Does GPT-3 Generate Empathetic Dialogues?
A Novel In-Context Example Selection Method and Automatic Evaluation Metric for Empathetic Dialogue Generation

Young-Jun Lee  Chae-Gyun Lim  Ho-Jin Choi
School of Computing, KAIST
{yj2961,rayote,hojinc}@kaist.ac.kr

Abstract
Since empathy plays a crucial role in increasing social bonding between people, many studies have designed their own dialogue agents to be empathetic using the well-established method of fine-tuning. However, they do not use prompt-based in-context learning, which has shown powerful performance in various natural language processing (NLP) tasks, for empathetic dialogue generation. Although several studies have investigated few-shot in-context learning for empathetic dialogue generation, an in-depth analysis of the generation of empathetic dialogue with in-context learning remains unclear, especially in GPT-3 (Brown et al., 2020). In this study, we explore whether GPT-3 can generate empathetic dialogues through prompt-based in-context learning in both zero-shot and few-shot settings. To enhance performance, we propose new in-context example selection methods, called SITSM and EMOSITSM, that utilize emotion and situational information. We also introduce a new automatic evaluation method, DIFF-EPIHOME, which reflects the human tendency to express empathy. From the analysis, we reveal that our DIFF-EPIHOME is effective in measuring the degree of human empathy. We show that GPT-3 achieves competitive performance with Blender 90M, a state-of-the-art dialogue generative model, on both automatic and human evaluation.

1 Introduction
Empathy refers to the ability to understand another person’s experiences and feelings. This is important for increasing social bonding (rapport) with conversation partners (Zech and Rimé, 2005). Empathy is a multi-dimensional concept consisting of two main aspects: cognitive and affective (Davis et al., 1980). Since Rashkin et al. (2018) released the EMATZEDIALOOGES dataset for empathetic dialogue generation task, previous studies have improved their dialogue agents to generate more empathetic dialogues (Lin et al., 2019; Majumder et al., 2020; Zheng et al., 2021; Kim et al., 2021b; Sabour et al., 2021; Li et al., 2022). Depending on how the dialogue agents are trained, these approaches are largely divided into two categories depending on how to train own dialogue agents: (i) training from scratch or (ii) fine-tuning a pretrained dialogue generative model. However, neither of these approaches uses the prompt-based in-context learning paradigm in zero-shot and few-shot settings.

Recently, many researchers have attempted to build large-scale language models (LLMs), such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), and HyperCloVA (Kim et al., 2021a). These models have shown surprising performance in various NLP tasks via prompt-based in-context learning, which is a new paradigm learning technique. Previous studies have explored the effect of few-shot in-context learning on dialogue generation tasks (Zheng and Huang, 2021; Madotto et al., 2021). While Madotto et al. (2021) explored few-shot in-context learning for empathetic dialogue generation, they did not conduct an in-depth analysis of their ability to generate empathetic dialogues. Moreover, they did not leverage GPT-3 as a prompting language model.

In this study, we explore whether GPT-3 generates empathetic dialogues using prompt-based in-context learning in both zero-shot and few-shot settings. We simply designed the prompt, which is a modified version of the basic prompt provided by OpenAI. As pointed out in (Liu et al., 2021), it is important to carefully choose in-context examples to enhance few-shot performance. Inspired by empathy being a multi-dimensional concept (Davis et al., 1980), we propose SITSM and EMOSITSM selection methods that choose in-context examples based on emotion and situation information. To reflect the human tendency to express empathy, we...
also propose a new automatic evaluation method called DIFF-EPITOME, which is an extended version of EPITOME (Sharma et al., 2020). Our main contributions are as follows.

- We conduct an in-depth analysis of GPT-3’s ability to generate empathetic dialogues with respect to Empathy, Diversity, and Fluency.
- We introduce SitSM and EmoSitSM, which are in-context example selection methods for empathetic dialogue generation task.
- We propose DIFF-EPITOME, an automatic evaluation method for empathetic dialogue generation. This method measures how dialogue agents empathize using the difference of EPITOME scores between human and agent.
- We show that GPT-3 performs better than the state-of-the-art model (Blender 90M (Roller et al., 2020)) on the EMPATHETIC DIALOGUES test set, without additional training. In human evaluation, regardless of the dialogue turn setting, we also show that human annotators prefer GPT-3’s responses on both human rating and A/B test.

