REG: Retrieval via Emotion Similarity for Guiding Empathetic Dialogue Generation

Xu Wang[1,2*], **Bo Wang**[1,†], **Yang Xiang**[2],
Yihong Tang[3,4], **Dongming Zhao**[5], **Zifei Yu**[6], **Yuexian Hou**[1]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[4]Shenzhen Loop Area Institute, Shenzhen, China
[5]AI Lab, China Mobile Communication Group Tianjin Co., Ltd.
[6]Tianjin Huizhixingyuan Information Technology Co., Ltd., Tianjin, China
{2123244001@tju.edu.cn, bo_wang@tju.edu.cn}

Abstract

Empathy relies on the cognitive capacity to relate to similar past experiences. Consequently, retrieval-based approaches utilize analogous exemplars to guide empathetic dialogue generation. However, existing methods prioritize semantic similarity over emotion characteristics, often leading to unempathetic responses. To address this, we propose REG, a framework that integrates four Emotion Attributes into the retrieval process to ensure explicit emotional alignment. Furthermore, to mitigate the noise and limited diversity caused by coarse-grained sentence-level attributes, we incorporate Token-level Retrieval for finer granularity and a Retrieval Candidate Augmentation strategy to enhance diversity. Empirical results on the EmpatheticDialogues dataset demonstrate that REG significantly outperforms baselines, offering a robust solution for empathetic generation.

1 Introduction

Developing dialogue systems' capability of demonstrating empathy – the ability to understand, potentially share, and react appropriately to human experiences and feelings – is increasingly recognized as crucial for enhancing user's interaction and satisfaction across diverse applications (Liu et al., 2021; Wang et al., 2021). Drawing inspiration from psychological models that distinguish between affective (representing emotions like happiness or sadness) and cognitive (reflecting experiences and realities) components of empathy (Mischel and Shoda, 1995), incorporating elements such as explicit emotion labels, emotion causes, and commonsense knowledge has been proven effective in improving systems' empathetic capabilities

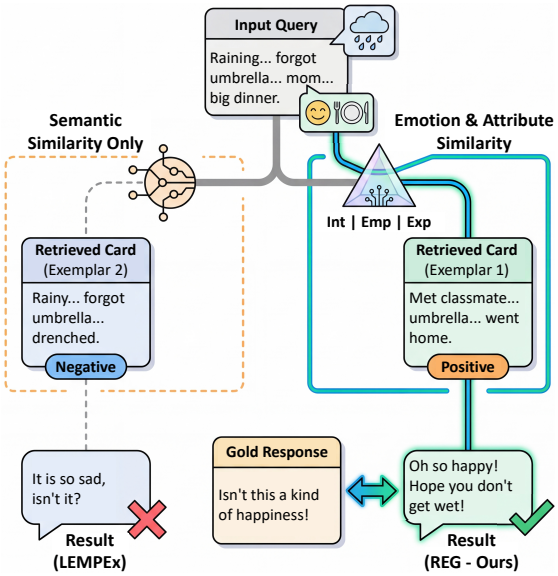

Figure 1: Comparison between semantic similarity and emotion characteristics similarity.

(Majumder et al., 2020; Sabour et al., 2022; Zhou et al., 2023; Gao et al., 2021; Wang et al., 2025).

Despite their accomplishments, empathy also fundamentally arises from the cognitive ability to draw upon analogous past scenarios and dialogues. Thus, models can further achieve empathetic responses by referencing similar dialogue contexts. For example, some researchers adopt the Dense Passage Retrieval(Karpukhin et al., 2020) to retrieve exemplars similar to the user's query (Majumder et al., 2021; Zhu et al., 2023; Xu et al., 2024). Then the exemplars are fed into the response generator as templates to guide the generator to produce empathetic and accurate responses. However, this kind of retrieval is solely based on semantic similarity, ignoring the similarity of emotion characteristics. As shown in Figure 1, the user query is that: *It was raining and I forgot to take an umbrella with me. After school, my mom and I went to the nearby supermarket to take shelter from the*

[*]Work done during internship at the Peng Cheng Laboratory.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18872–18883

rain and had a big dinner. Undoubtedly, the user's emotional state is positive. Following Majumder et al. (2021), the retrieved context is as follows: *User: It's raining, but I forgot to bring an umbrella, and my mom and I got soaked in the rain on our way to the supermarket in embarrassment. Answer: Oh, this is a little terrible!* Intuitively, this sample is semantically similar but inconsistent with the user's emotional state. Injecting such exemplars as auxiliary for generation misguides and impedes the model's understanding of the user's situations and emotional states, thereby generating potentially unempathetic or inappropriate responses.

