

# 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-EPITOME, which reflects the human tendency to express empathy. From the analysis, we reveal that our DIFF-EPITOME 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. Our code is available at <https://github.com/passing2961/EmgPT-3>.

## 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 EMPATHETICDIALOGUES 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 EMPATHETICDIALOGUES 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 EMPATHETICDIALOGUES 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  $p(y|x) = \prod_t p(y_t|x, y_1, \dots, y_{t-1})$ , 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 EMPATHETICDIALOGUES. 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:

$$p(y|x, C) = \prod_t^{ |y| } p(y_t|C, x, y_1, \dots, y_{t-1}),$$

where  $C = \{x_1, y_1, x_2, y_2, \dots, x_k, y_k\}$  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 ( $C = \emptyset$ ).

---

**Algorithm 1: SITSM In-Context Example Selection**

---

**Input:** A training dataset  $\mathcal{D} = \{(x_i, y_i, s_i, e_i)\}_{i=1}^N$ , number of training examples  $N$ , a sentence encoder  $f_\theta(\cdot)$ , number of in-context examples  $k$ , a test input  $(x_{test}, y_{test}, s_{test}, e_{test})$

**Output:** a prompt input  $P$  to GPT-3

```
/* Step 1: Prepare  $\mathcal{M}$  */
1  $\mathcal{M} \leftarrow$  empty list
2 for  $(x_i, y_i, s_i, e_i) \in \mathcal{D}$  do
3    $v_i = f_\theta(s_i)$ 
4    $\mathcal{M}.\text{append}(v_i)$ 
5 end
/* Step 2: Get similarity score */
6  $v_{test} = f_\theta(s_{test})$ 
7  $\mathcal{M}_{sim} \leftarrow$  empty list
8 for  $v_i \in \mathcal{M}$  do
9    $sim_i = \frac{v_{test} \cdot v_i}{\|v_{test}\|_2 \|v_i\|_2}$ 
10   $\mathcal{M}_{sim}.\text{append}(sim_i)$ 
11 end
/* Step 3: Construct prompt with
selected  $k$  examples */
12 Select  $k$  indices  $\mathcal{I} = \{idx_j\}_{j=1}^k$  from sorted  $\mathcal{M}_{sim}$ 
(in descending order)
13  $\mathcal{M}_{ctx} \leftarrow$  empty list
14 for  $idx_j \in \mathcal{I}$  do
15    $\mathcal{M}_{ctx}.\text{append}((x_{idx_j}, y_{idx_j}))$ 
16 end
17  $C = [x_{idx_k}; y_{idx_k}; \dots; x_{idx_1}; y_{idx_1}]$ 
18  $P = [C; x_{test}; y_{test}]$ 
```

---

### 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 EMOSITSM. 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 EMOSITSM.

#### 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<sup>1</sup> 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

<sup>1</sup>We use stsb-roberta-large version of SentenceBERT (Reimers and Gurevych, 2019)

---

**Algorithm 2: EMOSITSM In-Context Example Selection**

---

**Input:** A training dataset  $\mathcal{D} = \{(x_i, y_i, s_i, e_i)\}_{i=1}^N$ , number of training examples  $N$ , an emotion  $E = \{e\}_1^{32}$ , a sentence encoder  $f_\theta(\cdot)$ , number of in-context examples  $k$ , a test input  $(x_{test}, y_{test}, s_{test}, e_{test})$

