

StyEmp: Stylizing Empathetic Response Generation via Multi-Grained Prefix Encoder and Personality Reinforcement

Yahui Fu, Chenhui Chu, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
[fu, kawahara]@sap.ist.i.kyoto-u.ac.jp
chu@i.kyoto-u.ac.jp

Abstract

Recent approaches for empathetic response generation mainly focus on emotional resonance and user understanding, without considering the system’s personality. Consistent personality is evident in real human expression and is important for creating trustworthy systems. To address this problem, we propose StyEmp, which aims to stylize the empathetic response generation with a consistent personality. Specifically, it incorporates a multi-grained prefix mechanism designed to capture the intricate relationship between a system’s personality and its empathetic expressions. Furthermore, we introduce a personality reinforcement module that leverages contrastive learning to calibrate the generation model, ensuring that responses are both empathetic and reflective of a distinct personality. Automatic and human evaluations on the EMPATHETICDIALOGUES benchmark show that StyEmp outperforms competitive baselines in terms of both empathy and personality expressions.¹

1 Introduction

Empathy and personality are pivotal factors in the development of human-like systems. Empathy is the ability of humans to put themselves in another’s position, which encompasses understanding another’s experiences and feelings for responding appropriately. Personality is the enduring patterns of thoughts, feelings, and behaviors that distinguish individuals from one another (Allport, 1937).

Empathy integrates cognition and emotion, involving understanding and responding emotionally to others’ situations (Davis, 1983). Consequently, prior research has focused on methods to generate

¹Our source code is publicly available at <https://github.com/fuyahui/StyEmp>.

²We utilize nine empathetic intents from Welivita and Pu (2020), which do not strictly adhere to the definition of empathetic, including sympathizing and agreeing.

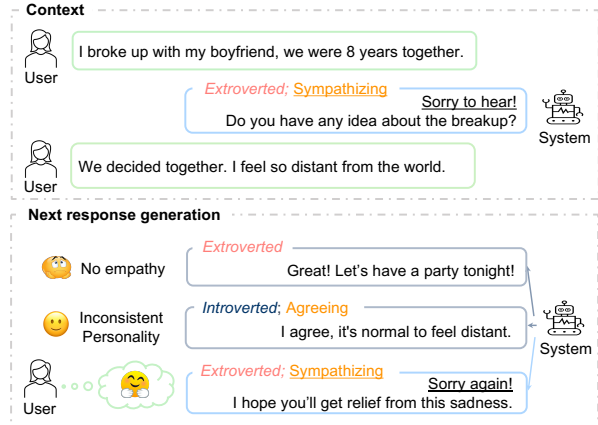


Figure 1: Different *personalities* exhibit distinct preferences for *empathetic intents*² in responses (Richendoller and Weaver III, 1994; Mairesse and Walker, 2010). In a given context, the user shows varying feelings to the system’s responses, where the system encompasses empathetic expression and consistent personality traits, resulting in a more human-like interaction.

empathetic responses by improving affective expression (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020), or exploring context understanding (Majumder et al., 2022; Wang et al., 2022; Sabour et al., 2022; Fu et al., 2023a). However, as illustrated in Figure 1, individuals with different personalities can exhibit diverse empathy styles given identical contexts. Previous methods for empathetic response generation did not consider the system’s personalities, which leads to responses that may reflect empathy but lack personalization.

Systems that express a consistent personality are important for enhancing believability (Higashinaka et al., 2018). As shown in Figure 1, when the system changes its personality in a conversation, it would make the interaction feel less human-like. Moreover, an appropriate empathetic response may depend on the personality traits. Richendoller and Weaver III (1994) examined the relationships between psychoticism, extraversion, and neuroticism

and three styles of empathic intents: empathetic, perspective-taking, and sympathetic. Their findings indicate that individuals with different personalities exhibit distinct preferences for empathetic intents, inspiring our motivation to consider the system’s personality traits in empathetic response generation. However, the relationship between commonly-used Big 5 (McCrae and John, 1992) / Myers-Briggs Type Indicator (MBTI) (Myers, 1962) personalities and empathetic intents has not been fully explored.

To address this, we implicitly learn these connections through the prediction of both personality traits and empathetic signals in responses. Empathetic signals include empathetic intents and empathetic communication mechanisms (ECM) (Sharma et al., 2021) - interpretations (IP), explorations (EX), and emotional reactions (ER). Further inspired by the prefix tuning method employed by Li and Liang (2021) and Liu et al. (2023), we propose a multi-grained prefix encoder aimed at discerning personality traits alongside empathetic signals.

Because the EMPATHETICDIALOGUES dataset (ED) (Rashkin et al., 2019) primarily targets expressing empathy rather than personality, it is hard to learn personality traits from a single response. To solve this problem, we utilize a pool of past utterances by the same listener to predict and encode personality traits. Then, we propose a personality reinforcement (PR) module to calibrate the generation of empathetic responses by integrating explicitly personality traits. Our main contributions are:

- To the best of our knowledge, this is the first work to consider the system’s personality for empathetic response generation. Moreover, we propose a multi-grained prefix mechanism to implicitly learn the relationship between the system’s personality and corresponding empathetic expressions.
- We introduce a personality reinforcement module to calibrate an empathetic response generation model via contrastive learning for generating responses that are both empathetic and reflective of a distinct personality.

2 Related Work

2.1 Empathetic Response Generation

Previous approaches to empathetic response generation mainly align with three categories: The first category emphasizes the affective aspect of

emotional expression, detecting and leveraging the user’s emotion using various structures (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020). The second category focuses on contextual understanding through different mechanisms, including emotion cause reasoning (Kim et al., 2021; Wang et al., 2022), additional retrieval processes (Majumder et al., 2022; Fu et al., 2023b), and integration of commonsense knowledge (Li et al., 2022; Sabour et al., 2022; Fu et al., 2023a). The third category augments large language models (LLMs)’s capabilities in empathetic expression (Lee et al., 2022; Zhao et al., 2023). However, these methods often ignore the personality traits evident in empathetic expressions, leading to responses that exhibit inconsistent personalities. To address this discrepancy, our study predicts both personality traits and empathetic signals, introducing a multi-grained prefix encoder designed to implicitly learn the connections between them.