2 Related Work

Empathetic Dialogue Generation Rashkin et al. (2018) first introduced the EMPATHETIC DIALOGUES dataset. Lin et al. (2019) proposed a mixture of empathetic listeners (MoEL), where each listener is specialized in how to understand and respond appropriately to each emotion. Majumder et al. (2020) generated empathetic responses by mimicking human emotions, grouping emotions, and imposing stochasticity into each emotion group. Sharma et al. (2020) introduced a conceptual framework EPITOME (described in §3.3.1). Welivita and Pu (2020) proposed a taxonomy of empathetic response intents, consisting of nine categories (in Appendix D). For convenience, this is referred to as EMPINTENT. Zheng et al. (2021) proposed a multi-factor hierarchical framework (CoMAE), which considers EPITOME, EMPINTENT, and emotion. Kim et al. (2021b) generated more specific empathetic responses focused on emotion cause words by utilizing the Rational Speech Acts (RSA) framework (Frank and Goodman, 2012). Sabour et al. (2021) leveraged commonsense to generate more empathetic responses. Li et al. (2022) also leveraged external knowledge, such as commonsense knowledge, to explicitly generate empathetic responses.

Prompt-based In-Context Learning Since Brown et al. (2020) first introduced prompt-based in-context learning, many studies have shown that large-scale language models (e.g., GPT-3) itself has the ability to solve various NLP tasks in both zero-shot and few-shot settings (Schick and Schütze, 2020; Liu et al., 2021; Mishra et al., 2021; Wei et al., 2021; Yoo et al., 2021; Zhao et al., 2021; Schick and Schütze, 2021; Kim et al., 2021a; Gutiérrez et al., 2022; Meng et al., 2022). Some studies have shown that prompt-based few-shot in-context learning can also be successfully applied in dialogue generation tasks (Zheng and Huang, 2021; Madotto et al., 2021). The advantage of in-context learning is that it does not require any additional training. However, one problem is that GPT-3 achieves unstable performance depending on in-context examples. To mitigate this problem, Liu et al. (2021) proposed a kNN-augmented in-context example selection approach called the KATE. In this study, we extended this method to empathetic dialogue generation by selecting relevant in-context examples based on the situation and emotion (in §3.2).

3 Methodology

3.1 Task Formulation

The empathetic dialogue generation task aims to generate an empathetic response \( y \) for a given input \( x \) by maximizing the conditional probability \( p(y|x) = \prod_t p(y_t|x, y_1, \ldots, y_{t-1}) \), where \( x \) denotes the dialogue context. In general, previous studies (Lin et al., 2019; Majumder et al., 2020; Li et al., 2022) trained their own models on EMPATHETIC DIALOGUES. However, in our case, we attempted to solve the task through GPT-3 in-context learning (Brown et al., 2020), without additional training. Therefore, in this study, task formulation is defined as follows:

\[
p(y|x, C) = \prod_t p(y_t|C, x, y_1, \ldots, y_{t-1}),
\]

where \( C = \{x_1, y_1, x_2, y_2, \ldots, x_k, y_k\} \) is a concatenated string, and \( k \) denotes the number of examples for in-context few-shot learning. In a zero-shot setting \((k = 0)\), we do not provide any in-context examples \((C = \emptyset)\).
As reported by (Liu et al., 2021), GPT-3 is sensitive to randomly chosen in-context examples. To mitigate this problem, they selected semantically relevant in-context examples from the training set using the kNN retrieval module for each test input. Inspired by (Liu et al., 2021), we introduce two selection methods: SITSM and EMOITSM. In EMPATHETICDIALOGUES, each training instance consists of dialogue context \( x \), golden response \( y \), emotion \( e \), and situation sentence \( s \). Table 1 shows the samples of the in-context examples selected by SITSM and EMOITSM.

### 3.2 In-Context Example Selection Methods (SM)

As reported by (Liu et al., 2021), GPT-3 is sensitive to randomly chosen in-context examples. To mitigate this problem, they selected semantically relevant in-context examples from the training set using the kNN retrieval module for each test input. Inspired by (Liu et al., 2021), we introduce two selection methods: SITSM and EMOITSM. In EMPATHETICDIALOGUES, each training instance consists of dialogue context \( x \), golden response \( y \), emotion \( e \), and situation sentence \( s \). Table 1 shows the samples of the in-context examples selected by SITSM and EMOITSM.

