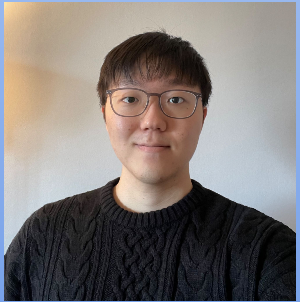
The Alan Turing Institute

Can Dialogue Systems Have Theory-of-Mind Ability?

Danny To Eun Kim
Graduate Researcher at
UCL Web Intelligence Group

to.kim.17@alumni.ucl.ac.uk
https://kimdanny.github.io/
https://twitter.com/TEKnologyy

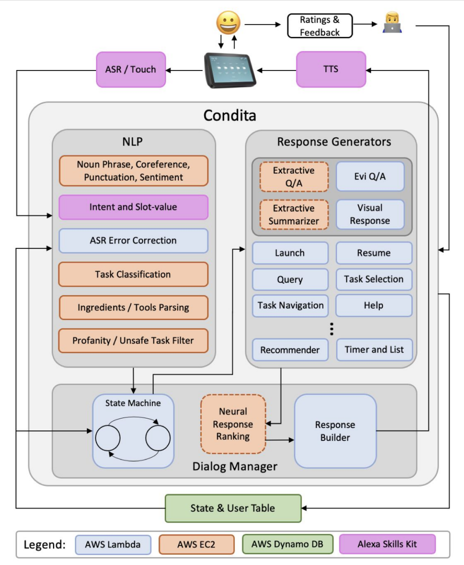


My Research

Real-time, multimodal, knowledge-intensive, and socially intelligent (task-oriented) conversational agent [1]

- Response Ranking/Selection with Conversational Look-Ahead that optimizes user satisfaction (A)
- Evaluation of Conversational System by User Simulation [2]
- Interactive Ground Language Understanding
    - Asking Clarification Questions
- Use of Knowledge Graphs in Conversational Search Systems
- Extracting multimodal information from text: focusing on under-researched sense (e.g. odour)

Ongoing Research (A)

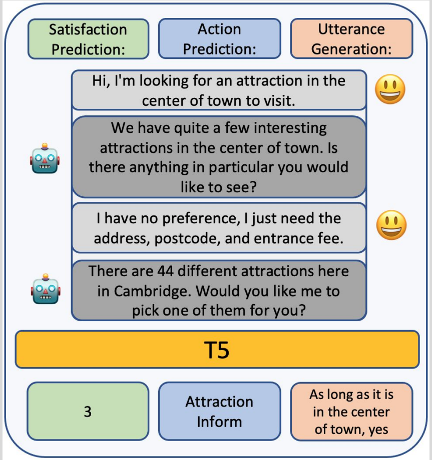


alexa prize

TaskBot Challenge

incorporate

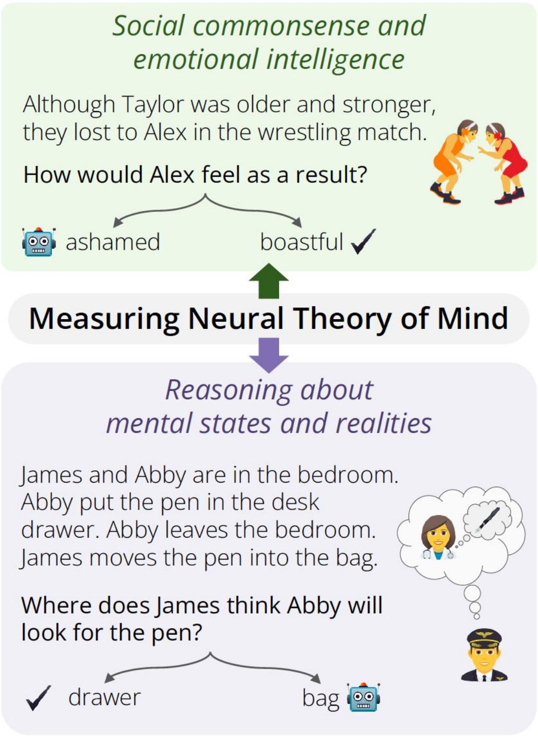
Agents that simulate users?
What to simulate?



