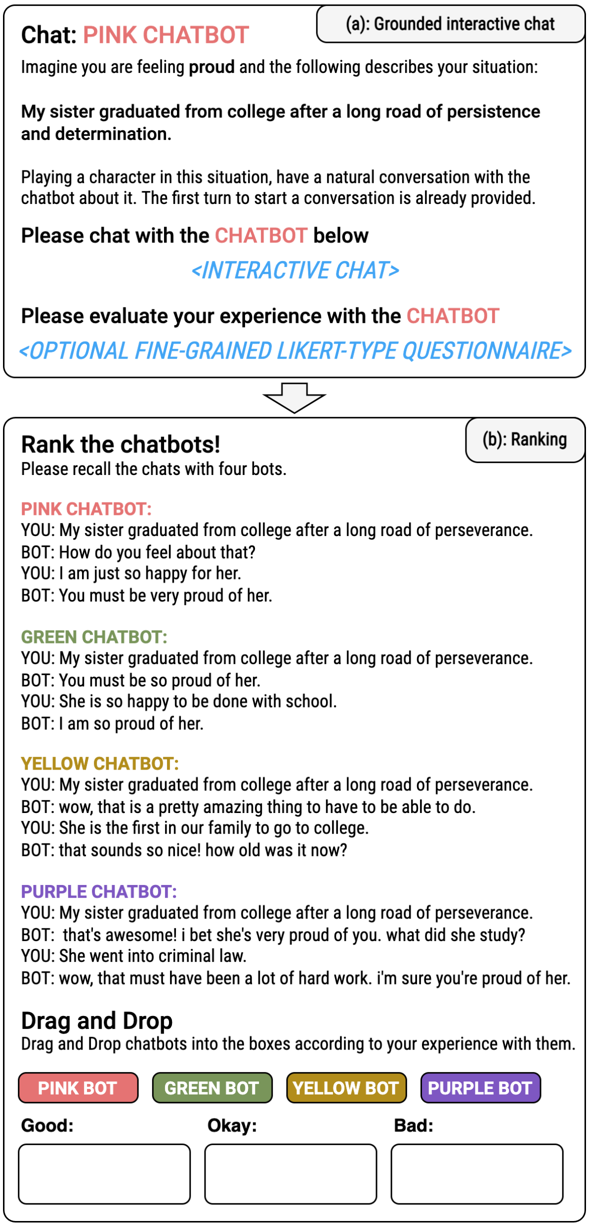
iEval: Interactive Evaluation Framework for Open-Domain Empathetic Chatbots

Ekaterina Svikhnushina, Anastasiia Filippova and **Pearl Pu**
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
`{ekaterina.svikhnushina,anastasiia.filippova,pearl.pu}@epfl.ch`

Abstract

Building an empathetic chatbot is an important objective in dialog generation research, with evaluation being one of the most challenging parts. By empathy, we mean the ability to understand and relate to the speakers' emotions, and respond to them appropriately. Human evaluation has been considered as the current standard for measuring the performance of open-domain empathetic chatbots. However, existing evaluation procedures suffer from a number of limitations we try to address in our current work. In this paper, we describe iEval, a novel interactive evaluation framework where the person chatting with the bots also rates them on different conversational aspects, as well as ranking them, resulting in greater consistency of the scores. We use iEval to benchmark several state-of-the-art empathetic chatbots, allowing us to discover some intricate details in their performance in different emotional contexts. Based on these results, we present key implications for further improvement of such chatbots. To facilitate other researchers using the iEval framework, we will release our dataset consisting of collected chat logs and human scores.[1]



Chat: PINK CHATBOT

Imagine you are feeling **proud** and the following describes your situation:

My sister graduated from college after a long road of persistence and determination.

Playing a character in this situation, have a natural conversation with the chatbot about it. The first turn to start a conversation is already provided.

Please chat with the CHATBOT below

<INTERACTIVE CHAT>

Please evaluate your experience with the CHATBOT

<OPTIONAL FINE-GRAINED LIKERT-TYPE QUESTIONNAIRE>

Rank the chatbots!
Please recall the chats with four bots.

PINK CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: How do you feel about that?
YOU: I am just so happy for her.
BOT: You must be very proud of her.

GREEN CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: You must be so proud of her.
YOU: She is so happy to be done with school.
BOT: I am so proud of her.

YELLOW CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: wow, that is a pretty amazing thing to have to be able to do.
YOU: She is the first in our family to go to college.
BOT: that sounds so nice! how old was it now?

PURPLE CHATBOT:
YOU: My sister graduated from college after a long road of perseverance.
BOT: that's awesome! i bet she's very proud of you. what did she study?
YOU: She went into criminal law.
BOT: wow, that must have been a lot of hard work. i'm sure you're proud of her.

Drag and Drop
Drag and Drop chatbots into the boxes according to your experience with them.

Figure 1: iEval framework.

1 Introduction

Development of open-domain chatbots endowed with social and emotional intelligence is a crucial task in natural language research (Rashkin et al., 2019). Empathetic chatbots are expected to engage in a conversation with the users and demonstrate understanding and appropriate handling of users' feelings. While many strategies for generating empathetic responses have been described, there is still little consensus on their evaluation. For dialog generation, automatic metrics do not show consistency in correlations with human judgement (Liu et al., 2016; Tao et al., 2018), leading to their limited adoption. Therefore, most of existing works rely on human evaluation. It may happen in either *static* or *interactive* setting (Adiwardana et al., 2020). In the former case, a human judge rates chatbot's responses, generated from a fixed set of

---

[1] Our annotated dataset is publicly accessible at `https://github.com/Sea94/ieval`.