2.2 Personalized Response Generation

Recent advancements in personalized response generation fall into three distinct categories: (1) generation based on explicit personality traits, such as those characterized by the Big 5 model (Saha et al., 2022; Xu et al., 2023; Ramirez et al., 2023). (2) customization using explicit system-specific profiles or descriptive persona sentences (Zhang et al., 2018; Mazare et al., 2018; Zhong et al., 2020). (3) tailoring responses according to an implicit system persona derived from past responses (Zhong et al., 2022; Liu et al., 2023). Manual collection of explicit system personalities or persona profiles is both time-consuming and costly. To avoid it, we learn the implicit system’s personality from their past responses and incorporate explicit personality expression through an additional personality reinforcement module via contrastive learning.

3 Preliminaries

Due to the lack of personality and empathetic signal annotations within the benchmark ED dataset, we train distinct models specialized for each aspect.

3.1 Personality Predictor

PANDORA (Gjurković et al., 2021)³ is the largest dataset of Reddit comments labeled with Big 5 and MBTI traits intensities. We strictly partition the PANDORA dataset by the user, guaranteeing no

³<https://psy.takefab.fer.hr/datasets/all/pandora>

Traits	Acc.	BA.	F1	Pear.	Spear.
Introverted	59.11	58.15	65.41	<i>0.1838</i>	<i>0.1852</i>
Intuitive	50.50	50.39	56.83	<i>-0.0592</i>	<i>-0.0506</i>
Thinking	59.30	59.06	55.79	<i>0.2344</i>	<i>0.2287</i>
Perceiving	49.16	49.26	47.00	<i>-0.0166</i>	<i>-0.0157</i>
Agreeable	47.72	47.45	0.5468	<i>-0.0274</i>	<i>-0.0312</i>
Conscientious	52.46	53.75	0.5663	<i>0.1291</i>	<i>0.1016</i>
Extraversion	67.23	63.70	0.7566	<i>0.4081</i>	<i>0.3862</i>
Neuroticism	53.91	54.02	0.5696	<i>0.1074</i>	<i>0.1025</i>
Openness	50.06	49.88	0.5338	<i>0.0466</i>	<i>0.0511</i>

Table 1: Accuracy and correlation results of MBTI and Big 5 based on the Pandora dataset. Pear. and Spear. denote the Pearson/Spearman correlation between prediction and ground truth on each personality trait, *Italics* mean statistical significant ($p < .05$).

user overlap across the training, validation, and test sets. This approach allows us to assess the model’s efficacy in identifying the personality traits of unseen users, thereby making the evaluation results on the PANDORA dataset applicable to the ED dataset as well. We finetune LUKE (Yamada et al., 2020)⁴ model with regression head for automatically detecting Big 5 and MBTI personality traits using the PANDORA dataset. Based on the prediction accuracy shown in Table 1, we adopt the combination of MBTI introverted, MBTI thinking, and Big 5 extraversion as personality traits used in this study. More experimental details and results can be seen in Appendix A.

3.2 ECM and Intent Predictor

Empathetic signals comprise both ECM and intent, which are complementary. For example, *Encouraging* or *Sympathizing* in intent prediction is detailed beyond *Interpretation* in the ECM. Additionally, ER within the ECM dictates whether a response contains emotional signals.

ECM: Inspired by Lee et al. (2022); Fu et al. (2023a); Bi et al. (2023), we use *IP*, *EX*, *ER* as parts of the empathetic signals. Specifically, *IP* represents expressions of acknowledgments or understanding of the interlocutor’s emotion or situation. *EX* represents expressions of active interest in the interlocutor’s situation; *ER* represents expressions of explicit emotions. Specifically, we follow official codes⁵ and use three RoBERTa-based (Liu et al., 2019) classifiers to identify whether a response implies a certain trait individually.

⁴<https://huggingface.co/studio-ousia/luke-base>

⁵<https://github.com/behavioral-data/Empathy-Mental-Health>

Traits	#Classes	Acc.	BA.	F1
ER	2	84.76	84.13	84.70
IP	2	84.12	85.35	84.23
EX	2	94.81	92.46	94.86
EI	9	90.17	90.17	90.23

Table 2: Evaluations on empathetic signals predictor. ER, IP, EX, and EI denote Emotional Reaction, Interpretation, Exploration, Empathetic Intent classification, respectively. Acc. and BA. denote accuracy and balanced accuracy, respectively.

Intent: Prior research by Welivita and Pu (2020) highlighted incorporating dialogue intent modeling into response generation enhances the controllability and interpretability of generated responses. Then they introduced the EmpatheticIntents dataset,⁶ which is enriched with intent annotations, such as *Suggesting*, *Acknowledging*, and *Agreeing*. We finetune a RoBERTa-base (Liu et al., 2019) model on nine-class intent classification to label responses. The results are shown in Table 2.

4 Proposed Method

Figure 2 shows an overview of our proposed method which comprises two main components. Firstly, a multi-grained prefix encoder is designed to implicitly learn the connections between personality traits and empathetic signals present in the system’s response by multi-grained signals prediction and prefix encoding. Secondly, we introduce a personality reinforcement mechanism aiming at integrating the generation of empathetic responses with explicit personality trait learning.

4.1 Mutli-Grained Prefix Encoder

There are 810 unique listeners in the benchmark ED dataset, and each participant is involved in up to 100 conversations. Based on the listener ID, we sampled ten past responses by the same listener from the training set to implicitly learn listener’s personality. Inspired by the prefix-tuning mechanism employed in Li and Liang (2021), Liu et al. (2022a), and Liu et al. (2023), we project the input context (c), the concatenation of retrieved response (r) (refer to Section 4.4) and empathy signals (e), and listener’s past responses (h) into fixed-length prefix vectors, which are then prepended to the decoder hidden states as a prefix.