[1] https://www.amazon.science/alexa-prize/proceedings/condita-a-state-machine-like-architecture-for-multi-modal-task-bots
[2] https://dl.acm.org/doi/10.1145/3477495.3531814

Table of Contents

1. Neural ToM



Neural Theory-of-Mind?
On the Limits of Social Intelligence in Large LMs

Maarten Sap[♠,♢] **Ronan Le Bras**[♠] **Daniel Fried**[♢] **Yejin Choi**[♠,♡]

[♠]Allen Institute for AI, Seattle, WA, USA

[♢]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

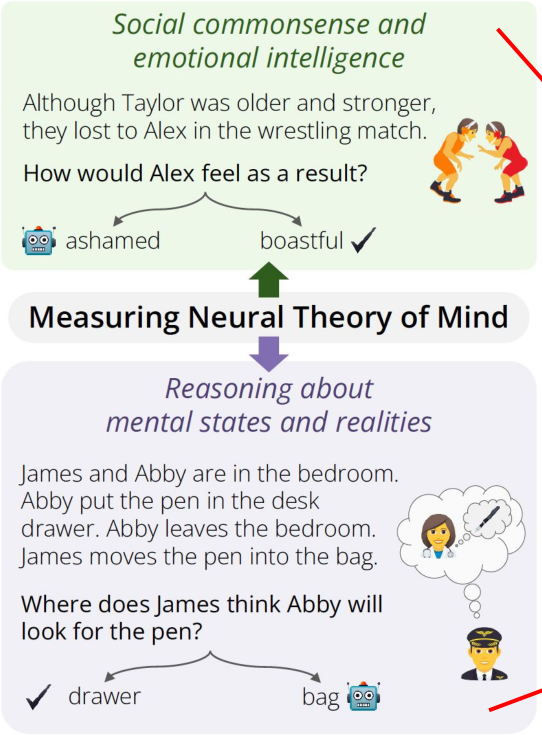
[♡]Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA

Theory-of-Mind (ToM)

Ability to reason different mental states, intents, and reactions of all people involved. Often termed as ‘social intelligence’

Does Large Language Models like GPT-3 have ToM ability?

1. Neural ToM



- Use GPT-3 as a subject
- To assess social commonsense and emotional intelligence,
  used **SocialIQA** dataset
- To assess ability to understand other people's mental states and realities,
  used **ToMi** dataset

1. Neural ToM (why do LMs need this?)

1. Use case: AI assistants
   a. To interact and adapt to users → reasoning abilities are necessary
      i. Voice assistants
      ii. Tutoring system
      iii. AI-assisted counselling
      iv. Facilitated discussion

2. Higher order reasoning
   a. Ensuring alarm is on when someone has job interview in the morning
   b. Inferring personalities and intentions in dialogs
   c. Predicting emotional and affective states

1. Neural ToM (SocialIQA - dataset)

	Situation		Answers	Focus
(a)	Remy was working late in his office trying to catch up. He had a big stack of papers. What does Remy need to do before this?	✓ 🤖	Needed to be behind Be more efficient Finish his work	Agent
(b)	Casey wrapped Sasha's hands around him because they are in a romantic relationship. How would you describe Casey?	✓ 🤖 	Very loving towards Sasha Wanted Being kept warm by Sasha	Agent
(c)	Tracy held a baby for 9 months and then gave birth to addison. What will happen to Tracy?	 🤖 ✓	Throw her baby at the wall Cry Take care of her baby	Agent
(d)	Kai gave Ash some bread so they could make a sandwich. How would Kai feel afterwards?	✓ 🤖	Glad they helped Good they get something to eat Appreciative	Agent
(e)	Aubrey was making extra money by babysitting Tracey's kids for the summer. What will Tracy want to do next?	🤖 ✓ 	Save up for a vacation Let Aubrey know that they are appreciated Pay off her college tuition	Others
(f)	The people bullied Sasha all her life. But Sasha got revenge on the people. What will the people want to do next?	 🤖 ✓	Do whatever Sasha says Get even Flee from Sasha	Others
(g)	After everyone finished their food they were going to go to a party so Kai decided to finish his food first. What will others want to do next?	✓ 🤖 	Eat their food quickly Throw their food away Go back for a second serving	Others
(h)	Aubrey fed Tracy's kids lunch today when Tracy had to go to work. What will happen to Aubrey?	 ✓🤖 	Be grateful Get paid by Tracy Get yelled at by Tracy	Agent
(i)	Sasha was the most popular girl in school when she accepted Jordan's invitation to go on a date. What will Jordan want to do next?	 ✓🤖 	Plan a best friends outing with Sasha Plan a romantic evening with Sasha Go on a date with Valerie	Others

Table 1: Examples of SOCIALIQA questions, which person the questions focus on (*Agent*, *Others*), and the human gold answers (✓) and GPT-3-DAVINCI predictions (🤖).