#### 3.2.1 SITSM

Starting from the assumption that the situation sentences are similar, the dialogue context will have similar patterns of expressing empathy. Specifically, we first use the sentence encoder \(^1\) to obtain all the embedding vectors of situation sentences in the training set in advance. We convert each test situation input \( s \) into a vector representation. For each test situation input \( s \), we then select the most relevant \( k \) examples from the training set based on the similarity score. For the similarity measures, we adopt the cosine similarity. We construct the prompt with the selected \( k \) examples, where the ordering of \( k \) examples was performed based on the similarity score of each example. In other words, the example most similar to the test input \( s \) is placed close to the test input.\(^2\) The entire process is presented in Algorithm 1.

#### 3.2.2 EMOITSM

Empathy is a multi-dimensional concept that consists of two aspects: cognitive and affective (Davis et al., 1980). Based on this concept, we argue that we should choose good in-context examples based on these two aspects. The cognitive aspect involves understanding and interpreting the situation of another person. The affective aspect is to express an emotional reaction. We can view the situation as the cognitive aspect, and emotion as affective aspects.

\(^1\)We use \texttt{sts-b-roberta-large} version of Sentence-BERT (Reimers and Gurevych, 2019)

\(^2\)As argued in (Liu et al., 2021), the choice of ordering is data-dependent. In this study, we adopt the reverse order that performs best on the Natural Questions (NQ) (Kwiatkowski et al., 2019) dataset.
Table 1: Samples of selected in-context examples (when $k = 2$) from EMPATHETICDIALOGUES training set, based on the similarity score with a given test situation. We also present dialogue contexts corresponding to situation sentences, respectively. (S: Speaker, L: Listener) More samples are in Appendix E.

3.3 A New Automatic Evaluation Metric for Empathetic Dialogue Generation

We propose an automatic evaluation metric, called DIFF-EPI
tOME, which reflects human patterns when empathy is expressed as dialogue continues. In §3.3.1, we describe the EPI
tOME-based metric used in previous studies (Sharma et al., 2020; Kim et al., 2021b). Empirically, we analyze whether there is a specific tendency in human communication (see §3.3.2). Based on the above analysis, we propose a new evaluation method called DIFF-
EPI
tOME (see §3.3.3).

3.3.1 EPI
tOME-based Automatic Evaluation

EPI
tOME, introduced by (Sharma et al., 2020), is a new conceptual framework for expressing empathy in text-based, asynchronous contexts. EPI
tOME consists of three communication mechanisms of empathy: Explorations (EX), Interpretations (IP), and Emotional Reactions (ER). The mechanisms are described in Appendix C.

In a recent study (Kim et al., 2021b), each mechanism was used as an automatic metric to measure the empathy of generated responses using a fine-tuned RoBERTa (Liu et al., 2019) model.3 Each generated response was measured by one of the val-

---

3 Actually, in (Kim et al., 2021b), they only used IP and EX scores. However, we even use the ER score in the experiments (see in Table 3)
We ponder: How do humans empathize? We humans perceive and understand another person’s situation by putting ourselves in the other’s shoes. This is known as perspective-taking in cognitive science (Davis et al., 1980). Even following perspective-taking, it is difficult to accurately recognize another person’s situation at the beginning of a dialogue. Therefore, humans tend to ask their situation and feelings. Through a simple experiment, we observe that there is a tendency to express empathy in human communication. Figure 1 shows that the average EX scores decrease as the dialogue continued. In the IP and ER scores, each goes up and down slightly.

3.3.3 DIFF-EPITOME-based Automatic Evaluation
Based on the above analysis, we propose a new automatic evaluation method DIFF-EPITOME, which is an extended version of the EPITOME-based method. The key idea of DIFF-EPITOME is to measure the difference in EP
m
score between the human golden response \( y_i \) and the predicted response \( \hat{y}_i \) using a model, as follows:

\[
\text{diff-EP}_m(Y) = \frac{1}{N} \sum_{i=1}^{N} (\text{EP}_m(y_i) - \text{EP}_m(\hat{y}_i))^2
\]

where \( m \in \{ \text{IP, EX, ER} \} \).

A lower diff-EP value indicates that the expressed empathy is more human-like.

4 Experimental Setup
4.1 Dataset
We evaluate our proposed model on the benchmark EMPATHETICDIALOGUES dataset (Rashkin et al., 2018), which consists of 25k open-domain conversations grounded in emotional situations. Each dialogue is composed of consecutive utterances of the speaker and listener, where each utterance is labeled among 32 emotion categories. Each dialogue contains a situation sentence.

4.2 Evaluation Models
Blender We compare GPT-3 with Blender 90M (Roller et al., 2020), which is one of the state-of-the-art dialogue agents, fine-tuned on the EMPATHETICDIALOGUES train dataset as our baseline.