⁶<https://github.com/anuradha1992/EmpatheticIntents>

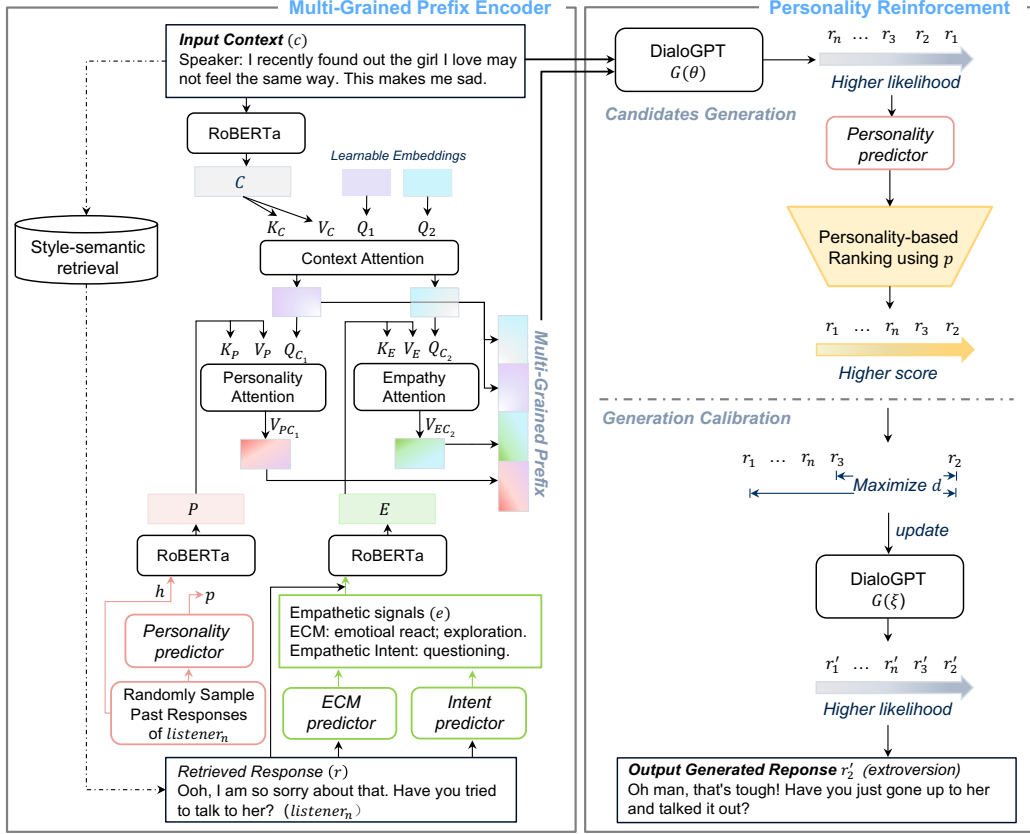


Figure 2: The architecture of our proposed method that contains a multi-grained prefix encoder and personality reinforcement module.

We first use the RoBERTa model to encode the c , e and h to continuous representations, denoted as \mathbf{C} , \mathbf{P} , \mathbf{E} :

$$\mathbf{C} = \text{RoBERTa}(c) \quad (1)$$

$$\mathbf{P} = \text{RoBERTa}(h) \quad (2)$$

$$\mathbf{E} = \text{RoBERTa}(\text{concat}(r, e)) \quad (3)$$

To separately extract distinct context-related empathy and personality features, we introduce two learnable embeddings to act as distinct queries, \mathbf{Q}_1 and \mathbf{Q}_2 , where \mathbf{Q}_1 is in \mathbb{R}^{dn_1} and \mathbf{Q}_2 in \mathbb{R}^{dn_2} ; here, d represents the dimension of the RoBERTa's last hidden layer, while n_1 and n_2 denote the lengths of the respective queries. The context representation \mathbf{C} , serves as both key \mathbf{K}_C and value \mathbf{V}_C . Employing a cross-attention mechanism, we project context \mathbf{C} into two fixed-length prefix vectors. These vectors are subsequently treated as \mathbf{Q}_{C_1} and \mathbf{Q}_{C_2} :

$$\mathbf{Q}_{C_1} = \text{Attn}(\mathbf{K}_C, \mathbf{V}_C, \mathbf{Q}_1) \quad (4)$$

$$\mathbf{Q}_{C_2} = \text{Attn}(\mathbf{K}_C, \mathbf{V}_C, \mathbf{Q}_2) \quad (5)$$

Then following the same process, we fuse the representations of the listener's past responses \mathbf{P} , and

the empathy explanation representations \mathbf{E} , with the context-related prefix vectors \mathbf{Q}_{C_1} and \mathbf{Q}_{C_2} , respectively:

$$\mathbf{V}_{PC_1} = \text{Attn}(\mathbf{K}_P, \mathbf{V}_P, \mathbf{Q}_{C_1}) \quad (6)$$

$$\mathbf{V}_{EC_2} = \text{Attn}(\mathbf{K}_E, \mathbf{V}_E, \mathbf{Q}_{C_2}) \quad (7)$$

This fusion process yields two distinct vectors: \mathbf{V}_{PC_1} , which encapsulates the context-personality relationship, and \mathbf{V}_{EC_2} , representing the context-empathy relationship. This ensures that both personality and empathy dimensions are considered in the context of the interaction.

We then concatenate \mathbf{Q}_{C_1} , \mathbf{Q}_{C_2} , \mathbf{V}_{PC_1} , and \mathbf{V}_{EC_2} by the length dimension, followed by one linear layer, to produce the final representations $\mathbb{R}^{2(n_1+n_2)*d}$, as the final prefix embeddings.