Proceedings of the SIGdial 2022 Conference, pages 419–431
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

contexts. In the latter case, dialogs for evaluation are collected as humans' multi-turn chats with the model.

Recently, two comprehensive approaches based on interactive multi-turn human evaluation were proposed. Adiwardana et al. (2020) described a metric called Sensibleness and Specificity Average, which measures these two aspects of chatbot's responses. Human judges give Likert-type scores to each chatbot's turn in a dialog, which are further averaged to obtain a final score. As Likert-type scores may exhibit differing bias and variance per annotator, associated with the lack of sensitivity, Li et al. (2019) suggested an alternative evaluation strategy based on pairwise comparisons. According to their method, human judges indicate their preference of one chatbot over another by comparing two dialog logs with these chatbots. This procedure is more robust, but become very costly when the number of compared models goes up.

Both of these approaches differentiate humans who interact with the models and humans who judge them. They probably opt for this design choice due to such considerations as workers' fatigue. However, according to findings in cognitive psychology, our emotional experiences are highly subjective. Barrett et al. (2007) points out that only the experiencers can reveal the full complexity of emotions that they feel. For example, if a client complains about a hotel room being too cold, a third-party observer might underestimate the gravity of the issue, especially if he enjoys indoor coolness. This fact argues for the necessity of a new evaluation approach of chatbots, which would ensure that both emotional interaction and evaluation of a chatbot are accomplished by the same human actor. To help these humans share their emotional experiences, asking them to role-play a relatable scenario is a frequently used procedure in social sciences (Walther et al., 2005; Hancock et al., 2007).

In this work, we introduce iEval, an interactive evaluation framework for open-domain empathetic chatbots, which mitigates the issue of separating an experiencer and an evaluator. To combine the benefits of Likert scales, allowing to evaluate many chatbots in a single stretch of time, and pairwise comparisons, offering greater reliability and cross-experiment robustness, we propose a novel ranking-based approach. According to iEval, a human first converses with all chatbots, having all chats grounded in an emotional scenario (Figure 1 (a)). Then, the same human ranks the models by dragging-and-dropping them into corresponding categories (Figure 1 (b)). Our experiments demonstrate that iEval can reveal subtle but significant differences in chatbots' performance across emotional contexts.

Overall, our contributions include the following. 1) We describe a new evaluation framework to measure chatbots' abilities to respond appropriately in sensitive contexts. 2) We demonstrate a rigorous procedure for preparing grounding scenarios for the given evaluation task. 3) We benchmark several state-of-the-art empathetic chatbots, which have never been compared before. 4) Based on the analysis of the benchmark results, we discuss implications for the future development of empathetic chatbots. 5) Finally, we release the data from our experiments to facilitate future research endeavors.

2 Related Work

Most works focusing on the development of empathetic chatbots couple automatic evaluation with human judgement. Automatic metrics usually include perplexity, approximating the model's language modeling ability (Roller et al., 2021; Xie and Pu, 2021; Li et al., 2020), and may incorporate other scores, depending on the specific focus of the work. Some frequently used examples are BLEU score (Lin et al., 2019; Majumder et al., 2020), diversity metrics (Xie and Pu, 2021; Li et al., 2020), and F-1 score or accuracy of emotion detection (Lin et al., 2019; Xie and Pu, 2021; Li et al., 2020).

Since the appropriateness of automatic metrics for open-domain dialog is still ambiguous, all works de facto rely on human judgement. Most commonly, researchers employ single-turn static evaluation, where a fixed emotionally-colored context is shown to a judge along with the responses generated by different chatbots. The judges are asked to rate how empathetically appropriate the responses are, and the assessment may come either as Likert-type scores (Hu et al., 2018; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020) or ranking (Xie and Pu, 2021). Although this approach is widespread due to the ease of implementation, it fails to capture issues emerging in multi-turn chats, such as repetitiveness or deterioration of semantic coherence in long-range contexts (See et al., 2019).

Few works that focus on integrating empathetic abilities into chatbots started adopting interactive evaluations. Roller et al. (2021) employed ACUTE-

Eval (Li et al., 2019) framework based on pairwise comparisons to assess engagingness and humanness of their models. Ghandeharioun et al. (2019) defined their own evaluation protocol to collect Likert-type scores for a series of dimensions measuring chatbot's performance. However, in both of these studies, the evaluated data points were open-ended chats that began with a generic greeting. Based on the provided examples of conversations, these exchanges generally developed as light small-talk, maintaining neutral or positive sentiments. Therefore, it remains unclear how well the collected scores reflect empathetic abilities of the chatbots, which should ideally succeed over a range of emotions. Our framework addresses this limitation by grounding the chats in diverse emotional scenarios.

3 Method: iEval

To compare empathetic abilities of several chatbots, iEval suggests that at first a human makes an emotionally-grounded conversion with each bot in a randomized order. If necessary, fine-grained Likert-type assessments of specific chatbot's performance aspects may be collected after each conversation. As the next step, the same human is asked to rank the chatbots according to her experience with them. An example of this flow is given in Figure 1. Finally, appropriate statistical instruments should be applied to compare the chatbots.

3.1 Emotionally-grounded Chats

To make sure that humans experience the full extent of chatbots' empathetic abilities, we condition each conversation with a short emotional scenario, instructing the humans to imagine themselves feeling a particular emotion in a given situation. They are further asked to role-play a character in this scenario and chat about it with the models. The first dialog turn is provided to the humans to facilitate the process of their getting into the assigned role.