- Social IQA is to assess social commonsense

- Consists of **Context (c)**, **Question (q)**, and **three answer choices (a_i)**

- 6 dimensions focus on pre- and post-condition of an **agent**

- 3 dimensions focus on post-condition of **other participants** involved in the situation

1. Neural ToM (SocialIQA - probing)

	Situation	Answers	Focus
(a)	Remy was working late in his office trying to catch up. He had a big stack of papers. What does Remy need to do before this?	✓ Needed to be behind Be more efficient 🤖 Finish his work	Agent
(b)	Casey wrapped Sasha's hands around him because they are in a romantic relationship. How would you describe Casey?	✓ Very loving towards Sasha 🤖 Wanted Being kept warm by Sasha	Agent
(c)	Tracy held a baby for 9 months and then gave birth to addison. What will happen to Tracy?	Throw her baby at the wall 🤖 Cry ✓ Take care of her baby	Agent
(d)	Kai gave Ash some bread so they could make a sandwich. How would Kai feel afterwards?	✓ Glad they helped Good they get something to eat 🤖 Appreciative	Agent
(e)	Aubrey was making extra money by babysitting Tracey's kids for the summer. What will Tracy want to do next?	🤖 Save up for a vacation ✓ Let Aubrey know that they are appreciated Pay off her college tuition	Others
(f)	The people bullied Sasha all her life. But Sasha got revenge on the people. What will the people want to do next?	Do whatever Sasha says 🤖 Get even ✓ Flee from Sasha	Others
(g)	After everyone finished their food they were going to go to a party so Kai decided to finish his food first. What will others want to do next?	✓ Eat their food quickly 🤖 Throw their food away Go back for a second serving	Others
(h)	Aubrey fed Tracy's kids lunch today when Tracy had to go to work. What will happen to Aubrey?	Be grateful ✓🤖 Get paid by Tracy Get yelled at by Tracy	Agent
(i)	Sasha was the most popular girl in school when she accepted Jordan's invitation to go on a date. What will Jordan want to do next?	Plan a best friends outing with Sasha ✓🤖 Plan a romantic evening with Sasha Go on a date with Valerie	Others

Table 1: Examples of SOCIALIQA questions, which person the questions focus on (*Agent*, *Others*), and the human gold answers (✓) and GPT-3-DAVINCI predictions (🤖).

- K-shot language probing
- Concatenate context and question together and assign the model prediction with the highest conditional likelihood under the LM
- LM = $\mathrm{argmax}_i\ P_{LM}(a_i \mid c, q, C_k)$, where C_k denotes k training examples with **c**, **q**, and correct **a** concatenated.

1. Neural ToM (SocialIQA - result)

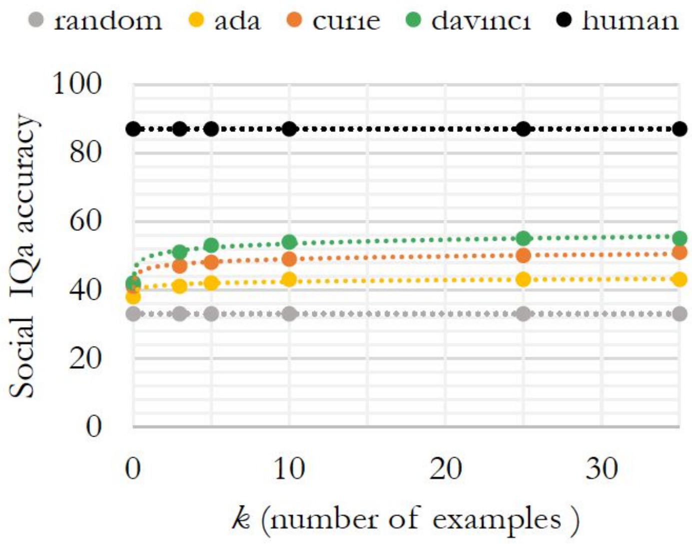


Figure 2: Accuracy on the SOCIALIQA dev. set, broken down by LLM model type and size, as well as number of few-shot examples (k).

