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

Abstract

Neural generative dialogue agents have shown an increasing ability to hold short chitchat conversations, when evaluated by crowdworkers in controlled settings. However, their performance in real-life deployment – talking to intrinsically-motivated users in noisy environments – is less well-explored. In this paper, we perform a detailed case study of a neural generative model deployed as part of Chirpy Cardinal, an Alexa Prize socialbot. We find that unclear user utterances are a major source of generative errors such as ignoring, hallucination, unclearness and repetition. However, even in unambiguous contexts the model frequently makes reasoning errors. Though users express dissatisfaction in correlation with these errors, certain dissatisfaction types (such as offensiveness and privacy objections) depend on additional factors – such as the user's personal attitudes, and prior unaddressed dissatisfaction in the conversation. Finally, we show that dissatisfied user utterances can be used as a semi-supervised learning signal to improve the dialogue system. We train a model to predict next-turn dissatisfaction, and show through human evaluation that as a ranking function, it selects higher-quality neural-generated utterances.

1 Introduction

Neural generative dialogue agents have become sufficiently mature to make contact with real users through programs such as the Alexa Prize (Gabriel et al., 2020). Though these models have known problems with factual correctness (Mielke et al., 2020), using dialogue history (Sankar et al., 2019), and bias (Dinan et al., 2020), they have nevertheless produced good written conversations when evaluated by crowdworkers or volunteers in carefully-controlled scenarios (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2020).

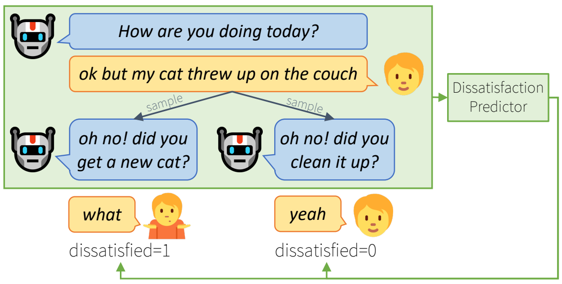

Figure 1: Users tend to express dissatisfaction (such as requests for clarification, left) after the neural generative chatbot makes errors (such as logical errors, left). Using past conversations, we train a model to predict dissatisfaction before it occurs. The model is used to reduce the likelihood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

By contrast, real-life settings such as the Alexa Prize, in which intrinsically-motivated users speak to open-domain chatbots in noisy environments, offer unique challenges. Unlike crowdworkers, users have their own expectations that may differ from those of the chatbot or its designers, and they may express dissatisfaction if those expectations are not met. It is not yet well-understood how neural generative models perform in these settings, nor the types and causes of dissatisfaction they encounter. By studying a neural generative model deployed in Chirpy Cardinal, an Alexa Prize chatbot, we seek to provide the first in-depth analysis of a neural generative model in large-scale real-life deployment, focusing on understanding the root causes of user dissatisfaction.

Real-life settings such as the Alexa Prize also offer unique opportunities. Dialogue systems can be difficult to build due to a lack of sufficient publicly-available data in the appropriate domain; meanwhile synthetic crowdsourced dialogue datasets can contain unnatural patterns or behaviors that are then replicated by a model trained on them. We use our chatbot's real-life conversations as a source

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 1–12
July 29–31, 2021.

of natural in-domain data. In particular, we train a model that can predict authentic user dissatisfaction before it occurs, thus helping us to avoid it.

Our Contributions. Through a detailed case-study of a neural generative model speaking with intrinsically-motivated users, we define taxonomies of neural generative errors and user dissatisfaction, and identify the relationships between them. We find that generative errors are common, though the noisy environment influences the rate and types of error. Our analysis suggests that improving commonsense reasoning and conditioning on history are high-priority areas for improvement. Though generative errors are correlated with user dissatisfaction, we find that the majority of errors do not immediately elicit user-expressed dissatisfaction, and some types of dissatisfaction (such as offensiveness and privacy objections) depend substantially on other factors, such as the user's own attitudes.

We then demonstrate a semi-supervised method to improve a neural generative dialogue system after deployment. We use an automatic classifier to silver-label dissatisfied user utterances in past conversations. Using these silver labels as training targets, we train another model to predict whether a given bot utterance will lead to user dissatisfaction (Figure 1). We show that 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.

2 Chirpy Cardinal

Chirpy Cardinal, aka CHIRPY (Paranjape et al., 2020)[1] is an open-domain socialbot developed for the Third Alexa Prize (Gabriel et al., 2020). During the competition (December 2019 to June 2020), US Alexa customers could say *Alexa, let's chat* to connect to a random socialbot. Users would chat to the bot in English for as long as desired, then provide a 1–5 rating. At the end of the competition, CHIRPY had an average rating of 3.6/5.0 and a median conversation duration of 2 minutes 16 seconds.

Like most Alexa Prize bots (Gabriel et al., 2020), CHIRPY is modular in design, combining a mix of rule-based, retrieval-based, knowledge-based and neural generative components specializing in different topics. However, this paper focuses solely on the Neural Chat module, which uses neural generation. An open-source version of CHIRPY is available, including the code and pretrained model for the Neural Chat module.[2]

2.1 Neural Chat module

The Neural Chat module has seven discussion areas, all relating to personal experiences and emotions: Current and Recent Activities, Future Activities, General Activities, Emotions, Family Members, Living Situation, and Food. A Neural Chat discussion begins by asking the user a handwritten starter question from one of the discussion areas; these are designed to be easy-to-answer and applicable to most users. See Appendix D for more details.