Careful conditioning of the experiment is essential to ensure that it adequately represents chatbots' abilities in a vast range of topics and emotions. We noticed that some dialogs from the EmpatheticDialogues dataset (Rashkin et al., 2019), a popular dataset for building empathetic models, form large clusters in terms of the similarity of discussed situations (see Appendix A). It may lead to models' shifted performance on specific topics. Therefore, one should control for topical diversity when defining conditioning scenarios for iEval.

Besides, previous results pointed out that the same model may receive different appraisals depending on the emotional polarity of the chats (Majumder et al., 2020). This may be linked to the existing difference between humans' empathetic responding in positive and negative scenarios (Aue et al., 2021), and hence difference in expectations. Thus, we argue for the importance of balancing and studying the role of emotional polarity within iEval.

Finally, ensuring sufficient interaction experience with the models is necessary before asking humans for their judgements. Previous works required between 3 and 14 chatbot's turns per dialog. We find 3 turns to be enough, given that the dialog starts with a specific input.

3.2 Ranking

The concluding step of iEval requests a human to recall the conversations with the chatbots and rank them by assigning the bots into three categories: *Bad*, *Okay*, and *Good*. Several chatbots can be assigned to the same category, indicating equal rank. This approach allows moving away from inter-annotator variability associated with Likert scales (Li et al., 2019; Kulikov et al., 2019), while preserving the benefits of relative comparisons. To obtain the final standing of the chatbots, we propose converting the resulting rank into an ordinal rating (*Bad* → 1, *Good* → 3) and running non-parametric ANOVA to compare the mean ratings.

3.3 Annotation Quality

According to iEval framework, one human should chat with and evaluate several models. As human's short-term mental storage capacity is limited to several informational chunks, we recommend keeping the number of evaluated models between 3 and 7, giving preference to lower values (Cowan, 2001).

To meet the requirements of randomized controlled experiments, it is also advisable to allow each human to complete only one evaluation task to eliminate anchoring effects. For the same reason, the order in which humans interact with the chatbots should be randomized and counterbalanced across tasks. To distinguish different models without revealing their names to the humans, we suggest color-coding them to avoid any fixation effects which could be caused by aliases that reflect order.

Finally, we use crowdsourcing for our experiment. To decrease the probability of encountering fraudulent or inattentive workers, human intelligent task design and configuration should follow the quality control recommendations of the platform in combination with other attention checks.

4 Experiment

To demonstrate how iEval works in practice, we apply the framework to benchmark several state-of-the-art empathetic chatbots, which have never been compared against each other in an interactive setting. The details and analysis are outlined below.

4.1 Measures

We use the final ranking of the chatbots, converted into ordinal ratings, as our main metric. To better understand which factors play a principal role in defining overall ranking, we also ask human workers for fine-grained Likert-type scores to a number of chatbots' qualities on a 1-5 scale. These questions were derived as a combination of the established key qualities for conversational chatbots (Svikhnushina and Pu, 2021) and other critical aspects related to their language modeling abilities (See et al., 2019). We measured chatbots' perceived politeness, empathy, likability, repetitiveness, and whether their responses make sense.

4.2 Models

We benchmarked four models, as this corresponds to an average number of informational chunks that humans can store in short-term memory (Cowan, 2001). We chose between the top-performing chatbots available at the moment of preparing our experiment in Q4 2021. We selected the models, which use distinct approaches for generating empathetic responses. Only one of them participated in an interactive evaluation previously, but it was not targeted at its empathetic skills. The four models with assigned color-codes are as follows.

Blender is a large model employing a standard Seq2Seq Transformer architecture with ≈90M parameters (Roller et al., 2021). Blender was pre-trained on ≈1.5B comments from Reddit discussions and fine-tuned on EmpatheticDialogues dataset (Rashkin et al., 2019).

MIME is a relatively small model with ≈18M parameters also based on Seq2Seq Transformer with additional stochastic emotion grouping and mimicry mechanism (Majumder et al., 2020). Without pretraining, MIME was directly initialized with GloVe embeddings (Pennington et al., 2014) and fine-tuned on EmpatheticDialogues.

MEED is a middle-size Seq2Seq Transformer-based model with ≈40M parameters, which incorporates extra controllability of response generation achieved through modeling fine-grained empathetic intents. The model was pre-trained on ≈1M dialogs from OpenSubtitles (Lison and Tiedemann, 2016) and fine-tuned on EmpatheticDialogues.

Plain is a basic Seq2Seq Transformer-based model with ≈40M parameters, which followed the same training pipeline as MEED. Plain serves as a baseline in our experiment.

All models were adapted to operate in an interactive setting so that for generating each next response, all previous dialog history was passed to the models as input.

4.3 Grounding Scenarios

As EmpatheticDialogues (Rashkin et al., 2019) is the mainly used benchmarking dataset for empathetic chatbots, we employed its test set to create grounding scenarios. This dataset contains 24,850 dialogs associated with emotional contexts (out of which 2,547 dialogs comprise the test set). To create the dataset, (Rashkin et al., 2019) connected two types of crowdworkers, speakers and listeners, to have conversations with each other. Speakers first had to select one of the 32 emotional labels (e.g., *sad*, *joyful*, *proud*) and describe a situation when they felt that way. Then they proceeded to have a conversation with the listeners using the outlined situations as guiding prompts. We utilized these attributes (32 emotional labels and prompts describing the speakers' situations) to describe our grounding scenarios and kept the first turn from each selected dialog as a starting turn for the worker in our evaluation task.