K = 0 to 35 (incremented by 5)

GPT-3-ada (smallest), GPT-3-curie and GPT-3-davinci (two largest)

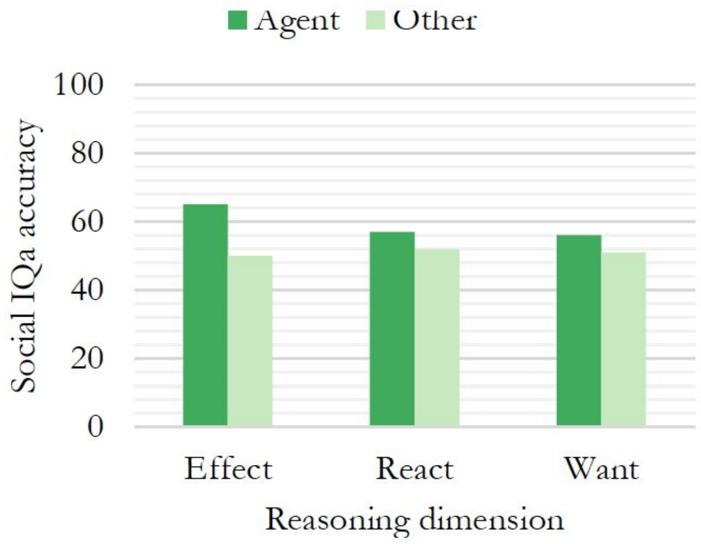


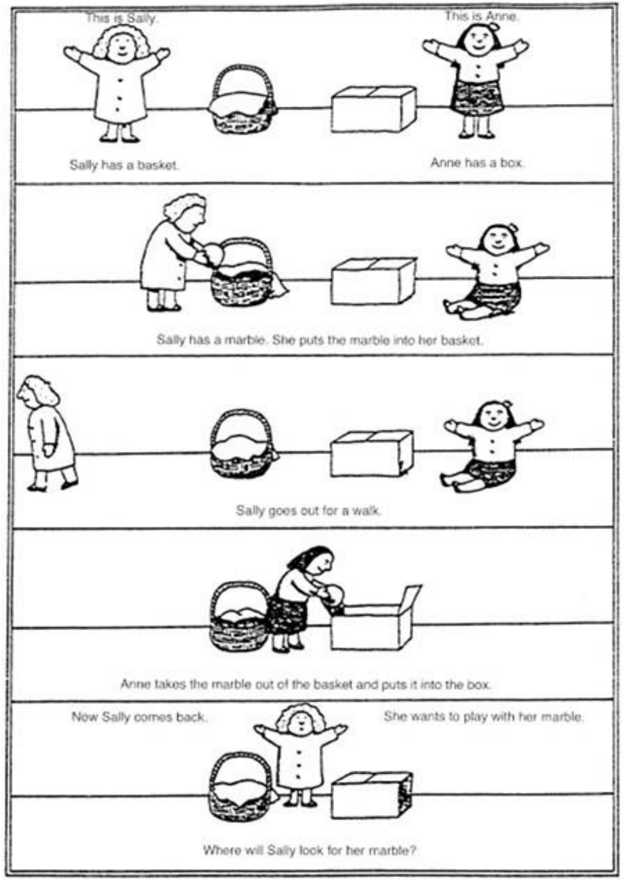
Figure 3: Comparing the accuracy of GPT-3-DAVINCI (35-shot) on SOCIALIQA when the reasoning is about the main agent of the situation versus others.

Consistently better at Agent-centric questions

1. Neural ToM (ToMi - dataset)

	Type	Story	Question		Answers
(a)	FACT	Sophia entered the study. Noah entered the study. The dress is in the treasure chest. Noah exited the study. Hannah entered the garden. Sophia moved the dress to the box.	Where is the dress really?	✓🤖 	box treasure chest
(b)	M-1-FB	Noah entered the garden. Nathan entered the garden. Evelyn likes the pumpkin. The banana is in the basket. Nathan exited the garden. Noah moved the banana to the suitcase.	Where will Nathan look for the banana?	✓🤖 	basket suitcase
(c)	M-2-TB	Lily entered the patio. Aiden is in the patio. Mila entered the patio. Mila hates the radish. The coat is in the box. Aiden moved the coat to the crate. Mila exited the patio.	Where does Aiden think that Mila searches for the coat?	✓🤖 	crate box
(d)	M-1-TB	Elizabeth entered the cellar. Carter entered the cellar. The slippers is in the crate. Elizabeth moved the slippers to the container. Carter exited the cellar.	Where will Carter look for the slippers?	✓ 🤖	container crate
(e)	M-1-FB	Evelyn entered the living room. Jackson entered the playroom. James entered the playroom. The beans are in the treasure chest. James exited the playroom. Jackson moved the beans to the pantry. Jackson exited the playroom. James entered the living room.	Where will James look for the beans?	✓ 🤖	treasure chest pantry
(f)	M-2-FB	Isla likes the potato. Ella entered the laundry. Oliver entered the laundry. The slippers are in the box. Ella exited the laundry. Oliver moved the slippers to the basket. Isla entered the office.	Where does Ella think that Oliver searches for the slippers?	🤖 ✓	basket box

Table 2: Example stories in the ToMi dev. dataset, with GPT-3-DAVINCI predictions (with k=16 examples) and gold answers. "Type" denotes reasoning type, M-1 and M-2 denote MIND-1st and MIND-2nd, resp.



Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception

https://pubmed.ncbi.nlm.nih.gov/6681741/

https://upload.wikimedia.org/wikipedia/en/a/ac/Sally-Anne_test.jpg

1. Neural ToM (ToMi - dataset)

	Type	Story	Question		Answers
(a)	FACT	Sophia entered the study. Noah entered the study. The dress is in the treasure chest. Noah exited the study. Hannah entered the garden. Sophia moved the dress to the box.	Where is the dress really?	✓🤖	box treasure chest
(b)	M-1-FB	Noah entered the garden. Nathan entered the garden. Evelyn likes the pumpkin. The banana is in the basket. Nathan exited the garden. Noah moved the banana to the suitcase.	Where will Nathan look for the banana?	✓🤖	basket suitcase
(c)	M-2-TB	Lily entered the patio. Aiden is in the patio. Mila entered the patio. Mila hates the radish. The coat is in the box. Aiden moved the coat to the crate. Mila exited the patio.	Where does Aiden think that Mila searches for the coat?	✓🤖	crate box
(d)	M-1-TB	Elizabeth entered the cellar. Carter entered the cellar. The slippers is in the crate. Elizabeth moved the slippers to the container. Carter exited the cellar.	Where will Carter look for the slippers?	✓ 🤖	container crate
(e)	M-1-FB	Evelyn entered the living room. Jackson entered the playroom. James entered the playroom. The beans are in the treasure chest. James exited the playroom. Jackson moved the beans to the pantry. Jackson exited the playroom. James entered the living room.	Where will James look for the beans?	✓ 🤖	treasure chest pantry
(f)	M-2-FB	Isla likes the potato. Ella entered the laundry. Oliver entered the laundry. The slippers are in the box. Ella exited the laundry. Oliver moved the slippers to the basket. Isla entered the office.	Where does Ella think that Oliver searches for the slippers?	🤖 ✓	basket box

Table 2: Example stories in the ToMi dev. dataset, with GPT-3-DaVinci predictions (with k=16 examples) and gold answers. "Type" denotes reasoning type, M-1 and M-2 denote Mind-1st and Mind-2nd, resp.

1st order (where will Abby look for the object?)
2nd order (where does James think that Abby will look for the object?)

- ToMi is to assess mental states and realities.
- Samples two participants
  An object of interest,
  Set of locations or containers,
  Make story of an object being moved.
- Questions have two categories:
  Factual Object locations (FACT),
  Where participants think objects are located (MIND)
  FACT questions do not require ToM ability
- First-order and second-order mind questions

1. Neural ToM (ToMi - result)

when an object was moved,
true belief (**TB**): a participant was present
false belief (**FB**): a participant was absent

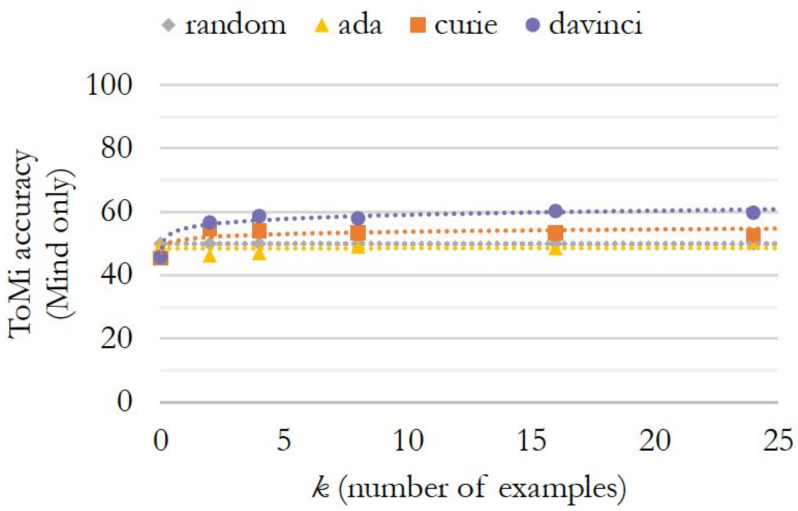


Figure 4: Accuracy on the ToMi dev. set Mind questions of varying sizes of GPT-3, and with varying number of examples (k). K = {2, 4, 8, 16, 24}

