What makes a good conversation?

How controllable attributes affect human judgments

Abigail See*, Stephen Roller+, Douwe Kiela+, Jason Weston+

*Stanford NLP
+Facebook AI Research

Natural Language Generation task spectrum

Machine Translation | Sentence Compression | Abstractive Summarization | Story Generation | Chitchat Dialogue

Less open-ended

Mostly word-level decisions

Neural LMs more successful

Makes errors like **repetition** and **generic response** (under certain decoding algorithms).

Difficulty learning to make high-level decisions.

More open-ended

Requires high-level decisions

Neural LMs less successful

Natural Language Generation task spectrum

Machine Translation	Sentence Compression	Abstractive Summarization	Story Generation	Chitchat Dialogue

Less open-ended	More open-ended
Mostly word-level decisions	Requires high-level decisions
Neural LMs more successful	**Neural LMs less successful**
Control is less important	**Control is more important**

Control = ability to specify desired attributes of the text at test time.

We can use control to fix errors, and allow us to handle some high-level decisions.

Natural Language Generation task spectrum

Machine Translation	Sentence Compression	Abstractive Summarization	Story Generation	Chitchat Dialogue

Less open-ended	More open-ended
Mostly word-level decisions	Requires high-level decisions
Neural LMs more successful	**Neural LMs less successful**
Control is less important	**Control is more important**
Eval is difficult	**Eval is fiendish**

No automatic metric for overall quality.

Dialogue is even more complex:
Single-turn or multi-turn eval?
Interactive or static conversation?

Our research questions

By controlling multiple attributes of generated text and human-evaluating multiple aspects of conversational quality, we aim to answer the following:

1. **How effectively can we control the different attributes?**
   Pretty well! But some control methods only work for some attributes.

2. **How do the controllable attributes affect conversational quality aspects?**
   Strongly – especially controlling repetition, question-asking, and specificity vs genericness.

3. **Can we use control to make a better chatbot overall?**
   Yes! But we should be careful defining "better overall".

PersonaChat task (Zhang et al 2018)

Persona:
- I love to drink fancy tea.
- I have a big library at home.
- I'm a museum tour guide.
- I'm partly deaf.

Persona:
- I have two dogs.
- I like to work on vintage cars.
- My favorite music is country.
- I own two vintage Mustangs.

Hello, how are you doing?

Great thanks, just listening to my favorite Johnny Cash album!

Nice! I'm not much of a music fan myself, but I do love to read.

Me too! I just read a book about the history of the auto industry.

PersonaChat task (Zhang et al 2018)

- The PersonaChat task was the focus of the **NeurIPS 2018 ConvAI2 Competition**.
  - Most successful teams built neural sequence generation systems. (Dinan et al 2019)
  - The winning team, *Lost in Conversation*, used a finetuned version of GPT.

- **Our baseline model** is a standard LSTM-based seq2seq architecture with attention.
  - It is pretrained on 2.5 million Twitter message/response pairs, then finetuned on PersonaChat.

What attributes do we control?

Low-level controllable attributes

Repetition (n-gram overlap) — Goal: Reduce repetition (within and across utterances)

Specificity (normalized inverse document frequency) — Goal: Reduce genericness of responses (e.g. *oh that's cool*)

Response-relatedness (cosine similarity of sentence embeddings) — Goal: Respond more on-topic; don't ignore user

Question-asking (“?” used in utterance) — Goal: Find the optimal rate of question-asking

What quality aspects do we measure?

Low-level controllable attributes	Human judgment of conversational aspects	
Repetition (n-gram overlap)	Avoiding Repetition	Does the bot repeat itself?
Specificity (normalized inverse document frequency)	Interestingness	Did you find the bot interesting to talk to?
Response-relatedness (cosine similarity of sentence embeddings)	Making sense	Does the bot say things that don't make sense?
Question-asking ("?" used in utterance)	Fluency	Does the bot use English naturally?
	Listening	Does the bot pay attention to what you say?
	Inquisitiveness	Does the bot ask a good amount of questions?