To address this limitation, we introduce a novel framework called REG (**R**etrieval via **E**motion Similarity for **G**uiding Empathetic Dialogue Generation). REG leverages 4 Emotion Attributes, **Sentiment, Emotional Presence, Interpretation, and Exploration**, which are recognized as key factors in empathy and human communication (Sharma et al., 2021, 2020; Watson et al., 2002) to guide the retriever towards emotionally similar exemplars. Emotion attributes are inherently sparse and coarse-grained. Relying solely on them can lead to the retrieval of exemplars that differ significantly in actual emotional expression and these exemplars may introduce emotional noise. Moreover, such coarse-grained retrieval also lacks finer distinctions among retrieved contents, limiting the diversity of retrieved results and ultimately hindering the stability and quality of the generated responses. Hence, we incorporate two key mechanisms for the above two issues: Token-level Retrieval, which enhances emotional alignment at a finer granularity between the exemplar and the user query; and Retrieval Candidate Augmentation, which introduces randomized exemplars during the retriever training process to improve the model's generalization to diverse emotional contexts.

In summary, our contributions are as follows:

(1) We propose REG, a novel framework that leverages four key **Emotion Attributes** to guide exemplar retrieval, enabling the model to capture emotional resonance beyond mere semantic similarity for enhancing empathetic dialogue generation.

(2) To address the challenges posed by coarse-grained Emotion Attributes, such as potential noise and limited diversity, we incorporate **Token-level Retrieval** and **Retrieval Candidate Augmentation** mechanisms to refine the emotionally-guided retrieval process.

(3) Comprehensive experiments on the EmpatheticDialogues dataset demonstrate the significant superiority of REG over existing baselines in generating more empathetic, coherent, and diverse responses.

2 Related work

Empathetic Dialogue Generation Expressing empathy aims to establish smooth and meaningful relationships during communication. Therefore, it is crucial to equip dialogue systems with genuinely human-like empathetic capabilities. Prior research can be broadly categorized into three directions: First, some studies focus on detecting the user's emotional state to support empathetic response generation (Majumder et al., 2020; Fu et al., 2023; Yuan et al., 2025). Second, with deeper investigation, some researchers argue that emotion is not the only determining factor in empathetic dialogue generation, as empathy involves both affective and cognitive dimensions (Davis, 1983). Accordingly, Sabour et al. (2022); Pan Gao (2023); Yang et al. (2024) incorporate commonsense knowledge to represent cognition, while Kim et al. (2021) introduce emotion causes to enhance empathetic understanding. Ma customizes user query to match LLMs preferences for better intent recognition(Ma et al., 2026), which can serve for empathetic dialogue generation. Zhao investigates social bias in large language models through self-reflection(Zhao et al.). Tang modeles role from personalized dialogue history by exploring and utilizing latent space(Tang et al., 2024) and enhances personalized dialogue generation with contrastive latent variables through combining sparse and dense persona(Tang et al., 2023). Third, differing from the above two ones, some work employs retrieval-based methods to find semantically similar exemplars, which serve as guidance for generating empathetic responses (Zhu et al., 2023; Xu et al., 2024), akin to providing hints when solving complex problems.

Unlike previous approaches that focus on emotion detection, cognitive features, or semantic retrieval, our work introduces the concept of emotional similarity and expand retrieval-based methods based on this.

Retrieval-based Dialogue Generation Recently, dialogue generation models have shown commendable performance. However, this also faces some challenges, such as relatively mundane(Guo et al.,

2018)(e.g, *I'm so sorry.*) or hallucinations(Shuster et al., 2021; Weston et al., 2018), which significantly undermine the user's experience. Retrieval-based models effectively address these issues, since they can provide similar exemplars for models as reference. For instance, in the realm of empathetic dialogue generation, some work leverages retrieval model to mine exemplars, enhancing model's empathetic ability (Zhu et al., 2023; Majumder et al., 2021; Chen et al., 2024). However, they mainly focus on semantic similarity, lacking the similarity of emotion characteristics, limiting model's generation quality. Thus, our **REG** model first introduces four **Emotion Attributes** to solve this issue.

