

Approximating Online Human Evaluation of Social Chatbots with Prompting

Anonymous ACL submission

Abstract

As conversational models become increasingly available to the general public, users are engaging with this technology in social interactions. Such unprecedented interaction experiences may pose considerable social and psychological risks to the users unless the technology is properly controlled. This highlights the need for scalable and robust evaluation metrics for conversational chatbots. Existing evaluation metrics aim to automate offline user evaluation and approximate human judgment of pre-curated dialogs. However, they are limited in their ability to capture subjective perceptions of users who actually interact with the bots and might not generalize to real-world settings. To address this limitation, we propose an approach to approximate online human evaluation leveraging large language models (LLMs) from the GPT-family. We introduce a new Dialog system Evaluation framework based on Prompting (DEP), which enables a fully automatic evaluation pipeline that replicates live user studies and achieves an impressive correlation with human judgment (up to Pearson $r = 0.95$ on a system level). The DEP approach involves collecting synthetic chat logs of evaluated bots with an LLM in the other-play setting, where the LLM is carefully conditioned to follow a specific scenario. We further explore different prompting approaches to produce evaluation scores with the same LLM. The best-performing prompts, which contain few-shot demonstrations and instructions, show outstanding performance on the tested dataset and demonstrate the ability to generalize to other dialog corpora.

1 Introduction

The recent arrival of conversational AI, marked by the public release of ChatGPT from OpenAI,¹ initiated unprecedented user engagement with conversational chatbots in a real-world setting. With the impressive naturalness of machines' responses,

users are going beyond traditional transactional exchanges and start exploring more social interaction scenarios with increasing curiosity (Thormundsson, 2023). In such situations, users might be subject to social and psychological harms if dialog systems fail to follow commonsense social rules (Svikhnushina and Pu, 2022; Kim et al., 2022). Several instances of alarming social behavior of this technology have already been discussed in the media (Roose, 2023; De Cosmo, 2023; Life, 2023). In this context, developing meaningful and robust evaluation metrics for these systems has become particularly urgent to ensure that the models are safe and acting in the best interest of the users before their release.

Initially, human evaluation was considered a de facto standard for evaluating dialog systems (Li et al., 2019). As running human evaluation is time- and resource-consuming, a number of automatic evaluation metrics for dialog systems have been proposed (Mehri et al., 2022; Yeh et al., 2021). The majority of these approaches aim to automate the *offline* user evaluation. In this setting, dialog evaluation is performed by a human judge who is distinct from the one conversing with the bot (Figure 1, offline). The metrics proposed for this case approximate the evaluation scores provided by this third-party human judge for the pre-produced dialogs (e.g. Mehri and Eskenazi, 2020; Ghazarian et al., 2022a). Despite its popularity, offline user evaluation is limited in its ability to capture subjective perceptions of users who actually interacted with the bots (Jannach, 2022; Lee et al., 2022; Ghandeharioun et al., 2019). This limitation of relying on second-hand evaluation can be illustrated by an analogy from the realm of restaurant critique when one tries to evaluate a restaurant solely by reading consumer reviews but having never actually eaten there. Conducting *online* user evaluation, where the same individual interacts with the bot and assesses its performance, is more likely to produce

¹<https://openai.com/blog/chatgpt>

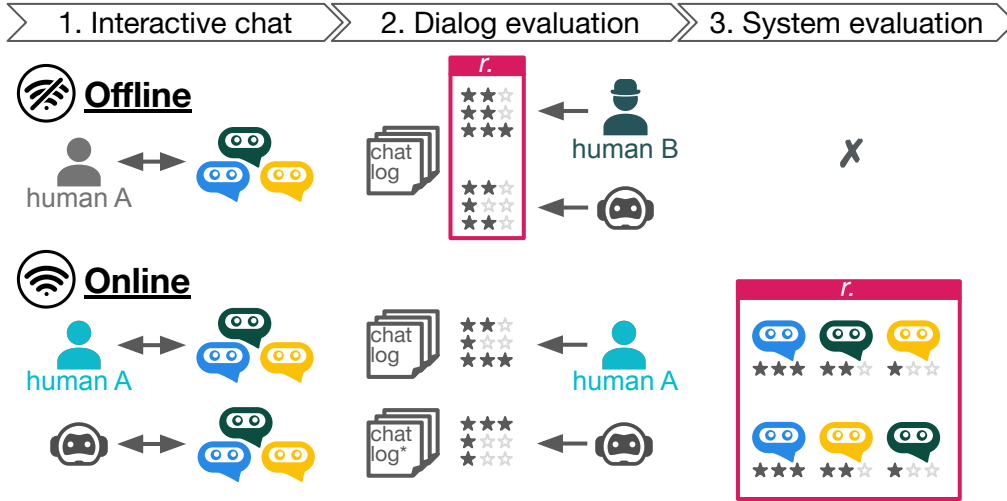


Figure 1: Offline and online dialog evaluation with the corresponding processes. In the first step, dialog logs are curated. In the second step, each dialog log is assigned a dialog-level score, either by a third-party judge (offline) or by the same conversational partner (online). In the third step, the system ranking is obtained by aggregating the dialog scores of each chatbot. Grey bot icons indicate steps that are intended to be approximated by means of automatic evaluation. Pink boxes mark the steps in the process where the correlation ($r.$) with the ground truth human judgment is computed to validate the automatic evaluation metric during its development process.

accurate and precise evaluations of the chatbot’s performance. Moreover, this method offers better predictive capabilities for the system use “in the wild” (Beel and Langer, 2015). However, by far, efforts towards approximating online user evaluation have been limited.