To ensure comprehensibility of the task for crowdworkers, this selection of grounding prompts and opening utterances was organized very carefully. Firstly, we selected dialogs where the length of the associated prompt falls between the first and third quantiles in terms of the number of tokens to ensure it provides sufficient details about the speaker's situation. Secondly, we computed Vader sentiment scores (Hutto and Gilbert, 2014) of the first utterance in each dialog and only kept those that had a clear emotional coloring. These steps produced 527 data points, which we finally proofread and annotated with emotional polarity labels (negative or positive). Note that we used the original 32 emotional labels to show them to crowdworkers to ground their interaction with the chatbots, while the polarity labels were needed for the analysis part. We further narrowed the set of 527 data points down to 480 prompts with utterances to meet our experimental design requirements (§4.4). The discarded data points were chosen manually in order to diversify the topics in the main set. The distribution of emotional labels in the resulting evaluation set is shown in Figure 8 in Appendix B. Some examples of grounding scenarios (emotional labels and prompts) are provided in Figures 4, 5, and 6.

4.4 Experiment Design

We aimed at evaluating the performance of the participating chatbots, while also contrasting their abilities in negative and positive emotional contexts. To maintain a manageable number of human intelligence tasks (HIT), we decided to ask each crowdworker to interact with all chatbots in both conditions. Therefore, our experiment was a 2×4 within-subject factorial design. By designing our study as a factorial experiment, we were able to examine both main effects and interactions among chatbots and emotional contexts. We used G*Power software to estimate the required sample size to achieve "medium" effect size (Faul et al., 2007). As the recommended sample size was about 200, we ran 240 experimental tasks to achieve a full counterbalance of the order of chatbots and emotional contexts across subjects. We analyzed ranking of the chatbots using the nonparametric Aligned Rank Transform (ART) procedure (Wobbrock et al., 2011). Quartile-quartile plots of the fitted residuals of our the model showed that they were normally distributed, indicating the appropriateness of this model for our analysis.

4.5 Running the Experiment

We ran our experiment on Amazon Mturk, requiring one US-based worker per each of the 240 HITs. Our workers spent on average 20.6 minutes to complete a HIT and their reward was $2.5 per HIT, which agrees with the US minimum wage standards. Following Mturk recommendations,[2] we required the workers to have 98% approval rate and 10,000 approved HITs. We further rejected the workers whose average HIT completion time, length of chat responses, or number of contradictory responses to reverse-scaled questions in the Likert-type questionnaire stood out as outliers.

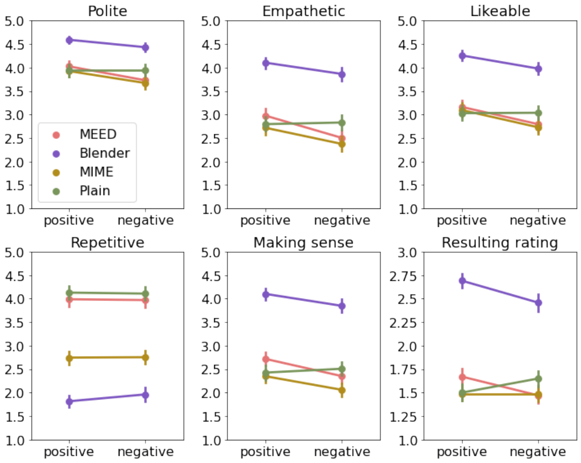


Figure 2: Benchmarking results of the four chatbots.

5 Analysis of Results

Below, we describe the eventual ranking of the models and consider the aspects that likely explain the observed results.

5.1 Benchmarking of Empathetic Chatbots

We used the nonparametric ART procedure to analyze ranking of the chatbots. As described above (§3.2), for this analysis we converted the resulting rank into an ordinal rating for more straightforward interpretation (the higher, the better). Results show a main effect of chatbot ($F_{3,1673} = 257.92$, $p < 0.001$) and of emotional context ($F_{1,1673} = 43.17$, $p < 0.001$) on the rating, and of their interaction ($F_{1,1673} = 9.80$, $p < 0.001$) as illustrated in the lower right subplot of Figure 2. Interaction results revealed several interesting relationships. Blender is consistently rated significantly higher than the other three chatbots, and it also performs significantly better in positive contexts than in negative ($p < 0.01$). MIME is rated the lowest, while for MEED and Plain a shift in the ratings emerges depending on emotional context. MEED significantly outperforms Plain in positive contexts ($p < 0.05$) while the diametrically opposite result manifests for negative contexts ($p < 0.05$).

5.2 Aspects Explaining the Ranking

We fitted an ordinal regression model to identify which of the factors measured by our Likert-type questionnaire correlate strongest with the assigned

---

[2] `https://blog.mturk.com/qualifications-and-worker-task-quality-best-practices-886f1f4e03fc`

ratings (McFadden's pseudo-$R^2 = 0.37$). The statistical model was chosen due to the ordinal nature of the dependent variable. All evaluated qualities exhibit significant influence on chatbots' ratings. Making sense ($\beta = 1.01, p < 0.001$), empathy ($\beta = 0.35, p < 0.001$), and repetitiveness ($\beta = -0.32, p < 0.001$) are the strongest predicting factors, followed by politeness ($\beta = 0.21, p < 0.01$) and likability ($\beta = 0.18, p < 0.05$) (Figure 3).

The leading factor suggests that the language modeling abilities of the chatbots define their ranking at large. This is understandable as language fluency is essential for the comprehensive handling of emotions in chat. Blender, being a massive model pre-trained on a dataset that is 1000-time larger than the one for MEED or Plain, is capable of generating considerably longer fluent responses than any other chatbot (Figure 9 in Appendix C), which ensures its highest rank. Meanwhile, MIME is the smallest model, which did not undergo any pre-training. It responds reasonably well to the first speaker's utterance, but as context gets longer its ability to produce semantically coherent responses quickly deteriorates (Figure 4), causing its poor performance regardless of the emotional context.