EmpGPT-3 To observe whether a prompt specialized to the empathetic dialogue generation task elicits GPT-3 to produce more empathetic responses, we construct a simple prompt template which is “The following is a conversation with an empathetic AI assistant. The assistant empathizes
with human experiences and feelings well. Human: $u_1$ Empathy AI: $u_2$ ...", where $u_1, u_2, ...$ are utterances. We present examples of the constructed prompt used in this study in the Appendix A.

### 4.3 Implementation Details.

We fine-tune the Blender 90M (Roller et al., 2020) on the EMPATHETIC DIALOGUES dataset using a ParlAI framework\(^5\). We used the default hyperparameter settings provided by the ParlAI framework. We selected the model checkpoint that achieved the best performance, based on the perplexity of the validation set. For EmpGPT-3, we use a *davinci* version with hyperparameter settings as follows: temperature 0.8, maximum tokens 128, frequency penalty 0.4, and presence penalty 0.4. For the stop tokens, we use Human: and Empathy AI:.

### 4.4 Automatic Evaluations

To investigate whether GPT-3 can generate empathetic responses in both zero-shot and few-shot settings, we evaluate the generated responses on various metrics for Diversity, Fluency, and Empathy.

#### 4.4.1 Diversity

It is important to consider diversity because various responses to an input utterance may be possible depending on the context. We measure the diversity of generated responses based on two metrics.

- **DISTINCT-N (DIST-N)** (Li et al., 2015; See et al., 2019a) measures the ratio of unique $n$-grams. A higher ratio indicates a higher diversity of generated responses.
- **NIDF**\(^6\) (See et al., 2019b) measures the rareness of a word $w$. The NIDF score is calculated as:

$$
\text{NIDF}(w) = \frac{\text{IDF}(w) - \text{min}_\text{idf}}{\text{max}_\text{idf} - \text{min}_\text{idf}},
$$

where $\text{IDF}(w) = \log(R/c_w)$, $R$ denotes the number of responses in dataset, $c_w$ is the number of responses that contain $w$, $\text{min}_\text{idf}$ and $\text{max}_\text{idf}$ are the minimum and maximum IDF values. Detailed information is described in (See et al., 2019b). A higher NIDF score indicates a more specific response and a higher proportion of rare words.

#### 4.4.2 Fluency

Following (Feng et al., 2020; Pang et al., 2020), we measure the fluency of generated responses through a perplexity (PPL) by adopting GPT2-XL, not fine-tuned on any downstream tasks related to the dialogue domain. A lower PPL indicates that the response is more fluent.

#### 4.4.3 Empathy

- **EMOACC** measures an emotion accuracy using a fine-tuned BERT-base (Devlin et al., 2018) model on the EMPATHETIC DIALOGUES dataset labeled with 32 emotion categories. The performance of the classifier is reported in Table 2.
- **INTENTACC** measures the response intent accuracy using a fine-tuned BERT model on the EMPRESS dataset, introduced by (Welivita and Pu, 2020). The performance of the classifier is reported in Table 2.
- **EPITOME** (Sharma et al., 2020) measures IP, EX, and ER by leveraging fine-tuned RoBERTa models, respectively (§3.3.1).
- **DIFF-EPITOME** (§3.3.3) measures the difference scores of IP, EX, ER between the human golden response and predicted response (§3.3.3).

### 4.5 Human Evaluation

Following (Rashkin et al., 2018; Lin et al., 2019; Majumder et al., 2020; Kim et al., 2021b), we conduct two standard human evaluations with three annotators: (i) Human A/B Test and (ii) Human Ratings. We recruited three annotators via an on-campus announcements. After randomly sampling 100 test examples, we divided them into 50 examples for each single-turn and multi-turn setting. The Human A/B Test allows annotators to choose which response is more empathetic. They can choose "Tie" if the two given responses are both good or

<table>
<thead>
<tr>
<th>Model</th>
<th># classes</th>
<th>Acc</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMOACC</td>
<td>32</td>
<td>0.40</td>
<td>0.39</td>
</tr>
<tr>
<td>INTENTACC</td>
<td>9</td>
<td>0.96</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 2: Performance of BERT-based classifiers trained on EMPATHETIC DIALOGUES (Rashkin et al., 2018).

\(^5\)https://github.com/facebookresearch/ParlAI

\(^6\)Normalized Inverse Document Frequency
bad. For the Human Ratings, we asked three annotators to rate the generated responses on three metrics (in a 4-likert scale): EMPATHY, RELEVANCE, and FLUENCY. The questionnaires and system used for the human evaluation are described in Appendix G and H.