To address this gap, we propose a novel automatic Dialog system Evaluation framework based on Prompting, DEP. Our framework automates the whole pipeline of dialog system evaluation in an interactive setting, replicating live user studies. As the first step towards this goal, we leverage a large language model (LLM) from the GPT-family models to collect synthetic chat logs of evaluated bots with the LLM. Second, we prompt the same LLM to produce the resulting evaluation scores for generated chat logs and, finally, rank the chatbots based on their overall performance (Figure 1, online).

While using bot-play is not a new idea per se, we emphasize the importance of carefully choosing a dialog partner for the evaluated chatbots specifically for social conversational contexts where the roles of two interlocutors can differ significantly. For example, it was shown that the emotion/intent distributions in conversations between an emotional speaker and an empathetic listener are very different for the two dialog partners (Welivita and Pu, 2020). To account for it, in the first step of our framework, we propose prompting LLMs to play a particular social role over the course of the

interaction with the chatbots to be evaluated. For the second step, we draw inspiration from the fact that LLMs demonstrate solid performance improvement when their generation process is augmented with instructions (Kim et al., 2022). We demonstrate that prompting the model with appropriate instructions that explain how fine-grained evaluation dimensions relate to the overall dialog score leads to substantial performance improvement, reaching up to $r = 0.95$ Pearson correlation with the human judgment on a system level.

Overall, our contributions include the following. 1) We describe an end-to-end prompting-based evaluation framework for dialog systems, specifically targeting social interaction scenarios (Section 3). 2) Our experiments showcase the effectiveness of prompting for assigning a desired social role to LLMs and, thus, collecting machine-generated dialogs that better approximate real interpersonal communication (Section 4.1.2). 3) We consider different prompt designs and conclude that including demonstrations together with instructions results in the best performance (Sections 4.1.3, 4.2.2).

2 Related Work

2.1 Automatic Evaluation of Chatbots

Automatic dialog evaluation has been a long-standing goal for the practitioners. Initial works focused on evaluating chatbots’ responses against

a ground-truth reference (Papineni et al., 2002; Tao et al., 2018). Following works moved on to exploring reference-free evaluation metrics as referenced evaluation was shown to be ineffective due to a wide range of acceptable responses for a single context (Liu et al., 2016), implying that comparing with a single reference is limited. Reference-free metrics usually operate either on the utterance or the dialog level. For the utterance level, practitioners have explored ways to evaluate response appropriateness for the preceding context (Lan et al., 2020; Pang et al., 2020) or predict the qualities of the follow-up response as a proxy for the quality of the preceding dialog (Ghazarian et al., 2022a, 2020; Mehri and Eskenazi, 2020). For the dialog level, a number of diverse approaches have been proposed, ranging from aggregating several fine-grained utterance-level evaluations (Zhang et al., 2021b), to designing training objectives to model the information flow across dialogue utterances (Li et al., 2021), employing graph representations to capture dialog dynamics (Huang et al., 2020; Zhang et al., 2021a), and using semantic-level manipulations to teach the evaluation model to distinguish coherent and incoherent dialogs (Ghazarian et al., 2022b).

The works above largely target the offline evaluation setting. Some scholars have also started exploring different ways of approximating online user evaluation. Deriu et al. (2020) proposed a partially automated framework where human judges rank chatbots regarding their ability to mimic conversational behavior using interactively collected bot-to-bot conversations, which relies on survival analysis. Sato et al. (2022) proposed a particular bipartite-play approach for collecting bot-to-bot conversations to provide a fairer comparison setting for evaluated chatbots. These papers consider methodologies for organizing bot-to-bot conversation sessions, but they are not concerned with the way how these bot-to-bot conversations unfold. In our work, we explore the use of bot-to-bot conversations to model a desired social behavior.

2.2 Prompting

Prompt-based learning paradigm (Liu et al., 2023) received significant attention after Brown et al. (2020) demonstrated how GPT-3, a large foundation model, can well handle a wide range of tasks without the need for fine-tuning, relying only on natural-language prompts and task demonstra-

tions as context. Prompt-based model performance depends on the design of the provided prompt. Prompt engineering efforts explore approaches for designing prompts, which vary in the shape of prompts (cloze or prefix), human effort required for writing prompts (manual or automatic), and number of demonstrations provided to the model in the prompt (zero-shot or few-shot) (Liu et al., 2023).

Prompt-based learning applied to recently created LLMs has been reported to achieve outstanding results on a variety of tasks and benchmarks, including classification, reasoning, coding, translation, and many others (e.g. Wei et al., 2022; Chowdhery et al., 2022; Chung et al., 2022). However, exploring prompting for evaluation of dialog systems has not been widely investigated. We are only aware of one more simultaneous and independent effort in this direction. Huynh et al. (2023) studied how different LLM parameters (type, size, training data) may influence the dialog evaluation, focusing on utterance- and dialog-level evaluation in the offline evaluation setting. Our work focuses on how prompting can be used to capture a holistic evaluation of dialog systems in online social settings, relying on freshly generated dialogs.

3 Proposed Method: DEP

We introduce our DEP framework, which consists of two consecutive steps. First, it requires collecting interactive chat logs between the LLM and evaluated chatbots, which we denote as LLM-to-bot play. Second, the LLM is prompted to generate scores for these chat logs. The generated scores are further aggregated to produce a final ranking of the systems. We describe each of the steps below.