3 Method

3.1 Task Formulation

Our goal is to generate an empathetic response R given a dialogue context $X = \{u_1, ..., u_n\}$ consisting of n utterances. To enhance the empathy and content richness, we retrieve a set of exemplars $T_q = \{t_1, ..., t_K\}$ with similar features from the training set T of EmpatheticDialogues (Rashkin et al., 2019). The model serves as a listener, generating responses that are not only contextually coherent but also empathetic to the user's situation.

3.2 Framework

As illustrated in Figure 2, our framework, REG, employs a **two-stage** training strategy to integrate **Emotion Attributes** into the retrieval-augmented generation process.

(1) Stage 1 (Generator Initialization): We first train an initial generator G_1 conditioned on dialogue contexts and their corresponding emotion attributes. This warm-up stage ensures the generator can recognize and utilize fine-grained emotional signals.

(2) Stage 2 (Retriever Optimization & Joint Training): We optimize the retriever by distilling knowledge from G_1. Specifically, we use token-level evaluation metrics from G_1 as reward signals to guide the retriever, mitigating the noise often introduced by coarse-grained emotion labels. Furthermore, we introduce **Retrieval Candidate Augmentation** to improve the diversity and robustness of the retrieved exemplars.

Finally, the retriever and a stronger generator G_2 are jointly trained to maximize performance.

3.2.1 Emotion Attributes

To move beyond mere semantic matching, we incorporate four key Emotion Attributes (Sharma et al., 2021; Watson et al., 2002)—*Sentiment, Emotional Presence, Interpretation, and Exploration*. These attributes characterize the emotional depth of a conversation:

- **Sentiment (Sent):** Indicates the emotional polarity, ranging from -1 (negative) to 1 (positive) using VADER (Hutto and Gilbert, 2014). Ideally, the sentiment should align with or appropriately react to the user's state.
- **Emotional Presence (Emp):** Categorizes whether emotion is *absent*, *implicitly present*, or *explicitly present*.
- **Interpretation (Int):** Assesses whether the response would correctly interpret the user's situation and if it includes shared self-experiences.
- **Exploration (Exp):** Determines the level of attempt (none, generic, or specific) to explore the user's emotions through questions.

For a pair of context X and gold response r, we extract the sentiment score via independent classifiers trained following Majumder et al. (2021):

$$\begin{aligned} Sent &= \text{VADER}(X) \\ Emp &= \text{CLS}_{\text{Emp}}(X), \\ Int &= \text{CLS}_{\text{Int}}(X), \\ Exp &= \text{CLS}_{\text{Exp}}(X), \end{aligned} \tag{1}$$

where each attribute is classified into discrete levels (Low/Mid/High). These distinct attribute embeddings serve as auxiliary inputs to the generator, providing explicit emotional guidance.

3.2.2 Generator Initialization

First, we train the initial generator G_1 to generate responses conditioned on the context X and the extracted emotion attributes $Attr$. The objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{G_1} = -\sum_{t=1}^{T} \log \mathcal{P}(r_t | r_{<t}, X, Attr), \tag{2}$$

where T is the length of the target response. This stage equips G_1 with the capability to understand the correlation between attributes and empathetic responses.

