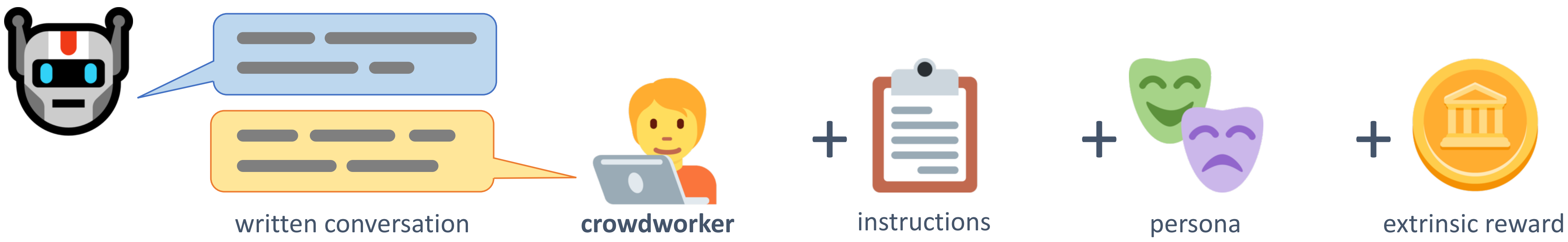
Understanding and predicting user dissatisfaction in a neural generative chatbot

Abigail See, Christopher D. Manning

Stanford NLP

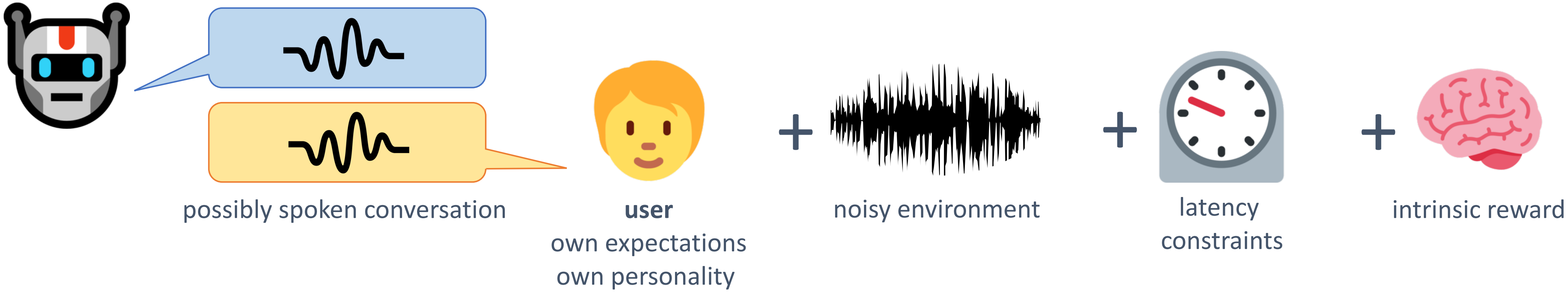
Typical dialogue evaluation setup



Neural generative models perform increasingly well in this environment

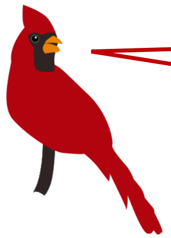
[Zhang et al 2019, Adiwardana et al 2020, Roller et al 2020]

Real-life dialogue settings

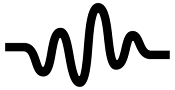


Neural generative models are less well-explored in this environment.

Chirpy Cardinal 1.0



For more details:
stanfordnlp.github.io/chirpycardinal

- An open-domain socialbot
- Won 2nd place in third Alexa Prize (2019-2020) 
- Alexa customers speak  to socialbots voluntarily 
- Bots must be fast  and manage the noisy home environment 

Neural Generation Meets Real People: Towards Emotionally Engaging Mixed-Initiative Conversations, Paranjape*, See* et al, Alexa Prize Proceedings, 2020

Neural Chat component

The Neural Chat component uses a neural generator to discuss the user's personal experiences and emotions.

How are you feeling today?

kind of bored to be honest

I'm sorry to hear that! Why are you bored?

stuck inside all day

Oh yeah, it can get stuffy being inside all day. Could you go outside for a bit?

it's raining

⋮

Discussion starts with a handwritten starter question.
Topics include: hobbies, food, everyday activities, family/friends.