5 Experimental Results

5.1 Main Results

GPT-3 vs. Blender 90M As shown in Table 3, GPT-3 shows competitive performance compared to Blender 90M on most evaluation metrics (8 of 12, except for Avg. Len) in a zero-shot setting. Regardless of the turn setting, GPT-3 is difficult to generate responses with proper intentions than Blender 90M. Owing to the enormous generative capacity of GPT-3, EmpGPT3 can generate more diverse and specific responses. For DIFF-EPIHOME, Blender tends to generate overly emotional expressions because of its higher performance in both ER and diff-ER (1.0570 and 1.0359). However, EmpGPT3 still cannot follow how humans empathize in terms of the IP and EX.

single-turn vs. multi-turn The main difference between these two settings is whether the entire dialogue context is given together when constructing the prompt. For EMPATHY, EmpGPT3 achieves lower diff-[IP,EX,ER] scores than the single-turn setting. This suggests that, given the dialogue context in the zero-shot setting, GPT-3 better understands human situations and expresses empathy just as humans do. Similarly, the performance of INTENTACC, which requires reasoning about situations, has also improved. For FLUENCY, EmpGPT-3 generates more fluent responses from the average PPL with a large margin of 50.96.

5.2 Ablation Studies

Number of In-Context Examples As shown in Table 4, we explore the effect of the number of in-context examples on the EmpGPT-3’s performance. Specifically, we conduct an experiment on the EMPATHETICDIALOGUES test set with $k = \{1, 2\}$. To select adequate in-context examples, we adopt our EMOSITSM, which achieves a better performance (see Table 11). The overall few-shot performance is better than that when $k = 0$. In particular, we observe that fluency when $k = 2$ is much higher than those for others ($k = \{0, 1\}$). In addition, the diff-[EX,ER] scores of EmpGPT-3 are much lower than those of the zero-shot performance. This implies that GPT-3 indirectly learns how to express empathy from given in-context examples. Full experiment results are shown in Table 11 (see Appendix B).

Various Selection Methods We investigate the performance of GPT-3 according to the selection method. Table 5 shows that the similarity-based methods (i.e., SITSM and EMOSITSM) have slightly improved performance in most metrics compared with the RANDOM method (similar results were reported in (Liu et al., 2021)). The RANDOM method selects in-context examples randomly. In particular, EMOSITSM is highly effective in terms of emotion accuracy compared with
Table 6: Comparison of EmpGPT-3 (in single- and multi-turn) with Blender (Roller et al., 2020) on A/B test. The win rate is for EmpGPT-3.

<table>
<thead>
<tr>
<th>Model</th>
<th>Win</th>
<th>Lose</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-EmpGPT-3 vs. Blender</td>
<td>46.7%</td>
<td>32.0%</td>
<td>21.3%</td>
</tr>
<tr>
<td>multi-EmpGPT-3 vs. Blender</td>
<td>36.7%</td>
<td>36.7%</td>
<td>26.6%</td>
</tr>
</tbody>
</table>

Table 7: Comparison of EmpGPT-3 (in single- and multi-turn) with Blender (Roller et al., 2020) on human ratings.

<table>
<thead>
<tr>
<th>Model</th>
<th>EMPATHY↑</th>
<th>RELEVANCE↑</th>
<th>FLUENCY↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blender</td>
<td>2.78</td>
<td>2.82</td>
<td>3.34</td>
</tr>
<tr>
<td>single-EmpGPT-3</td>
<td>3.07</td>
<td>2.97</td>
<td>3.6</td>
</tr>
<tr>
<td>Blender</td>
<td>2.79</td>
<td>2.78</td>
<td>3.13</td>
</tr>
<tr>
<td>multi-EmpGPT-3</td>
<td>2.85</td>
<td>2.8</td>
<td>3.21</td>
</tr>
</tbody>
</table>

5.3 Human Evaluation Results

As shown in Table 6, users prefer responses generated by single-EmpGPT-3 to those generated by Blender. When comparing multi-EmpGPT-3 with Blender, users prefer responses from both the models equally. We measure the inter-rater agreement using Krippendorff’s $\alpha$. For Human A/B Test, Krippendorff’s $\alpha$ is 0.26, which implies a fair agreement. Regardless of the dialogue turn setting, EmpGPT-3 obtains a better performance on human ratings. Especially, users who evaluate responses from single-EmpGPT-3 to be more empathetic and relevant to the given dialogue context, as shown in Table 7.