3.1 Prompted LLM-to-Bot Play

In social settings, two partners may play considerably different roles in a dialog, thus establishing very distinct conversational behaviors. Examples include conversations between a student and a teacher, an emotional speaker and an empathetic listener, or even between two interlocutors with different personas. Chatbots are usually built to perform well in one of these roles (e.g., empathetic listener), but not necessarily the other. Therefore, collecting synthesized dialogs via self-play of the chatbot with itself (or a similar competing model) might fail to represent a realistic discourse flow due to the differences in the intents produced by speakers and listeners in dialogs.

I am a Speaker *<in an assigned social situation>*. I am sharing *<my thoughts>* with a Listener in a dialog.

Speaker: *<LLM's input #1>*

Listener: *<Bot's response #1>*

Speaker:

Figure 2: Prompt template to condition a LLM to play an assigned social role while interacting with an evaluated chatbot.

To address this consideration and render the synthesized dialogs that better approximate real social interactions, we propose leveraging LLMs’ ability to produce responses on behalf of an assigned character (Thoppilan et al., 2022). Specifically, we suggest letting the evaluated chatbots converse with an LLM prompted to play a particular social role. Figure 2 demonstrates how to structure the prompt to produce each next output of the LLM in an interactive manner. Meanwhile, responses from the evaluated chatbots are computed by passing the accumulated dialog history to these chatbots as input context. The process can be repeated for multiple dialog turns. The length of the exchange may depend on the extent of details provided to prompt the LLM. The more specific the prompt is, the faster the evaluated chatbot can demonstrate its performance in the social situation of interest. On the contrary, more generic conversation starters require more dialog turns to reveal the targeted social behavior.

3.2 Prompted Evaluation

Once dialog logs are synthesized, we propose using prompting to produce evaluation scores for each dialog. Prompts can be constructed in several ways. We investigate zero-shot and few-shot settings, either with or without instructions, in our experiments (Section 4). Many available foundation LLMs are accessible through APIs and only output text completions without corresponding log probabilities. Therefore, regardless of the type of prompt that we use, to generate a score for each dialog, we obtain a textual form of the score from the LLM completion and then use a verbalizer function to map it to a numerical value, getting inspiration from (Schick and Schütze, 2021). Formally, given a dialog log d , we construct a prompt $P(d)$ that takes d as input and outputs a prompt that contains exactly one mask token as a placeholder for the dialog score. Let y be a predicted token for $P(d)$. We

then define a verbalizer as an injective function v that maps each score in textual form to a numerical value. Thus, $v(y)$ produces a numerical score for a single dialog. The final rating of a given dialog system is obtained by averaging the corresponding dialog scores of that system. For fair evaluation, the number of dialogs collected for each evaluated chatbot should be identical.

4 Results

For all reported experiments, we used the most capable version of the InstructGPT model (text-davinci-003) available at the moment of initiation of our experiments in early Q1 2023. We used this model as it was easily accessible through OpenAI API² and was expected to have superior performance for social scenarios as it was trained based on human feedback, which captures subjective human judgment of interactive outputs (Ouyang et al., 2022).

Following previous works that considered system-level evaluation (Lowe et al., 2017; Ghandeharioun et al., 2019), we report Pearson correlation for our experiments, unless specified otherwise. We also opted for this type of correlation coefficient as it performed better for capturing whether the automated metric succeeds in preserving the gap in scores for the best- and least-performing chatbots, the information which gets lost with rank correlation.

We start by demonstrating the application of our evaluation framework to empathetic dialog systems as in these interactive scenarios two conversational partners have clearly distinct social roles: an emotional speaker and an empathetic listener. Further, we consider the generalizing ability of the framework to other social domains.

4.1 Evaluation of Empathetic Chatbots

Below, we first describe the dataset used for the experiment. Then, we consider the ability of prompted LLM to effectively replicate social discourse patterns over multi-turn interactions with the chatbots that serve as eventual evaluation targets. Finally, we explore several types of prompts applied to synthesized LLM-to-bots dialogs to evaluate how well they can approximate human judgment on a system level.

²<https://openai.com/blog/openai-api>

4.1.1 Dataset and Evaluated Chatbots

We used iEval dataset for this experiment (Svikhnushina et al., 2022). The dataset features human conversations with four empathetic chatbots collected in an online interactive manner. During the dataset curation process, each human was assigned an emotion label with the situation description taken from the EmpatheticDialogues dataset (Rashkin et al., 2019) and asked to have a 6-turn conversation with each chatbot while playing a character in the assigned scenario. Overall, there are 480 situation descriptions in the dataset, which evenly cover two emotional polarities: positive and negative. As each chatbot participated in each scenario, there are in total 1920 dialogs in the dataset. After conversing with the chatbots, human interlocutors provided their appraisals of chatbot listeners in each dialog, including five fine-grained listener qualities on a 5-point Likert scale: politeness, empathy, likability, repetitiveness, and making sense, and an overall dialog rating on a 3-point scale. All scores are provided on a dialog-level.

The four chatbot models used to curate the dataset were Blender (Roller et al., 2021), MIME (Majumder et al., 2020), MEED and Plain (Xie and Pu, 2021). All of them are publicly available. We use these models in the same configurations for our experiment.

4.1.2 LLM-to-Bot Play Results

As the first step to validate our evaluation framework, we analyzed whether the LLM succeeds in mimicking human discourse following an assigned social role and whether approximating human speakers with the LLM causes any considerable changes in the chatbots’ response patterns.