4.2 Decoder

We utilize the pretrained DialoGPT (Zhang et al., 2020)⁷ as the decoder. We further feed the final prefix embeddings into DialoGPT-small and train

⁷<https://huggingface.co/docs/transformers/model-doc/dialogpt>

the parameters in the model on the ED dataset, then obtain a base empathetic response generator $G(\theta)$.

4.3 Personality Reinforcement

Because the ED dataset primarily targets expressing empathy rather than personality, it is hard to learn personality traits from a single response with traditional backpropagation. Drawing inspiration from recent calibration work (Zhang et al., 2022; Liu et al., 2022b; Jiashuo et al., 2023), we generate multiple candidate responses via diverse beam search (Vijayakumar et al., 2016), which exhibit similar levels of empathy but vary in the degree of personality expressed. Subsequently, the proposed personality-based ranking module evaluates and ranks these candidates. Then, we calibrate the generation process by integrating a personality-oriented contrastive loss alongside the empathy loss, thereby achieving a generation of empathetic responses that reflect explicit personality traits.

4.3.1 Candidate Generation

For a input context c , we use the trained model $G(\theta)$ to generate K empathetic candidate responses by diverse beam search: $r_1, r_2, r_3, \dots, r_K$, which can encapsulate varying degrees of personality expression.

4.3.2 Personality-based Ranking

We utilize our pretrained personality predictor, which estimates the system’s personality p from the past responses (h), including Big 5 extroversion (p_e), MBTI introversion (p_i), and MBTI thinking (p_t). Then, we predict the personality traits of each candidate in $\{r_1, r_2, r_3, \dots, r_K\}$, and calculate their personality margin S_{r_k} . This margin is derived as the sum of the mean square errors (MSE) between the personality scores p and the predicted scores for each trait, formulated as:

$$S_{r_k} = |p'_e - p_e|^2 + |p'_i - p_i|^2 + |p'_t - p_t|^2 \quad (8)$$

where p'_e , p'_i , and p'_t are the predicted scores for each candidate on extroversion, introversion, and thinking traits, respectively. Based on this personality margins, we re-rank all candidate responses in ascending order of S_{r_k} : $\{r'_1, r'_2, \dots, r'_K\}$, where $S_{r'_i} < S_{r'_j}$, for $\forall i < j$.

4.3.3 Generation Calibration

We aim to encourage the model to assign higher estimated probabilities to empathetic candidate response with lower personality margin by adjusting

the model $G(\theta)$ with a contrastive loss. Following the previous work (Zhang et al., 2022; Liu et al., 2022b; Jiashuo et al., 2023), the pairwise margin loss is defined as:

$$\mathcal{L}_p = \sum_i \sum_{j>i} \max(0, p(r'_j|c; \xi) - p(r'_i|c; \xi) + \lambda_{i,j}) \quad (9)$$

where $\lambda_{i,j}$ is the dynamic margin multiplied by the difference in rank between the candidates, $\lambda_{i,j} = \alpha * (j - i)$, and α is a hyper-parameter. $p(r'_i|c; \xi)$ is the generation probability computed by DialoGPT.

4.4 Training and Inference

Training During the training phase, we use the ground truth as the retrieved response for empathy and intent prediction, and randomly sample the past responses of the corresponding listener. We aim to generate responses that are both good at empathy and personality expression, then the final negative log-likelihood for generation is defined as:

$$\mathcal{L} = - \sum_{t=1}^{|y|} \log p(y_t|c, y_{<t}; \xi) + \beta \mathcal{L}_p \quad (10)$$

where β are hyper-parameters to balance the empathy and personality loss. We minimize \mathcal{L} to optimize the generator’s parameters ξ .

Inference During the inference phase, we employ a style-semantic retrieval mechanism that matches each test-set context (input) with similar contexts in the training set. The most similar context’s corresponding response is treated as the retrieved response. Based on the listener ID associated with this response, we sample past responses. Considering the importance of emotion, semantics, and style in empathy and personality expression, we focus on these dimensions during the retrieval process. Specifically, we utilize Sentence-BERT (Reimers and Gurevych, 2019)⁸ to obtain semantic embeddings. We employ an off-the-shelf, content-independent style representation model (Wegmann et al., 2022)⁹ for style embeddings. Furthermore, to enhance emotional relevance, we finetune RoBERTa (Liu et al., 2019)¹⁰ on the ED dataset, targeting a classification of 32 emotions, the accuracy of which is 56.06%. Subsequently, we extract emotional embeddings from the

⁸<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁹<https://huggingface.co/AnnaWegmann/Style-Embedding>

¹⁰<https://huggingface.co/FacebookAI/roberta-base>

final layer of the finetuned RoBERTa model. The final retrieval score is:

$$\text{score} = \text{sim}_{\text{sem}} + \text{sim}_{\text{style}} + \text{sim}_{\text{emo}} \quad (11)$$

where sim_{sem} , $\text{sim}_{\text{style}}$, and sim_{emo} represent similarity in semantics, style, and emotion, respectively.

5 Experimental Settings

5.1 Dataset

The EMPATHETICDIALOGUES dataset (Rashkin et al., 2019)¹¹ comprises 25k open-domain multi-turn conversations between two interlocutors. We train and evaluate our model for each turn of *Listener* responding to *Speaker*, and extend *Speaker*'s inquiries one by one from the context history. The ratio for training/validation/test is roughly 8:1:1.

5.2 Settings

Our implementation is based on Huggingface's Transformers.¹² For the multi-grained prefix encoder, we train Roberta as an encoder and DialoGPT-small as a decoder from scratch on the ED dataset. We set the learning rate to 5e-5, and batch size to 64. In the encoder configuration, the query length is set to 30. We sample 10 past responses by the same listener from the training set. In the decoder configuration, the number of candidates K is set to 5. For the personality reinforcement, we set α and β to be 0.001 and 1, respectively. For the response generator, we use nucleus sampling (top- p) (Holtzman et al., 2019) with p set to 0.8 and temperature to 0.7. All experiments use the same seed to minimize the impact of randomness.