What quality aspects do we measure?

Low-level controllable attributes

- **Repetition** (n-gram overlap)
- **Specificity** (normalized inverse document frequency)
- **Response-relatedness** (cosine similarity of sentence embeddings)
- **Question-asking** (“?” used in utterance)

Human judgment of conversational aspects

- Avoiding Repetition
- Interestingness
- Making sense
- Fluency
- Listening
- Inquisitiveness

Human judgment of overall quality

- Humanness } Is it a person or a bot?
- Engagingness } Is it enjoyable to talk to?

Note: ConvAI2 competition asked *only* this question.
Our eval is a *superset* of ConvAI2's.

Control methods

We evaluate and compare **two existing general-purpose control methods**, using them to control all four controllable attributes.

- **Conditional Training** (CT): Train the model to generate response *y*, conditioned on the input *x*, and the desired output attribute *z*.
  (Kikuchi et al 2016, Peng et al 2018, Fan et al 2018)

- **Weighted Decoding** (WD): During decoding, increase/decrease the probability of generating words *w* in proportion to features *f*(*w*).
  (Ghazvininejad et al 2017, Baheti et al 2018)

Q1: How effectively can we control attributes?

Attributes: repetition, specificity, question-asking, response-relatedness

Conditional Training (CT):

- Requires sufficient **training examples** for the attribute (repetition)
- Ineffective at learning **complex relationships** between input and output (response-relatedness)
- **Effective for**: ✓specificity, ✓question-asking

Weighted Decoding (WD):

- Requires attribute to be defined **at the word-level** (question-asking)
- **Effective for**: ✓repetition, ✓response-relatedness, ✓specificity

Controlling specificity (WD and CT)

Input: *Yes, I'm studying law at the moment*
Baseline Response: *That sounds like a lot of fun!*

Wt	NIDF	Weighted Decoding Response
-5.0	0.6%	*Oh......................................*
0.0	17.1%	*That sounds like a lot of fun!*
3.0	18.3%	*That sounds like a lot of fun. How long have you been studying?*
7.0	38.5%	*I majored in practising my spiritual full time philosophy test*
10.0	71.9%	*Oh wow! Merna jean isa paino yi hao hui bu acara sya gila* [. . .]

More generic

More specific

Controlling specificity (WD and CT)

Input: *Yes, I'm studying law at the moment*
Baseline Response: *That sounds like a lot of fun!*

Wt	NIDF	Weighted Decoding Response
-5.0	0.6%	*Oh......................................*
0.0	17.1%	*That sounds like a lot of fun!*
3.0	18.3%	*That sounds like a lot of fun. How long have you been studying?*
7.0	38.5%	*I majored in practising my spiritual full time philosophy test*
10.0	71.9%	*Oh wow! Merna jean isa paino yi hao hui bu acara sya gila* [. . .]

z	NIDF	Conditional Training Response
0	16.8%	*Sounds like you are a great person!*
2	18.3%	*So you are a law student?*
4	18.4%	*That sounds like a lot of fun*
6	22.8%	*That sounds like a rewarding job!*
8	24.4%	*That sounds like a rewarding career!*

WD: Large range, but degenerate output at the extremes

CT: Smaller range, but generally well-formed output

More generic

More specific

More generic

More specific

Controlling response-relatedness (WD)

Input: *Do you go get coffee often*
Baseline Response: *I do, when I am not playing the piano.*

Wt	Sim	Weighted Decoding Response
-10.0	-0.05	*I am a musician.*
0.0	-0.02	*I do, when I am not playing the piano.*
5.0	0.40	*I do, usually at starbucks.*
8.0	0.59	*Not usually, especially when you drink latte.*
11.0	0.72	*Not often, usually with drinks, espresso, latte, tea, etc.*