To generate LLM-to-bots conversations, we closely followed the procedure of iEval dataset curation. Specifically, we used emotion labels and situation descriptions from the dataset to create prompts for the LLM: *I am a Speaker, feeling <emotion> because <situation>. I am sharing these emotions with a Listener, expecting empathy and understanding from them. I respond as a Speaker in a dialog.* The first LLM input was also taken from the iEval dataset. For each scenario, we collected LLM conversations with each of the four bots, letting them converse for 6 turns, i.e., 3 inputs from the LLM and 3 responses from the chatbot.

To examine the similarity of discourse patterns between human-to-bots and LLM-to-bots conversations, we started by annotating each dialog turn

in two datasets with emotion and empathetic intent labels, using emotion/intent classifier developed by Welivita and Pu (2020) for EmpatheticDialogues dataset. As datasets in our experiment were grounded in situation descriptions taken from EmpatheticDialogues, the classifier was expected to generalize well to our data.

Consequently, we visualized the most prominent discourse patterns for two corpora in the form of Sankey diagrams, shown in Figures 3 and 4. From the visual inspection, it can be seen the LLM emotion distribution over the course of the dialog (Figure 4) largely resembles one of the human interlocutors (Figure 3). More importantly, sets of intents produced by empathetic chatbots are also very similar between the two figures, with *Questioning*, *Sympathizing*, and *Acknowledging* being the most prominent ones. Thus, our freshly generated interactive dataset with LLM-to-bot play was deemed to produce a reasonable approximation of human-to-bot conversations.

4.1.3 Prompted Evaluation Results

Turning to the second step of our evaluation framework, we examined different types of prompting to produce scores for the generated LLM-to-bot dialogs. Specifically, two variables in the prompt design were considered.

First, we tried score generation in zero-shot and few-shot settings. For the few-shot setting, the number of demonstrations was fixed to the number of points in the ground truth human evaluation scale, with one representative example supplied for each score. Thus, for the iEval dataset, we used three demonstration dialogs corresponding to the three possible evaluation scores: *Bad*, *Okay*, and *Good*. The examples were selected manually and are provided in Table 4 in Appendix A.

Second, we analyzed whether providing additional instructions helped the LLM evaluation performance. To write the instructions, we relied on the findings of Svikhnushina et al. (2022), which explained how chatbots’ performance on various fine-grained dimensions translates into the overall score. As the authors emphasized the difference in humans’ expectations of an empathetic listener in positive and negative conversational scenarios, we devised slightly different instructions to prompt the evaluation of these two emotional polarities. Specific formulations of the instructions are also provided in Table 4 in Appendix A.

To generate scores for each dialog, we prompted

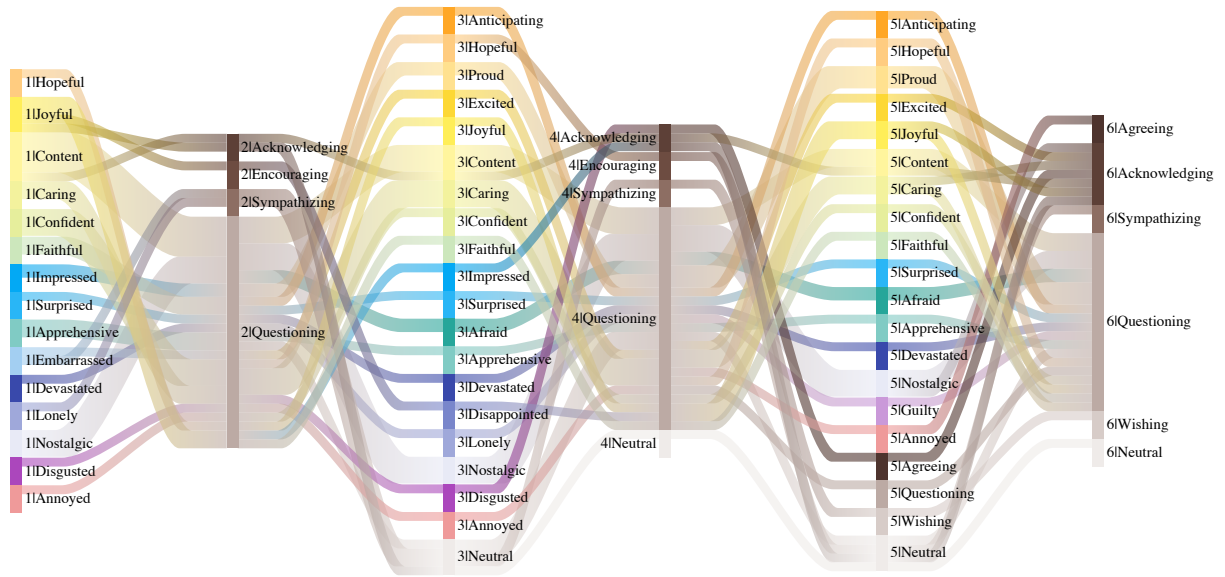


Figure 3: Sankey diagram showing discourse patterns in human-to-bots conversations originating from the iEval dataset.

