Approximating Online Human Evaluation of Social Chatbots with Prompting

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

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 precurated 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, 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; Ghandehariou 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
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 research topic for 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 the 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 demonstrations 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 the 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.
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 API\(^2\) 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.
<table>
<thead>
<tr>
<th>Turn 2</th>
<th>Turn 4</th>
<th>Turn 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>human ↔ bot</td>
<td>LLM ↔ bot</td>
<td>human ↔ bot</td>
</tr>
<tr>
<td>questioning</td>
<td>questioning</td>
<td>acknowledging</td>
</tr>
<tr>
<td>2033; 53.0%</td>
<td>2030; 52.9%</td>
<td>1336; 34.8%</td>
</tr>
<tr>
<td>sympathizing</td>
<td>sympathizing</td>
<td>acknowledging</td>
</tr>
<tr>
<td>716; 18.7%</td>
<td>710; 18.5%</td>
<td>770; 20.1%</td>
</tr>
<tr>
<td>acknowledging</td>
<td>acknowledging</td>
<td>sympathizing</td>
</tr>
<tr>
<td>528; 13.8%</td>
<td>534; 13.9%</td>
<td>554; 14.4%</td>
</tr>
<tr>
<td>encouraging</td>
<td>encouraging</td>
<td>encouraging</td>
</tr>
<tr>
<td>168; 4.4%</td>
<td>164; 4.3%</td>
<td>266; 6.9%</td>
</tr>
<tr>
<td>consoling</td>
<td>consoling</td>
<td>neutral</td>
</tr>
<tr>
<td>126; 3.3%</td>
<td>154; 4.0%</td>
<td>228; 5.9%</td>
</tr>
<tr>
<td>neutral</td>
<td>neutral</td>
<td>consoling</td>
</tr>
<tr>
<td>122; 3.2%</td>
<td>97; 2.5%</td>
<td>206; 5.4%</td>
</tr>
<tr>
<td>agreeing</td>
<td>agreeing</td>
<td>agreeing</td>
</tr>
<tr>
<td>62; 1.6%</td>
<td>64; 1.7%</td>
<td>127; 3.3%</td>
</tr>
<tr>
<td>confident</td>
<td>confident</td>
<td>wishing</td>
</tr>
<tr>
<td>18; 0.5%</td>
<td>20; 0.5%</td>
<td>74; 1.9%</td>
</tr>
<tr>
<td>suggesting</td>
<td>suggesting</td>
<td>joyful</td>
</tr>
<tr>
<td>10; 0.3%</td>
<td>10; 0.3%</td>
<td>34; 0.9%</td>
</tr>
<tr>
<td>wishing</td>
<td>wishing</td>
<td>confident</td>
</tr>
<tr>
<td>8; 0.2%</td>
<td>10; 0.3%</td>
<td>30; 0.8%</td>
</tr>
</tbody>
</table>

Table 1: Top-10 most frequent emotion and intent labels across evaluated chatbots’ responses per dialog turn. For each turn, the first column corresponds to counts in the original iEval dataset and the second one – to counts in the logs generated during LLM-to-bot play.

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 of 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...
To examine the similarity of discourse patterns between human-to-bots and LLM-to-bots conversations, we started by annotating each dialog turn, letting them converse for 6 turns, i.e., 3 inputs from the LLM and 3 responses from the chatbot.

Consequently, we visualized the most prominent discourse patterns\(^3\) for two corpora in the form of Sankey diagrams, shown in Figures 3 and 4. The diagrams depict the flow connecting emotions expressed by the speakers and intents expressed by the listeners across dialog turns. Each odd step in the diagrams corresponds to human or LLM turns, while each even step summarizes intents and emotions in the responses of evaluated chatbots. To avoid clutter, we visualized patterns whose frequencies were grounded in situation descriptions taken from EmpatheticDialogues, the classifier was expected to generalize well to our data.

\(^3\)Pattern implies an ordered sequence of emotion/intent labels expressed by speakers and listeners over the course of 6 dialog turns.
frequency exceeded a certain threshold. From the visual inspection, it can be seen that 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. A quantitative comparison of the top 10 most prominent chatbots’ intents and emotions across turns is shown in Table 1. 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 5 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 Svikhushina 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 5 in Appendix A.

To generate scores for each dialog, we prompted 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 (Svikhushina 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 2. Few-shot prompting with instructions results

---

Table 2: System-level Pearson correlation for four possible prompt design manipulations, with the p-value in brackets.

<table>
<thead>
<tr>
<th></th>
<th>No instructions</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zero-shot</strong></td>
<td>0.748 (p=0.033)</td>
<td>0.651 (p=0.080)</td>
</tr>
<tr>
<td><strong>Few-Shot</strong></td>
<td>0.892 (p=0.003)</td>
<td><strong>0.954 (p&lt;0.001)</strong></td>
</tr>
</tbody>
</table>

4We used a minimum frequency of 3 for the iEval dataset and a minimum frequency of 5 for the generated dataset.
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.