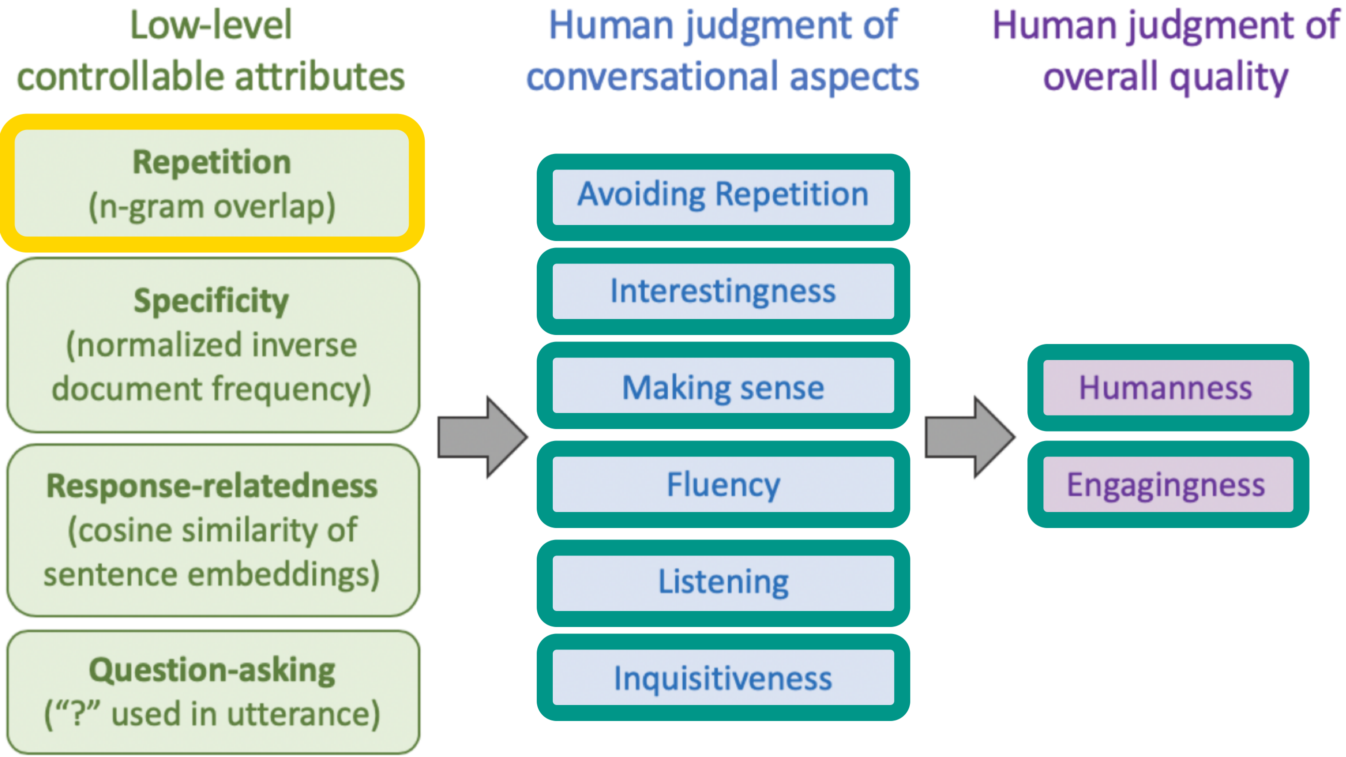
Less related

More related

Output is degenerate