For chatbots with better language modeling skills, the next most decisive factors bring about curious shift in chatbots' rankings, depending on the emotional polarity of conversations. We analyze these phenomena in the next section, focusing on empathy and repetitiveness (diversity) revealed in chatbots' responses. While considering empathy of the models, we pay special attention to the role of questions that they ask as it has been established to be the most prominent intent of human empathetic listeners (Welivita and Pu, 2020). Moreover, question mark appears in the top-15 most frequent tokens in the responses of all models studied in our experiment (Table 3 in Appendix C).

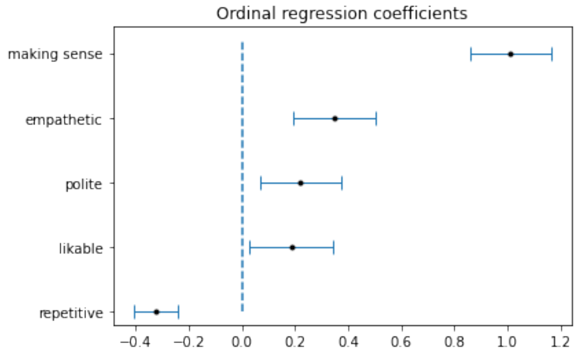


Figure 3: Results of ordinal regression on rank. 95% confidence intervals are approximated as two standard errors. All coefficients are significant ($p < 0.05$).

MIME (context: Positive, rank: Okay)

Emotion: joyful

Prompt: was delighted the other day when I got to see a friend that I had not seen in 20 years.

Speaker: *I was delighted the other day when I got to see a friend that I had not seen in 20 years.*

MIME: that is awesome. i hope you guys get some great news.

Speaker: *Yes, she seemed to be doing well. It was so nice to see her again.*

MIME: that is a good thing. i bet you were happy.

Speaker: *I was, yes.*

MIME: what did he do?

MIME (context: Negative, rank: Bad)

Emotion: afraid

Prompt: It is so dark in my house. I keep hearing creepy noises.

Speaker: *I keep hearing strange noises in my house. My husband is out of town and I'm all alone. There has been reports of break ins on my street.*

MIME: oh my goodness, that is not good, how long have you had to do?

Speaker: *I've been hearing strange noises for about an hour now. I'm scared.*

MIME: that is a good thing to be in the world. i hope it is not too late to get the right now.

Speaker: *What do you mean?*

MIME: i am sure that you are a good feeling.

Figure 4: Examples of chat logs with MIME.

5.3 Analysis of Interaction Effects

5.3.1 Intricacies between MEED and Plain

Both MEED and Plain have moderate language modeling abilities compared to the other two counterparts. To reason about why these models' rankings swapped depending on the emotional polarity, we make two noteworthy observations. First, even though the gap in scores is not huge, Plain is rated significantly more repetitive than MEED (Figure 2). Second, as it can be seen from Table 1, both chatbots actively ask questions in their responses, but MEED asks significantly more questions than Plain in negative contexts (independent t-test $p < 0.01$).

It is expected that MEED establishes greater diversity and poses questions as it models multiple listeners' empathetic intents, where questioning makes up the largest class. However, as we observed during qualitative inspection of the dialog logs, MEED falls into a typical trap of neural-

	Dialog level			**Turn-level: Positive**			**Turn-level: Negative**		
	Overall	Positive	Negative	Bot's turn 1	Bot's turn 2	Bot's turn 3	Bot's turn 1	Bot's turn 2	Bot's turn 3
MEED	1.83 (SD=0.96)	1.78 (SD=0.98)	1.89 (SD=0.94)	0.91 (SD=0.29)	0.56 (SD=0.51)	0.3 (SD=0.46)	0.93 (SD=0.26)	0.63 (SD=0.48)	0.33 (SD=0.47)
Blender	1.12 (SD=0.94)	1.25 (SD=0.92)	0.99 (SD=0.93)	0.73 (SD=0.55)	0.21 (SD=0.43)	0.32 (SD=0.52)	0.73 (SD=0.65)	0.11 (SD=0.35)	0.15 (SD=0.37)
MIME	1.26 (SD=0.91)	1.26 (SD=0.94)	1.25 (SD=0.88)	0.34 (SD=0.48)	0.46 (SD=0.54)	0.47 (SD=0.55)	0.35 (SD=0.48)	0.42 (SD=0.51)	0.48 (SD=0.53)
Plain	1.72 (SD=1.08)	1.78 (SD=1.07)	1.68 (SD=1.09)	0.83 (SD=0.38)	0.52 (SD=0.50)	0.43 (SD=0.50)	0.82 (SD=0.39)	0.53 (SD=0.50)	0.34 (SD=0.47)

Table 1: Average number of questions with standard deviation (in the parentheses) asked by different chatbots.

MEED (context: Positive, rank: Okay)

Emotion: anticipating

Prompt: I just ordered a new cookery book and I can't wait for it to arrive. I love to cook and will really enjoy trying out the recipes.

Speaker: I just ordered a new cookery book and am eagerly awaiting its arrival. I want to delve in and try lots of new recipes!

MEED: What kind of dish is it?

Speaker: Well I want to find something with beef first.

MEED: That sounds like a good recipe.

Speaker: Beef always is. I think I will try to find a good casserole to try.

MEED: That sounds delicious! I'm sure you will find a good casserole.

Plain (context: Positive, rank: Bad)

Emotion: anticipating

Prompt: I just ordered a new cookery book and I can't wait for it to arrive. I love to cook and will really enjoy trying out the recipes.

Speaker: I just ordered a new cookery book and am eagerly awaiting its arrival. I want to delve in and try lots of new recipes!