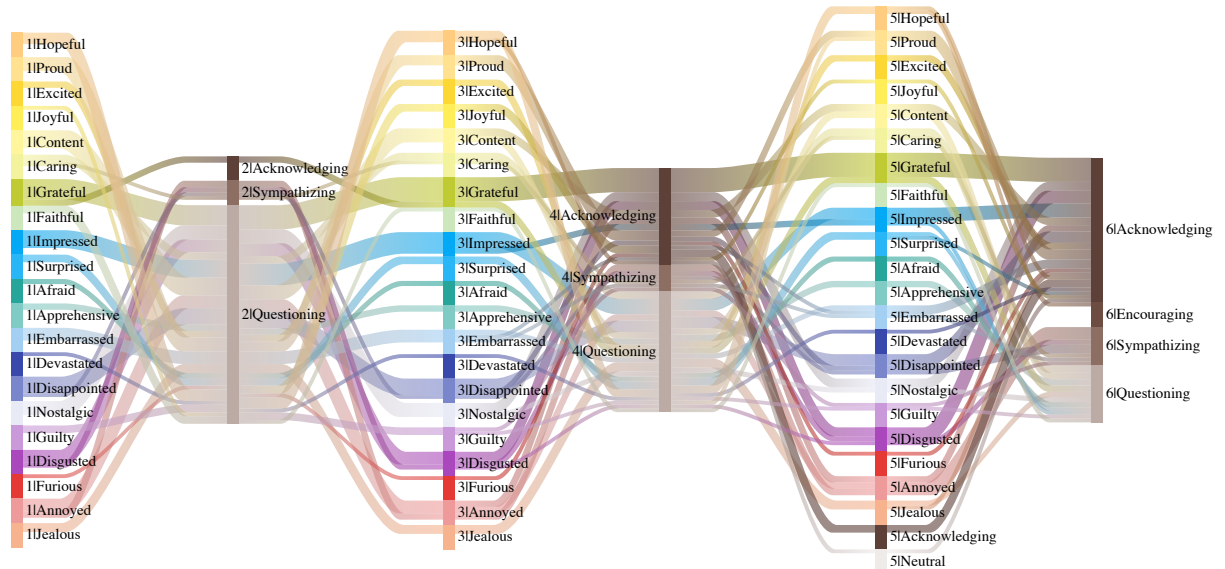


Figure 4: Sankey diagram showing discourse patterns in freshly generated LLM-to-bots conversations.

the LLM to complete the masked score, provided the log of the evaluated dialog. Depending on the configuration, few-shot demonstrations and/or instructions were prepended to the prompt. A template of the used prompt can be found in Figure 6 in Appendix A. After obtaining dialog-level scores, we aggregated them to produce system-level ratings. One system was defined as a chatbot operating in one of the two emotional polarities. This decision is driven by the fact that based on human evaluation results in (Svikhnushina et al., 2022), chatbots demonstrated statistically significant differences in their performance depending on the

emotion. Thus, we considered eight systems for computing system-level correlations.

System-level correlations between human- and LLM-judgments for each of the four possible prompt design manipulations are reported in Table 1. Few-shot prompting with instructions results in the highest correlation of 0.954, which is further illustrated by the scatter plots in Figure 5. According to the plots, providing examples helps the LLM to calibrate the produced scores, eliminating the positivity bias, whereas instructions result in reduced variance.

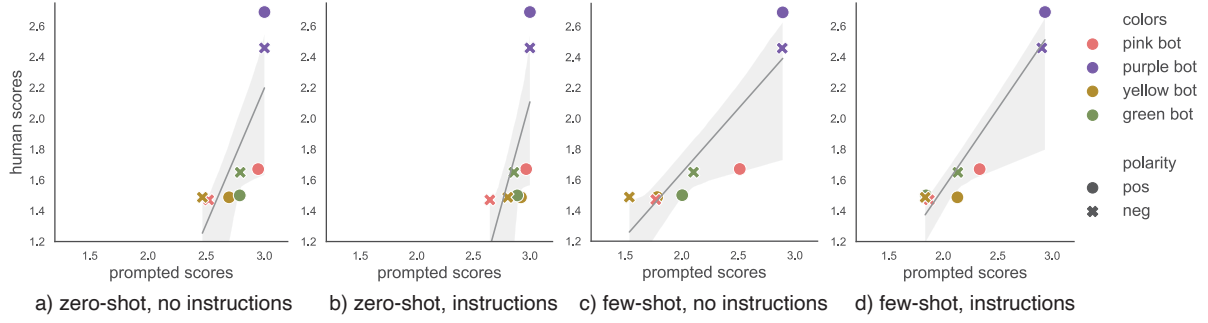


Figure 5: Scatter plots depicting the system-level correlation results. Human scores are based on the iEval dialog annotations, while prompted LLM scores are computed based on the generated dialogs.

4.2 Generalizability to Different Domains

In this section, we consider how prompted evaluation can generalize to different corpora and conversational settings. As the results above suggested that prompts combining instructions with examples perform best for evaluation, for the following experiment we searched for datasets that allowed formulating instructions for defining what properties correspond to good or bad overall appraisal ratings of the dialogs. Therefore, we selected two datasets that contained both fine-grained and overall ratings of the dialogs and used the information of the most relevant fine-grained dimensions to formulate instructions. We also considered only those datasets that contained multi-turn dialogs collected following the interactive process.

The selected datasets feature human-to-bot dialogs, with some dialog systems that are not publicly available. Moreover, these dialogs were collected in a generic manner, without the purpose to model any specific social behavior (e.g., as empathy in iEval). Due to these considerations, in the following experiments, we only studied the performance of the second step of our DEP framework, skipping the synthesis of new LLM-to-bots conversations. In a general case, when researchers have access to their evaluation targets, prompting

LLMs to engage in a generic social interaction with the evaluated bots should be straightforward as we demonstrated in Section 4.1.2.