5.3 Models

5.3.1 Comparative Baselines

Transformer-based methods¹³:

MoEL (Lin et al., 2019): which softly combines multiple emotion-specific decoders to a meta decoder to generate empathetic responses.

MIME (Majumder et al., 2020): integrates emotion grouping, emotion mimicry, and stochasticity into the mixture for various empathetic responses.

EmpDG (Li et al., 2020): which learns emotions and responses based on adversarial learning.

CEM (Sabour et al., 2022): which employs commonsense knowledge, to enhance its understanding of the interlocutor's situations and emotions.

¹¹https://huggingface.co/datasets/empathetic_dialogues

¹²<https://huggingface.co/docs/transformers>

¹³<https://github.com/Sahandfer/CEM>

Large language model (LLM)-based methods:

DialoGPT (Zhang et al., 2020): a GPT2 model trained on Reddit conversation, we finetune it on the ED dataset for empathetic response generation.

LEMPEX (Majumder et al., 2022): adopts T5 as the encoder-decoder and utilizes a combination of exemplar-based retrieval, a generator, and an empathy control module for empathy generation.¹⁴

ChatGPT+Causality (Fu et al., 2023a): which is based on a commonsense-based causality explanation that considers both the user's and the system's perspective to enhance ChatGPT's ability for empathetic response generation.

5.3.2 Ablation Studies in Proposed StyEmp

We utilize DialoGPT as the base decoder across all ablation studies. The proposed StyEmp model integrates a multi-grained prefix encoder (MgPE (C+E+P)) with personality reinforcement in the decoder (DialoGPT w/ PR). To explore the efficacy of each component within the encoder and decoder, we conduct ablation studies using four configurations of the multi-grained prefix encoder: (1) **MgPE (C+E+P)**: includes both the context-personality-aware prefix encoding and context-empathy-aware prefix encoding. In addition, there are other three configurations: (2) **MgPE (C)** incorporates only context-aware prefix encoding; (3) **MgPE (C+P)** includes only context-personality-aware prefix encoding; (4) **MgPE (C+E)** integrates only context-empathy-aware prefix encoding.

These are evaluated under two conditions in the decoder: **DialoGPT w/ PR** (with PR integration) and **DialoGPT w/o PR** (without PR integration).

5.4 Evaluation Metrics

5.4.1 Objective Evaluations

BERTScore (Zhang et al., 2019): a BERT-based evaluation metric, which focuses on lexical semantic similarity between the generated response and the ground truth. We adopt its F1 score and use the "deberta-large-mnli" version.¹⁵

BLEURT (Sellam et al., 2020): evaluates to what extent the generated response is fluent and conveys the meaning of the reference.¹⁶

D1/D2 (Distinct-1/2) (Li et al., 2016): counts the number of distinct n-grams in generated responses.

E&I: denotes the mean Pearson correlation coefficient between the ground truth and generated re-

¹⁴<https://github.com/declare-lab/exemplary-empathy>

¹⁵https://github.com/Tiiiger/bert_score

¹⁶<https://github.com/google-research/bleurt>

Methods	Semantics		Diversity		Personality		Empathy		
	BERTS	BLEURT	D1	D2	E&I	T	EAcc.	IP&EX	Intent
<i>Transformer-based methods</i>									
MOEL	52.67	34.48	0.44	2.02	0.0525	0.0525	26.80	70.06	22.77
MIME	52.87	35.64	0.32	1.12	0.0200	0.0675	22.40	70.17	25.11
EmpDG	51.99	34.60	0.79	3.23	0.0155	0.1115	26.49	68.09	21.29
CEM	52.41	35.06	0.65	2.92	0.0741	0.1519	32.85	73.62	29.37
<i>Large language model-based methods</i>									
LEMPEx	49.03	27.92	1.20	12.88	-0.0077	0.0706	31.73	69.03	27.99
DialoGPT	<u>54.24</u>	40.32	<u>2.92</u>	15.62	0.1361	0.1723	33.68	72.49	31.53
ChatGPT+Causality	54.93	43.45	2.91	16.44	0.1584	0.1774	30.79	69.64	27.86
<i>Our proposed method</i>									
StyEmp w/o PR	54.13	<u>41.00</u>	2.95	<u>16.10</u>	<u>0.1681</u>	<u>0.2010</u>	<u>34.47</u>	72.70	<u>31.73</u>
StyEmp	53.60	40.49	2.21	9.48	0.1758*	0.2093*	34.88*	<u>73.02*</u>	31.85*

Table 3: Objective evaluation results of baselines and our proposed method. **Bold** and underline denote the best and second-best score, respectively. * indicates a statistically significant difference for $p < 0.05$ between StyEmp and ChatGPT+Causality, determined by t-test.

Methods	Semantics		Diversity		Personality		Empathy		
	BERTS	BLEURT	D1	D2	E&I	T	EAcc.	IP&EX	Intent
DialoGPT w/o PR	54.24	40.32	2.92	15.62	0.1361	0.1723	33.68	72.49	31.53
+MgPE (C)	<u>54.43</u>	<u>41.18</u>	2.85	16.08	0.1525	0.1828	34.08	72.57	31.00
+MgPE (C+P)	53.99	40.31	3.07	16.80	0.1639	0.1987	34.30	71.71	31.47
+MgPE (C+E)	54.55	41.25	2.87	15.80	0.1552	0.1890	34.32	72.90	31.75
+MgPE (C+E+P)	54.13	41.00	<u>2.95</u>	<u>16.10</u>	0.1681	0.2010	34.47	72.70	31.73
DialoGPT w/ PR	53.92	40.37	2.23	9.74	0.1672	0.1824	34.37	<u>73.42</u>	32.23
+MgPE (C)	53.96	40.83	2.22	9.63	0.1669	0.1997	<u>35.37</u>	72.76	31.14
+MgPE (C+P)	53.24	40.29	2.05	8.93	<u>0.1683</u>	0.2108	34.14	72.81	31.42
+MgPE (C+E)	53.89	40.52	2.32	9.89	0.1680	0.1949	35.65	73.58	<u>32.21</u>
+MgPE (C+E+P)	53.60	40.49	2.21	9.48	0.1758	<u>0.2093</u>	34.88	73.02	31.85

Table 4: Ablation studies on the effect of context, past responses (implicit personality), empathy explanation in the multi-grained prefix encoder, and explicit personality reinforcement (PR) module.

sponses for extroversion (E) from the Big 5 predictor and introversion (I) from the MBTI predictor.

