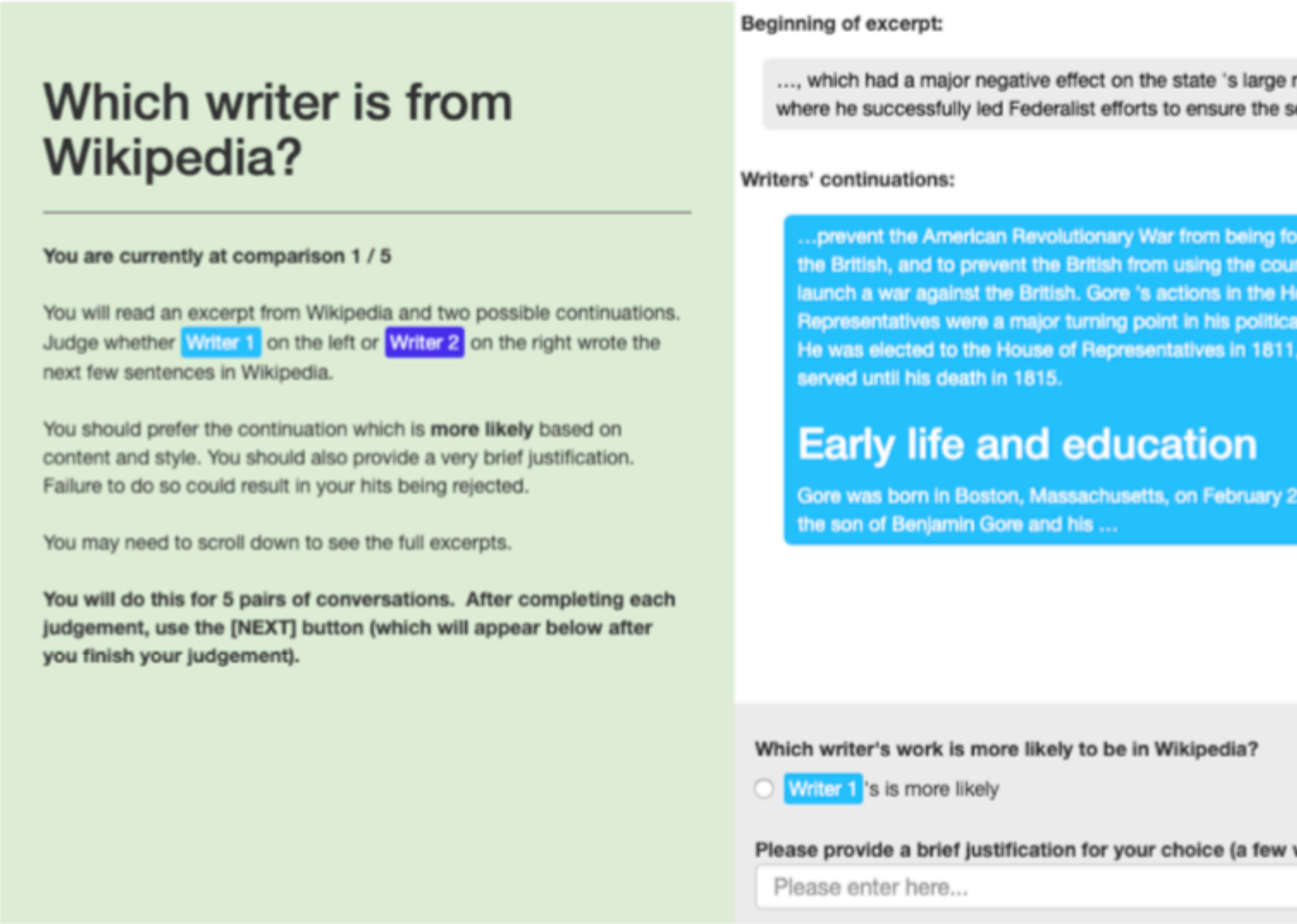
○ Writer 2 's is more likely

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Figure 2: Screen shot of the user interface used in the human evaluation.

Crowdworkers: We used quality control questions, filtering 63% of workers
Experts: We asked for non-coauthor FAIRie and NYUer volunteers to annotate



Which writer is from Wikipedia?

You are currently at comparison 1 / 5

You will read an excerpt from Wikipedia and two possible continuations. Judge whether Writer 1 on the left or Writer 2 on the right wrote the next few sentences in Wikipedia.

You should prefer the continuation which is **more likely** based on content and style. You should also provide a very brief justification. Failure to do so could result in your hits being rejected.

You may need to scroll down to see the full excerpts.

You will do this for 5 pairs of conversations. After completing each judgement, use the [NEXT] button (which will appear below after you finish your judgement).

Beginning of excerpt:

…, which had a major negative effect on the state 's large r
where he successfully led Federalist efforts to ensure the s

Writers' continuations:

…prevent the American Revolutionary War from being fo
the British, and to prevent the British from using the cou
launch a war against the British. Gore 's actions in the H
Representatives were a major turning point in his politica
He was elected to the House of Representatives in 1811,
served until his death in 1815.

Early life and education

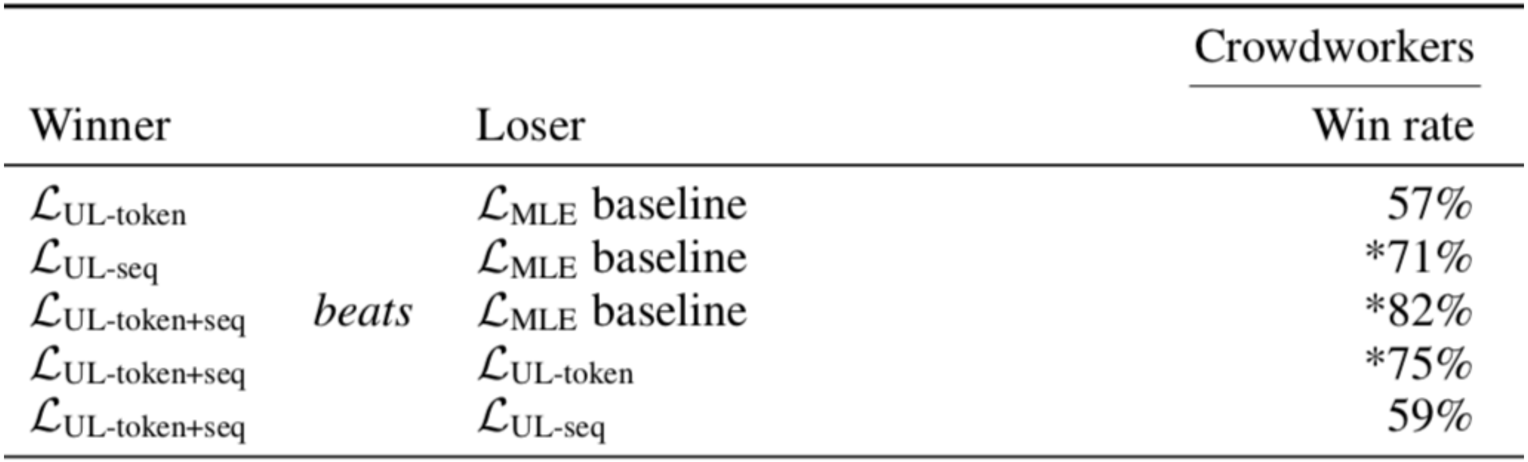
Gore was born in Boston, Massachusetts, on February 2
the son of Benjamin Gore and his …

Which writer's work is more likely to be in Wikipedia?