Subsequent responses are generated by a GPT2-medium model finetuned on EmpatheticDialogues (Rashkin et al, 2019).

Decoding algorithm: top-p sampling

Towards empathetic open-domain conversation models: A new benchmark and dataset, Rashkin et al, 2019

Types of user dissatisfaction

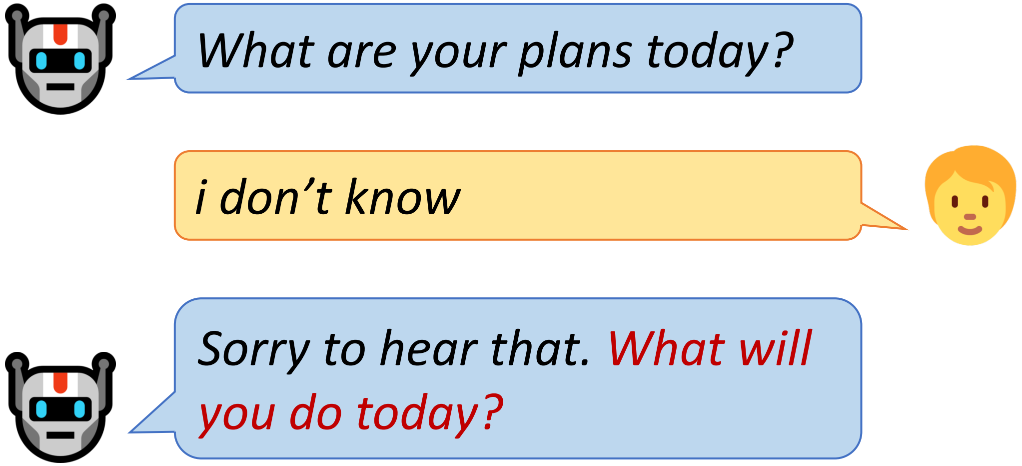
Dissatisfaction type	Example
Clarification	*what do you mean*
Misheard	*that's not what i said*
Privacy	*none of your business*
Repetition	*you already said that*
Criticism	*you’re bad at this*
Offensive	*will you talk dirty*
Negative Navigation	*change the subject*
Stop	*goodbye*
Any	Any of the above

Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

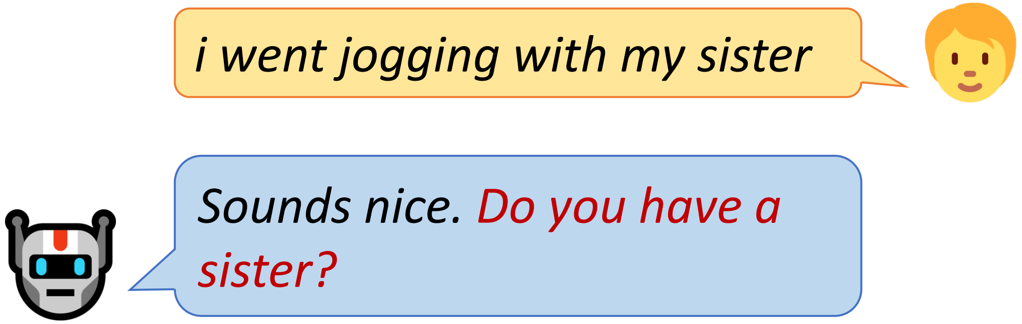
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



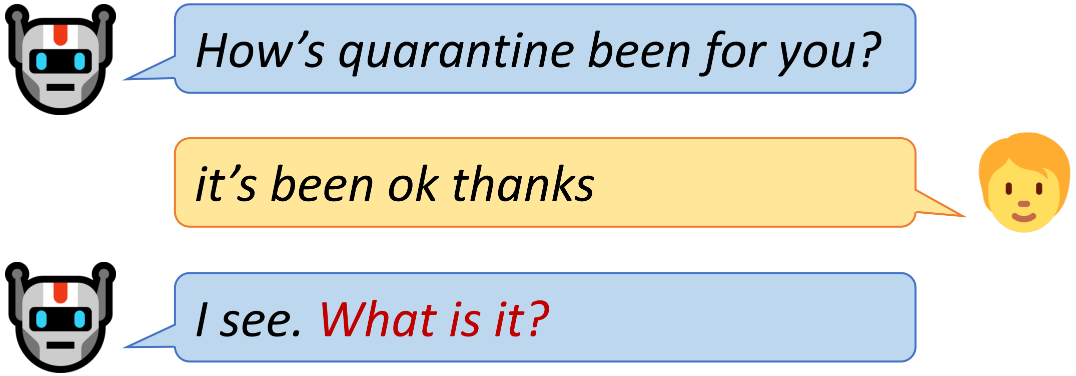
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