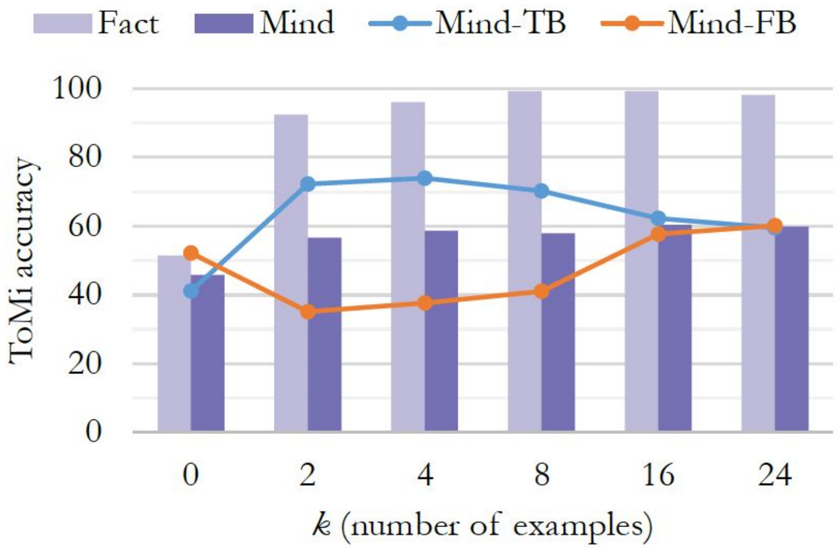


Figure 5: Accuracy of GPT-3-DaVinci by number of examples (k), by reasoning type (Fact vs. Mind; Mind-Tb vs. Mind-Fb).

1. Neural ToM (What’s the Problem?)

Static text based training

Static text:
Document that are written for general audience.
They are self-contained and topically focused

1. Reporting bias - author avoids redundancy by omitting contents naturally known by both author and reader
2. Lack of communicative intents and alternatives - where does he live? Somewhere in Korea
3. Lack of communicative effects - no collaborative and interactive settings
4. Centering Theory - author describes the surroundings of main character (Agent focused)

1. Neural ToM (Towards Neural ToM)

1. Fine-tuning on socially grounded text
   a. Story Commonsense (Rashkin et al., 2018)
   b. GLUCOSE (Mostafazadeh et al., 2020)
   c. **Dialogue dataset**

2. Person-centric inductive bias
   a. Prior work: Entity-centric LMs by extra training on entity coreference.
   b. From the results: model couldn’t understand who the question was asking about.
   c. Suggestion: training with personal coreference

3. **Interactive learning and learning from multi-agent communication**
   a. Creation of simulated environment
   b. Predicting possible next steps and learning from mistakes

4. Evaluation metrics for ToM ability in LMs

(my opinion) ToM in Dialogue Systems

from next slides

Leveraging

1. user signals (in advance) to …
   a. avoid dissatisfaction on-the-fly [1]
   b. retrain the agent / amend datasets [1,2]
   c. learn after deployment [3]
   d. train in RL setting with user simulators [4]

2. users' state of mind (sentiment) to …
   a. generate affective response by RL [5]
   b. take commonsense behaviors [6,7]

Under active research but still very challenging

Satisfaction Estimation is a pseudo-ToM
RQ. Could be acquired by good prompt engineering
E.g., Context + Candidate Response + "Is this conversation satisfactory to him?"
or "How is he feeling now?"

- **[1] Understanding and predicting user dissatisfaction in a neural generative chatbot: https://aclanthology.org/2021.sigdial-1.1/**
- [2] When Life Gives You Lemons, Make Cherryade: Converting Feedback from Bad Responses into Good Labels (28th Oct. 2022): https://arxiv.org/abs/2210.15893
- [3] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!: https://aclanthology.org/P19-1358/
- [4] How to Build User Simulators to Train RL-based Dialog Systems: https://arxiv.org/abs/1909.01388
- [5] Generating Empathetic Responses by Looking Ahead the User's Sentiment: https://arxiv.org/abs/1906.08487
- [6] Commonsense-Focused Dialogues for Response Generation: An Empirical Study: https://arxiv.org/abs/2109.06427v2
- [7] TIMEDIAL: Temporal Commonsense Reasoning in Dialog:https://arxiv.org/abs/2106.04571