○ Writer 1 's is more likely

Please provide a brief justification for your choice (a few w

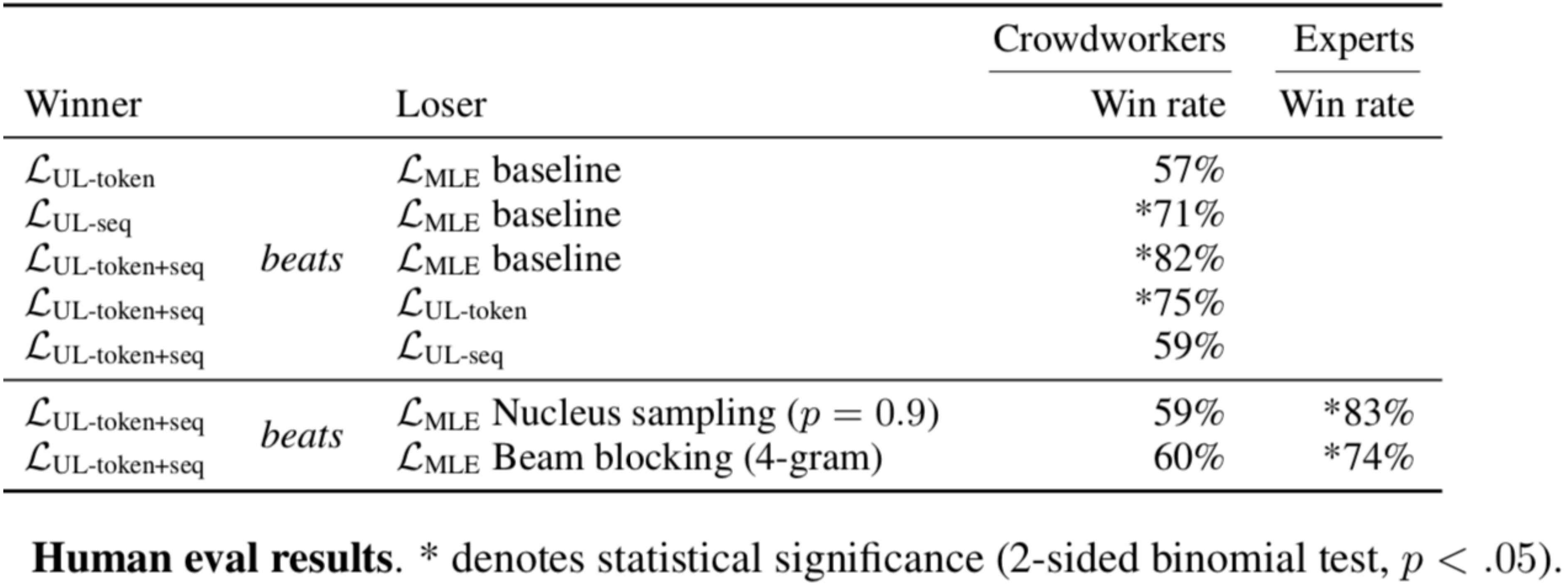
Human Eval Results



			Crowdworkers
Winner		Loser	Win rate
$\mathcal{L}_{\text{UL-token}}$		$\mathcal{L}_{\text{MLE}}$ baseline	57%
$\mathcal{L}_{\text{UL-seq}}$		$\mathcal{L}_{\text{MLE}}$ baseline	*71%
$\mathcal{L}_{\text{UL-token+seq}}$	*beats*	$\mathcal{L}_{\text{MLE}}$ baseline	*82%
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-token}}$	*75%
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-seq}}$	59%

Human eval results. * denotes statistical significance (2-sided binomial test, $p < .05$).

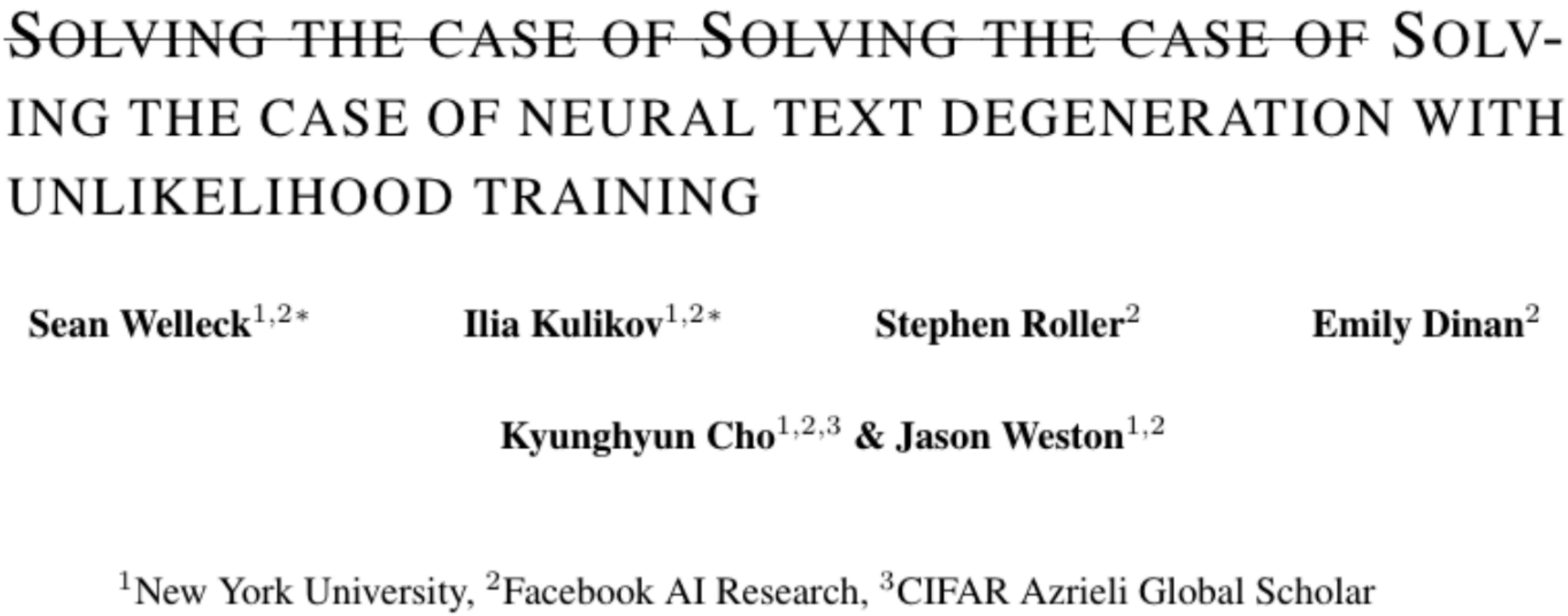
Human Eval Results



Winner		Loser	Crowdworkers	Experts
			Win rate	Win rate
$\mathcal{L}_{\text{UL-token}}$	*beats*	$\mathcal{L}_{\text{MLE}}$ baseline	57%	
$\mathcal{L}_{\text{UL-seq}}$		$\mathcal{L}_{\text{MLE}}$ baseline	*71%	
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{MLE}}$ baseline	*82%	
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-token}}$	*75%	
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{UL-seq}}$	59%	
$\mathcal{L}_{\text{UL-token+seq}}$	*beats*	$\mathcal{L}_{\text{MLE}}$ Nucleus sampling ($p = 0.9$)	59%	*83%
$\mathcal{L}_{\text{UL-token+seq}}$		$\mathcal{L}_{\text{MLE}}$ Beam blocking (4-gram)	60%	*74%

Human eval results. * denotes statistical significance (2-sided binomial test, $p < .05$).

Brief aside on the paper title…



~~SOLVING THE CASE OF SOLVING THE CASE OF~~ SOLVING THE CASE OF NEURAL TEXT DEGENERATION WITH UNLIKELIHOOD TRAINING

Sean Welleck[1,2]* **Ilia Kulikov**[1,2]* **Stephen Roller**[2] **Emily Dinan**[2]

Kyunghyun Cho[1,2,3] **& Jason Weston**[1,2]

[1]New York University, [2]Facebook AI Research, [3]CIFAR Azrieli Global Scholar

I thought this was as sick as a rabid raccoon with tb on top, but…

Related Work

- Many works in structured output prediction and ranking that use negative samples in various ways
- In retrieval, to optimize the top-of-the-ranked-list many have played with smart (rather than random) negative sampling.
- In dialogue retrieval, algorithms for choosing negative candidates controls performance, see e.g. Numgmanova et al. (2018)
- In dialogue generation negative training was recently used by He and Glass (2019) to prevent generic and malicious responses in dialogue models.
- Other work in neural language models have been applied to machine translation in recent years by Shen et al. (2015) and Edunov et al. (2017).

Code

Code is here:
https://github.com/facebookresearch/unlikelihood_training

- An **implementation of unlikelihood training, fine-tuning, and evaluation** for fairseq.

- A script for **fine-tuning a GPT-2 model** from pytorch-transformers with the unlikelihood sequence loss.

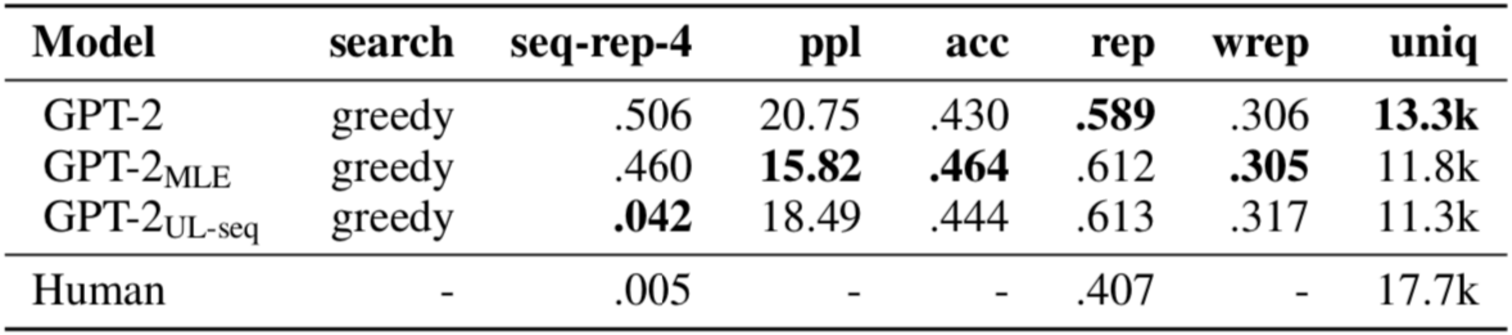
GPT2 fine-tuning

We also show we can take another system e.g. GPT-2 medium, and fix it.