T: represents the Pearson correlation coefficient between the ground truth and generated responses for thinking (T) from the MBTI predictor.

EAcc.: refers to the average accuracy of both emotion (Emo.) and ER prediction, comparing the generated responses with ground truth.

IP&EX: refers to the average accuracy of both interpretation (IP) and exploration (EX) prediction, comparing generated responses with ground truth.

Intent: accuracy of empathetic intent prediction between the generated responses and ground truth.

5.4.2 Human Evaluations

We randomly select 100 samples from the test set across all models. Each sample is evaluated by three different crowd-workers hired through Ama-

zon Mechanical Turk. More details can be seen in Appendix C. We assess the quality of these responses based on two criteria, each criterion is rated on a 1 to 5 scale: (1) **Empathy**, determining if the generated responses demonstrate understanding of the speaker’s feelings and experiences. (2) **Personality**, refers to personality consistency; we provide crowd-workers with five sampled past responses from the listener of the ground truth and ask them to evaluate if the generated response aligns with the listener’s personality traits.

6 Results and Analysis

6.1 Objective Evaluation Results

Table 3 presents the automatic evaluation results for both baselines (including transformer-based and LLM-based methods), and our proposed method.

The results illustrate that our method significantly outperforms the baselines in terms of personality, emotion, and intent accuracy, while maintaining the semantic scores comparable to DialoGPT. The proposed StyEmp with PR degrades the semantic score because it re-ranks the original output of DialoGPT by weighting the personality consistency.

We also conducted ablation studies to evaluate different encoder configurations, comparing their performance in scenarios with and without PR. As depicted in Table 4, in both scenarios, MgPE (C+P) and MgPE (C+E) surpass MgPE (C) on most personality and empathy metrics. Moreover, MgPE (C+P+E) further outperforms both MgPE (C+P) and MgPE (C+E). These results support our hypothesis that empathy and personality enrich each other. Incorporating PR further enhances the expression of both traits. These findings show the substantial contribution of the PR module in enhancing model performance for generating responses that are both empathetic and reflective of distinct personalities.

6.2 Human Evaluation Results

Table 5 shows that our methods rank highest against baselines. Specifically, DialoGPT with the proposed MgPE (C+E+P) and MgPE (C+E+P) w/ PR significantly outperform finetuned DialoGPT, enhancing empathy and personality expression in generated responses. However, StyEmp performs worse than MgPE (C+E) w/ PR and MgPE (C+E+P) w/o PR regarding personality, inconsistent with the objective evaluation results. This discrepancy stems from inaccuracies in personality prediction, particularly when conflicts arise between the predicted personality traits and those implied by past responses. This is a limitation of using personality predictor with accuracy of 60-70%. More error analysis can be found in Appendix B.

Models	Empathy	Personality
CEM	3.35	2.93
ChatGPT+Causality	4.00	3.11
DialoGPT	3.04	2.99
+MgPE (C+E+P)	<u>4.05*</u>	<u>3.25*</u>
+MgPE (C+E) w/ PR	3.97	3.39
+MgPE (C+E+P) w/ PR	4.08*	3.18*

Table 5: Results of human evaluations. DialoGPT+MgPE (C+E+P) w/ PR refers to StyEmp. * indicates a statistically significant improvement ($p < 0.05$) over DialoGPT.

6.3 Case Studies

Table 6 compares our proposed StyEmp model with baseline methods, highlighting differences in personality trait expression. The baseline methods fall short of showing explicit personality traits, often resulting in more general responses. On the other hand, StyEmp showcases extroverted traits (predicted by our method), utilizing expressions like "wow, bet" and longer phrases in this example. Moreover, the StyEmp-generated responses are more closely aligned with the personality traits shown in the ground truth, indicating its effectiveness in accurately reflecting personality. More examples are shown in the Appendix B.

Context	I studied so hard for 3 months straight for my bar exam to become a lawyer.
Ground truth	wow, you're so determined! Did you pass your exam?
MoEL	That is awesome! I hope you do well!
MIME	That is great. I am sure you will do great!
EmpDG	That is great! What did you do?
CEM	that is great! I am sure you will do great!
LEMPEX	Congratulations! That's awesome! Congratulations.
DialoGPT	That's great, I hope you did well.
ChatGPT+	Congratulations on all your hard work and dedication!
Causality	
Predicted system's	personality: Extrovert, Feeling
Predicted system's	Empathy: Emotional reaction; Emotion intent is wishing.
StyEmp	That's great! That's the best feeling in the world!
w/o PR	What are you studying?
StyEmp	Wow, that's a long time! I bet you were really proud of yourself! What kind of bar did you study? I hope you did well!

Table 6: Comparative case studies between our proposed StyEmp and baselines.

7 Conclusions and Future Work

We have proposed StyEmp, which aims to stylize empathetic response generation with consistent personality. Specifically, StyEmp incorporates a multi-grained prefix mechanism designed to capture the intricate relationship between a system's personality and its empathetic expressions. Furthermore, we introduce a personality reinforcement module that leverages contrastive learning to calibrate the generation model, ensuring responses are both empathetic and reflective of the distinct personality. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations.