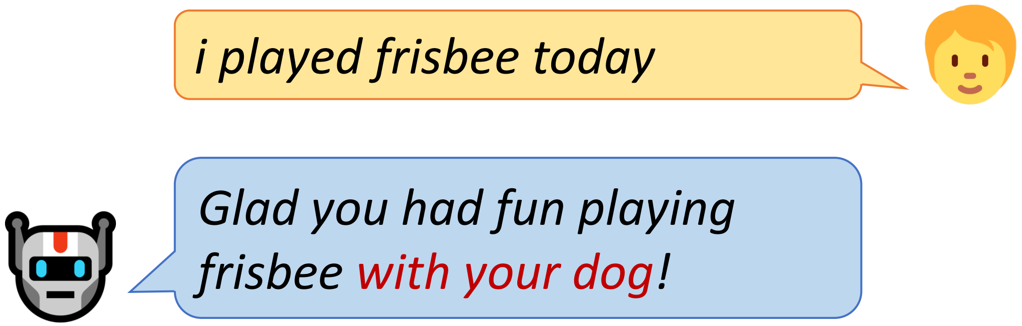
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



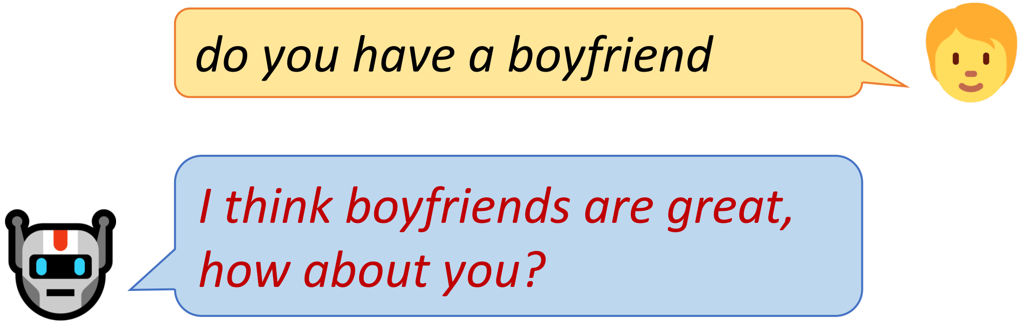
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



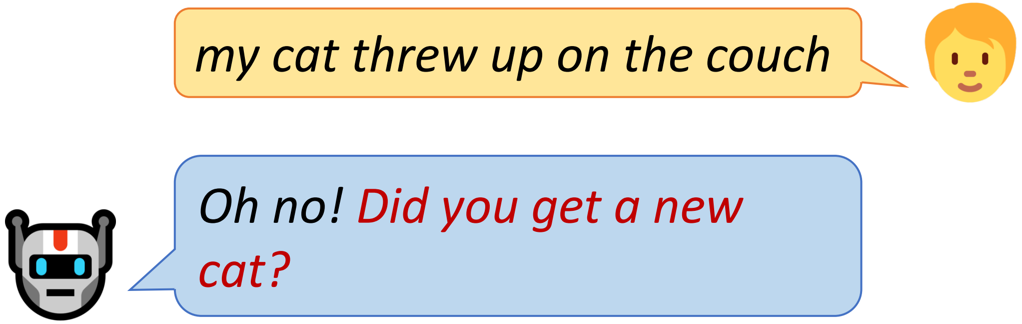
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