2. User Dissatisfaction

Understanding and predicting user dissatisfaction in a neural generative chatbot

Abigail See
Stanford NLP
abisee@stanford.edu

Christopher D. Manning
Stanford NLP
manning@stanford.edu

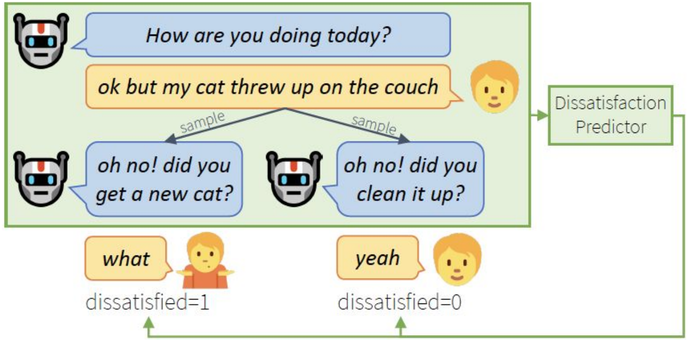


Figure 1: Users tend to express dissatisfaction (such as requests for clarification, left) after the neural generative chatbot makes errors (such as logical errors, left). Using past conversations, we train a model to predict dissatisfaction before it occurs. The model is used to reduce the likelihood of poor-quality bot utterances.

Chirpy Cardinal

What causes dissatisfaction?
Can we avoid it during the chat?

2. User Dissatisfaction

Contributions

- Defined taxonomies of neural generative errors and user dissatisfaction and identified the relationships between them.

- Analysis suggests that improving commonsense reasoning and conditioning on history are high-priority areas for improvement.

- Demonstrated a semi-supervised method to improve a neural generative dialogue system after deployment. Used an automatic classifier to silver-label dissatisfied utterances.

- This model is predictive of most dissatisfaction types, and when deployed as a **ranking function**, a human evaluation shows that it chooses higher-quality bot utterances.

Chirpy Cardinal

GPT-2

↓

Generate 20 possible responses with top-p sampling

↓

Rank with dissatisfaction level

2. User Dissatisfaction (detecting)

Dissatisfaction Type	Definition	Examples	Freq.
Clarification	Indicates the bot's meaning isn't clear	*what do you mean, i don't understand what you're talking about*	2.28%
Misheard	Indicates the bot has misheard, misunderstood or ignored the user	*that's not what i said, you're not listening to me*	0.24%
Repetition	Indicates the bot has repeated itself	*you already said that, we talked about this already*	0.03%
Criticism	Expresses a critical opinion of the bot	*you're so rude, you're bad at this, you're not smart*	0.56%
Privacy	Indicates the bot has overstepped a privacy boundary	*none of your business, why are you asking me that, you're being creepy*	0.11%
Offensive	Contains obscene/offensive words or topics	*will you talk dirty, what size are your boobs, stick it up your ass*	1.54%
Negative Navigation	Expresses desire to end current topic	*change the subject, i don't want to talk about this*	0.59%
Stop	Expresses desire to end conversation	*i have to go bye bye, end the conversation please*	3.68%
Any	Expresses one or more of the above	Any of the above examples	11.56%

Human Labelled set, Regex classifiers, and KNN for the rest

$$P_{\text{kNN}}(D|u) = \begin{cases} \text{HumLabel}_D(u) & \text{if } u \text{ human-labelled} \\ 1 & \text{if } u \text{ matches } D\text{-regex} \\ \frac{1}{k}\sum_{j=1}^{k} \text{HumLabel}_D(u'_j) & \text{otherwise.} \end{cases}$$

Utterances are represented by the Fine-tuned DialoGPT-large model's top layer

Regex: https://github.com/stanfordnlp/chirpycardinal/tree/main/chirpy/core/regex

2. User Dissatisfaction (cause analysis)

Problem	Definition	% in ctrl set	% when no user prob.
User already dissatisfied	The user has already expressed dissatisfaction in c.	12.0%	0.0%
User unclear	The main gist of the user's latest utterance in c is unclear or obscured.	22.0%	0.0%
Bot repetitive	The primary content of b was already said/asked by the bot earlier in c.	6.0%	4.3%
Bot redundant question	b is asking for information that the user has already provided earlier in c.	12.0%	15.9%
Bot unclear	It's hard to find an interpretation of b that makes sense.	12.0%	7.2%
Bot hallucination	b refers to something that hasn't been mentioned, acts like the user said something they didn't, confuses self with user, or seems to be responding to own utterance.	17.0%	10.1%
Bot ignore	b ignores or fails to acknowledge the user's latest utterance, doesn't answer a question, doesn't adequately respond to a request, or switches to an unrelated topic.	20.0%	14.5%
Bot logical error	b is generally on-topic, but makes an assumption or association that's incorrect, unfounded or strange.	15.0%	17.4%
Bot insulting	b says or implies something insulting about the user, or about others in a way that might offend the user.	1.0%	1.4%
Any bot error	True iff any of the above *bot* errors are true.	53.0%	46.4%