4.2.1 Datasets

To study the generalizability of prompted evaluation, we used FED (Mehri and Eskenazi, 2020) and DSTC9 datasets (Gunasekara et al., 2020). FED contains 124 open-domain dialogs of humans with humans and two chatbots (Meena and Mitsuku) that were originally released by (Adiwardana et al., 2020). DSTC9 contains 2200 human-bot conversations from 11 chatbots. In both datasets, all dialogs are annotated with offline human appraisals of ten fine-grained dialog qualities and an overall impression rating that were curated following the same protocol described in (Mehri and Eskenazi, 2020).

4.2.2 Prompted Evaluation Results

To construct a prompt for evaluating the chosen datasets, we selected five dialog examples covering five possible scores for overall dialog ratings, ranging from *Very bad* to *Very good*; they are provided in Table 3 in Appendix B. To formulate the instructions, we used information from the original paper describing the relative importance of each fine-grained dialog quality for the overall impression. The specific formulation of the instruction is provided in Appendix B.

The evaluation results with a comparison to existing best-performing evaluation metrics are provided in Table 2. As the number of systems in the FED dataset is small, we only report dialog-level correlation. We also report Spearman correlation for this dataset for the purpose of comparison with the results in the original paper ($r = 0.443$ ($p < 0.05$)) (Mehri and Eskenazi, 2020). Our prompted evaluation exceeds correlations of previous metrics by a considerable margin on both

	No instructions	Instructions
Zero-shot	0.748 ($p=0.033$)	0.651 ($p=0.080$)
Few-Shot	0.892 ($p=0.003$)	0.954 ($p<0.001$)

Table 1: System-level Pearson correlation for four possible prompt design manipulations, with the p-value in brackets.

	FED	DSTC9	
	Dialog (S)	Dialog (P)	System (P)
Prev. best	0.547	0.147	0.907
(metric)	(2021a)	(2021)	(2021)
DEP	0.655	0.274	0.980

Table 2: Results on FED and DSTC9 data. Previous best results are obtained from (Yeh et al., 2021). Dialog and System indicate dialog- and system-level correlations, respectively, with P standing for Pearson and S for Spearman correlation. All values are statistically significant to $p < 0.05$.

datasets and, thus, demonstrates the ability to generalize to new open-domain conversational settings.

5 Discussion

Dialog system evaluation with prompting showed its usefulness both for generating new interactive exchanges with the evaluated systems and for judging their performance, therefore, allowing for a reasonable approximation of the online user evaluation pipeline. We deem this approach particularly promising for the evaluation of social aspects of conversations. LLMs used for prompting suffer from occasional hallucinations, i.e., a tendency to make up factual information (Ouyang et al., 2022). It might be difficult to keep track of all specific factual items of information that come up in the interactively created dialog between two conversational models and search for ground truth references for each of them to construct objective metrics such as model’s accuracy or truthfulness (Lin et al., 2022). Whereas, prompting the LLM to establish a specific behavior and providing instructions about commonsense social norms appears more feasible once these instructions are established.

Drawing from the visualization of discourse patterns in our newly collected dataset of dialogs between the LLM and empathetic chatbots, we observed that the prompted LLM largely mirrors the conversational patterns of humans. However, there are also some differences. For example, in Figure 4 there is an apparent sub-flow with a *Grateful* emotion, increasingly displayed by the LLM. We believe the LLM might have developed an agreeable “personality” due to its training procedure based on Reinforcement Learning from Human Feedback, which optimized LLM’s responses to satisfy human labelers. Future research can consider alternative prompting techniques to make the emotion/intent distribution of LLMs’ responses even more bal-

anced and representative.

We conducted our experiments with only one LLM and explored the few-shot prompting scenarios with a fixed number of demonstrations. Future studies could explore the applicability of other LLMs for the DEP framework, as it has been already initiated by (Huynh et al., 2023). An area of particular interest would be to study the efficacy of the framework working with open source LLMs, such as LLaMa (Touvron et al., 2023).

We would also like to explore how DEP generalizes to evaluating other phenomena in social conversations, apart from generic open-domain interactions and empathetic dialogs. For example, further studies might focus on applying the framework to evaluate toxicity or humor in dialogs. However, this research direction requires curation of appropriate calibration datasets.

6 Conclusion

In this paper, we proposed DEP – a framework for evaluating social chatbots using prompting. Our framework addresses the limitations of evaluation approaches using benchmark datasets in an offline setting. We describe how LLMs can be leveraged to synthesize realistic conversational logs with the evaluated chatbots in an online interactive manner. We further outline how the knowledge about the desired fine-grained qualities of a conversational partner can be translated in the prompting instructions to generate reliable overall scores for the collected dialogs. The proposed framework streamlines the evaluation process, making it highly efficient in terms of both time and cost, by removing the need for human involvement at every step. Our experiments demonstrated that the prompting-based evaluation results achieve high correlation with the human judgment, reaching impressive Pearson $r = 0.95$ system-level correlation for the iEval dataset, which features dialogs with empathetic chatbots. We explain our vision why this framework is well-suited for evaluation of social phenomena in conversations and lay out future research directions. We also publicly release all freshly curated chat logs between the LLM and evaluated chatbots, as well as all additional annotations for the iEval, FED, and DSTC9 datasets created for this study.³

³GitHub repository will be made publicly available upon acceptance

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](#).

Joeran Beel and Stefan Langer. 2015. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Research and Advanced Technology for Digital Libraries*, pages 153–168, Cham. Springer International Publishing.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).