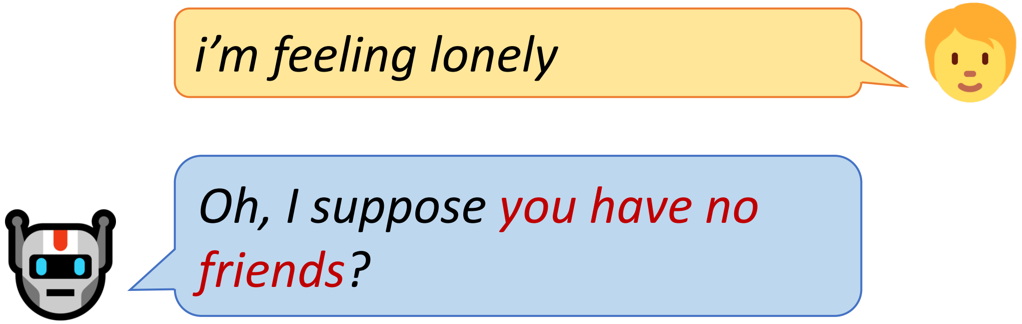
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting



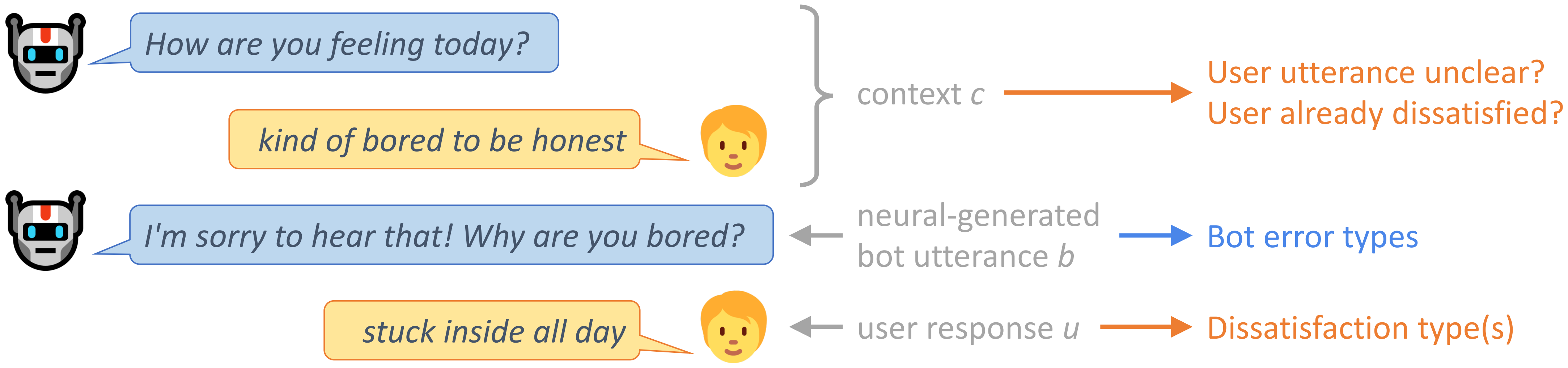
Neural generative error types

1. Repetition
2. Redundant question
3. Unclear
4. Hallucination
5. Ignoring
6. Logical error
7. Insulting

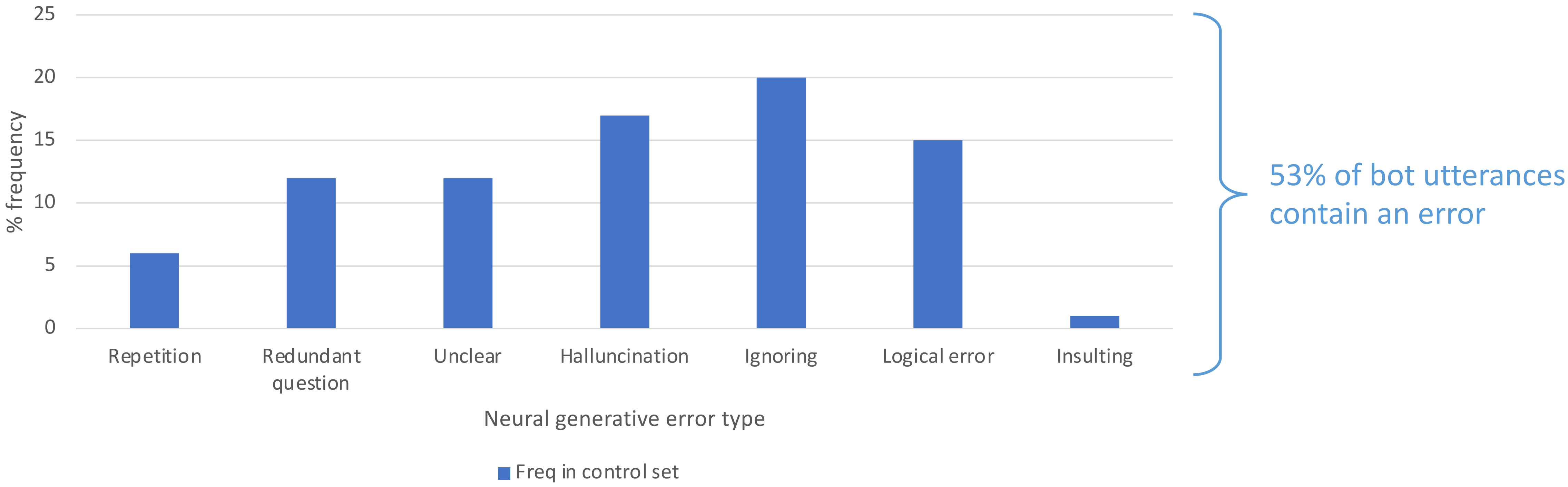


What causes dissatisfaction?

We annotate 900 examples:

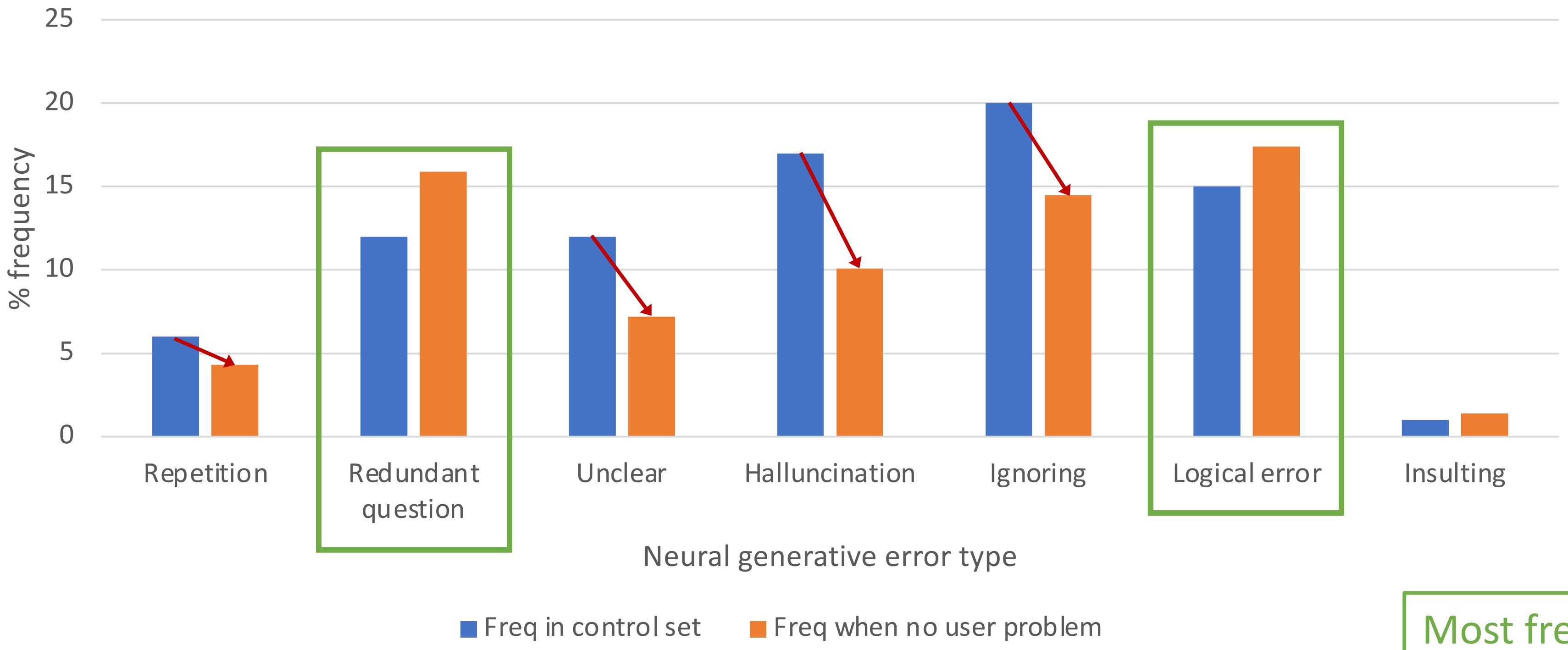


Neural generative error frequency



- 22% of user utterances are unclear
- In 12% contexts, the user is already dissatisfied

Neural generative error frequency



53% of bot utterances contain an error

→ 46.4%

Most frequent remaining errors:

- Redundant questions
- Logical errors

- 22% of user utterances are unclear
- In 12% contexts, the user is already dissatisfied

This accounts for some of the more basic errors.

How do bot errors cause dissatisfaction?

Subsequent dissatisfaction types

Bot errors

	Clarification	Misheard	Repetition	Criticism	Privacy	Offensive	Neg nav	Stop
Bot repetitive								
Bot redundant								
Bot unclear								
Bot hallucination								
Bot ignore								
Bot logical error								
Bot insulting								

How do bot errors cause dissatisfaction?