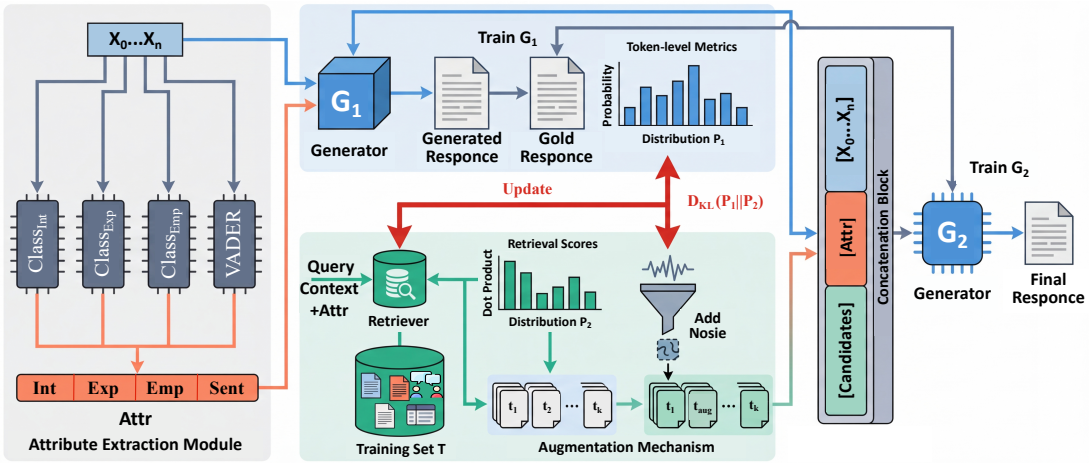

Figure 2: An overview of our proposed empathetic dialogue generation framework. The left part illustrates how we extract and concatenate four types of Emotion Attributes (Sentiment, Interpretation, Exploration, Emotional presence) from the gold response and dialogue context using pre-trained classifiers and VADER, serving as input for initial generator training (G_1). In the middle, we show the use of token-level metrics (e.g., BLEU) to train a fine-grained retriever via a token-level retrieval strategy. The right part shows the retriever optimization via retrieval candidate augmentation, enabling the retriever to avoid local optima and improve exemplar diversity. The last two parts will be jointly trained with the generator G_1 to obtain G_2.

3.2.3 Retriever Optimization

We utilize Contriever (Izacard et al., 2021) as the backbone to encode the query q and candidate exemplars. The initial retrieval is based on the dot-product similarity of their embeddings. However, standard dense retrieval often relies on semantic overlap and may miss emotional nuances. Relying solely on coarse emotion attributes for filtering can also introduce noise. To address this, we refine the retriever using two strategies:

Token-level Retrieval (Metric-driven Distillation) Since the generation process is non-differentiable, we cannot directly backpropagate generation loss to the retriever. Instead, we use the pre-trained G_1 to score the quality of retrieved candidates. We convert token-level metrics (e.g., ROUGE-L, which reflects fine-grained alignment) into a supervision distribution. For a query X, given K retrieved candidates $\{t_1, ..., t_K\}$, the probability of selecting candidate t_i is defined as:

$$p_i = \frac{\exp\left(\frac{1}{\tau}F(r, \text{Dec}(G_1, (t_i, A_{t_i}, X, A_x)))\right)}{\sum_{j=1}^{K}\exp\left(\frac{1}{\tau}F(r, \text{Dec}(G_1, (t_j, A_{t_j}, X, A_x)))\right)}, \tag{3}$$

where $F(\cdot)$ is the utility metric (e.g., ROUGE-L) measuring the quality of the generated text against the gold response r. $Dec(\cdot)$ denotes the decoding process of G_1 enhanced by the exemplar t_i and its attributes. τ is a temperature hyperparameter. Here, a higher p_i indicates that exemplar t_i helps G_1 generate a more accurate response with less noise. We treat $P = \{p_i\}_{i=1}^{K}$ as the target distribution to supervise the retriever's output similarity scores S_q. The loss is the KL divergence:

$$\mathcal{L}_R = \text{KL}(P||\sigma(S_q)), \tag{4}$$

where σ is the Sigmoid function.

Retrieval Candidate Augmentation To prevent the retriever from collapsing into local optima (i.e., retrieving a narrow set of semantically close but emotionally irrelevant samples), we introduce a randomization mechanism. During training, we replace a retrieved candidate t_i with a randomly selected candidate t_i^{aug} from the corpus with a probability k:

$$t_i^{\text{aug}} = \begin{cases} \text{Random}(T), & \text{with probability } k \\ t_i. & \text{otherwise} \end{cases} \tag{5}$$

This forces the retriever to rank truly useful exemplars higher than random noise, enhancing robustness. The similarity scores for the augmented set S_q^{aug} are optimized using the same KL divergence loss as Eq. 4.