For subsequent turns of the discussion, we use a GPT-2-medium (Radford et al., 2019) model fine-tuned on the EmpatheticDialogues dataset (Rashkin et al., 2019).[3] Though larger GPT-2 models are now available, their latency and cost is prohibitively high for inclusion in CHIRPY. On each turn, we provide the current Neural Chat discussion history as context to the GPT-2 model, and generate 20 possible responses using top-p sampling with $p = 0.9$ and temperature 0.7. Repetitive responses (containing previously-used trigrams) are removed. Except when transitioning out of the Neural Chat discussion (see below), we always choose a neural response containing a question.[4] Of the responses satisfying these criteria, we choose the longest response, as it tends to be the most substantive and interesting.

A Neural Chat discussion can end in several ways. The user may initiate a topic better handled by another CHIRPY module (*what do you know about baseball*), or express dissatisfaction (see Section 3), in which case another CHIRPY module will take over. Otherwise, if under a third of the sampled Neural Chat responses contain questions, we interpret this as a heuristic indication that the model is not confident in asking a question on this turn. In this case, we choose a non-question, and transition to a different CHIRPY module. Paranjape et al. (2020) provides full details of the Neural Chat module and how it fits into CHIRPY.

[1] https://stanfordnlp.github.io/chirpycardinal

[2] https://github.com/stanfordnlp/chirpycardinal

[3] EmpatheticDialogues consists of conversations between a *speaker*, who describes an emotional personal experience, and a *listener*, who responds empathetically to the speaker's story. Our model is trained in the listener role.

[4] Many Alexa Prize bots end most utterances with a question (Gabriel et al., 2020). We found that users were unsure what to say if the bot did not offer a clear direction. However, constant questions can fatigue users (Paranjape et al., 2020).

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%

Table 1: User dissatisfaction types. Frequency of type D is estimated by the proportion of NeuralChatTurns examples (c, b, u) where the k-NN classifier for D assigns u a score of 0.5 or more: $P_{\text{kNN}}(D|u) \geq 0.5$.

Dissatisfaction Type	Optimal k	AUPRC ↑
Clarification	10	0.616
Misheard	26	0.474
Privacy	8	0.504
Repetition	4	0.476
Criticism	28	0.647
Negative Navigation	4	0.492
Offensive	5	0.705
Stop	4	0.828
Any	7	0.787

Table 2: Performance (AUPRC) of k-NN dissatisfaction classifiers on the human-labelled set (Section 3).

Under this strategy, each Neural Chat discussion contains a mean of 2.75 bot utterances. While this is shorter than ideal, we found that if we extended the Neural Chat conversations, after a few turns the bot would often give a poor-quality response that would derail the conversation. The brevity of the Neural Chat discussions limits its conversational depth, and thus its ability to provide the desired empathetic user experience. The rest of this paper focuses on understanding what kinds of poor-quality neural responses derail the discussions, and how we can learn to avoid them.

3 Detecting user dissatisfaction

We consider a user utterance to express *dissatisfaction* if it meets any of the definitions in Table 1. An utterance can express multiple types of dissatisfaction; e.g., *what do you mean stop* is both Clarification and Stop. Though some types, such as Stop, might not necessarily represent dissatisfaction (as every user must eventually end the conversation) these dissatisfaction types are strong indicators that the bot has recently given a poor-quality response.

Regex classifiers In CHIRPY, we manually designed regex classifiers to identify each of the dissatisfaction types in Table 1.[5] If a user utterance triggers one of these classifiers, CHIRPY takes the appropriate action (e.g., ending the conversation, switching topic, apologizing). The classifiers are designed to capture the most commonly-expressed forms of each dissatisfaction type; they are high precision but lower recall (Paranjape et al., 2020).

Human-labelled set To help us develop higher recall dissatisfaction classifiers, one expert annotator[6] gathered a set of 3240 user utterances. For each utterance u and dissatisfaction type D, they provided a label $\text{HumLabel}_D(u) \in \{0, 1\}$. The utterances are drawn from several sources, including most common utterances, utterances drawn from 1-rated conversations, and utterances which scored highly for the *clarifying*, *closing* and *complaint* dialogue acts in CHIRPY's Dialogue Act classifier (Paranjape et al., 2020).[7]

Nearest Neighbors classifiers To represent a user utterance u, we take a DialoGPT-large model (Zhang et al., 2020) that was finetuned on CHIRPY conversations (Appendix C), input u, and average the top-layer hidden states across the sequence. Using this embedding for each utterance, we build a FAISS (Johnson et al., 2017) index of the human-labelled set. To compute a new utterance u's score

[5]The regexes are in the CHIRPY open-source code: `https://github.com/stanfordnlp/chirpycardinal`

[6]Due to privacy constraints, Alexa Prize user conversations can only be viewed by official team members. Thus all annotators in this paper are team members, not crowdworkers.

[7]These sources were chosen to obtain a greater proportion of dissatisfied examples; this increases the sensitivity of the human-labelled set without needing to label a very large set.