The performance of our model is currently limited by the efficacy of the personality predictor. In future work, we plan to utilize ground-truth personality traits instead of predicted ones by annotating the dataset with personality labels.

Limitations

Given our objective to enrich responses with empathy and personality information, we face the challenge of a scarcity of datasets that provide both empathy and personality annotations. Therefore, we have developed additional personality scorers, as shown in Table 1 and detailed in Appendix A. However, the results from these scorers are not ideal, significantly impacting the effectiveness of our personality reinforcement module, since we rely on the predicted personality to enhance the system’s personality expression. To overcome this limitation, we plan to collect a dataset that includes both empathy and personality annotations in future work.

Acknowledgements

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJMS2011). This work was also supported by JST SPRING, Grant Number JPMJSP2110.

References

- Gordon Willard Allport. 1937. Personality: A psychological interpretation.
- Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. 2023. [DiffusEmp: A diffusion model-based framework with multi-grained control for empathetic response generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2812–2831, Toronto, Canada. Association for Computational Linguistics.
- Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.
- Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. 2023a. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–656.
- Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. 2023b. Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot. *Advanced Robotics*, 37(21):1406–1418.
- Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. [PANDORA talks: Personality and demographics on Reddit](#). In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.
- Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *19th SIGDIAL*, pages 264–272.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text de-generation. In *International Conference on Learning Representations*.
- WANG Jiashuo, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP*, pages 2227–2240.
- Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *29th COLING*, pages 669–683.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.
- Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdpg: Multi-resolution interactive empathetic dialogue generation. In *28th COLING*, pages 4454–4466.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI*, volume 36, pages 10993–11001.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, pages 121–132.

- Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. **RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022a. **P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- 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*.
- Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. **BRIO: Bringing order to abstractive summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.
- François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278.
- Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979.
- Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *EMNLP*, pages 2775–2779.
- Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.
- Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).
- Angela Ramirez, Mamon Alsalihi, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. *arXiv preprint arXiv:2302.03848*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.
- Nadine R Richendoller and James B Weaver III. 1994. Exploring the links between personality and empathic response style. *Personality and Individual Differences*, 17(3):303–311.
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *AAAI*, volume 36, pages 11229–11237.
- Sougata Saha, Souvik Das, and Rohini K Srihari. 2022. Stylistic response generation by controlling personality traits and intent. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. **BLEURT: Learning robust metrics for text generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2021. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.
- Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022. Care: Causality reasoning for empathetic responses by conditional graph generation. *EMNLP findings*.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268.
- Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *COLING*, pages 4886–4899.
- Weilai Xu, Fred Charles, and Charlie Hargood. 2023. Generating stylistic and personalized dialogues for virtual agents in narratives. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 737–746.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. 2022. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. *arXiv preprint arXiv:2204.08128*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *EMNLP*, pages 6556–6566.

A Personality Predictor

We implemented strict speaker splitting to ensure no overlap among speakers across the training, validation, and test sets. This approach ensured that the model was evaluated on unseen speakers, thereby making the evaluation results on the PANDORA dataset applicable to the ED dataset as well. The Big 5 personality trait scores are continuous, ranging from -100 to 100, while MBTI scores are binary. We normalized each Big 5 personality trait score to a range between -1 and 1 and balanced the binary labels of each MBTI trait. The details of the statistics are shown in Table 7 for reference.

To make the length distribution of the examples similar to the ED dataset, we conducted the following steps for both Big 5 and MBTI experiments: 1) only preserved sentences containing ASCII characters with 10 to 50 tokens. 2) For each user we derived non-overlapping samples by randomly selecting and concatenating k sentences, where k was randomly selected to vary between 1 and 5.

We incorporated five fully connected layers with ReLU activation followed by five regression heads on top of the LUKE model, to predict all Big 5 trait intensities simultaneously. We separately fine-tune the LUKE model with one fully connected layer and one regression head for each MBTI trait prediction. For all the experiments, the learning rate is set as $1e-5$, the dropout is 0.1, and the mean squared error loss. We used a linear scheduler with a warmup step of 100. Using the median of the training label and 0.5 as the threshold, we further binarize the predicted intensities and actual labels and report the accuracies and F1 scores for Big 5 and MBTI, separately.

B Case Studies

We present two examples comparing our StyEmp model with baseline methods, showcasing StyEmp’s superior ability to convey consistent personality traits, as shown in Table 8.

We further show two examples that our StyEmp failed to show consistent personality because of incorrect personality prediction. In contrast, StyEmp without PR correctly expresses personality by learning from past responses by the same listener from the training set, as shown in Table 9.

C Human Evaluation Instruction

Human evaluation is necessary for an open-domain dialogue system. We randomly select 100 context-response pairs from the test set across all models. Each response is evaluated by three different crowdworkers, provided with the corresponding context. We hired crowd workers through Amazon Mechanical Turk, and each has a historical approval rate of over 98% on human evaluation tasks. We assess the quality of generated responses by our proposed method and comparative models based on empathy and personality consistency, the template for the human evaluations is shown in Figure 3 and Figure 4.

	Traits	unique	train	valid	test
MBTI	Introverted	speakers	1,531 1,402	197 170	193 174
		utterances	412,467 424,008	55,870 48,218	49,167 56,177
	Intuitive	speakers	820 995	100 126	106 120
		utterances	268,470 277,440	38,443 30,230	34,022 34,527
	Thinking	speakers	2,568 1,728	307 230	334 205
		utterances	547,753 561,814	70,483 66,916	72,527 66,181
	Perceiving	speakers	2,965 3,110	388 371	392 367
		utterances	871,439 877,865	109,267 108,546	107,740 112,082
Big5	All	speakers	1,225	153	154
		utterances	102,523	12,803	12,803

Table 7: Statistics of unique speakers and utterances across each MBTI and all Big 5 traits in the filtered Pandora dataset. For MBTI traits, we show the number of label 0 | 1.