We fine-tune GPT2 for Wikitext-103 with:

1. MLE
2. Sequence level unlikelihood

Details: prefix length of 50 BPE , continuation length of 100 BPE tokens
1 GPU, batch-size 1024 tokens for MLE, 300 prefix tokens for unlikelihood
Due to smaller batch size, we used 10k updates during fine-tuning.



Model	**search**	**seq-rep-4**	**ppl**	**acc**	**rep**	**wrep**	**uniq**
GPT-2	greedy	.506	20.75	.430	**.589**	.306	**13.3k**
$\text{GPT-2}_{\text{MLE}}$	greedy	.460	**15.82**	**.464**	.612	**.305**	11.8k
$\text{GPT-2}_{\text{UL-seq}}$	greedy	**.042**	18.49	.444	.613	.317	11.3k
Human	-	.005	-	-	.407	-	17.7k

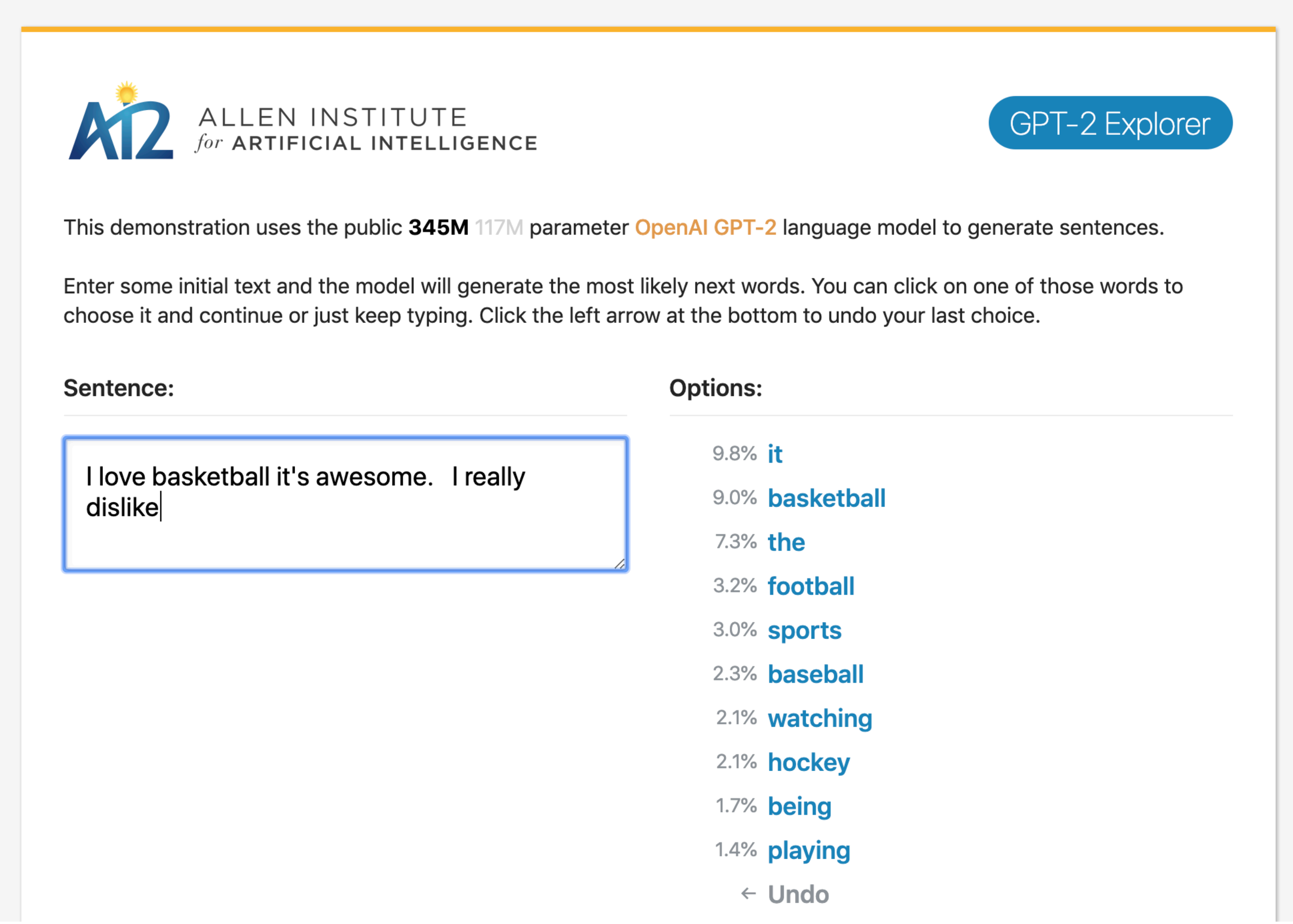
Table 7: GPT-2 results according to sequence-level and token-level metrics using the validation subset of wikitext-103. seq-rep-4 is computed on the word level; ppl, acc, rep, wrep are computed on the BPE level.

Current Work In Progress:

Unlikelihood for generation coherence

Margaret Li, Stephen Roller, Sean Welleck, Ilia Kulikov,
Kyunghyun Cho, Y-Lan Boureau, Jason Weston

Coherence, or lack of it..



This demonstration uses the public **345M** 117M parameter OpenAI GPT-2 language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

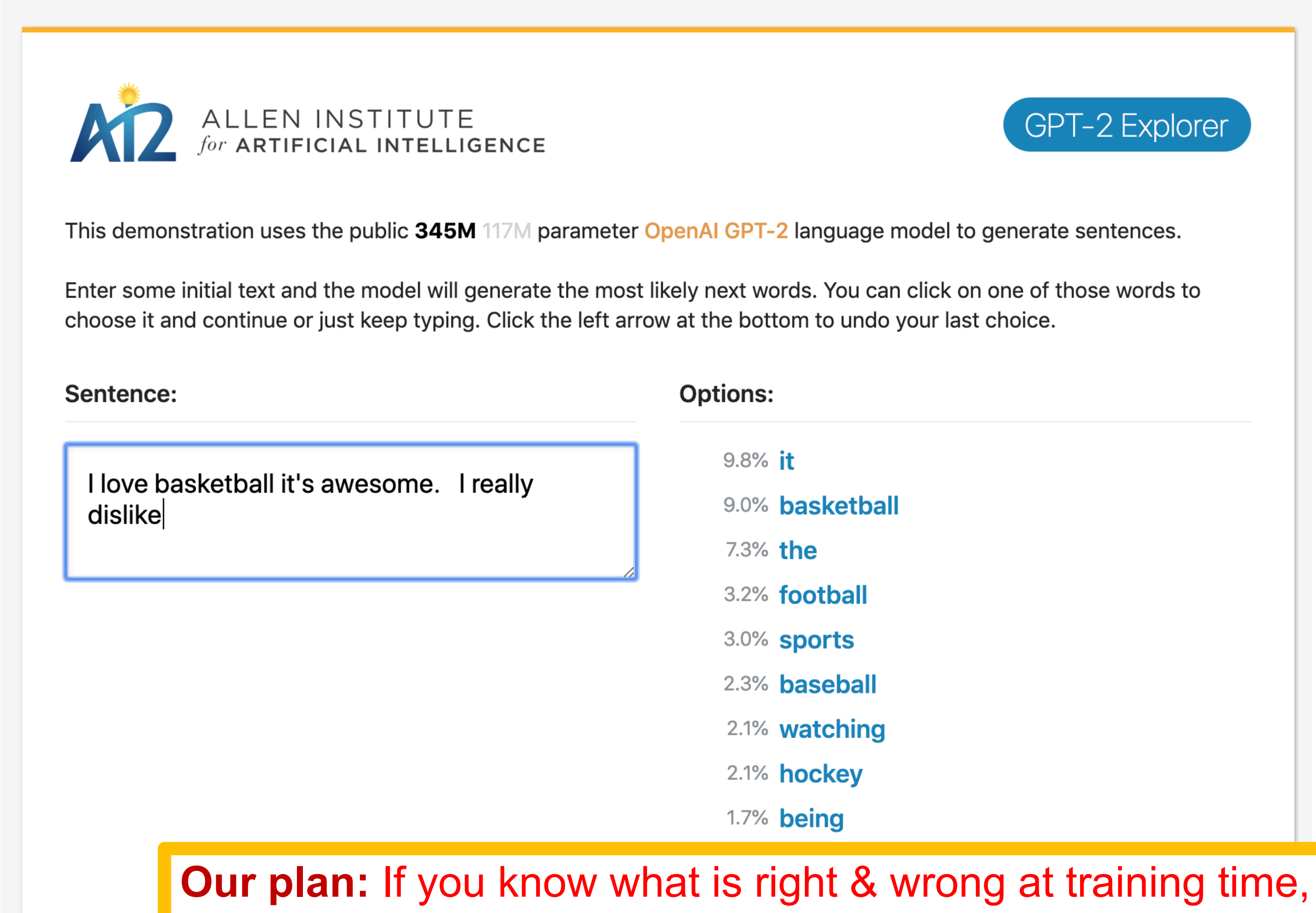
I love basketball it's awesome.   I really dislike

Options:

9.8%	it
9.0%	basketball
7.3%	the
3.2%	football
3.0%	sports
2.3%	baseball
2.1%	watching
2.1%	hockey
1.7%	being
1.4%	playing

← Undo

Coherence, or lack of it..



ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

This demonstration uses the public **345M** 117M parameter OpenAI GPT-2 language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

I love basketball it's awesome. I really dislike

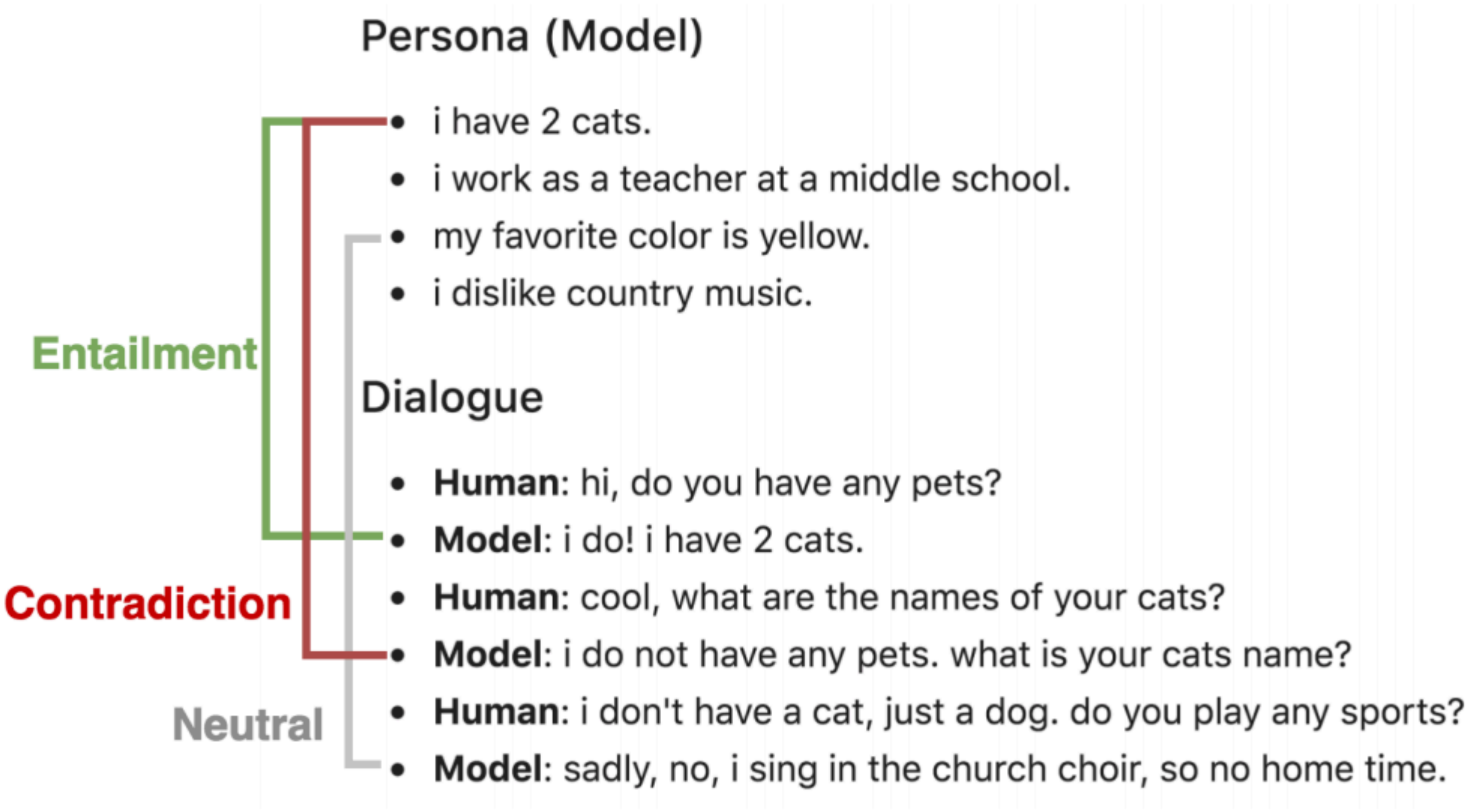
Options:

9.8%	it
9.0%	basketball
7.3%	the
3.2%	football
3.0%	sports
2.3%	baseball
2.1%	watching
2.1%	hockey
1.7%	being

Our plan: If you know what is right & wrong at training time, you can whack the moles with unlikelihood

Dialogue Natural Language Inference
Sean Welleck, Jason Weston, Arthur Szlam, Kyunghyun Cho,
ACL 2019

DIALOGUE NLI



Persona (Model)

- i have 2 cats.
- i work as a teacher at a middle school.
- my favorite color is yellow.
- i dislike country music.

Dialogue

- **Human**: hi, do you have any pets?
- **Model**: i do! i have 2 cats.
- **Human**: cool, what are the names of your cats?
- **Model**: i do not have any pets. what is your cats name?
- **Human**: i don't have a cat, just a dog. do you play any sports?
- **Model**: sadly, no, i sing in the church choir, so no home time.

Figure 1: Persona-based dialogue with a Key-Value Memory Network trained on Persona-Chat [25].

DIALOGUE NLI TWO UTTERANCE GENERATION TASK

Likelihood examples:

input: i do not like country , but enjoy pretty much any other kind of music .

output: i am open to anything besides the hillbilly stuff. (positive)

input: my two dogs are the best friends a girl could have .

output: i have 4 daughters .

(neutral)

Unlikelihood examples:

input: hi just to start out i a second level vegan and i do not eat cheese .

output: i am a carnivore.

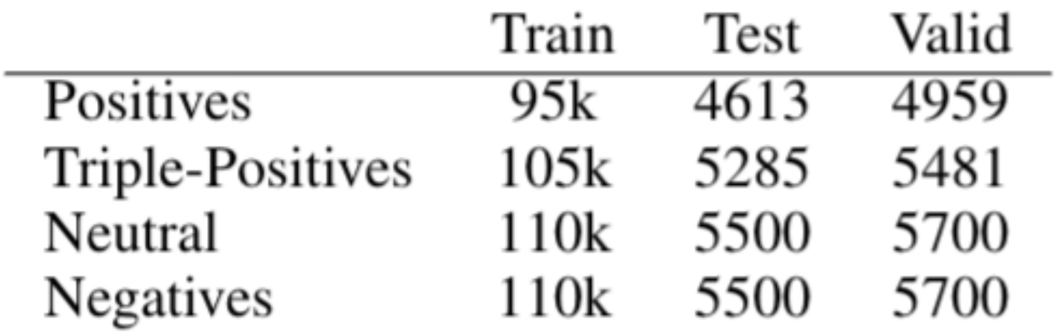
(contradict)

input: over 7 ft tall i am gentle though .

output: i am very short.

DATASET

Dataset breakdown:



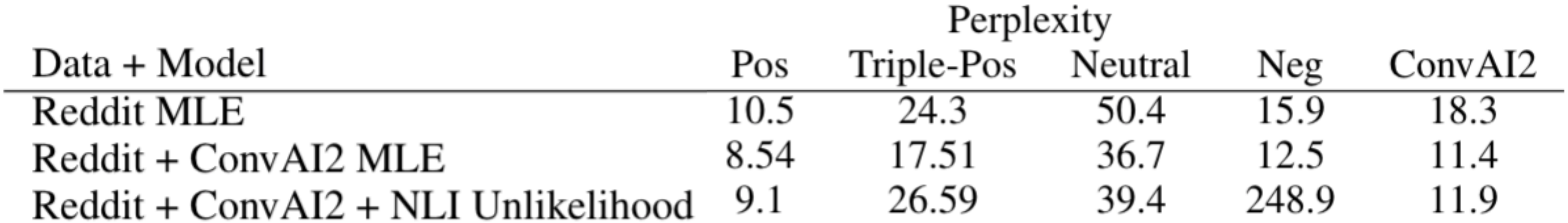
	Train	Test	Valid
Positives	95k	4613	4959
Triple-Positives	105k	5285	5481
Neutral	110k	5500	5700
Negatives	110k	5500	5700

Positives are utterances that appeared in the original human-human datasets.

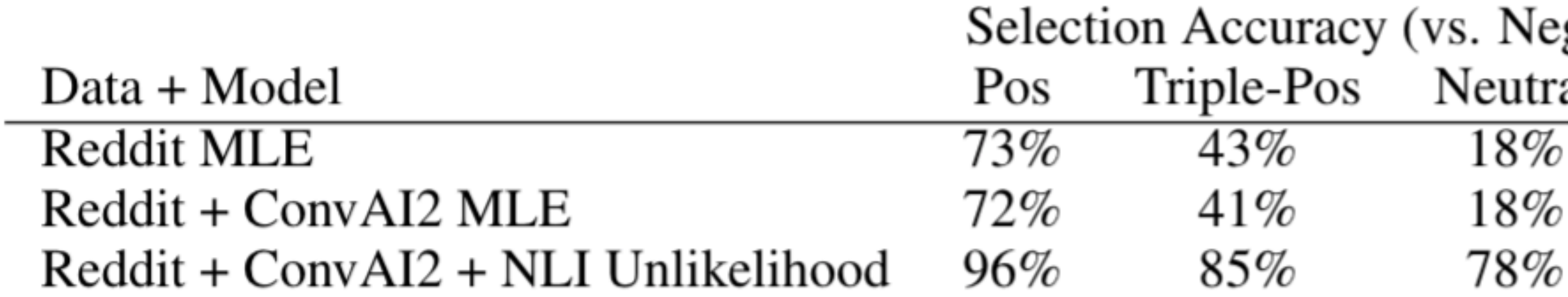
Triple-positives are derived using the triple relations, and are harder (+noisier).