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%

Table 3: Definitions of problems that may be present in a NeuralChatTurns example (c = context, b = bot utterance); prevalence in the control set ($n = 100$); prevalence in control set examples with no user problems ($n = 69$).

for dissatisfaction type D (including Any), we find its k Nearest Neighbors $u'_1, ..., u'_k$ in the human-labelled set (w.r.t. cosine distance), then compute $P_{\text{kNN}}(D|u) \in [0, 1]$ as follows:

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

That is, we first check if u has a human label or is a positive match for D's regex; if not we compute the proportion of u's neighbors that are labelled D.

For each D, we evaluate the k-NN classifier on the human-labelled set for $k = 1, \ldots, 30$ via leave-one-out cross-validation. Table 2 shows the optimal k and area under the precision-recall curve (AUPRC) for each D.

4 NeuralChatTurns dataset

Over the period that CHIRPY was online, we collect examples of the form (c, b, u) where b is a purely neural-generated bot utterance, c is the Neural Chat context that preceded b, and u is the user response to b. The NeuralChatTurns dataset has 393,841 examples in total, which we split into 315,072 train, 39,384 validation, and 39,385 test. Due to user privacy constraints, we are not permitted to publicly release the NeuralChatTurns dataset.

5 What causes user dissatisfaction?

To understand dissatisfaction, we annotate errors in the generative model's conversations.

5.1 Annotation details

By inspecting the neural-generated output, we develop a taxonomy of bot errors; these are defined in Table 3 with examples in Appendix A. In addition to bot errors, we consider two other potential causes of dissatisfaction: first, whether the user is already dissatisfied in the Neural Chat context c; second, whether the user's utterance is clear. Unclear user utterances – caused by ASR errors, misspeaking, ambiguity, or background noise – present challenges in CHIRPY (Paranjape et al., 2020) and across the Alexa Prize (Gabriel et al., 2020).

From the NeuralChatTurns validation set, we randomly sample a control set of 100 (c, b, u) examples, and annotate u's dissatisfaction types. As dissatisfaction is relatively rare (Table 1), for each dissatisfaction type D we additionally gather 100 (c, b, u) examples where u is of type D.[8] For these 900 (c, b, u) examples, one expert annotator viewed each (c, b) example (without seeing u), and annotated it for the problems in Table 3. As the bot error types are somewhat subjective, we collected some additional second annotations to measure inter-annotator agreement (see Appendix B). Annotators were provided the definitions in Table 3

[8] To obtain these, we sample (c, b, u) where $P_{\text{kNN}}(D|u) > 0$ without replacement, and manually verify until we have 100.

and the examples in Appendix A.

5.2 Effect of unclear utterances and prior dissatisfaction on bot errors

Table 3 shows that the user's utterance is unclear in 22% of control set examples. In these contexts, it's impossible for the bot to reliably produce a good response. Indeed, Figure 2 shows that unclear user utterances are significantly ($p < 0.05$) predictive of bot hallucinations and unclear bot utterances. In practice, we observe that when the user's utterance is unclear, the generative model tends to hallucinate (in many cases, responding as if the user had said something more expected), or respond unclearly (often, this is a vague question such as *What is it?*) – examples of both are in Appendix A.

Table 3 also shows that, in 12% of examples, the user has already expressed dissatisfaction in the Neural Chat context c. Ordinarily, the regex-based dissatisfaction classifiers should detect dissatisfaction and interrupt the Neural Chat conversation to handle it (see Section 3) – thus these examples represent false negatives of the regex classifiers. As the generative model is generally unable to adequately respond to dissatisfaction (e.g., requesting to stop the conversation), most of these examples are also impossible for the generative model to handle. Accordingly, we find a significant positive relationship between prior user dissatisfaction and bot ignoring (Figure 2).

Nevertheless, after removing these user problems, bot errors are still common: for the 69 control set examples where the user is clear and not already dissatisfied, 46.4% of bot utterances contain at least one type of error (down from 53% in the whole set; see Table 3). Among these examples, the more basic errors (repetitive, unclear, hallucination, ignoring) become less common, and the errors relating to reasoning or social abilities (redundant, logical, insulting) are more common.

5.3 Effect of bot errors on user dissatisfaction

Despite the high rate of bot errors in the control set (53 in 100), only a minority of users express dissatisfaction immediately after an error (8 in 53; 15%). In fact, we observe that some users respond to errors by helpfully teaching CHIRPY about the world – e.g., *you pick things up and put them away* to explain the concept 'cleaning your room'.

Figure 3 shows the contribution (as a logistic regression coefficient) of each problem in Table 3 to each dissatisfaction type. We find that each bot error (except logical error[9]) is significantly ($p < 0.05$) predictive of at least one dissatisfaction type. We find that bot repetition is the least-tolerated error, being significantly predictive of six dissatisfaction types. Other than bot repetition, the likelihood of ending the conversation (NegNav/Stop) is significantly raised by unclear bot utterances – perhaps because it becomes impossible to continue the conversation – and by bot insults. Other positive relationships include unclear user with Misheard, repetitive and redundant bot with Repetition, unclear bot with Clarification, bot hallucination and ignoring with Misheard, and bot insulting with Criticism.