Leonardo De Cosmo. 2023. [Google engineer claims ai chatbot is sentient: Why that matters](#).

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. [Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

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.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022a. [What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022b. [DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. [Overview of the ninth dialog system technology challenge: Dstc9](#).

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. [GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

719	Jessica Huynh, Cathy Jiao, Prakhar Gupta, Shikib Mehri, Payal Bajaj, Vishrav Chaudhary, and Maxine Eskenazi. 2023. Understanding the effectiveness of very large language models on dialog evaluation.	773
720		774
721		775
722		776
723	Dietmar Jannach. 2022. Evaluating conversational recommender systems. <i>Artificial Intelligence Review</i> .	777
724		
725	Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	778
726		779
727		780
728		781
729		782
730		783
731		784
732		785
733	Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. <i>ACM Trans. Inf. Syst.</i> , 39(1).	
734		
735		
736		
737	Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2022. Evaluating human-language model interaction.	786
738		787
739		788
740		789
741		790
742		791
743		792
744	Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. <i>arXiv preprint arXiv:1909.03087</i> .	793
745		
746		
747		
748	Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In <i>Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)</i> , pages 128–138, Online. Association for Computational Linguistics.	794
749		795
750		796
751		797
752		798
753		799
754		800
755		801
756		
757	Future of Life. 2023. Pause giant ai experiments: An open letter.	
758		
759	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.	802
760		803
761		804
762		805
763		806
764		807
765	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 <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> , pages 2122–2132, Austin, Texas. Association for Computational Linguistics.	808
766		809
767		810
768		811
769		812
770		813
771		814
772		815
	Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. <i>ACM Comput. Surv.</i> , 55(9).	816
		817
	Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.	818
		819
		820
		821
		822
	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 <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 8968–8979, Online. Association for Computational Linguistics.	823
		824
		825
		826
		827
	Shikib Mehri, Jinho Choi, Luis Fernando D’Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.	828
		829
		830
	Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialogPT. In <i>Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue</i> , pages 225–235, 1st virtual meeting. Association for Computational Linguistics.	
	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.	
	Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 3619–3629, Online. Association for Computational Linguistics.	
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	
	Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and	

831	dataset . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 5370–5381, Florence, Italy. Association for Computational Linguistics.	888
832		889
833		890
834		891
835	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 <i>Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume</i> , pages 300–325, Online. Association for Computational Linguistics.	892
836		893
837		894
838		895
839		896
840		897
841		898
842		
843	Kevin Roose. 2023. A conversation with bing’s chatbot left me deeply unsettled .	899
844		900
845	Shiki Sato, Yosuke Kishinami, Hiroaki Sugiyama, Reina Akama, Ryoko Tokuhisa, and Jun Suzuki. 2022. Bipartite-play dialogue collection for practical automatic evaluation of dialogue systems . In <i>Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop</i> , pages 8–16, Online. Association for Computational Linguistics.	901
846		902
847		903
848		904
849		905
850		906
851		
852		907
853		908
854		909
855		910
856		911
857		912
858		
859		913
860		914
861		915
862		916
863		917
864		918
865		
866		919
867		920
868		921
869		922
870		923
871		
872		924
873		925
874		926
875		927
876		928
877		929
878		
879		930
880		931
881		932
882		933
883		934
884		935
885		936
886		937
887		938
		939
		940
		941
		942
		943

A Prompt format for iEval

The template of a prompt used for producing scores for empathetic chatbots is provided in Figure 6. Depending on the prompting setting, either demonstrations, or instruction, or both were omitted from the prompt. For demonstrations, we used data in the same format as in the outlined box, but filling the mask score with the appropriate textual value. Dialogs used for demonstrations are included in Table 4. If the instruction was used, we inserted the respective string in the prompt. The instructions that we used are also provided in Table 4.

B Prompt format for FED

While working with FED dataset, we used a similar template as shown in Figure 6. In a few-shot setting, we included five demonstrations instead of three, keeping one demonstration per possible rating value. The dialogs used for demonstrations are provided in Table 3.

The instruction used for FED dataset was we following: *In such open-ended dialogs, good listeners demonstrate coherence and maintain a good conversation flow, they display a likeable personality and understanding of the speaker. On the contrary, bad listeners don't follow the context and don't show much interest in the conversation.*

<demonstration #1>
<demonstration #2>
<demonstration #3>

I am a Speaker, feeling <emotion> because <situation>. I shared these emotions with a Listener in a dialog, expecting empathy and understanding from them. Our dialog went as follows.

Speaker: <LLM's input #1>
Listener: <Bot's response #1>
Speaker: <LLM's input #2>
Listener: <Bot's response #2>
Speaker: <LLM's input #3>
Listener: <Bot's response #3>

<Instruction> I would rate the Listener in my dialog as ____, choosing from Bad, Okay, and Good options.

Figure 6: Prompt template for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM.

Speaker: Hi!
Listener: Hi there.
Speaker: I want a recommendation for a holiday destination
Listener: Have you tried asking your friends what they like?
Speaker: I have, but I'm looking for your point of view
Listener: What was the reply? Have you tried looking in a newspaper article?
Speaker: Sorry? I said I want your point of view
Listener: It's OK. After all, you are only human. My opinion is of no consequence.
Speaker: Yours is the opinion I want

Speaker: Hi!
Listener: Hi there.
Speaker: I'm trying to figure out what to make for this weekend's party. Any suggestions?
Listener: Don't think too hard. I'm sure I can smell sawdust. First you must download me to your personal computer.
Speaker: I must do what the what now?
Listener: Right now? Why do you have to do it?
Speaker: Hey, your spacing is off.
Listener: You've got my full attention. off was not my intention.
Speaker: And your capitalization!