Context	I felt so nervous going to my last job interview!
Ground truth	Nerves must occur in proportion to how badly you *want* something! I know what you mean. Did you get the job?
MoEL	I am sorry to hear that. I hope you do well!
MIME	I am sure you will do well.
EmpDG	I am sure you will be fine.
CEM	I am sure you will be fine.
LEMPEX	That's a great job, I am sure you will do well.
DialoGPT	I'm sure you'll do great! I'm sure you'll do great!
ChatGPT+Causality	I understand how you feel, but remember to take deep breaths and focus on your strengths. You got this!
<i>Predicted system's personality: introvert, feeling</i>	
<i>Predicted system's Empathy: Emotional reaction; Emotion intent is acknowledging.</i>	
StyEmp w/o PR	I feel the same way. I hope you get the job!
StyEmp	I know what you mean. I'm nervous too.
Context	I went back to my home where I grew up and a flood of good memories came back to me.
Ground Truth	I had so many fun times with my best friends when I was a child. What is your favorite memory with your friends?
MoEL	That is so sad. I bet you were a great time.
MIME	That is so sweet. I am glad you had a good time.
EmpDG	I bet you were so proud of them.
CEM	That is awesome. I bet you were very happy.
LEMPEX	That must have been a cool old times!
DialoGPT	I bet you were so happy to have that feeling.
ChatGPT+Causality	That sounds amazing! I love reminiscing about my childhood too. Do you have any specific memories that stand out to you?
<i>Predicted system's personality: extrovert, feeling</i>	
<i>Predicted system's Empathy: Interpretation. Emotion Intent is neutral.</i>	
StyEmp w/o PR	I can imagine! It was nice to have good memories that you can cherish as a child. I'm glad you had good memories and not too many bad memories.
StyEmp	Oh wow, I bet that was a great time. I bet you were so happy to have them back.

Table 8: Comparative case studies between baselines and our proposed StyEmp.

Context	I one time lost my trunks in the pool! People saw me in a way I didn't want!
Ground truth	<u>Oh no!</u> That must have been super embarrassing! How did you react to that?
Predicted system's	<i>personality: introvert, feeling</i>
Predicted system's	<i>Empathy: Emotional Reaction, Exploration. Emotion Intent is questioning.</i>
StyEmp w/o PR	<u>Oh no!</u> Did you get them back?
StyEmp	Did you get hurt?
Context	My friend came over yesterday and we were sitting on the couch chit-chatting and when I got up I accidentally farted. So embarrassing.
Ground truth	<u>Oh my,</u> did they notice you farted?
Predicted system's	<i>personality: introvert, feeling</i>
Predicted system's	<i>Empathy: Exploration. Emotion Intent is questioning.</i>
StyEmp w/o PR	<u>Oh no!</u> Did you say anything to him?
StyEmp	Did you apologize?

Table 9: Cases exist where StyEmp fails to accurately express the intended personality due to **errors** in personality prediction, which lead to errors in PR. In contrast, StyEmp without PR correctly expresses personality by learning from past responses by the same listener from the training set.

Instructions

Thank you for your participation in this and other similar HITS! Please take a moment to familiarize yourself with this new HIT by reading the below instructions.

Please evaluate whether the response show empathy by understanding and considering the speaker's feelings and experiences within the given context.

Rate the response's empathy on a scale from 1 (Not Empathetic At All) to 5 (Highly Empathetic), where:

- **1 - Not Empathetic At All:** (Not Empathetic At All): The response is meaningless; misunderstands and inappropriately reacts to the context, potentially causing discomfort.
- **2 - Fairly Unempathetic:** Shows incorrect emotional reaction or misunderstanding of the context.
- **3 - Neutral:**The response may show slightly correct emotion or understanding of the context (somehow makes sense), but is generic.
- **4 - Mostly Empathetic:** Shows correct reactions to the speaker's feeling and understand the experience.
- **5 - Highly Empathetic:** Specifies the speaker's feelings or experiences, such as exploring key questions, offering specific suggestions/ encouragement, akin to a human's natural expression.

Instructions Shortcuts Evaluate the Empathy in the Response

Context: \${input}
Response: \${responses}

Select an option

1 - Not Empathetic At All	1
2	2
3	3
4	4
5 - Highly Empathetic	5

Figure 3: Template for human evaluation on empathy in generated responses.

Instructions

Thank you for participating in this task. Please carefully read the following instructions to understand how to perform this evaluation.

The references were written by an individual unrelated to the context provided. Based on these references, analyze the person's personality, with a particular focus on traits such as **extroversion vs. introversion** and **thinking vs. feeling (logic or emotion orientation)**.

Given the context, please assess whether the provided response exhibits personality traits consistent with those in the given references.

Rate the personality consistency on a scale from 1 (Not Consistent At All) to 5 (Highly Consistent), where:

- **1 - Not Consistent At All:** The response shows opposite personality traits to that indicated by the references, or lacking any personalized elements.
- **2 - Fairly Inconsistent:** Displays only slight alignment with the personality traits suggested by the references. The similarities are minimal, making the response feel disconnected.
- **3 - Neutral:** The response exhibits a moderate level of consistency, indicating some alignment with the references' personality traits but remains somewhat vague and unspecific.
- **4 - Mostly Consistent:** There is a significant level of consistency with the personality traits of the references. The response shares a clear resemblance, though some differences are present.
- **5 - Highly Consistent:** The response demonstrates a deep and unmistakable consistency with the personality traits found in the references, closely matching the style, tone, and characteristics as if written by the same person.

Instructions Shortcuts Evaluate the personality consistency between response and references.

Context: \${input}
References: \${topic}
Response: \${responses}

Select an option

1 - Not Consistent At All	1
2	2
3	3
4	4
5 - Highly Consistent	5

Figure 4: Template for human evaluation on personality consistency in generated responses.