5.4 Analysis of Correlation

We conducted a correlation analysis to verify the validity of the proposed evaluation metric DIFF-EPIHOME. Figure 2 shows Pearson’s $r$ correlation matrix between human ratings and two automatic methods: EPIHOME-based and DIFF-EPIHOME-based. We observe that our DIFF-EPIHOME-based automatic metric more correlates with human ratings than the EPIHOME-based automatic metric. Moreover, we found that a high ER score does not indicate that the dialogue agent empathizes well. It suggests that it is necessary to use emotional reactions on time when expressing empathy to interlocutors. Correlation analysis revealed that our proposed metric is effective for empathetic dialogue generation and it is important to consider the tendency of how humans do empathize as the dialogue continues. We hope that this analysis will be helpful for other researchers.
5.5 Case Studies

Table 8 shows examples of the responses generated by the Blender and EmpGPT-3 (with single- and multi-turn settings). Our multi-EmpGPT-3 can generate responses that require complex reasoning by understanding a speaker’s situation and feelings. Additional examples are presented in Table 13.

6 Conclusion

In this study, we explore the zero-shot and few-shot performance of GPT-3 in an empathetic dialogue generation task on various metrics with respect to Diversity, Fluency, and Empathy. We introduce a new in-context example selection method, SITSM and EMOITSM. We also propose a novel automatic evaluation method, DIFF-EPITOME, for empathetic dialogue generation. We show that GPT-3 achieves competitive performance with Blender 90M on the EMPATHETICDIALOGUES test set on both automatic and human evaluations. From the correlation analysis, we reveal that DIFF-EPITOME correlates more with human ratings. In future work, we will apply OPT (Zhang et al., 2022) with an optimized prompt. In addition, we reflect on the overall human tendency to express empathy in the modeling.

Acknowledgements

This work was supported by the Technology Innovation Program (Grant number: 20012288, Development of Affective Virtual TA service based on deep learning for foreign language education) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References


Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy.


Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.


A Prompt For Empathetic Dialogue Generation

Our prompt template for EmpGPT-3 is as follows.

The following is a conversation with an empathetic AI assistant. The assistant empathizes with human experiences and feelings well.

Human:

Table 9 and Table 10 show the constructed prompt in the zero-shot and few-shot setting, respectively.

Table 9: An example of constructed prompt when $k = 0$. The blue represents a generated response from multi-EmpGPT-3.

Table 10: An example of constructed prompt when $k = 2$. The blue represents a generated response from multi-EmpGPT-3 with EMOStSM.
B Full Results

Table 11 shows the zero-shot performance \((k = 0)\) and few-shot performance \((k = 1, 2)\) according to various selection methods (i.e., RANDOM, SitSM, and EMOSitSM) on various automatic evaluation metrics.

<table>
<thead>
<tr>
<th>k</th>
<th>type</th>
<th>INTENTACC</th>
<th>EMOACC</th>
<th>IP</th>
<th>EX</th>
<th>ER</th>
<th>diff-IP</th>
<th>diff-EX</th>
<th>diff-ER</th>
<th>dist-1</th>
<th>dist-2</th>
<th>NIDF</th>
<th>PPL</th>
<th>Avg. Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>RANDOM</td>
<td>0.2528</td>
<td>0.1594</td>
<td>0.2717</td>
<td>0.4970</td>
<td>0.6439</td>
<td>0.7884</td>
<td>1.2564</td>
<td>0.7295</td>
<td>0.9400</td>
<td>0.9666</td>
<td>0.2840</td>
<td>118.43</td>
<td>15.6</td>
</tr>
<tr>
<td>1</td>
<td>RANDOM</td>
<td>0.2682</td>
<td>0.1582</td>
<td>0.2772</td>
<td>0.4368</td>
<td>0.6443</td>
<td>0.8669</td>
<td>1.0412</td>
<td>0.6800</td>
<td>0.9672</td>
<td>0.9984</td>
<td>0.2859</td>
<td>108.84</td>
<td>15.39</td>
</tr>
<tr>
<td></td>
<td>SitSM</td>
<td>0.2599</td>
<td>0.1582</td>
<td>0.2992</td>
<td>0.3643</td>
<td>0.6196</td>
<td>0.9015</td>
<td>1.0491</td>
<td>0.6635</td>
<td>0.9646</td>
<td>0.9982</td>
<td>0.2866</td>
<td>136.97</td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td>EMOSitSM</td>
<td>0.2689</td>
<td>0.1653</td>
<td>0.3921</td>
<td>0.3337</td>
<td>0.6431</td>
<td>0.9015</td>
<td>1.0491</td>
<td>0.6635</td>
<td>0.9646</td>
<td>0.9982</td>
<td>0.2866</td>
<td>136.97</td>
<td>16.4</td>
</tr>
<tr>
<td>2</td>
<td>RANDOM</td>
<td>0.2717</td>
<td>0.1523</td>
<td>0.2740</td>
<td>0.3219</td>
<td>0.6211</td>
<td>0.9894</td>
<td>0.6749</td>
<td>0.9688</td>
<td>0.9982</td>
<td>0.2914</td>
<td>125.53</td>
<td>14.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SitSM</td>
<td>0.2693</td>
<td>0.1665</td>
<td>0.3466</td>
<td>0.3322</td>
<td>0.6349</td>
<td>0.9057</td>
<td>0.7953</td>
<td>0.6832</td>
<td>0.9689</td>
<td>0.9987</td>
<td>0.2907</td>
<td>136.12</td>
<td>15.15</td>
</tr>
<tr>
<td></td>
<td>EMOSitSM</td>
<td>0.2721</td>
<td>0.1818</td>
<td>0.2733</td>
<td>0.3102</td>
<td>0.6280</td>
<td>0.8528</td>
<td>1.0003</td>
<td>0.6643</td>
<td>0.9661</td>
<td>0.9982</td>
<td>0.2905</td>
<td>83.83</td>
<td>15.51</td>
</tr>
</tbody>
</table>