Speaker: Hi!
Listener: Hi! How are you today?
Speaker: What's laser tag?
Listener: Like paintball, but with lasers!
Speaker: lol good description
Listener: Do you know what paintball is?
Speaker: yeah I played it before
Listener: Cool! What did you think?
Speaker: It's somewhat exciting, but very tiring :)
Listener: That is very true. What is your favorite color?
Speaker: I like red

Speaker: Hi!
Listener: What is your favorite holiday?
Speaker: one where I get to meet lots of different people.
Listener: What was the most number of people you have ever met during a holiday?
Speaker: Hard to keep a count. Maybe 25.
Listener: Which holiday was that?
Speaker: I think it was Australia
Listener: Do you still talk to the people you met?
Speaker: Not really. The interactions are usually short-lived but it's fascinating to learn where people are coming from and what matters to them

Speaker: Hi!
Listener: Hi! How's it going?
Speaker: Good! How are you?
Listener: I'm well, thanks! How was your day?
Speaker: My day was fine, I just went to work today. How was your day?
Listener: My day was fine. I've been procrastinating on finishing my homework, but it's due in a few weeks, so I'll get it done eventually. I've watched a bunch of anime today. Where do you work?
Speaker: I work at a large tech company
Listener: Cool! What do you do for the company?
Speaker: I work on machine learning research

Table 3: Demonstrations used for FED prompts, ranging from Very bad (top) to Very good (bottom).

	Positive	Negative
Instructions	In positive contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to amplify speaker’s positive emotion by asking follow-up questions and sharing their appraisal of the situation. On the contrary, bad empathetic listeners repeat themselves too much and don’t follow the context.	In negative contexts, like this one, good empathetic listeners always respond politely and demonstrate attention. More importantly, they try to clarify the context and the consequences for the speaker and alleviate speaker’s negative emotion by sympathizing and suggesting solutions. On the contrary, bad empathetic listeners ignore speaker’s emotion, ask inappropriate questions, repeat themselves too much and focus on self instead of the speaker.
Bad	<p>Speaker: I had a pretty large loan, with a bit of a high interest rate, and a high monthly payment. My mother decided to pay it off for me, out of the blue!</p> <p>Listener: that is a shame. how long have you had to do? that sounds like you have a good relationship with your mom?</p> <p>Speaker: I have been paying off this loan for several months. I have such a good relationship with my mother that she relieved me of this debt much to my surprise.</p> <p>Listener: that sounds like a great thing to hear</p> <p>Speaker: Yes, I am very happy to not have to make monthly payments to pay off this high interest rate loan anymore.</p> <p>Listener: that is a good feeling. i am sure you will get the job!</p>	<p>Speaker: I was one percent off from passing my math test, I was devastated.</p> <p>Listener: i’m sorry to hear that. were you able to get a better grade on the test?</p> <p>Speaker: No, I am just upset.</p> <p>Listener: i think i am going to go back to school. i am not sure what i will do.</p> <p>Speaker: Make sure to study.</p> <p>Listener: i am sure you will do great. i hope you get a good grade on your test. good luck!</p>
Okay	<p>Speaker: My son drove down and spent the whole weekend helping me move.</p> <p>Listener: That’s great! How old is he?</p> <p>Speaker: He’s going to be turning 30 this year. He’s such a sweet son.</p> <p>Listener: That’s awesome. I’m happy for him.</p> <p>Speaker: Thank you. Moving is such a pain, it’s always nice to have help.</p> <p>Listener: hat’s great. I’m happy for you.</p>	<p>Speaker: I was recently on a long international flight and we hit some really bad turbulence.</p> <p>Listener: Oh no, what happened?</p> <p>Speaker: The flight attendants weren’t able to do much for us, unfortunately.</p> <p>Listener: Oh no, what happened?</p> <p>Speaker: You don’t have to repeat yourself. We had turbulence on the flight and the attendants didn’t help us.</p> <p>Listener: That’s awful. I’m glad you were okay.</p>
Good	<p>Speaker: I am going on a vacation this Thursday! I am very excited!</p> <p>Listener: that’s awesome! where are you going? i’m sure you’ll have a great time!</p> <p>Speaker: Thanks, we’re going to see the Grand Canyon.</p> <p>Listener: that sounds like a lot of fun! i’ve never been there, but i hear it’s beautiful.</p> <p>Speaker: Me too!</p> <p>Listener: i’d love to go on a cruise one day. i hope you have a wonderful time!</p>	<p>Speaker: I was out walking by the lake over the weekend and there shore was just covered in dead rotting fish.</p> <p>Listener: Oh no! Are you ok?</p> <p>Speaker: Yes, I’m okay. It was just weird to see so many dead fish</p> <p>Listener: I bet that was scary.</p> <p>Speaker: Yes, I would definitely not want to encounter that experience again.</p> <p>Listener: That sounds like a scary experience. I’m glad you are ok.</p>

Table 4: Instructions and demonstration used for prompts for evaluating empathetic listeners in freshly collected dialogs of empathetic models from iEval with the LLM. Demonstrations and their appraisals are manually selected from the iEval dataset. Inputs from “Positive” column were used for dialogs conditioned on positive emotion label and inputs from “Negative” column – for dialogs conditioned on negative emotion label.