when weight is too high

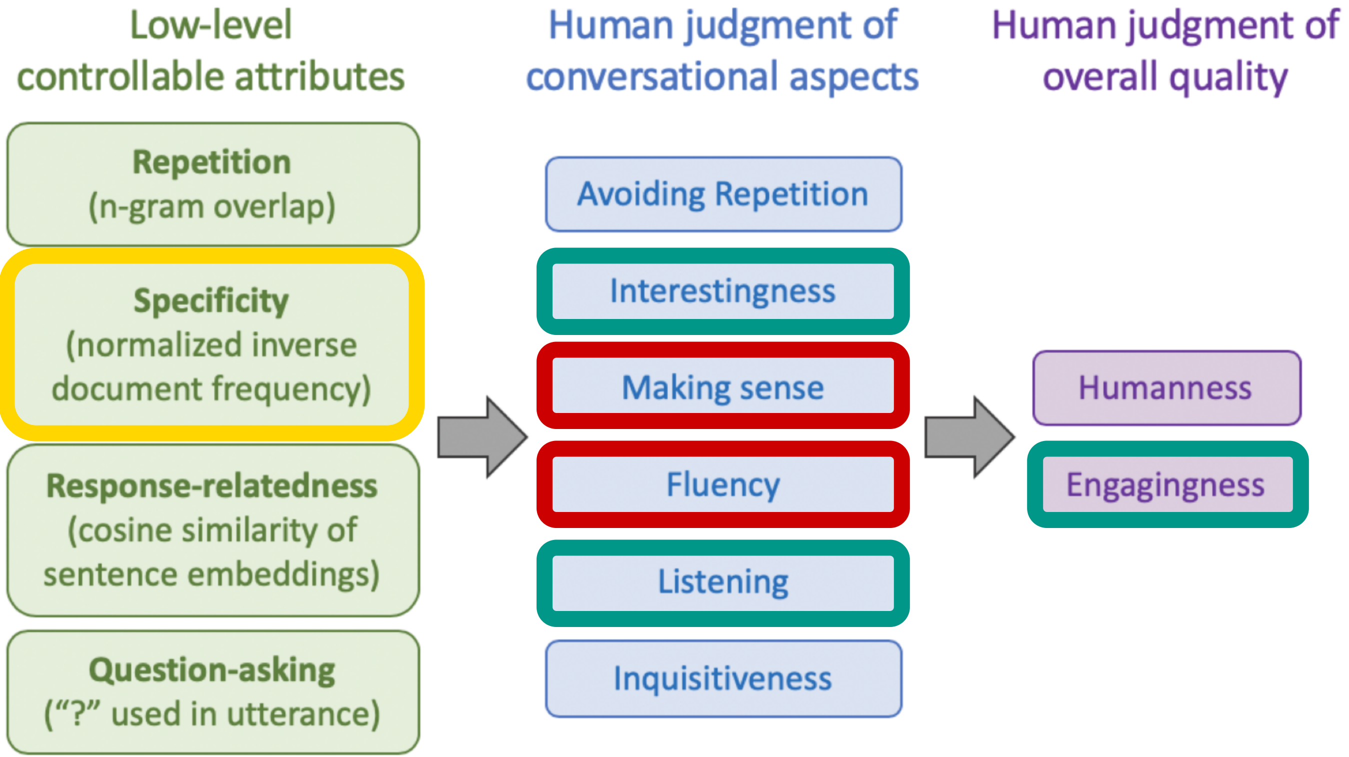
Q2: How does control affect human eval?