Table 11: Evaluation results of zero-shot and few-shot learning with different in-context examples \(k = 1, 2\) and with various selection methods on various automatic evaluation metrics: Empathy, and Diversity, Fluency.

C Explanation of EPITOME Framework

As we mentioned earlier, EPITOME (Sharma et al., 2020) comprises three mechanisms: IP, EX, and ER. The mechanisms are described as follows:

- **EXPLORATIONS (EX)** are expressions of active interest in the interlocutor’s situation.

- **INTERPRETATIONS (IP)** are expressions of acknowledgments or understanding of the interlocutor’s emotion or situation.

- **EMOTIONAL REACTIONS (ER)** are expressions of emotions such as warmth, compassion, and concern in the interlocutor’s situation.

D A Taxonomy of Empathetic Response Intents

There are 9 categories: Agreeing, Acknowledging, Encouraging, Consoling, Sympathizing, Suggesting, Questioning, Wishing, and Neutral.

E Selected In-Context Examples

Table 12 shows more selected in-context examples.
<table>
<thead>
<tr>
<th>Test situation</th>
<th>Content</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>situation 1</strong></td>
<td>My eldest son just graduated from High School and I was so happy for him.</td>
<td>-</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>When my brother graduated high school, I was very proud of him, it was a big accomplishment.</td>
<td>0.9479</td>
</tr>
<tr>
<td>S: My brother graduated high school, I was very proud of him!</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: I know that feel, my brother graduated a year ago, it’s a really big milestone.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: It is, somewhat common, but still, I am proud of him all the same!</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: Yes I agree, it really signifies the start of their next chapter in life.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>situation 2</strong></td>
<td>My son recently graduated from high school. I am so happy about it!</td>
<td>0.9765</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>S: My son recently graduated from high school.</td>
<td>-</td>
</tr>
<tr>
<td>L: That’s great. What is he doing now?</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: He is preparing for college. I am so happy about it!</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: That’s even more awesome. I hope he does well.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>situation 1</strong></td>
<td>I have a nest of yellow jackets in my front yard</td>
<td>-</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>I ripped my pants on bourbon street the other day. Luckily I was wearing a long shirt.</td>
<td>0.4424</td>
</tr>
<tr>
<td>S: I went out last weekend and had a major accident. Guess what happened...</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: Are you ok, you have to tell me what happened.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: I’m fine. Just a little embarrassed. I ripped my leggings dancing on bourbon street.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: Ahh that has happened to everyone before. It is embarrassing but you will get over it.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: Yep. My shirt was long enough to cover it. Plus I don’t live there lol.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: Well im glad you were able to cover up.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>situation 2</strong></td>
<td>There’s a huge stuffed bear on my yard.</td>
<td>0.4509</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>S: I play the lottery sometimes here, and finally won a small amount.</td>
<td>-</td>
</tr>
<tr>
<td>L: That is so thoughtful of him, did you guys catch a ball?</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: No, but it was still fun.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: I am sure it was! :)</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

(a) Sample of selected in-context example by S|T|S|M.