**Output:** a prompt input  $P$  to GPT-3

```
/* Step 1: Group  $\mathcal{D}$  by 32 emotions
and prepare  $\mathcal{M}$  */
1  $\mathcal{M} \leftarrow$  empty dict
2 for  $e \in E$  do
3   for  $(x_i, y_i, s_i, e_i) \in \mathcal{D}$  do
4      $v_i = f_\theta(s_i)$ 
5      $\mathcal{M}[e].\text{append}((x_i, y_i, s_i, e_i, v_i))$ 
6   end
7 end
/* Step 2: Get similarity score */
8  $e_{test} \leftarrow$  emotion of test input
9  $v_{test} = f_\theta(s_{test})$ 
10  $\mathcal{M}_{sim} \leftarrow$  empty list
11 for  $(x_i, y_i, s_i, e_i, v_i) \in \mathcal{M}[e_{test}]$  do
12    $sim_i = \frac{v_{test} \cdot v_i}{\|v_{test}\|_2 \|v_i\|_2}$ 
13    $\mathcal{M}_{sim}.\text{append}(sim_i)$ 
14 end
/* Step 3: Construct prompt with
selected  $k$  examples */
15 Select  $k$  indices  $\mathcal{I} = \{idx_j\}_{j=1}^k$  from sorted  $\mathcal{M}_{sim}$ 
(in descending order)
16  $\mathcal{M}_{ctx} \leftarrow$  empty list
17 for  $idx_j \in \mathcal{I}$  do
18    $\mathcal{M}_{ctx}.\text{append}((x_{idx_j}, y_{idx_j}))$ 
19 end
20  $C = [x_{idx_k}; y_{idx_k}; \dots; x_{idx_1}; y_{idx_1}]$ 
21  $P = [C; x_{test}; y_{test}]$ 
```

---

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.<sup>2</sup> The entire process is presented in Algorithm 1.

#### 3.2.2 EMOSITSM

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* as-

<sup>2</sup>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.

	Content	Score
<b>Test situation</b>	I'm starting a new job next week, and I am super nervous.	-
<b>situation 1</b>	I start my new job tomorrow. I am extremely nervous about it	0.9741
Dialogue Context	S: I start my new job tomorrow.	
	L: Nice! What will you be doing?	
	S: I will be working as a consultant, I am extremely nervous	-
	L: Nothing to be nervous about, I'm sure you'll do fine!	
<b>situation 2</b>	I am starting a new job next week. I am so nervous.	0.9937
Dialogue Context	S: MY new job starts next week.	
	L: Awesome! What will you be doing?	
	S: Im transferring to another store and becoming a supervisor. Im nervous cause of the position and its all new people.	-
	L: That is awesome. They chose you for a reason. You will do great!	

(a) Sample of selected in-context example by SITSM.

	Content	Score
<b>Test situation</b>	I remain loyal to my wife always	-
<b>situation 1</b>	I've always been loyal to my wife.	0.9716
Dialogue Context	S: I've always been faithful as a husband, I pride myself on that.	
	L: Nice, there is not a lot of people like you	
	S: Well, I don't know about that, but I actually wanted to marry my wife.	-
	L: Good	
<b>situation 2</b>	i am loyal to my wife and i'll always be	0.9786
Dialogue Context	S: i am loyal to my wife and i'll always be.	
	L: That's an amazing attitude. Not many people are like that nowadays.	
	S: i don't get it why people cheat and hurt each other,	-
	but i suppose because they are never happy with what they have and only feel grateful for it after they lose it.	
	L: Crazy world we live in.	

(b) Sample of selected in-context example by EMOSITSM when "faithful" emotion.

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.

pect. Fortunately, EMPATHETICDIALOGUES contains situation and emotion information for each instance. To this end, we propose EMOSITSM, which selects in-context examples based on emotions and situations. Specifically, we first group all training instances by 32 emotion types and encode each situation sentence  $s$  into a vector representation by using the sentence encoder (same model in SITSM) simultaneously. We then calculate the similarity score between the test input and all instances in the group with the same emotion as that of the test input. Finally, based on the similarity scores, we select  $k$  examples that are closest to the test input. Similar to SITSM, we also consider the ordering of  $k$  selected examples when constructing the final prompt. The algorithm EMOSITSM is presented in Algorithm 2.