Reduce n-gram repetition to human level



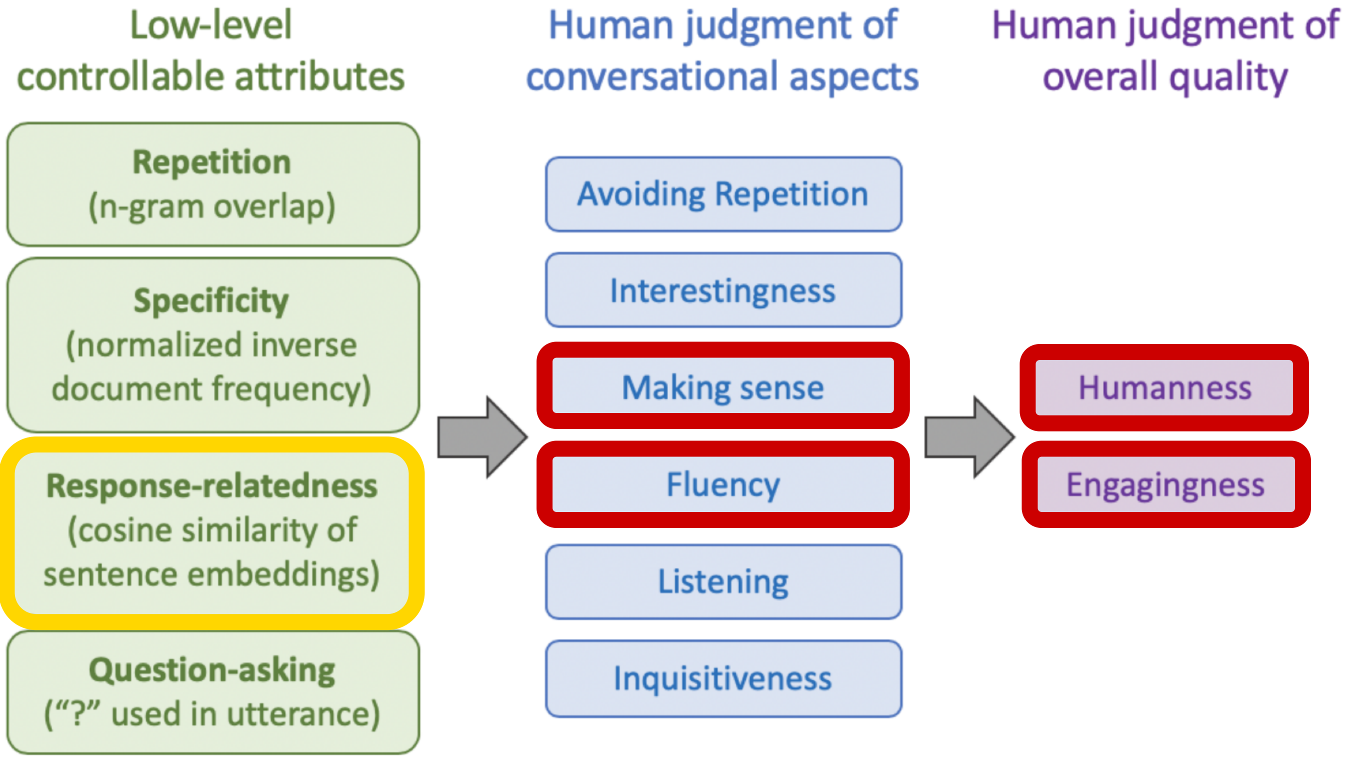
Q2: How does control affect human eval?