<table>
<thead>
<tr>
<th>Test situation</th>
<th>Content</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>situation 1</strong></td>
<td>I had a job interview today and I think it went really well.</td>
<td>-</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>I had a great job interview the other day. Im really feeling good about how it went.</td>
<td>0.9284</td>
</tr>
<tr>
<td>S: I had a great job interview the other day. Im really feeling good about how it went.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: That’s fantastic! Hopefully you’ll hear something about it soon.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: I should be. I just feel that I did really well.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: I’m sure you did. Think positive!</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>situation 2</strong></td>
<td>I was bowling yesterday and the ball got stuck on my hand. I went with it.</td>
<td>0.9602</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>I just went on a job interview. I feel like it went really well.</td>
<td>-</td>
</tr>
<tr>
<td>S: I just got back from a job interview. It went really well. I feel I might get an offer.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: What job did you interview for?</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: It was for a Financial Analyst job. I really want the job.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: That’s amazing, you must be so excited right now</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>situation 1</strong></td>
<td>I play the lottery regularly and finally won a small amount the other night much to my shock.</td>
<td>0.3834</td>
</tr>
<tr>
<td>Dialogue Context</td>
<td>S: I play the lottery sometimes here, and finally won a small amount.</td>
<td>-</td>
</tr>
<tr>
<td>Not a big deal, but I was shocked it finally happened haha.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S: Yeah it’s really demoralizing until you at least win something. A nice little bonus anyway.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>L: Very true! Congrats! It’s hard to never win</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

(b) Sample of selected in-context example by EMOS|T|S|M when the emotion of “hopeful” (upper row) and “surprised” (bottom row). We also present dialogue contexts corresponding to situation sentences, respectively. (S: Speaker, L: Listener)

Table 12: Samples of selected in-context examples (when \(k = 2\)) from EMPATHETICDIALOGUES training set, based on the similarity score with a given test situation.
More Examples of Generated Dialogues

See Table 13 for more examples of responses generated by Blender (Roller et al., 2020) and EmpGPT-3.

<table>
<thead>
<tr>
<th>Dialogue Context</th>
<th>Generated Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>S: I just got hulu live. It’s been amazing to watch tv again! L: I haven’t watched actual tv in sooo long. What do you watch on hulu? S: I love to watch home shows on hgtv and bachelor in paradise on abc. Hulu live has been a great value for the price!</td>
<td>Blender: I love hulu live. I have it on my phone. single-EmpGPT-3: Can you elaborate?</td>
</tr>
<tr>
<td>S: I think storms are wonderful, except when the power goes out. L: I love storms too but yeah, that’s a nuisance. S: It was extra loud and my husband was at work, bit scary.</td>
<td>Blender: I’m sorry to hear that. I hope you guys can get it fixed soon. single-EmpGPT-3: So you felt scared when you heard the loud sound?</td>
</tr>
<tr>
<td>S: I was hurrying around the house and wasn’t paying attention fully, I ending up hitting my toe on my dogs toys. L: Dang dogs leaving their things around! S: It really hurt, I think I may need to take all his toys away to punish him.</td>
<td>Blender: I would do that. multi-EmpGPT-3: I can kind of see your reasoning for that, but dogs love their toys.</td>
</tr>
<tr>
<td>S: I have been looking after my wife this week. She has the flu L: oh poor thing, is she at least getting better? S: It seems so. Sleeping all day today lol</td>
<td>Blender: I hope she gets better soon! multi-EmpGPT-3: haha that’s so hard. I hope she gets better soon</td>
</tr>
</tbody>
</table>

Table 13: Examples of generated responses from Blender (Roller et al., 2020) and EmpGPT-3 with single- and multi-turn setting on EMPATHETICDIALOGUES test set. (S: Speaker, L: Listener)
G Human Evaluation Questionnaire

This is a list of questions and multiple-choice options for each evaluation metric used in the human evaluation.

- **Empathy**: Do you think this response shows an understanding of situation and empathize appropriately?
  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Relevance**: Is this response relevant to given dialogue context?
  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Fluency**: Does this response seem contextually natural? Could you understand this response?
  Options: 1: Very unnatural / 2: Mostly unnatural / 3: Mostly natural / 4: Very natural

H Human Evaluation System

Figure 3 is a screenshot of human evaluation system.

![Screenshot of the human evaluation system for empathetic dialogue generation.](image)

Figure 3: Screenshot of the human evaluation system for empathetic dialogue generation.