### 3.3 A New Automatic Evaluation Metric for Empathetic Dialogue Generation

We propose an automatic evaluation metric, called DIFF-EPITOME, which reflects human patterns when empathy is expressed as dialogue continues.

In §3.3.1, we describe the EPITOME-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-EPITOME (see §3.3.3).

**3.3.1 EPITOME-based Automatic Evaluation**  
EPITOME, introduced by (Sharma et al., 2020), is a new conceptual framework for expressing empathy in text-based, asynchronous contexts. EPITOME 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.<sup>3</sup> Each generated response was measured by one of the val-

<sup>3</sup>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)



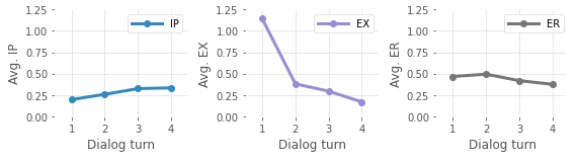


Figure 1: Analysis of EMPATHETICDIALOGUES train set w.r.t. EPITOME, as the conversation continues. The x-axis represents the index of dialogue turn and the y-axis represents the average score of IP, EX, ER for each dialogue turn. Each score was predicted by the RoBERTa model (described in §3.3.1).

ues (0, 1, or 2) predicted from the model. Higher values indicate stronger empathy. Following the author’s official code<sup>4</sup>, we fine-tuned three RoBERTa-base models to measure the IP, EX, and ER scores. The EPITOME-based metric is formulated as

$$EP_m(y) = \text{RoBERTa}_m(x, y) \quad (1)$$

$$EP_m(Y) = \frac{1}{N} \sum_{i=1}^N EP_m(\hat{y}_i) \quad (2)$$

where  $m \in \{\text{IP}, \text{EX}, \text{ER}\}$ , the number of test examples  $N$ , and  $Y = \{y_1, \dots, y_N\}$ .

We ask: "Can we argue that empathetic dialogue agents empathize well with high scores measured by EPITOME-based metrics?" To validate our hypothesis, we check the results of Blender 90M (Roller et al., 2020) trained on EMPATHETICDIALOGUES. Given the input utterance, "I feel like deepening my connection to god is the most important thing in my life," Blender produces a response "I'm sorry to hear that. What is your connection to god?", which is measured using the EPITOME-based method with an IP of 0, EX of 2, and ER of 2. This indicates more exploration of the situation of a partner and expressing emotional reactions. On the other hand, the IP, EX, and ER scores of the golden response "I think that is a lovely thing" are all measured to be 0. According to the EPITOME-based method, Blender expresses stronger empathy, which is regarded as an empathetic dialogue agent. However, in this case, humans do not express empathy. Therefore, it is necessary to develop a new automatic evaluation method.

### 3.3.2 How do humans empathize?

We ponder: *How do humans empathize?* We humans perceive and understand another person’s situation by putting ourselves in the other’s shoes.

<sup>4</sup><https://github.com/behavioral-data/Empathy-Mental-Health>

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:

$$\text{diff-EP}_m(Y) = \frac{1}{N} \sum_{i=1}^N (EP_m(y_i) - EP_m(\hat{y}_i))^2 \quad (3)$$

where  $m \in \{\text{IP}, \text{EX}, \text{ER}\}$

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

Model	# classes	Acc	Macro F1
EMOACC	32	0.40	0.39
INTENTACC	9	0.96	0.90

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

with human experiences and feelings well. Human:  $u_1$  Empathy AI:  $u_2 \dots$ ", where  $u_1, u_2, \dots$  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 EMPATHETICDIALOGUES dataset using a ParlAI framework<sup>5</sup>. 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<sup>6</sup>** (See et al., 2019b) measures the rareness of a word  $w$ . The NIDF score is calculated as:

$$\text{NIDF}(w) = \frac{\text{IDF}(w) - \text{min\_idf}}{\text{max\_idf} - \text{min\_idf}}, \quad (4)$$