		Subsequent dissatisfaction types							
		Clarification	Misheard	Repetition	Criticism	Privacy	Offensive	Neg nav	Stop
Bot errors	Bot repetitive	✓		✓		✓	✓	✓	✓
	Bot redundant			✓					
	Bot unclear	✓						✓	
	Bot hallucination		✓						
	Bot ignore		✓						
	Bot logical error								
	Bot insulting				✓			✓	

✓ indicates positive Logistic Regression coefficient with feature significance ($p<0.05$) using Likelihood Ratio Test

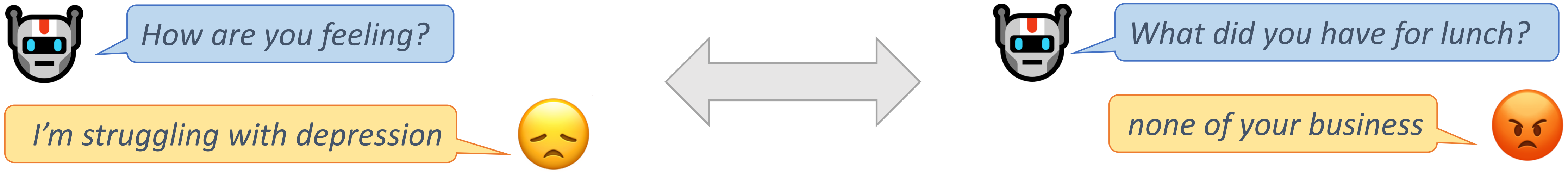
How do bot errors cause dissatisfaction?

Subsequent dissatisfaction types

Bot errors	Clarification	Misheard	Repetition	Criticism	Privacy	Offensive	Neg nav	Stop
Bot repetitive	✓		✓		✓	✓	✓	✓
Bot redundant			✓					
Bot unclear	✓						✓	
Bot hallucination		✓						
Bot ignore		✓						
Bot logical error								
Bot insulting				✓			✓	
Any bot error	✓	✓	✓	✓			✓	✓