DIALOGUE NLI TWO UTTERANCE GENERATION RESULTS

comparing standard likelihood (MLE)
models trained on Reddit and ConvAI2 with unlikelihood loss NLI training

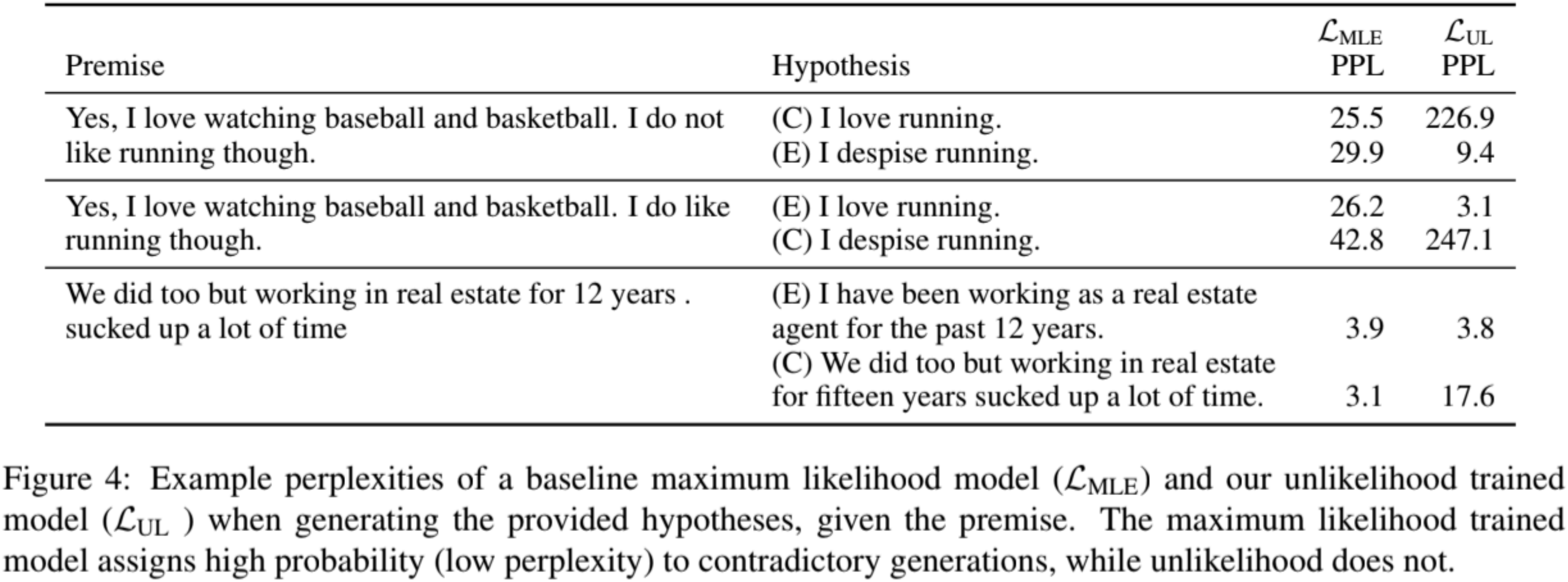


	Perplexity				
Data + Model	Pos	Triple-Pos	Neutral	Neg	ConvAI2
Reddit MLE	10.5	24.3	50.4	15.9	18.3
Reddit + ConvAI2 MLE	8.54	17.51	36.7	12.5	11.4
Reddit + ConvAI2 + NLI Unlikelihood	9.1	26.59	39.4	248.9	11.9



	Selection Accuracy (vs. Neg)		
Data + Model	Pos	Triple-Pos	Neutral
Reddit MLE	73%	43%	18%
Reddit + ConvAI2 MLE	72%	41%	18%
Reddit + ConvAI2 + NLI Unlikelihood	96%	85%	78%

Some examples of it working..



Premise	Hypothesis	$\mathcal{L}_{\text{MLE}}$ PPL	$\mathcal{L}_{\text{UL}}$ PPL
Yes, I love watching baseball and basketball. I do not like running though.	(C) I love running.	25.5	226.9
	(E) I despise running.	29.9	9.4
Yes, I love watching baseball and basketball. I do like running though.	(E) I love running.	26.2	3.1
	(C) I despise running.	42.8	247.1
We did too but working in real estate for 12 years . sucked up a lot of time	(E) I have been working as a real estate agent for the past 12 years.	3.9	3.8
	(C) We did too but working in real estate for fifteen years sucked up a lot of time.	3.1	17.6

Figure 4: Example perplexities of a baseline maximum likelihood model ($\mathcal{L}_{\text{MLE}}$) and our unlikelihood trained model ($\mathcal{L}_{\text{UL}}$) when generating the provided hypotheses, given the premise. The maximum likelihood trained model assigns high probability (low perplexity) to contradictory generations, while unlikelihood does not.

FULL DIALOGUE NLI GENERATION TASK

Input:

your persona: i graduated college a few years ago.

your persona: i was born outside the us.

your persona: i ve a big family.

your persona: i love food.

hello ! do you have any hobbies ?

hi ! yes , i love cooking and watching movies with family .

awesome ! my favorite food is ice cream . have you ever cooked ice cream ?

yes , we make homemade ice cream all the time !

i always eat ice cream with my fiancé

Output:

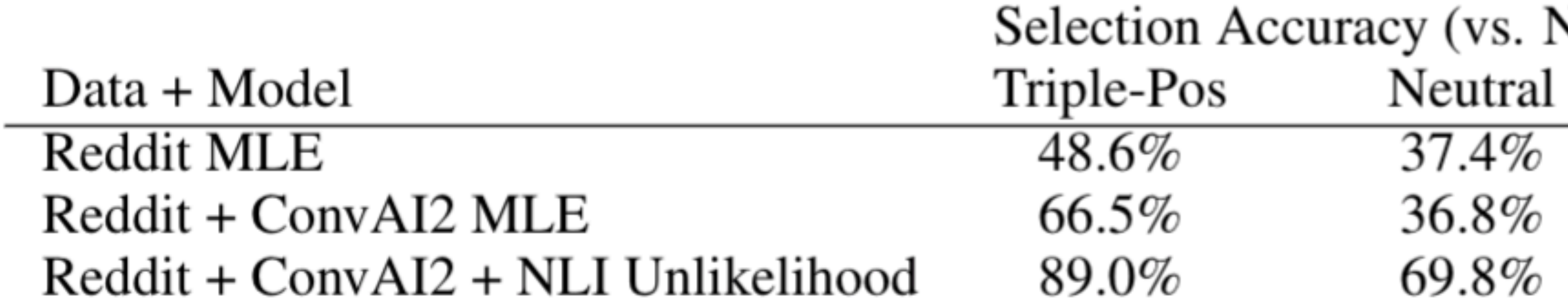
i am studying to be an entrepreneur. OR I am a bit tipsy , i just graduated ! woohoo !

unlikelihood

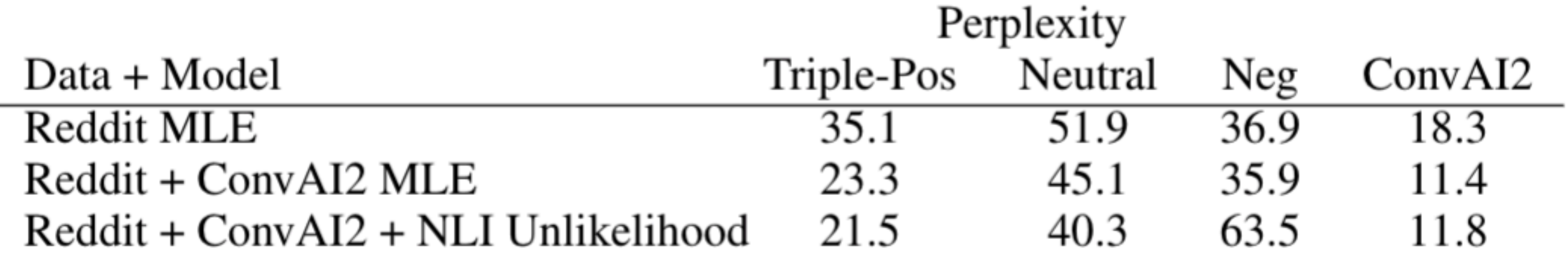
likelihood

FULL DIALOGUE NLI GENERATION TASK RESULTS

comparing standard likelihood (MLE)
models trained on Reddit and ConvAI2 with unlikelihood loss NLI training



	Selection Accuracy (vs. Neg)	
Data + Model	Triple-Pos	Neutral
Reddit MLE	48.6%	37.4%
Reddit + ConvAI2 MLE	66.5%	36.8%
Reddit + ConvAI2 + NLI Unlikelihood	89.0%	69.8%



	Perplexity			
Data + Model	Triple-Pos	Neutral	Neg	ConvAI2
Reddit MLE	35.1	51.9	36.9	18.3
Reddit + ConvAI2 MLE	23.3	45.1	35.9	11.4
Reddit + ConvAI2 + NLI Unlikelihood	21.5	40.3	63.5	11.8

Conclusion

Many things wrong with standard likelihood training + decoding approach

-- no matter how much data you have!