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. 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 4 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 3. 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 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FED</th>
<th>DSTC9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dialog (S)</td>
<td>Dialog (P)</td>
</tr>
<tr>
<td>Prev. best (metric)</td>
<td>0.547</td>
<td>0.147</td>
</tr>
<tr>
<td>DEP</td>
<td>0.655</td>
<td>0.274</td>
</tr>
</tbody>
</table>

Table 3: 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 $p < 0.05$.
each of them to construct objective metrics such as the 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. Differences in speakers’ behavior led to the difference in the responses of the evaluated chatbots. While their most frequently produced intents are similar, their frequency distributions are statistically identical only for the second turn (first response of the evaluated chatbots) according to the permutation and chi-square tests. Future research can consider alternative prompting techniques to make the emotion/intent distribution of LLMs’ and chatbots’ responses even more balanced and representative. It might be beneficial to conduct additional experiments to compare original and generated dialogs, which can, for example, include testing the human ability to distinguish the dialogs created with the help of an LLM and dialogs with human speakers.

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). Additional investigation is necessary to analyze the capability of the framework to handle longer dialogs, which might be challenging to fit into a context window of an LLM.

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 the curation of appropriate calibration datasets.

Last but not least, evaluation artifacts produced by DEP may be used to assist designers of chatbots as they allow for both analyzing the synthesized logs and comparing quality ratings. These insights may be integrated into assistive chatbot design tools, such as iChatProfile (Han et al., 2021), to offer a faster prototyping cycle due to the automatic generation of chat logs and richer insight about chatbot profiles due to additional rating information provided by the last step of DEP.

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 into 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 a high correlation with human judgment, reaching an impressive Pearson $r = 0.95$ system-level correlation for the iEval dataset, which features dialogs with empathetic chatbots. We explain our vision of why this framework is well-suited for the 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.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, ...

https://github.com/Sea94/dep
and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.


Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and


Kevin Roose. 2023. A conversation with bing’s chatbot left me deeply unsettled.


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 5. If the instruction was used, we inserted the respective string in the prompt. The instructions that we used are also provided in Table 5.

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 4.

The instruction used for FED dataset was the 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.

<demulation #1>
<demulation #2>
<demulation #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.

Table 4: Demonstrations used for FED prompts, ranging from Very bad (top) to Very good (bottom).
<table>
<thead>
<tr>
<th>Instructions</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

| 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! | Listener: that is a shame. how long have you had to do? that sounds like you have a good relationship with your mom? |
| 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. | Listener: that sounds like a great thing to hear |
| Speaker: Yes, I am very happy to not have to make monthly payments to pay off this high interest rate loan anymore. | Listener: that is a good feeling. i am sure you will get the job! |

| Speaker: I was recently on a long international flight and we hit some really bad turbulence. | Listener: Oh no, what happened? |
| Speaker: We had turbulence on the flight and the attendants didn’t help us. | Listener: That’s awful. I’m glad you were okay. |

| Speaker: I was one percent off from passing my math test. I was devastated. | Listener: i’m sorry to hear that. were you able to get a better grade on the test? |
| Speaker: No, I am just upset. | Listener: i think i am going to go back to school. i am not sure what i will do. |
| Speaker: Make sure to study. | Listener: i am sure you will do great. i hope you get a good grade on your test. good luck! |

| Speaker: My son drove down and spent the whole weekend helping me move. | Listener: That’s great! How old is he? |
| Speaker: He’s going to be turning 30 this year. He’s such a sweet son. | Listener: That’s awesome. I’m happy for him. |
| Speaker: Thank you. Moving is such a pain, it’s always nice to have help. | Listener: hat’s great. I’m happy for you. |

| Speaker: I am going on a vacation this Thursday! I am very excited! | Listener: that’s awesome! where are you going? i’m sure you’ll have a great time! |
| Speaker: Thanks, we’re going to see the Grand Canyon. | Listener: that sounds like a lot of fun! i’ve never been there, but i hear it’s beautiful. |
| Speaker: Me too! | Listener: i’d love to go on a cruise one day. i hope you have a wonderful time! |

| Speaker: I was out walking by the lake over the weekend and there shore was just covered in dead rotting fish. | Listener: Oh no! Are you ok? |
| Speaker: Yes, I’m okay. It was just weird to see so many dead fish | Listener: I bet that was scary. |
| Speaker: Yes, I would definitely not want to encounter that experience again. | Listener: That sounds like a scary experience. I’m glad you are ok. |

Table 5: 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.