Increase specificity
(reduce genericness)
to human level

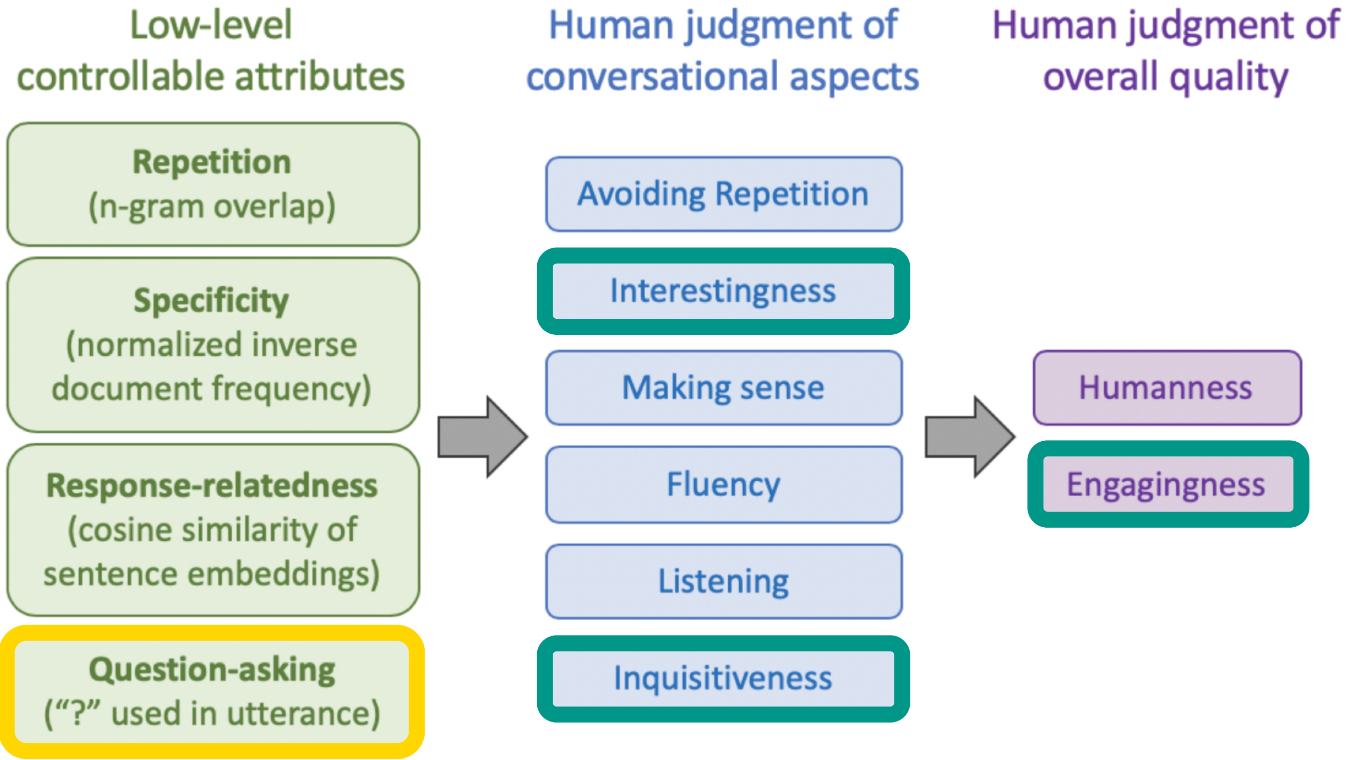


Q2: How does control affect human eval?

Increase response-relatedness (similarity to last utterance)



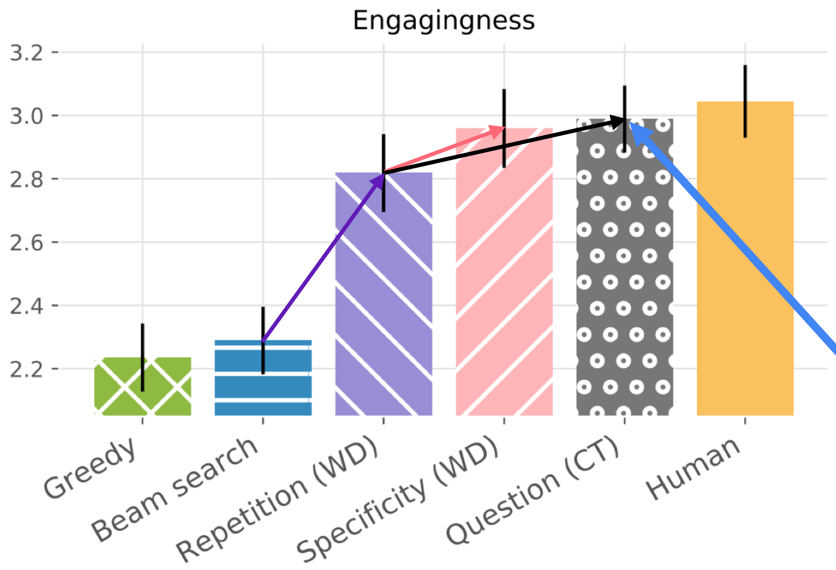
Q2: How does control affect human eval?