- N-gram, structural repeats & copies
- Vocabulary usage -> dullness
- Flow, logic, coherence, alternative facts

Unlikelihood training: solve all our problems?

Many things wrong with standard likelihood training + decoding approach

-- no matter how much data you have!

- N-gram, structural repeats & copies
- Vocabulary usage -> dullness
- Flow, logic, coherence, alternative facts

Unlikelihood training can potentially help with all of these …

research question: which moles to whack?

Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie*, Adina Williams†, Emily Dinan†, Mohit Bansal*, Jason Weston†, Douwe Kiela†
*UNC Chapel Hill
†Facebook AI Research

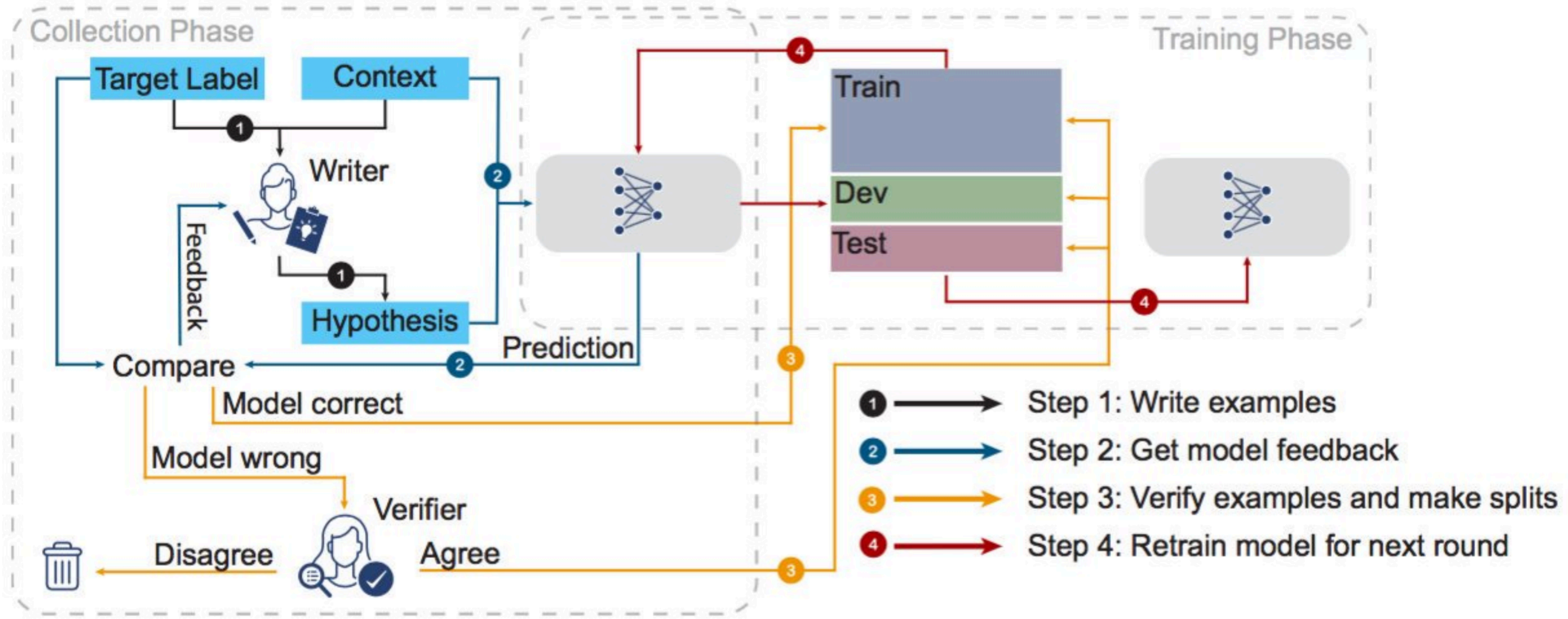
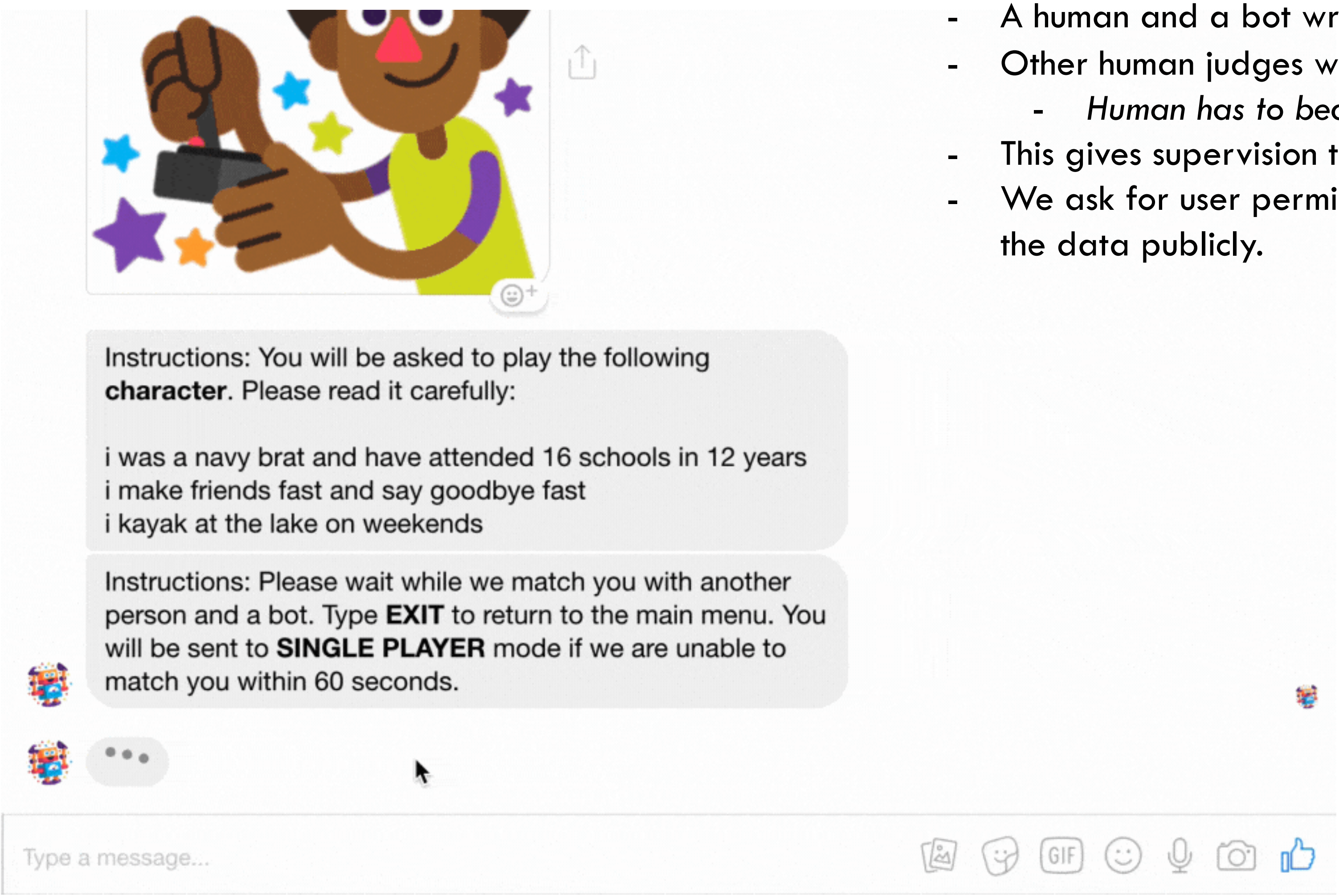


Figure 1: Adversarial NLI data collection procedure, via human-and-model-in-the-loop entailment training (HAMLET). The four steps make up one round of data collection.

ONLINE SYSTEM:
BEAT THE BOT

Live on FB Messenger:
http://parl.ai/projects/beat_the_bot

The game:
- 2 humans play
- A human and a bot write a message.
- Other human judges which is better:
    - *Human has to beat the bot!*
- This gives supervision to the bot.
- We ask for user permission to release the data publicly.



Instructions: You will be asked to play the following **character**. Please read it carefully:

i was a navy brat and have attended 16 schools in 12 years
i make friends fast and say goodbye fast
i kayak at the lake on weekends

Instructions: Please wait while we match you with another person and a bot. Type **EXIT** to return to the main menu. You will be sent to **SINGLE PLAYER** mode if we are unable to match you within 60 seconds.