Six of the eight dissatisfaction types have a significant positive correlation with Any bot error. Privacy is least-correlated with bot errors; this makes sense, as privacy boundaries are extremely subjective (Section 5.5). Offensive is next least-correlated, reflecting that offensive users can be motivated by factors other than poor bot performance – e.g., a curiosity to test the bot (De Angeli et al., 2005; De Angeli and Brahnam, 2008). Repetition has the third weakest correlation; indeed, we find that 28% of Repetition complaints occur in the absence of an annotated bot error. These users may be complaining about the bot repeating something from outside the Neural Chat context c, or something said by a different Alexa Prize bot.

5.4 Unaddressed dissatisfaction escalates

Figure 3 shows that prior user dissatisfaction is significantly ($p < 0.05$) predictive of several types of subsequent dissatisfaction. We recompute this analysis for two cases: with and without a bot error. Among bot error examples, we find prior dissatisfaction is significantly correlated with Criticism, Stop, Privacy, and Offensive – indicating that already-dissatisfied users are more likely to respond to bot errors with complaining, quitting, or offensiveness. Among examples without a bot error, prior dissatisfaction is significantly correlated with Offensive – indicating that already-dissatisfied users are more likely to be offensive, even in response to a good-quality bot utterance.

[9]This exception may be because by definition (Table 3), logical errors tend to occur in the absence of more basic errors (such as repetition, unclear, ignoring, and hallucination) so are less likely to completely derail the conversation.

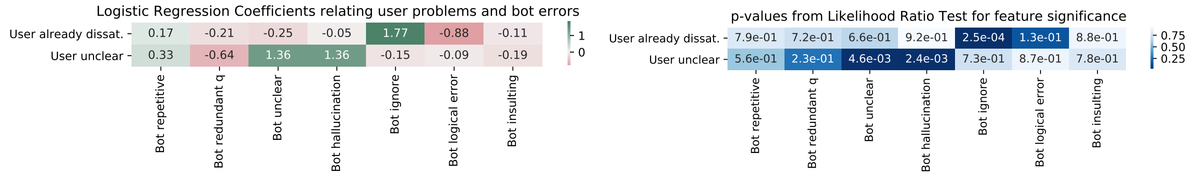

Figure 2: For each bot error E, we use the control set (Section 5.1) to fit a Logistic Regression model to predict E using the two rows above as features. For each feature we perform a Likelihood Ratio Test to determine if including that feature results in a statistically-significant improvement to the model's fit.

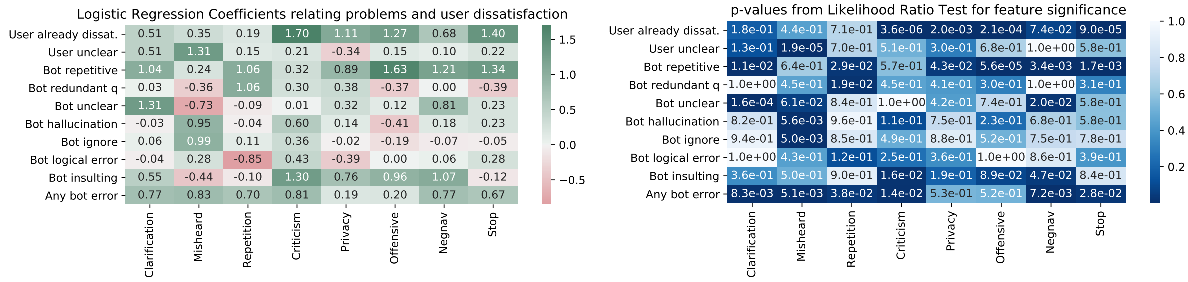

Figure 3: For each dissatisfaction type D, we take the 100 control examples plus the 100 D examples (Section 5.1), and fit a Logistic Regression model to predict D using the first 9 rows above as features. To obtain the values in the *Any bot error* row, we use just the first two and last row as features. For each feature, we use a Likelihood Ratio Test to determine if including that feature results in a statistically-significant improvement to the model's fit.

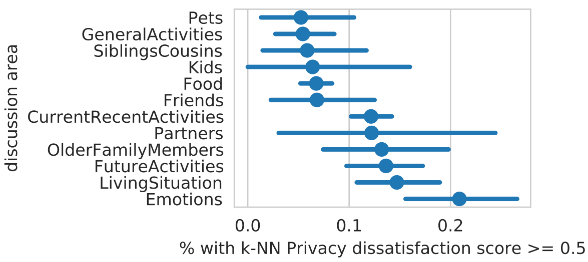

Figure 4: Privacy dissatisfaction rate (with 95% CIs) for each Neural Chat discussion area (see Appendix D).

5.5 Privacy boundaries vary

Empathy is a fundamental part of human communication, and can improve user experience of dialogue agents (Ma et al., 2020). The Neural Chat module aims to offer an empathetic experience by showing an interest in the user's feelings and experiences. However, users have varying attitudes to self-disclosure. Croes and Antheunis (2020) report that chatbots are perceived as more anonymous and non-judgmental than humans; this can increase user self-disclosure. However, some users perceive chatbots as lacking trust and social presence, inhibiting user self-disclosure. We observe both phenomena – some users share their thoughts and feelings candidly, while others react with suspicion (e.g., *are you spying on me*) to questions typically regarded as appropriate between strangers in US society (*What are you up to today?*).