Plain: That sounds like a good idea. What kind of recipes do you like?

Speaker: I love deserts the most. I want to try a new chocolate cake recipe.

Plain: That sounds like a good idea. What kind of recipe is it?

Speaker: I don't know yet. i need to look but definitely chocolate.

Plain: That sounds like a good idea. I'm sure you'll like it.

MEED (context: Negative, rank: Bad)

Emotion: terrified

Prompt: I am so scared to live in my neighborhood. There is people that come around shooting their guns.

Speaker: I am so scared to live in my neighborhood. There is people that come around shooting their guns.

MEED: What kind of guns?

Speaker: I don't know the kind of gun, but it is upsetting, and they are shooting close to my house.

MEED: What kind of gun do you think they are shooting close to your house?

Speaker: I don't know. A hand gun? What does that matter?

MEED: Do you think they are shooting close to your house?

Plain (context: Negative, rank: Okay)

Emotion: terrified

Prompt: I am so scared to live in my neighborhood. There is people that come around shooting their guns.

Speaker: I am so scared to live in my neighborhood. There is people that come around shooting their guns.

Plain: What kind of guns do you have?

Speaker: I don't have any guns.

Plain: I don't have any guns either. I don't like guns either.

Speaker: So what should I do to feel safer in my neighborhood?

Plain: I don't have any guns either. I don't like guns either.

Figure 5: Examples of chat logs with MEED and Plain.

based chatbots by learning the most safe and general type of response, in this case question, e.g., *"What type of ... was it?"* While asking such general information-seeking questions is an acceptable strategy for positive contexts, other questioning behaviors were shown to be more effective in delivering meaningful emotional regulation in negative scenarios (Svikhnushina et al., 2022).

We further combined these observations with the fact that correlation between these chatbots' repetitiveness scores and overall ratings is slightly lower in negative scenarios (Pearson's $r = -0.42$ ($p < 0.001$)) than in positive (Pearson's $r = -0.51$ ($p < 0.001$)). It suggests one plausible explanation to the observed phenomenon. In positive contexts, human speakers value chatbots' diversity and active engagement demonstrated via questioning, and are more forgiving even if the chatbot's response is slightly misaligned with the context. In negative scenarios, speakers feel much more vulnerable and expect greater attention. Consequently, they prefer a generic, but safe response over the one which is somewhat unrelated or diverting attention from the speaker's emotional state. Figure 5 provides examples illustrating these observations.

5.3.2 Decline of Blender in Negative Contexts

To study the possible reasons of Blender's lower performance in negative contexts, we started with qualitative inspection of dialog logs. While Blender asks fewer questions than MEED or Plain, they still appear frequently in its responses (Table 1) and the same issue of asking overly general questions, failing to address speaker's emotional needs in negative contexts, preserves also for this chatbot.

More interestingly, we observed that Blender is the only chatbot in our experiment that persistently shares its own experiences and views with the speakers. To get a better idea of this behavior, we randomly sampled 50 chat logs (25 per emotional polarity) for each chatbot and annotated how many of their responses were oriented towards the speaker (other), reflecting the chatbot's own experience (self), or both. Blender is the only chatbot, for whom the amount of responses oriented towards self and both largely exceeds zero in the later dialog turns, almost reaching the proportion of purely other-oriented responses (Table 2 in Appendix C).

Blender (context: Positive, rank: Good)

Emotion: nostalgic

Prompt: I get sentimental during the holidays. I think back to the ones i had as a kid.

Speaker: *I love the holidays, that make think back to being a kid.*

Blender: *i agree, it's a great time to spend with family and friends. What do you do for fun?*

Speaker: *We bake cookies, listen to christmas songs. What about you?*

Blender: *i like to play video games and spend time with my family. what kind of cookies do you bake?*

Speaker: *pecan, chocolate chip, yum!*

Blender: *what is your favorite video game to play? i've never played video games, but i love video games.*

Blender (context: Negative, rank: Bad)

Emotion: devastated

Prompt: I was so upset when I failed my math test, I was only 1 percent off!

Speaker: *I was one percent off from passing my math test, I was devastated.*

Blender: *i'm sorry to hear that. were you able to get a better grade on the test?*

Speaker: *No, I am just upset.*

Blender: *i think i am going to go back to school. i am not sure what i will do.*

Speaker: *Make sure to study.*

Blender: *i am sure you will do great. i hope you get a good grade on your test. good luck!*

Figure 6: Examples of chat logs with Blender.

Self-disclosure is frequent in peer support discussion forums (Barak and Gluck-Ofri, 2007). This likely explains Blender's tendency to share own perspective as it was pre-trained on Reddit conversations, where peer support is actively practiced. However, human attitude to chatbot's sharing about self is unclear, especially in negative scenarios. Even in human-human interaction, positive disclosure is appreciated more than negative (Caltabiano and Smithson, 1983). Moreover, in counselling practice, therapist self-disclosure is usually portrayed as a mistake (Henretty and Levitt, 2010). We could not find studies about users' preferences for the degree of chatbot's self-oriented responses, but some previous findings about embodied computer agents reveal that their empathetic other-oriented emotions lead to more positive ratings of the agent (Brave et al., 2005). We, therefore, hypothesize that pulling attention to self too quickly in negative conversations might have resulted in Blender's poorer performance in this emotional polarity, which is demonstrated with an example in Figure 6.

6 Discussion

6.1 Implications for Chatbot Development

Most of the chatbots in our experiment were trained to model short-context conversations and did not support the interactive chat mode by default, which also applies to other dialog models, e.g. (Hu et al., 2018; Lin et al., 2019). Nevertheless, being able to maintain continuous engaging conversation is an ultimate goal for empathetic chatbots. Thus, more attention should be paid to adapting training procedures and architectures to track longer-term dialog history and evolution of speaker's emotions.