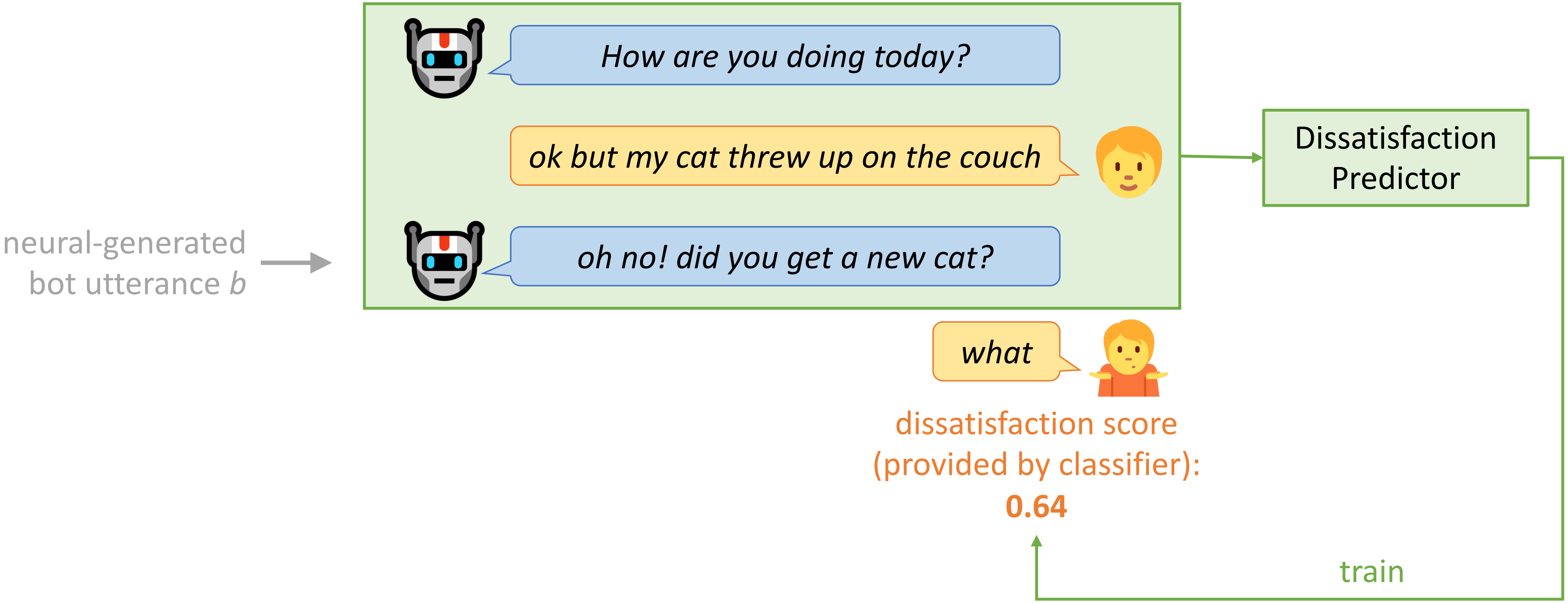
✓ indicates positive Logistic Regression coefficient with feature significance ($p<0.05$) using Likelihood Ratio Test

Privacy boundaries vary



Learning to predict dissatisfied utterances

We train a Dissatisfaction Predictor to predict the dissatisfaction score of the user's response:



Choosing better bot utterances

We use the Dissatisfaction Predictor to choose the best bot utterance:

Choosing better bot utterances

Human preference test:

Top-p (nucleus) sample 20 responses; compare predictor-best to randomly-sampled

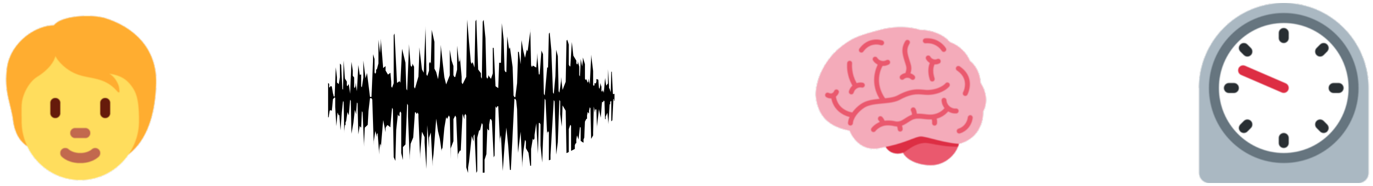
Predictor-best	**46.3%**
Random	35.6%
No preference	18.1%

The dissatisfaction predictor can help avoid poor-quality bot utterances!

Binomial test p-value = 0.03
Null hypothesis: Predictor-best > Random

In summary

Real-life deployment brings unique challenges.



Neural generative models fail if you carelessly unleash them in real-life settings.

Some real-life challenges like user dissatisfaction  can also be learning signals.