<sup>5</sup><https://github.com/facebookresearch/ParlAI>

<sup>6</sup>Normalized Inverse Document Frequency

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  $w$ ,  $\text{min\_idf}$  and  $\text{max\_idf}$  are the minimum and maximum IDF. 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 EMPATHETICDIALOGUES 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 EMPINTENT 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

Model	Empathy					Diversity			Fluency				
	INTENTACC	EMOACC	IP	EX	ER	diff-IP	diff-EX	diff-ER	dist-1	dist-2	NIDF	PPL	Avg. Len
Blender (single-turn)	<b>0.3084</b>	0.1593	0.2057	0.3423	<b>1.0570</b>	<b>0.7202</b>	<b>1.1934</b>	1.0359	0.9541	0.9824	0.2454	166.86	12.58
EmpGPT3 (single-turn)	0.2211	<b>0.1683</b>	<b>0.2780</b>	0.3118	0.8142	0.8625	1.5364	0.9239	<b>0.9614</b>	<b>0.9975</b>	<b>0.2860</b>	169.39	<b>16.03</b>
EmpGPT3 (multi-turn)	0.2528	0.1594	0.2717	<b>0.4970</b>	0.6439	0.7884	1.2564	<b>0.7295</b>	0.9400	0.9966	0.2840	<b>118.43</b>	15.6

Table 3: Comparison of the zero-shot performance of EmpGPT-3 with Blender 90M (Roller et al., 2020) on EMPATHETICDIALOGUES test set. In a single-turn setting, we inject only the last utterance with the prompt template, not including the whole dialogue context, into GPT-3. In contrast, in a multi-turn setting, we consider the whole dialogue context when constructing the prompt.

$k$	Empathy			Diversity		Fluency	
	INTENTACC	EMOACC	diff-IP	diff-EX	diff-ER	NIDF	PPL
0	0.2528	0.1594	<b>0.7884</b>	1.2564	0.7295	0.2840	118.43
1	<b>0.2650</b>	<b>0.1622</b>	0.8418	<b>0.9925</b>	0.6655	0.2896	175.71
2	0.2623	0.1614	0.8481	0.9988	<b>0.6514</b>	<b>0.3045</b>	<b>74.39</b>

Table 4: Ablation study on the number of in-context examples  $k$  in EmpGPT-3 prompts. Evaluation results are conducted on the EMOSITSM.

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-EPITOME, Blender tends to generate overly emotional expressions because of its higher performance in both ER and diff-ER (1.0570 and 1.0359). However, EmpGPT-3 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**, EmpGPT-3 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

type	Empathy							
	INTENTACC	EMOACC	IP	EX	ER	diff-IP	diff-EX	diff-ER
RANDOM	0.2603	0.1390	<b>0.2670</b>	0.3204	0.6141	0.8905	0.9894	0.6592
SITSM	0.2587	0.1461	0.2387	<b>0.3298</b>	<b>0.6282</b>	<b>0.8025</b>	<b>0.9706</b>	0.6678
EMOSITSM	<b>0.2623</b>	<b>0.1614</b>	0.2599	0.3094	0.6219	0.8481	0.9988	<b>0.6514</b>

Table 5: Comparison of EMPATHY performance of EmpGPT-3 with various selection methods when  $k = 2$  and multi-turn setting.

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

Model	Win	Lose	Tie
<i>single</i> -EmpGPT-3 vs. Blender	<b>46.7%</b>	32.0%	21.3%
<i>multi</i> -EmpGPT-3 vs. Blender	<b>36.7%</b>	<b>36.7%</b>	26.6%

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.