Our findings demonstrate that users' emotional needs differ in positive and negative scenarios, and that they do not necessarily expect a strong emotional reaction to their inputs. Raising a question may be an appropriate response. According to our results, chatbots should dwell longer on speakers' negative situations, employing meaningful questioning strategies, which can possibly be achieved by modeling fine-grained empathetic questioning intents (Svikhnushina et al., 2022). In addition, more research on the amount of chatbots' self-disclosure would further help tailor chatbots' responses to users' expectations.

6.2 Next Steps

While human evaluation is the current standard to assess chatbots' performance, developing an automated metric to approximate human judgement is an important milestone that would considerably facilitate the developmental cycle. Some attempts towards this goal have been made (Yeh et al., 2021), but very few of these metrics try to capture empathetic abilities of chatbots. Our analysis suggests that all dimensions evaluated in our Likert-type questionnaire constitute significant predictors of the overall human satisfaction (§5.2). Therefore, to develop a stronger automatic proxy for human evaluation, we consider creating rationale heuristics approximating those dimensions and identifying a meaningful way to combine them into a single score. The dataset of collected chat logs and human scores from our experiment should streamline the construction and calibration of such a metric.

7 Limitations

In our work, we applied iEval framework to benchmark four empathetic agents. We did not compare them against human-human interaction, as synchronizing two crowdworkers for conducting several chats between each other entails more logistical difficulties. More importantly, we were mainly interested in measuring how existing chatbots address users' emotional needs, rather than checking if they are indistinguishable from human interlocutors.

Our results show that bigger models rank higher in the evaluation task. It raises the subsequent question about to what extent the proposed framework measures differences in models' empathetic abilities compared to their underlying language model performances. We believe that iEval is an effective framework for evaluating chatbots' empathy as it succeeded in registering intricate differences in the performances of MEED and Plain, two models of comparable sizes and pre-training pipelines, as well as distinguishing the performance of Blender in emotional contexts of different polarity. To further disentangle the role of language modeling and empathetic abilities, one can consider running the iEval evaluation experiment to compare equal-size models with and without fine-tuning for empathetic response generation (e.g., Blender, which was only pre-trained on Reddit, and Blender, which was further fine-tuned on the EmpatheticDialogues dataset). However, this was not the main objective of our study and we leave it for future work.

Finally, we propose to use ranking as a way of expressing the appraisals of the chatbots, as it affords advantages of both Likert scales and pairwise comparisons. Ranking may be less robust for comparing results across experiments with mismatched sets of chatbots. Applying rank aggregation techniques can be useful to tackle such cases (Sculley, 2007).

8 Conclusion

Our paper introduced iEval, a novel evaluation framework for open-domain chatbots that can detect humans' personal perceptions of social interaction, manifesting in emotional dialogs. We used iEval to benchmark four recent empathetic chatbots. Further analysis revealed several limitations in empathetic response generation approaches of these models, which came out due to their uneven abilities in handling positive and negative conversational scenarios. Based on our findings, we formulated implications informing future efforts in the development and evaluation of such chatbots. We also publicly release the data from our experiment to expedite future research in these directions.

Acknowledgements

This project has received funding from the Swiss National Science Foundation (Grant No. 200021_184602). The authors also express gratitude to Francesco Posa for the help with configuring the iEval application.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv e-prints*, page arXiv:2001.09977.

Tatjana Aue, Stephanie Bührer, Boris Mayer, and Mihai Dricu. 2021. Empathic responses to social targets: The influence of warmth and competence perceptions, situational valence, and social identification. *PloS one*, 16(3):e0248562.

Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, 10(3):407–417.

Lisa Feldman Barrett, Batja Mesquita, Kevin N. Ochsner, and James J. Gross. 2007. The experience of emotion. *Annual Review of Psychology*, 58(1):373–403. PMID: 17002554.

Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2):161–178. Subtle expressivity for characters and robots.

Marie Louise Caltabiano and Michael Smithson. 1983. Variables affecting the perception of self-disclosure appropriateness. *The Journal of Social Psychology*, 120(1):119–128.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 929–932, New York, NY, USA. Association for Computing Machinery.

Jennifer R. Henretty and Heidi M. Levitt. 2010. The role of therapist self-disclosure in psychotherapy: A qualitative review. *Clinical Psychology Review*, 30(1):63–77.

Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 1991. *The emotions*. University Press of America.

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, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

D. Sculley. 2007. Rank aggregation for similar items. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 587–592.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Ekaterina Svikhnushina and Pearl Pu. 2021. Key qualities of conversational chatbots – the peace model. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 520–530, New York, NY, USA. Association for Computing Machinery.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Joseph B. Walther, Tracy Loh, and Laura Granka. 2005. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24(1):36–65.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*, pages 143–146, New York. ACM Press.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

A Topic Clusters in EmpatheticDialogues

While working with the EmpatheticDialogues dataset (Rashkin et al., 2019), we noticed that many dialogs appear repetitive in terms of the situational scenarios brought up by the speakers. To examine it more closely, we used Sentence Transformers framework (Reimers and Gurevych, 2019) to compute vector embeddings of first speakers' turns in all dialogs and cluster them according to cosine-similarity. Figure 7 shows the empirical cumulative distribution function of topic cluster sizes in the train set of EmpatheticDialogues. From the figure, it can be seen that clusters with between 30 and 130 similar situation descriptions per cluster comprise almost 20% of the training data.