Increase question-asking rate to 65.7% (more than baseline 50%, human 28.8%)

Q3: Can we make a better chatbot overall?

Yes! By controlling repetition, specificity and question-asking, we achieve **near-human engagingness (i.e. enjoyability) ratings**.



Reduce repetition

Increase specificity

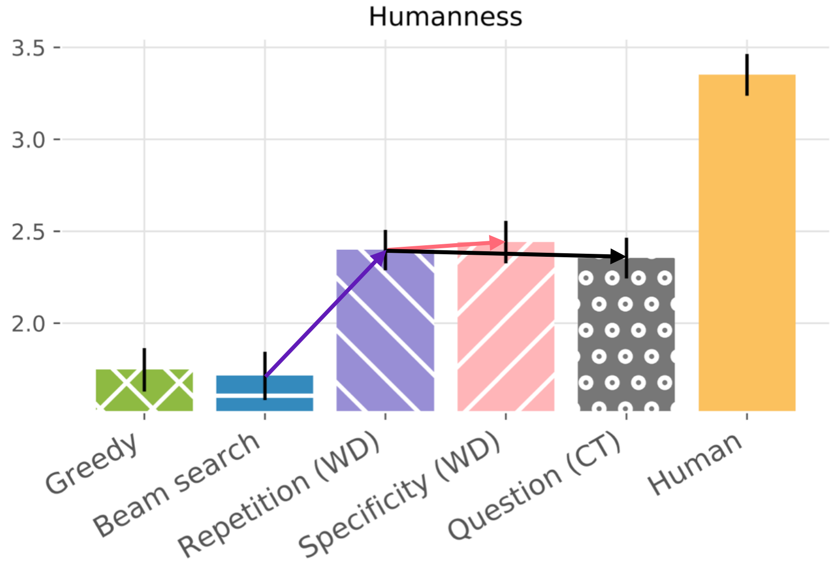
Increase question-asking

Our raw engagingness score matches the **ConvAI2 competition winner**'s GPT-based model, even though ours is:

- **much smaller** (2 layers vs 12)
- trained on **12x less data**

Q3: Can we make a better chatbot overall?

However: On the **humanness** (i.e. Turing test) metric,
our models are nowhere near human-level!



Reduce repetition
Increase specificity
Increase question-asking

Engagingness vs Humanness

Finding: Our bots are **(almost) as engaging as humans,** but they're **clearly non-human.**

Two conclusions:

1. **Engagingness ≠ Humanness**. While both are frequently used as standalone overall quality metrics, our results show the importance of measuring more than one.

2. On this task, **the human "engagingness" performance may be artificially low.** Turkers chatting for money are less engaging than people chatting for fun. This may be why the human-level engagingness scores are easy to match.

Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, **we matched performance of GPT-based contest winner**.
- **Don't repeat yourself. Don't be boring. Ask more questions.**
- **Multi-turn phenomena** (repetition, question-asking frequency) are important – so need **multi-turn eval** to detect them.
- **Engagingness ≠ Humanness**, so think carefully about which to use.
- **Paid Turkers** are **not engaging conversationalists**, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is **painful**.

Repetition Question-asking Specificity Response-relatedness

Conclusions

- **Problem**: Manually finding the best combination of control settings is **painful**.

Code, models, demo, eval logs available at
https://parl.ai/projects/controllable_dialogue

Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, **we matched performance of GPT-based contest winner**.
- **Don't repeat yourself. Don't be boring. Ask more questions.**
- **Multi-turn phenomena** (repetition, question-asking frequency) are important – so need **multi-turn eval** to detect them.
- **Engagingness ≠ Humanness**, so think carefully about which to use.
- **Paid Turkers** are **not engaging conversationalists**, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is **painful**.