Model	EMPATHY $\uparrow$	RELEVANCE $\uparrow$	FLUENCY $\uparrow$
Blender	2.78	2.82	3.34
<i>single</i> -EmpGPT-3	<b>3.07</b>	<b>2.97</b>	<b>3.6</b>
Blender	2.79	2.78	3.13
<i>multi</i> -EmpGPT-3	<b>2.85</b>	<b>2.8</b>	<b>3.21</b>

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

other methods. However, SITSM shows a better performance in diff- $\{IP, EX\}$ , demonstrating that SITSM better understands and explores situations. We report full experiment results in Table 11 (Appendix B).

### 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-EPITOME. Figure 2 shows Pearson’s  $r$  correlation matrix between human ratings and two automatic methods: EPITOME-based and DIFF-EPITOME-based. We observe that our DIFF-EPITOME-based automatic metric more correlates with human ratings than the EPITOME-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 in-

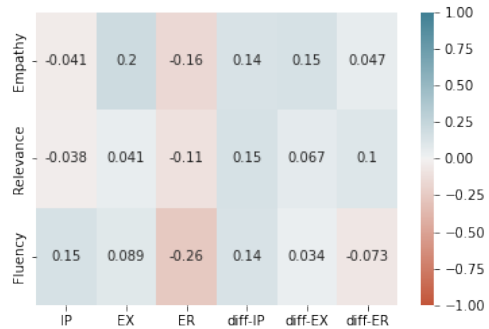


Figure 2: Pearson’s  $r$  correlation matrix between human ratings and EPITOME- and DIFF-EPITOME-based automatic metrics, respectively. The degree of correlation increases from red to blue.

Dialogue Context	
S:	Do you even know how crazy it is to skydive? I must have been absolutely nuts when I did it.
L:	Oh I don’t think I could do that. I have a fear of falling from high places.
S:	It gave me the biggest rush that’s for sure. But on the way down I was saying my prayers.
L:	I think I would pass out from fear lol.
S:	You should do it sometime. It’s fun to take chances.
Generated Responses	
<b>Blender:</b>	I have never skydive. I have heard it is pretty scary.
<b><i>single</i>-EmpGPT-3:</b>	No, I cannot say that I do. But I think I can imagine.
Dialogue Context	
S:	I have some great friends. One of them drove me to the airport tomorrow
L:	That was really nice of them. I bet you’re appreciative.
S:	Yes I am, they are great
Generated Responses	
<b>Blender:</b>	I’m glad you have some good friends.
<b><i>multi</i>-EmpGPT-3:</b>	I’m sure they’ll like to hear that.

Table 8: Examples of generated responses from Blender (Roller et al., 2020) and EmpGPT-3 with single- and multi-turn setting. (S: Speaker, L: Listener)

terlocutors. 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 EMOSITSM. 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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genau: Data

augmentation for finetuning text generators. *arXiv preprint arXiv:2010.01794*.

- Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.
- Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*.
- Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021a. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650*.
- Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021b. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.
- 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. *arXiv preprint arXiv:2010.01454*.
- Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing instructional prompts to gptk’s language. *arXiv preprint arXiv:2109.07830*.
- Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, Kewei Tu, et al. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.
- Timo Schick and Hinrich Schütze. 2020. It’s not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.
- Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019a. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019b. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.
- Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.
- Emmanuelle Zech and Bernard Rimé. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 12(4):270–287.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.
- Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.
- Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: a multi-factor hierarchical framework for empathetic response generation. *arXiv preprint arXiv:2105.08316*.

## 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.

---

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

Human: I'm so excited because I'm finally going to visit my parents next month! I didn't see them for 3 years

Empathy AI: 3 years is a long time. How come?

Human: I moved to the us and they stayed in my homecountry

Empathy AI: Oh, yeah. That is a long time. I can imagine how excited you are to go home.

---

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

---

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

Human: When I was eight my parents got me a dog for Christmas.

Empathy AI: A dog is an awesome gift for a little kid. I bet you were super excited.

Human: To this day I remember it as the best Christmas of my life.

Empathy AI: I bet! Maybe I should get my kids a dog for Christmas, I know it would be a great choice for them.