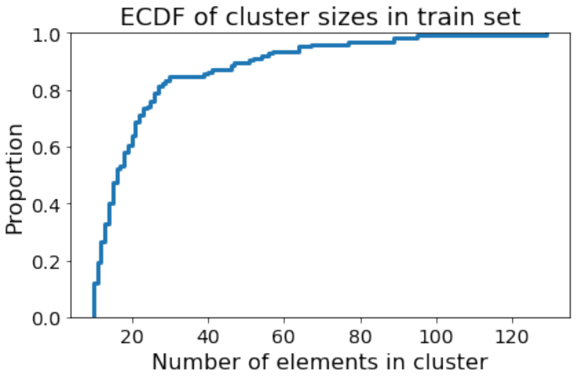


Figure 7: Empirical cumulative distribution function of topical cluster sizes in the train set of EmpatheticDialogues dataset (Rashkin et al., 2019).

	Pos: Other			Pos: Self			Pos: Both			Neg: Other			Neg: Self			Neg: Both		
	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3	t-1	t-2	t-3
MEED	25	24	24	0	0	0	0	1	1	25	25	25	0	0	0	0	0	0
Blender	22	16	11	0	3	4	3	6	10	24	14	15	0	4	6	1	7	4
MIME	22	22	20	2	1	1	1	2	4	25	24	22	0	0	1	0	1	2
Plain	24	20	20	1	4	4	0	1	1	25	24	23	0	0	2	0	1	0

Table 2: Counts of orientation of chatbots' responses (other-, self-, or both) in 50 sampled chat logs (25 for positive and 25 for negative contexts). Prefixes "Pos" and "Neg" stand for positive and negative contexts respectively.

B Emotion Distribution in Grounding Scenarios

Figure 8 shows the distribution of original emotional labels from the EmpatheticDialogues dataset (Rashkin et al., 2019) in 480 grounding scenarios used for our benchmarking experiment. To demonstrate the even coverage of the whole emotional spectrum, we mapped 32 emotions from the dataset to 14 emotions from Plutchik's wheel (Plutchik, 1991) (8 basic and 6 intermediate emotions) and color-coded the bars in Figure 8 according to these 14 categories.

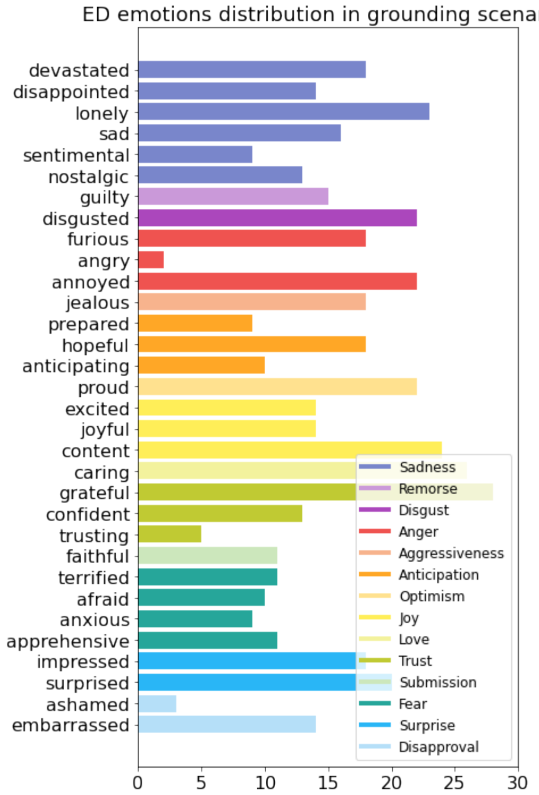


Figure 8: Distribution of emotional labels from EmpatheticDialogues dataset in grounding scenarios. The legend shows the mapping between the colors and 14 emotional categories from Plutchik's wheel (Plutchik, 1991) (8 basic and 6 intermediate emotions).

C Additional Details about Chatbots' Responses

Figure 9 depicts the average number of tokens in chatbots' responses over three dialog turns.

Table 3 shows the top-15 most frequent tokens for each of the four chatbots. As it can be noticed, question marks appear in the list of tokens of each model, pinpointing their tendency to ask questions.

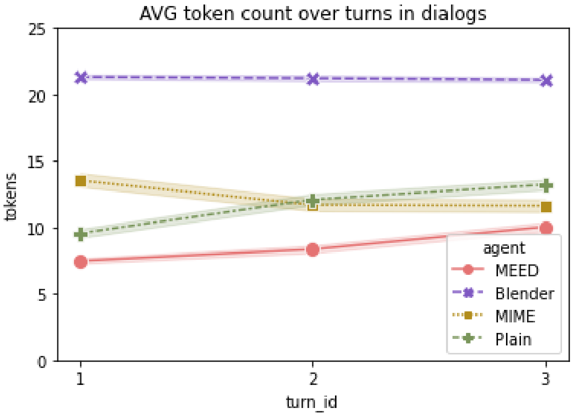


Figure 9: Counts of average number of token in chatbots' responses over three dialog turns with 95% confidence intervals.

Table 2 demonstrates the counts of orientation of chatbots' responses (other-, self-, or both) in 50 sampled chat logs (25 positive and 25 negative) over the dialog turns.

MEED	**Blender**	**MIME**	**Plain**
?	.	that	i
you	i	i	.
that	you	.	you
.	to	is	?
what	that	you	that
of	it	a	to
it	's	to	!
!	a	?	sorry
a	of	am	so
i	do	!	it
's	?	good	hear
kind	!	what	what
did	have	have	did
is	the	do	am
sounds	'm	,	of

Table 3: Top-15 most frequent tokens for each chatbot in order of decreasing frequency.