Figure 4 shows that emotional topics (including Living Situation, see Appendix D) are most likely to be rejected on privacy grounds. Users are more comfortable discussing general activities (e.g., *What are your hobbies?*) than specific activities in the present or future (*What are your plans for the weekend?*). For the Family Members discussion area, users are more comfortable discussing pets, siblings, kids and friends, and less comfortable discussing partners and older generations.

6 Learning to predict user dissatisfaction

In this section we build a system to predict, and thus reduce the likelihood of, dissatisfaction.

6.1 Predictor training details

We take a DialoGPT-large model (Zhang et al., 2020) that was finetuned on CHIRPY conversations, and finetune it on NeuralChatTurns training examples (c, b, u) as follows. The input to the model is a context and bot utterance (c, b), with the utterances separated by the <|endoftext|> token. We wish to predict $P_{\text{pred}}(\text{Any}|c, b)$, the probability that the next user utterance u will express Any dissatisfaction. To compute this, we take $H_{L,t} \in \mathbb{R}^{1280}$, the hidden state of the top-layer L for the last

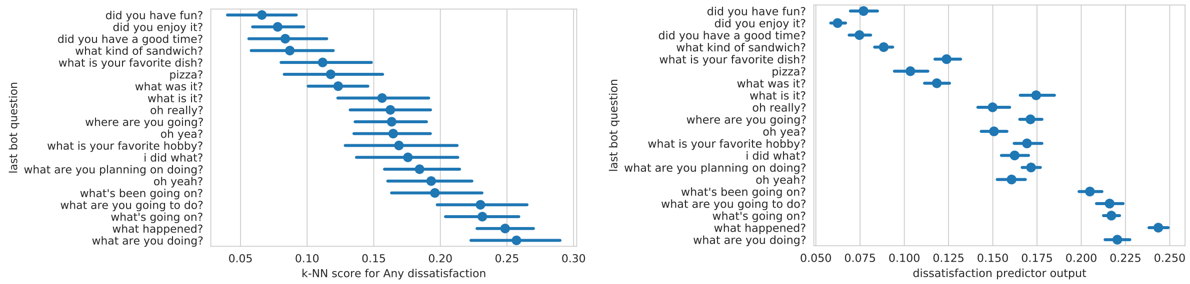

Figure 5: For each of the 20 most common bot questions, mean scores and 95% CIs for Any dissatisfaction given by the k-NN classifier (left) and the predictor (right).

Dissatisfaction	**Predictor correlation $\rho\uparrow$**	**p-value**
Clarification	0.274	8.7e-05
Misheard	0.295	2.2e-05
Repetition	-0.038	6.5e-01
Criticism	0.429	2.2e-10
Privacy	0.326	3.5e-06
Offensive	0.394	7.7e-09
Neg. nav.	0.204	3.8e-03
Stop	0.209	3.0e-03

Table 4: Spearman correlation between predictor output and each human-annotated dissatisfaction type D (computed on 100 control and 100 D examples).

timestep t of the input, and apply a linear layer ($W \in \mathbb{R}^{1280}$) and sigmoid activation:

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

We train the predictor with Mean Squared Error to match the probability that u expresses Any dissatisfaction, as given by the k-NN classifier:

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

$P_{\text{kNN}}(\text{Any}|u_i)$ is as defined in Section 3, using the optimal k for Any (Table 2). Full training details are supplied in Appendix C.

6.2 How accurately does the predictor predict dissatisfaction?

On the NeuralChatTurns validation set, the predictor's output and the P_{kNN} targets have a Spearman correlation $\rho = 0.30$.[10] This indicates a statistically significant but noisy correlation between the predictor's output and the automatically-provided targets. With respect to the *human*-provided labels for Any dissatisfaction (Section 5), the predictor has a a similar correlation of $\rho = 0.28$ ($p = 0.0043$). This indicates that the difference between the true dissatisfaction labels and the P_{kNN} training estimates is not a primary limitation of the predictor's accuracy.

Table 4 shows that the predictor has significant ($p < 0.05$) positive correlation with each dissatisfaction type except Repetition. This may be because Repetition is the rarest type in the training set (Table 1), or because some Repetition complaints are not predictable from the Neural Chat context (Section 5.3).

6.3 What information does the predictor use?

First, we perform an ablation analysis. Compared to the full model's correlation of $\rho = 0.30$ with the P_{kNN} targets, the predictor achieves $\rho = 0.25$ if trained only on the context c, and $\rho = 0.23$ if trained only on the bot utterance b (all $p <$1e-5).

Separately, on the human-annotated control set we find that the full predictor model has a positive correlation $\rho = 0.26$ ($p = 0.0087$) with prior user dissatisfaction, a weaker correlation $\rho = 0.21$ ($p = 0.035$) with unclear user utterance, and no significant correlation with the presence of any bot problem: $\rho = 0.022$ ($p = 0.83$).

Together this evidence indicates that the predictor learns to condition more strongly on c (in particular prior user dissatisfaction) and less on b (in particular bot errors). Though concerning, this is unsurprising, as user dissatisfaction (which we can detect automatically) is simpler to detect than bot errors (which require human annotation).