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

Human: Im already getting pumped up for christmas this year...

Empathy AI: I love Christmas!

Human: Me too especially this year.. new apartment for us new state. Whats your fav part of the holiday?

Empathy AI: Seeing my kids open their gifts. I never make it to Christmas because I'm so excited to see them lol

Human: Its such a good feeling right, the best! :)

Empathy AI: I never get anything but that's fine

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

Human: I couldn't wait to go to the concert.

Empathy AI: Which Concert?

Human: The U2 concert. Tickets were really expensive and I never thought we would be able to go, but somehow we did!!!

Empathy AI: It was awesome! I love U2.

---

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

## 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.

k	type	Empathy						Diversity			Fluency			
		INTENTACC	EMOACC	IP	EX	ER	diff-IP	diff-EX	diff-ER	dist-1	dist-2	NIDF	PPL	Avg. Len
0	-	0.2528	0.1594	0.2717	0.4970	0.6439	0.7884	1.2564	0.7295	0.9400	0.9966	0.2840	118.43	15.6
1	RANDOM	0.2682	0.1582	0.2772	0.3683	0.6443	0.8669	1.0412	0.6800	0.9672	0.9984	0.2859	108.84	15.39
	SITSM	0.2599	0.1582	0.2992	0.3643	0.6196	0.9015	1.0491	0.6635	0.9646	0.9982	0.2866	136.97	16.4
	EMOSITSM	0.2689	0.1653	0.2921	0.3337	0.6431	0.8433	0.9941	0.6710	0.9649	0.9980	0.2855	180.95	16.3
2	RANDOM	0.2717	0.1523	0.2740	0.3219	0.6211	0.9046	0.9894	0.6749	0.9688	0.9982	0.2914	125.53	14.84
	SITSM	0.2693	0.1665	0.2466	0.3322	0.6349	0.8057	0.9753	0.6832	0.9689	0.9987	0.2907	136.12	15.15
	EMOSITSM	0.2721	0.1818	0.2733	0.3102	0.6280	0.8528	1.0003	0.6643	0.9661	0.9982	0.2905	83.83	15.51

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.



	Content	Score
<b>Test situation</b>	My eldest son just graduated from High School and I was so happy for him.	-
<b>situation 1</b>	When my brother graduated high school, I was very proud of him, it was a big accomplishment	0.9479
Dialogue Context	S: My brother graduated high school, I was very proud of him! L: I know that feel, my brother graduated a year ago, it's a really big milestone. S: It is, somewhat common, but still, I am proud of him all the same! L: Yes I agree, it really signifies the start of their next chapter in life.	-
<b>situation 2</b>	My son recently graduated from high school. I am so happy about it!	0.9765
Dialogue Context	S: My son recently graduated from high school. L: That's great. What is he doing now? S: He is preparing for college. I am so happy about it! L: That's even more awesome. I hope he does well.	-
<b>Test situation</b>	I have a nest of yellow jackets in my front yard	-
<b>situation 1</b>	I ripped my pants on bourbon street the other day. Luckily I was wearing a long shirt.	0.4424
Dialogue Context	S: I went out last weekend and had a major accident. Guess what happened... L: Are you ok, you have to tell me what happened. S: I'm fine. Just a litte embarassed. I ripped my leggings dancing on bourbon street. L: Ahh that has happened to everyone before. It is embarrassing but you will get over it. S: Yep. My shirt was long enough to cover it. Plus I don't live there lol. L: Well im glad you were able to cover up.	-
<b>situation 2</b>	There's a huge stuffed bear on my yard.	0.4509
Dialogue Context	S: There's a huge stuffed bear on my yard. L: That sounds creepy S: Agree. Not sure what I should do with it. L: I guess ignore it for now	-

(a) Sample of selected in-context example by SITSM.