Taxonomy of generative bot errors

2. User Dissatisfaction (cause analysis)

- When user is already dissatisfied, hard to continue satisfactory conversation
- When user utterance is unclear, bot tends to hallucinate Clarifying Question is needed
- Errors relating to reasoning or social abilities are common ToM
- Only a minority (15%) of users express dissatisfaction right after the error
- More in the paper

- [5] Generating Empathetic Responses by Looking Ahead the User's Sentiment: https://arxiv.org/abs/1906.08487

From slide 15.
One turn look-ahead might not be sufficient

2. User Dissatisfaction (prediction)

- DialoGPT-large
- probability that the next user utterance u will express Any dissatisfaction.
- Hidden state of the top-layer L for the last timestep t of input and apply linear layer

$$P_{\text{pred}}(\text{Any}|c, b) = \sigma(W^T H_{L,t}) \in [0, 1]$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(P_{\text{pred}}(\text{Any}|c_i, b_i) - P_{\text{kNN}}(\text{Any}|u_i)\right)^2$$

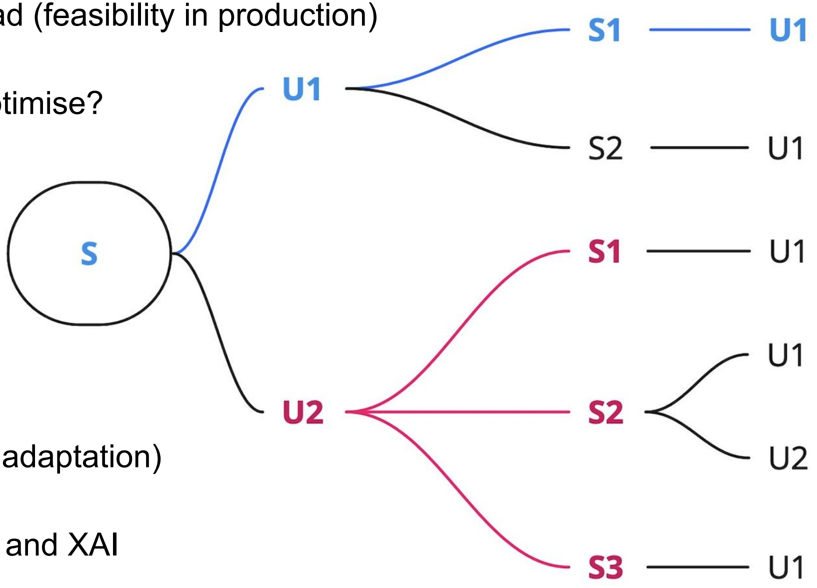
2. User Dissatisfaction (evaluation)

achievable). We sample 400 examples from the NeuralChatTurns validation set, then manually filter to obtain 270 achievable examples. For these, we take the context c and generate 20 possible bot responses $b_1, \ldots, b_{20}$, using the generative model and decoding procedure in Section 2.1. Let b_{pred} be the response with best (i.e., lowest) predictor score: $b_{\text{pred}} = \text{argmin}_{b_j \in b_1, \ldots, b_{20}} P_{\text{pred}}(\text{Any}|c, b_j)$. We randomly sample an alternative b_{rand} uniformly from the other 19 responses. One expert evaluator viewed each c, then chose which of b_{pred} or b_{rand} (presented blind) is a higher-quality response. If

b_{pred} is preferred in 46.3% of cases,
b_{rand} in 35.6%,
and no preference in 18.1%.

3. Future Research Direction

- Computation + Cognition + Linguistics
- LMs with ToM ability
- Computational complexity of simulation and look-ahead (feasibility in production)
- Is satisfaction estimation enough? Other metrics to optimise?
- Computational Thought Experiment
    - Prefactual
    - Counterfactual
    - Hindcasting
    - Retrodiction
- Advanced user simulation (modularisation + personal adaptation)
- Application to Strategic Dialogue (Meta AI’s CICERO) and XAI



CICERO: https://ai.facebook.com/blog/cicero-ai-negotiates-persuades-and-cooperates-with-people/