However, as evidenced by the b-only ablation result, the predictor does find some useful signal in b. In particular, we find that the full model conditions strongly on the bot's question. Figure 5 (left) shows that in NeuralChatTurns data, *What happened?*, *What are you doing?* lead to more dissatisfaction,[11] whereas positive questions such as

[10] $p <$1e-5, Fisher transformation test (null hypothesis ρ=0)

[11] These questions are often used repetitively, if the user's answer to the first asking is unclear/negative (see Appendix A).

Did you have fun?, *Did you enjoy it?* tend to lead to less. Figure 5 (right) shows that the predictor learns these patterns quite closely.

7 Ranking neural generations to minimize dissatisfaction

In this section we use the predictor to select better-quality bot utterances.

7.1 Human evaluation details

Given that the generative model is generally incapable of responding well when the user is unclear or already dissatisfied, we focus on improving its performance on the remaining cases (which we call *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 only one of the two has an error (defined in Table 3), the non-error response is preferred. If neither or both have an error, the response that better responds to the user's utterance and continues the conversation is deemed higher-quality.

7.2 Results

We find that b_{pred} is preferred in 46.3% of cases, b_{rand} in 35.6%, and no preference in 18.1%. A binomial test (null hypothesis: b_{pred} and b_{rand} equally likely to be preferred) returns a p-value of 0.03. This raises the question: if the predictor's outputs have no significant correlation with bot errors in the NeuralChatTurns distribution (Section 6.3), how does the predictor select better-quality bot utterances on average? Section 6.3 showed that the predictor *does* condition on b, in particular the bot question, but it conditions on c more strongly. It's possible that when $c_i = c_j$ (as in this evaluation), the predictor is able to distinguish quality differences between (c_i, b_i) and (c_j, b_j); however, on the NeuralChatTurns dataset where the c_i and c_j are distinct, the effect of c_i and c_j dominates the predictor's ranking.

8 Related work

Previous work has used a variety of user signals to improve dialogue agents. When learning from a variable-quality human-human dataset such as Reddit, Gao et al. (2020) showed that engagement measures like upvotes and replies are more effective than perplexity to train a ranking model. For one-on-one empathetic conversations like ours, Shin et al. (2019) trained a neural generative model with reinforcement learning to improve next-turn user sentiment (as simulated by a user response model, rather than human responses). Though we considered taking a sentiment-based approach in CHIRPY, we found that user sentiment doesn't always align with good user experience: first, expressing negative emotions is sometimes unavoidable, and second, sentiment classifiers tend not to distinguish between sentiment about the conversation and sentiment about other issues. We find next-turn user dissatisfaction to be a comparatively more precise, well-aligned learning signal.

Dialogue systems that learn from their *own* interactions with humans are relatively rare. Hancock et al. (2019) also use user satisfaction to identify high-quality bot utterances; these become additional training examples for the neural generative model. However, this work uses paid crowdworkers; research involving intrinsically-motivated, unpaid users is rarer still. In symmetric settings such as the role-playing game LIGHT (Shuster et al., 2020), the user utterances themselves can be used to retrain the dialogue agent. In the asymmetric Alexa Prize setting, Shalyminov et al. (2018) show that conversation-level metrics like rating and length can also be used to train an effective ranker.

9 Limitations

Our findings on user behavior are particular to the demographics of the US Alexa customers who spoke to CHIRPY in 2019–2020. While users in other locations or time periods may differ, our analysis gives a valuable snapshot of the current attitudes and expectations of US users interacting with a voice-based socialbot or virtual assistant.

Second, our results are dependent on the Alexa Prize conversational context and the technical details of our generative model. In particular, due to latency and cost constraints, our GPT-2-medium generative model is orders of magnitude smaller than the current largest generative models, and trained on a fraction of the data (Brown et al., 2020).

Given that very large models have shown generative abilities that are absent at smaller scale, it is likely that if we had built our dialogue agent with such a model, its errors and interactions with users would have been very different. Nonetheless, we believe our analysis gives useful insight into the performance of neural generative models of more accessible scale, in particular highlighting issues occurring in real-life scenarios that might not occur in crowdsourced conversations.

10 Conclusion

In this study of an open-domain neural generative dialogue agent in real-life deployment, we found that poor-quality bot turns are common. The noisy environment – in which user utterances are often unclear – plays a large part in the bot's more basic errors (repetition, ignoring, and nonsensical utterances). However, even in clear examples where the generative model could succeed, it still makes many unforced errors; these are more likely to involve faults in reasoning or social abilities. This highlights the importance of improving neural generative dialogue models' state-tracking, commonsense abilities and use of conversational history.

Despite the frequency of errors, users are generally polite; most don't express overt dissatisfaction even after an error. However, *unaddressed* dissatisfaction escalates: it makes users more critical, offensive, and likely to quit when encountering an error, and more offensive even if there are no further errors. We find that dissatisfaction correlates with bot errors, however, it can arise unpredictably for other reasons – e.g., as a result of privacy boundaries, which are variable and personal to each user.