	Content	Score
<b>Test situation</b>	I had a job interview today and i think it really well.	-
<b>situation 1</b>	I had a great job interview the other day. Im really feeling good about how it went.	0.9284
Dialogue Context	S: I had a great job interview the other day. Im really feeling good about how it went. L: That's fantastic! Hopefully you'll hear something about it soon. S: I should be. I just feel that I did really well. L: I'm sure you did. Think positive!	-
<b>situation 2</b>	I just went on a job interview. I feel like it went really well.	0.9602
Dialogue Context	S: I just got back from a job interview. It went really well. I feel I might get an offer. L: What job did you interview for? S: It was for a Financial Analyst job. I really want the job. L: That's amazing, you must be so excited right now	-
<b>Test situation</b>	I went bowling yesterday and the ball got stuck on my hand. I went with it.	-
<b>situation 1</b>	I was at home and at the last minute my father took me to the Yankee game.	0.3834
Dialogue Context	S: I was at home and last minute my dad took me to the Yankees game. L: That is so thoughtful of him, did you guys catch a ball? S: No, but it was still fun. L: I am sure it was! :)	-
<b>situation 2</b>	I play the lotto regularly and finally won a small amount the other night much to my shock.	0.4339
Dialogue Context	S: I play the lottery sometimes here, and finally won a small amount. Not a big deal, but I was shocked it finally happened haha. L: I would be too, I never win those things! S: Yeah it's really demoralizing until you at least win something. A nice little bonus anyway. L: Very true! Congrats! It's hard to never win	-

(b) Sample of selected in-context example by EMOSITSM 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.

## F More Examples of Generated Dialogues

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

Dialogue Context
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!
Generated Responses
<b>Blender:</b> I love hulu live. I have it on my phone. <b>single-EmpGPT-3:</b> Can you elaborate?
Dialogue Context
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.
Generated Responses
<b>Blender:</b> I'm sorry to hear that. I hope you guys can get it fixed soon. <b>single-EmpGPT-3:</b> So you felt scared when you heard the loud sound?
Dialogue Context
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.
Generated Responses
<b>Blender:</b> I would do that. <b>multi-EmpGPT-3:</b> I can kind of see your reasoning for that, but dogs love their toys.
Dialogue Context
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
Generated Responses
<b>Blender:</b> I hope she gets better soon! <b>multi-EmpGPT-3:</b> haha that's so hard. I hope she gets better soon

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.

The screenshot shows a web-based interface for human evaluation. It is divided into three main sections:

- User Assignments:** Contains a text input for 'E-mail (\*Unique ID)' with the value 'passing2961@gmail.com' and a 'Change Email..' button. Below it is a 'List of Conversations' with a scrollable list of conversation IDs from 'conv\_m\_031' to 'conv\_m\_045'. A 'Selected File' section shows 'conv\_m\_032'.
- Chat History:** Displays a conversation log for 'Log file: conv\_m\_032'. It shows a sequence of messages: a yellow bubble 'I really love my life right now.', a blue bubble 'Thats really good to hear. Whats going on?', a yellow bubble 'Nothing really! I just love that I can work at home and have time to clean or cook whenever I want.', a blue bubble 'I work from home too. Its so nice to be able to do laundry and things on my breaks. Im so productive around the house', a yellow bubble 'Right? And no having to interact with others unless we choose too.', and a separator line '- The Candidates of Response -'. Below the separator are two blue bubbles: 'Response A: Exactly. Makes me really happy to have that flexibility.' and 'Response B: Exactly. I love my job. I love to cook and clean'.
- Score Annotation:** Contains two sections for scoring responses. 'Score for Response A' and 'Score for Response B' each have three rows of radio buttons for 'Empathy', 'Relevance', and 'Fluency', with options 1, 2, 3, and 4. Below these is an 'A/B Pair Test' section with radio buttons for 'A', 'B', and 'Tie'. A 'Move to the Next File' button is at the bottom.

Conduct an annotation task for each response sentence by generating an pre-trained model.

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