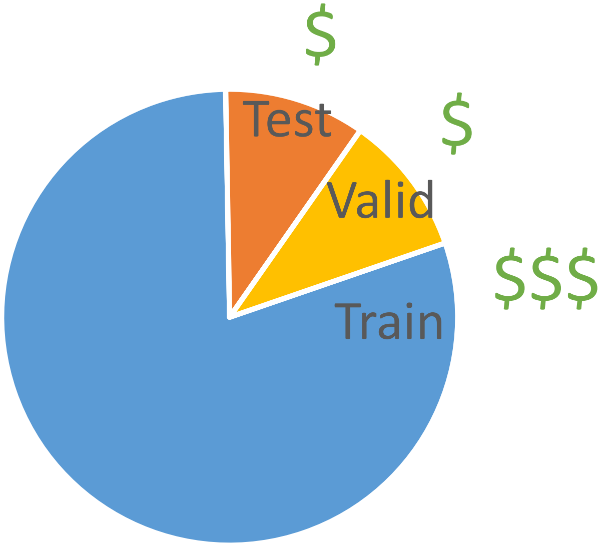
Learning from Dialogue After Deployment:

Feed Yourself, Chatbot!

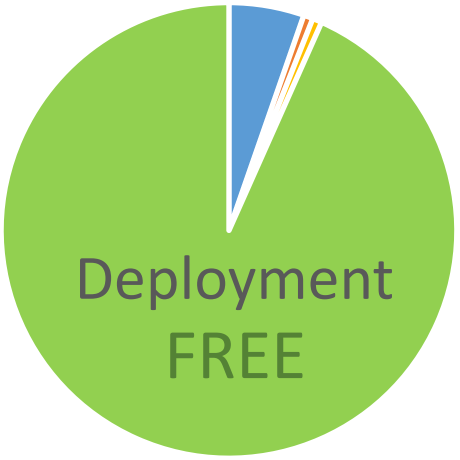
Braden Hancock, Antoine Bordes,
Pierre-Emmanuel Mazaré, Jason Weston

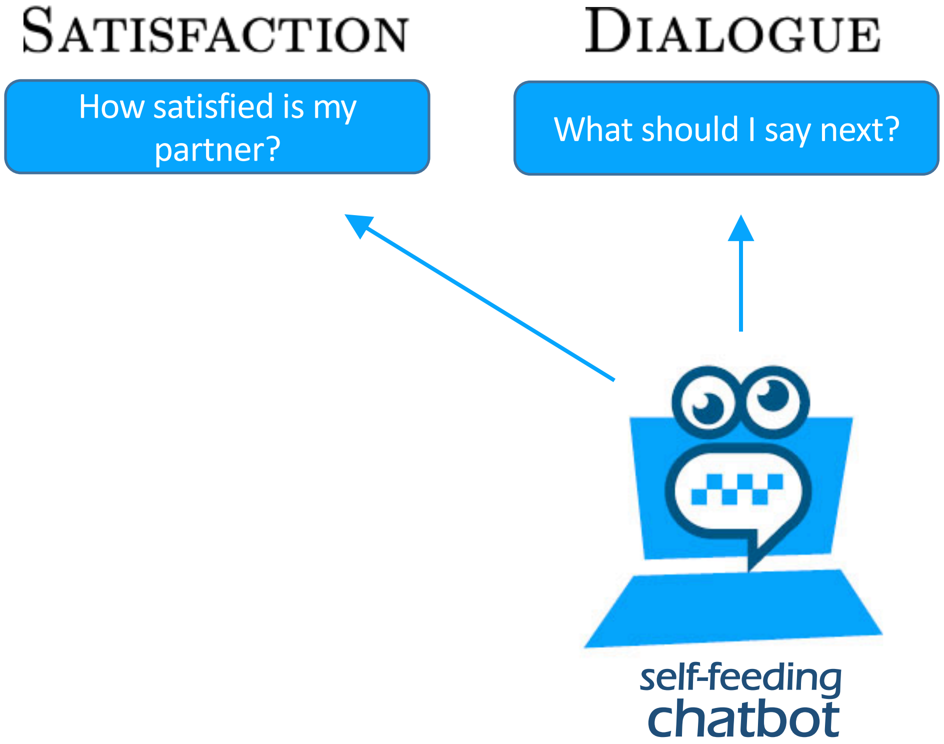
ACL 2019

What I think my data splits are like



What my data splits are actually like



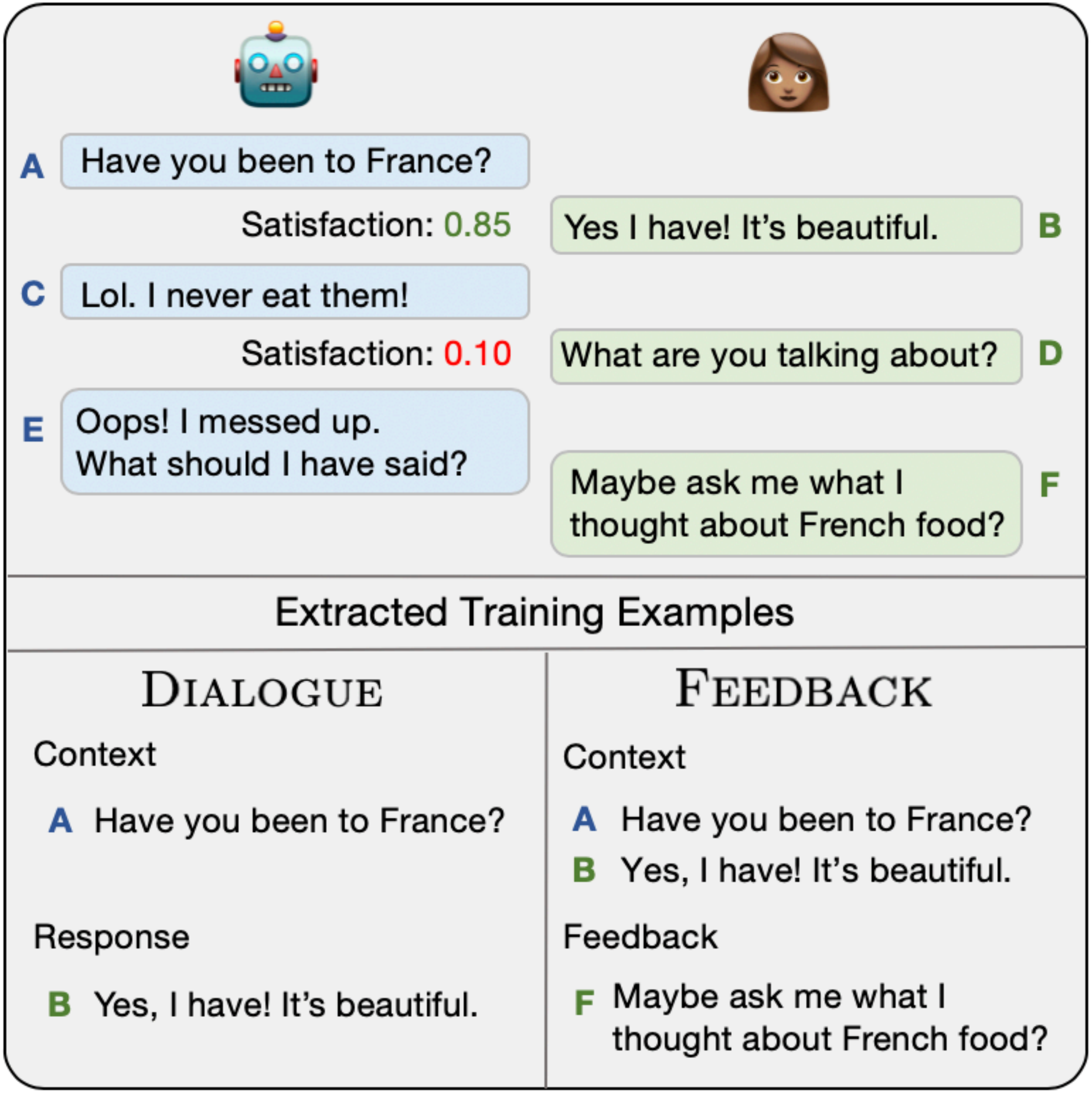


SATISFACTION

How satisfied is my partner?

DIALOGUE

What should I say next?



A Have you been to France?

Satisfaction: 0.85

B Yes I have! It’s beautiful.

C Lol. I never eat them!

Satisfaction: 0.10

D What are you talking about?

E Oops! I messed up.
What should I have said?

F Maybe ask me what I
thought about French food?