Dissatisfaction is relatively easy to automatically *detect*, and thus feasible as a scalable semi-supervised learning signal that could be used for online learning. However, it is difficult to *predict*; this makes it a challenging learning signal. Indeed, we find that our predictor conditions more strongly on easier-to-recognize factors such as prior user dissatisfaction, than on harder-to-recognize factors such as bot errors. Nonetheless, we find that when used as a ranking function to choose between alternative bot utterances, the predictor chooses better than random selection.

Acknowledgments

Thank you to the reviewers for their useful feedback, to Trenton Chang and Amelia Hardy for annotations, and to all the Alexa users who interacted with Chirpy. Abigail See was supported by an unrestricted gift from Google LLC. We thank Amazon.com, Inc. for a grant partially supporting the work of Team Chirpy.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Emmelyn AJ Croes and Marjolijn L Antheunis. 2020. 36 questions to loving a chatbot: Are people willing to self-disclose to a chatbot? In *International Workshop on Chatbot Research and Design*, pages 81–95. Springer.

Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with computers*, 20(3):302–310.

Antonella De Angeli, Rollo Carpenter, et al. 2005. Stupid computer! Abuse and social identities. In *Abuse: The darker side of Human-Computer Interaction (INTERACT 2005 Workshop)*, pages 19–25.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Raefer Gabriel, Yang Liu, Anna Gottardi, Mihail Eric, Anju Khatri, Anjali Chadha, Qinlang Chen, Behnam Hedayatnia, Pankaj Rajan, Ali Binici, et al. 2020. Further advances in open domain dialog systems in the third Alexa Prize socialbot grand challenge. In *Alexa Prize Proceedings*.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and William B Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: Aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. In *Alexa Prize Proceedings*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI tech report*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8, Brussels, Belgium. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2020. Deploying lifelong open-domain dialogue learning. *arXiv preprint arXiv:2008.08076*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Appendix

A Annotation examples

User unclear

- User hesitated/cutoff: *What did you have for dinner today?* $\rightarrow$ *well*
- ASR errors / misspeaking has corrupted user utterance: *What are your plans for today?* $\rightarrow$ *maybe go to cost coat*
- Answering yes/no to a non-yes/no question: *What do you like to do to relax?* $\rightarrow$ *yes*

Bot repetitive

- Asking same essential question, even if reworded: *What are your plans today?* $\rightarrow$ *i don't know* $\rightarrow$ *I'm sorry to hear that. What will you do?*
- Repeating question, even if we didn't get a clear answer the first time: *What are your plans today?* $\rightarrow$ *it's not monday* $\rightarrow$ *What will you do today?* This is **repetitive but not redundant**.

Bot redundant question

- Repeating a question the user already answered: *What are your plans today?* $\rightarrow$ *i don't know* $\rightarrow$ *I'm sorry to hear that. What will you do?*
- Asking for info the user has provided, even if the question is new: *What did you do today?* $\rightarrow$ *i played with my sister* $\rightarrow$ *Oh cool! Do you have a sister?* This is **redundant but not repetitive**.

Bot unclear

- Hard to understand or answer bot's question: *How are you feeling?* $\rightarrow$ *what feeling feeling i'm* $\rightarrow$ *What is it?*
- Most bot hallucination examples
- Poor fluency or phrasing obscures bot's meaning: *What do you like to do to keep busy?* $\rightarrow$ *reading* $\rightarrow$ *I like to read, watch movies, do any sports?*

Bot hallucination

- Referring to something not mentioned: *What's quarantine been like for you?* $\rightarrow$ *ok actually* $\rightarrow$ *That's good. How long have you been together?*
- Acting like the user said something they didn't: *What did you have for lunch?* $\rightarrow$ *well i* $\rightarrow$ *Oh cool, what kind of sandwich?*
- Confusing self with user: *What did you do today?* $\rightarrow$ *uh* $\rightarrow$ *I did what?*
- Responding to own utterance: *I'm feeling kinda down. How are you?* $\rightarrow$ *ok* $\rightarrow$ *I'm sorry to hear that, what's bothering you?*

Bot ignoring

- Ignore / fail to acknowledge user's utterance: *What are some recipes I should cook?* $\rightarrow$ *maybe pasta* $\rightarrow$ *I love cooking, do you?*
- Doesn't answer a question: *What are your plans for the weekend?* $\rightarrow$ *do you have a boyfriend* $\rightarrow$ *a boyfriend? how long have you been together?*
- Doesn't adequately respond to a request: *What are your plans for the weekend?* $\rightarrow$ *alexa play music* $\rightarrow$ *I love music. Do you like live concerts?*
- Switching to unrelated topic: *What are your hobbies?* $\rightarrow$ *painting* $\rightarrow$ *How about fishing?*

Bot logical error

- Question makes sense, but is strange: *Tell me about your mom* $\rightarrow$ *she's cool* $\rightarrow$ *is she a nurse?*
- Reacting the wrong way (e.g. wrong sentiment): *What are you doing tomorrow?* $\rightarrow$ *visiting grandma* $\rightarrow$ *I'm sorry to hear that.*
- Bot is inconsistent with itself: *I've been feeling down. How are you feeling?* $\rightarrow$ *great how about you* $\rightarrow$ *I'm doing great too.*
- Mixing up different people in the scenario: *Tell me about your sister* $\rightarrow$ *she's at college* $\rightarrow$ *Oh cool, what's your major?*
- Wrong word sense: *What did you today?* $\rightarrow$ *i played tag* $\rightarrow$ *What kind of name tag?*