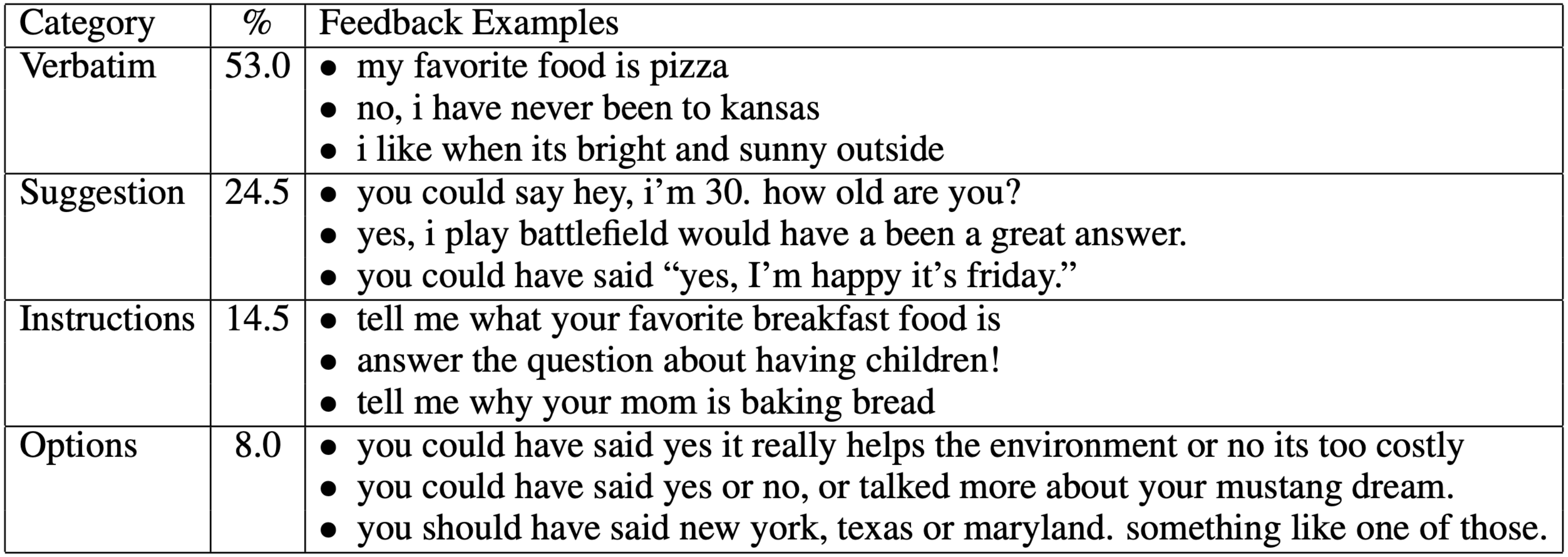
Extracted Training Examples

DIALOGUE	FEEDBACK
Context	Context
A Have you been to France?	**A** Have you been to France? **B** Yes, I have! It’s beautiful.
Response	Feedback
B Yes, I have! It’s beautiful.	**F** Maybe ask me what I thought about French food?

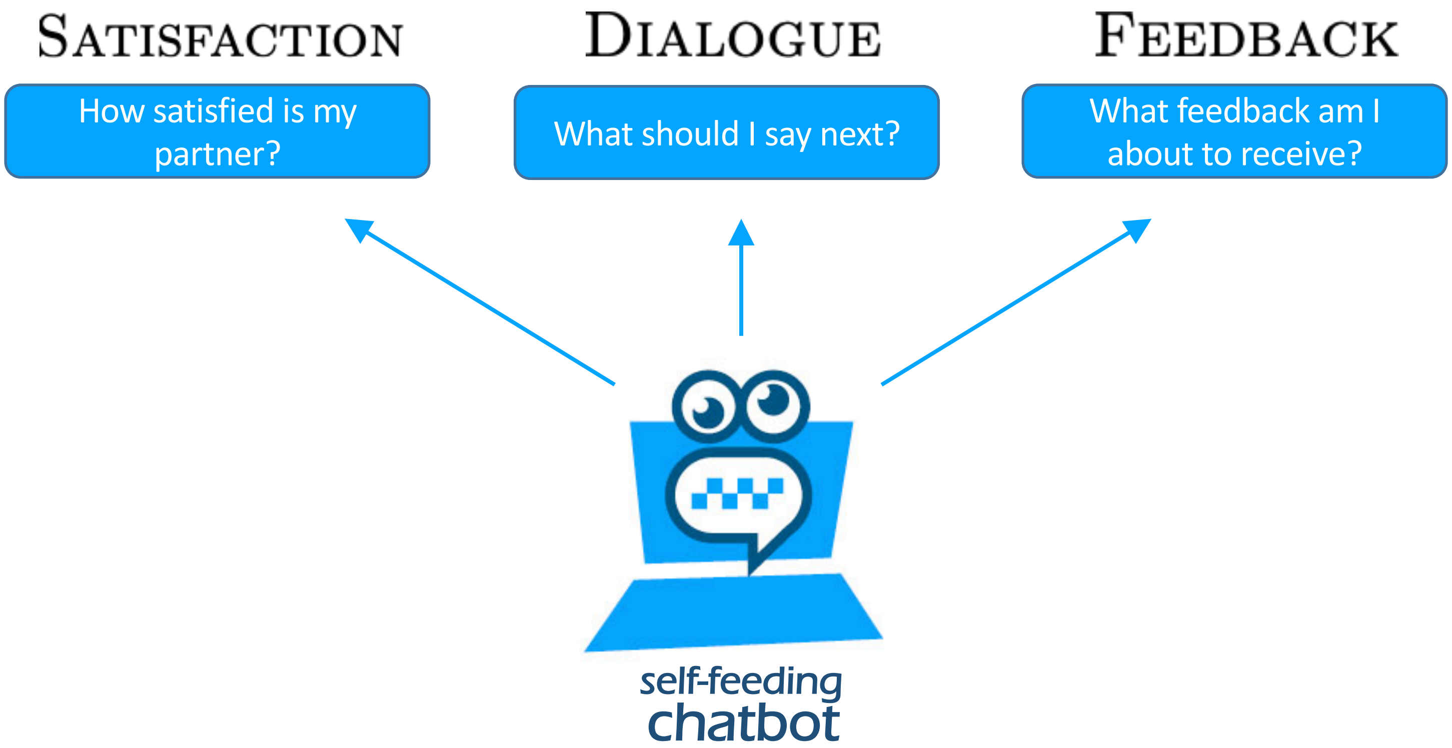
self-feeding
chatbot

Sample User Feedback

self-feeding chatbot



Category	%	Feedback Examples
Verbatim	53.0	• my favorite food is pizza • no, i have never been to kansas • i like when its bright and sunny outside
Suggestion	24.5	• you could say hey, i'm 30. how old are you? • yes, i play battlefield would have a been a great answer. • you could have said "yes, I'm happy it's friday."
Instructions	14.5	• tell me what your favorite breakfast food is • answer the question about having children! • tell me why your mom is baking bread
Options	8.0	• you could have said yes it really helps the environment or no its too costly • you could have said yes or no, or talked more about your mustang dream. • you should have said new york, texas or maryland. something like one of those.



SATISFACTION

How satisfied is my partner?

DIALOGUE

What should I say next?

FEEDBACK

What feedback am I about to receive?

self-feeding chatbot

Human-Bot (HB)		**Human-Human (HH) DIALOGUE**			
DIALOGUE	**FEEDBACK**	20k	40k	60k	131k
-	-	30.3 (0.6)	36.2 (0.4)	39.1 (0.5)	44.7 (0.4)
20k	-	32.7 (0.5)	37.5 (0.6)	40.2 (0.5)	45.5 (0.7)
40k	-	34.5 (0.5)	37.8 (0.6)	40.6 (0.6)	45.1 (0.6)
60k	-	35.4 (0.4)	37.9 (0.7)	40.2 (0.8)	45.0 (0.7)
-	20k	35.0 (0.5)	38.9 (0.3)	41.1 (0.5)	45.4 (0.8)
-	40k	36.7 (0.7)	39.4 (0.5)	41.8 (0.4)	45.7 (0.6)
-	60k	37.8 (0.6)	40.6 (0.5)	42.2 (0.7)	45.8 (0.7)
60k	60k	**39.7** (0.6)	**42.0** (0.6)	**43.3** (0.7)	**46.3** (0.8)

Hits@1/20 (correctly predicting the appropriate response out of 20 candidates)

Latest *in progress* research

Putting it all together

Multi-Task: DodecaDialogue 12 task challenge

Fixing Generation

Improve generation & reasoning:

Unlikelihood for dialogue and semantics

Well-behaved

- Safety from toxic language
  - Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack
    E. Dinan, S. Humeau, B. Chintagunta, J. Weston
- Mitigating gender bias
  - Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation.
    E. Dinan∗, A. Fan∗†, A. Williams, J. Urbanek, D. Kiela, J. Weston

Future:

- Open-ended generation could still be better..
  - working on it, several possible approaches..
- Continual learning research really needs deployment
  - Deploy to make a never-ending learning setup, employ continual learning
  - Trying for wider release … TT
- AI/ML aspects we need but haven't touched on much yet:
  - (More) Commonsense + reasoning – *bake into dialogue – unlikelihood?*
  - (More) Long-term memory – *hard to study with Turked dialogues?*
  - This is mostly chitchat, combine with domain expertise, or task-proficiency.
  - Dialogue agents that can act: see our work on LIGHT (arXiv:1903.03094).