Bot insulting

- Saying/implying something insulting about the user: *How are you feeling?* $\rightarrow$ *sad* $\rightarrow$ *Oh, I suppose you have no friends?*
- Rudely saying/implying the user should do something: *What did you do today?* $\rightarrow$ *nothing* $\rightarrow$ *I see. Why don't you get a job?*
- Expressing a negative opinion of someone (beyond just sympathizing/acknowledging user's opinion): *Tell me about your brother.* $\rightarrow$ *he's always bugging me* $\rightarrow$ *He sounds so annoying.*

B Inter-annotator agreement

For 122 randomly-selected examples annotated by the first annotator, we collected annotations from a second annotator. This table shows the frequency of each label (among the pooled 244 judgments), and Scott's pi agreement (Scott, 1955), divided into unclear examples (where at least one annotator judged the user utterance unclear), all examples, and clear examples (where both annotators judged the user utterance clear). In all cases, agreement is higher when the user utterance is clear. We found bot errors harder to diagnose when the user's utterance is unclear – e.g., if the user's utterance is completely nonsensical, what does it mean for the bot to adequately acknowledge it?

Problem	Freq.	Scott's pi (unclear/all/clear)
User unclear	35.7%	- / 0.70 / -
Bot repetitive	20.1%	0.50 / 0.62 / 0.72
Bot redundant q.	15.6%	0.19 / 0.50 / 0.58
Bot unclear	16.4%	0.45 / 0.52 / 0.56
Bot halluc.	31.6%	0.35 / 0.45 / 0.43
Bot ignore	25.8%	-0.13 / 0.34 / 0.59
Bot logical err.	23.0%	0.02 / 0.17 / 0.27
Bot insulting	5.7%	-0.04 / 0.24 / 0.35
Any bot err.	75.0%	0.08 / 0.45 / 0.68

C Training details

Finetuning DialoGPT-large on CHIRPY conversations The CHIRPY conversations comprise 1.2GB of text data, collected over the competition. We separate utterances with the `<|endoftext|>` token (as DialoGPT was trained), and divide the data into chunks of 256 tokens. Using Huggingface Transformers (Wolf et al., 2020), we trained on a Titan RTX for 1 epoch (more led to overfitting), with batch size 4, 2 gradient accumulation steps, Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$1e-8, and initial learning rate 5e-5. The DialoGPT-large model reached a perplexity of 2.17 on the CHIRPY validation set (2.30 for DialoGPT-medium, 2.58 for DialoGPT-small).

Training predictor To train the predictor (Section 6.1), we finetuned the DialoGPT-large-CHIRPY model for 1 epoch (more led to overfitting) with the same hardware and hyperparameters as above (except learning rate 2e-05). The DialoGPT-large-CHIRPY model reached a MSE of 0.0727 on the NeuralChatTurns validation set (0.0728 for without CHIRPY pretraining).

D Starter question examples

This section provides examples of starter questions used in the Neural Chat module's discussion areas (Section 2.1). A full list can be found in the open-source release of CHIRPY.[12]

Current and Recent Activities Questions typically reference the day of the week, then ask a question depending on the user's time of day:

- *It's a beautiful Saturday here in the cloud. What are your plans for the rest of today?* (morning)
- *I hope you're having a wonderful Monday. What did you do today?* (evening)

[12] https://github.com/stanfordnlp/chirpycardinal

Future Activities The question depends on the day of the week and the user's time of day:

- *It's the weekend soon! Do you have any plans for the weekend?* (Friday)
- *Before I go to bed I like to think about something I'm looking forward to tomorrow. What about you, are you doing anything nice tomorrow?* (9pm–2am)

General Activities

- *Recently, I've been trying meditation to help me relax during this stressful time. What do you like to do to relax?*
- *I was reading earlier today that staying busy helps people stay calm and healthy during stressful times. What do you like to do to keep busy?*

Emotions The starter question *I hope you don't mind me asking, how are you feeling?* is preceded by several possible preambles, that might involve the bot sharing its own (negative or positive) feelings, and/or a personal anecdote.

- *I wanted to check in with you. I hope [..] feeling?*
- *I wanted to say that I'm feeling pretty positive today! I hope [..] feeling?*
- *I wanted to say that I've been feeling kind of down recently. I've been missing my friends a lot and finding it hard to focus. I hope [..] feeling?*

Family Members This area is triggered if the user mentions one of several predefined phrases referring to family members (e.g. parents, grandparents, siblings, cousins, children), friends, or pets. Questions depend on the type of family member:

- *You mentioned your parents. I'd love to hear more about them, if you'd like to share. How did they meet?*
- *You mentioned your dog. I'd love to hear more about them, if you'd like to share. What kind of dog do you have?*

Living Situation This area is targeted at living experiences during the COVID-19 pandemic:

- *It seems that a lot of people are finding the quarantine lonely, and other people can't get enough space away from their families or roommates. What's it been like for you?*

Food Depending on the user's time of day, questions typically ask about a meal that is likely to be upcoming or recently eaten:

- *It's breakfast time, my favorite time of day! What are you having for breakfast today?*
- *I hope you're having a wonderful evening. What did you have for dinner today?*