3.2.4 Input-augmented Generator

With the optimized retriever, we aim to train the final generator G_2. We employ the Fusion-in-Decoder (FiD) architecture (Izacard and Grave, 2021). Each retrieved exemplar t_i is concatenated with the context X and encoded independently. The decoder then attends to all encoded representations to generate the final response. The generation

loss is:

$$\mathcal{L}_{G_2} = -\sum_{t=1}^{L} \log \mathcal{P}(r_t | r_{<t}, \mathbf{z}), \tag{6}$$

where $\mathbf{z}$ represents the aggregated encoder outputs of the context and retrieved exemplars.

3.2.5 Joint Training

In the final stage, we update both the retriever and the generator to ensure mutual adaptation. The total objective function combines the retrieval distillation loss (with augmentation) and the generation loss:

$$\mathcal{L} = \alpha \mathcal{L}_R^{\text{aug}} + (1 - \alpha)\mathcal{L}_{G_2}, \tag{7}$$

where α balances the two tasks. We set $\alpha = 0.2$ to prioritize generation quality while maintaining retrieval stability.

4 Experiments

4.1 Experimental Setup

Dataset & Retrieval Corpus We conduct experiments on the **EmpatheticDialogues** dataset (Rashkin et al., 2019), utilizing an $8:1:1$ split for training, validation, and testing. To construct the retrieval corpus, we employ the training set T as the knowledge source. Both the dataset and the retrieval corpus are augmented with the four Emotion Attributes described in Section 3.2.1 to facilitate emotionally aligned retrieval.

Implementation Details We implement our framework using PyTorch and HuggingFace Transformers. To verify the generalization of our method, we apply REG across three backbone architectures: (1) Transformer: A standard Transformer trained from scratch. (2) PLM: We utilize **T5-Small** (Raffel et al., 2020) initialized with pre-trained weights. (3) LLM: We employ Llama-3.1-8B-Instruct. The retriever is initialized with Contriever (Izacard et al., 2021). Computations are performed on 4 NVIDIA P100 GPUs. For the T5-based model (our primary analysis focus), the two-stage training takes approximately 30 and 36 hours respectively, with a learning rate of 1×10^{-4}. Please refer to the published project for additional details, which is publicly available.[1]

4.2 Baselines

We compare REG with competitive baselines across three categories:

Transformer-based Models: **(1) Transformer** (Raffel et al., 2020): Standard sequence-to-sequence model. **(2) MIME** (Majumder et al., 2020): Uses emotional grouping and mimicry. **(3) KEMP** (Li et al., 2020): Integrates commonsense knowledge graph. **(4) CASE** (Zhou et al., 2023): Aligns affection with commonsense cognition. **(5) CAB** (Pan Gao, 2023): Aligns affection with commonsense cognition. **(6) IAMM** (Yang et al., 2024): Model internal affect memory. **(7) ReflectDiffu** (Yuan et al., 2025): Generate empathetic responses via emotion-intent fusion.

PLM-based Models: **(1) DialoGPT** (Zhang et al., 2019) & **BlenderBot** (Roller et al., 2020): Large-scale pre-trained dialogue models. **(2) LEMPEx** (Majumder et al., 2021) & **Exemplar-Empathy** (Zhu et al., 2023): Exemplar-guided models relying on semantic similarity. **(3) EmpGPT-3** (Lee et al., 2022): Prompt-based in-context learning. **(4) PECER** (Cai et al., 2024): Focuses on dynamic personality and emotional reasoning.

LLM-based Models: **(1) GPT-4o** (Qian et al., 2023): State-of-the-art closed-source LLM in empathetic dialogue. **(2) Llama-3.1-8B** (Grattafiori et al., 2024): Evaluated with supervised fine-tune (SFT) and Chain-of-Thought (CoT) prompting.

4.3 Main Results

We leverage widely used metrics including Perplexity (**PPL**), **BLEU-1/2** (Papineni et al., 2002), Distinct-1/2 (**Dist-1/2**) (Li et al., 2016), and Emotion Classification Accuracy (**Acc**). Table 1 summarizes the performance of REG against baselines across three backbone architectures.

Effectiveness on T5 (Main Results) Our primary implementation, T5+REG, achieves State-of-the-Art performance. It records the lowest PPL (16.67), indicating high fluency and semantic plausibility. Crucially, REG demonstrates exceptional improvements in diversity mechanism, with Dist-1/2 scores significantly surpassing baselines (e.g., Dist-2 reaches 25.41 vs. 16.83 for PECER). This suggests that our emotionally guided retrieval introduces diverse yet relevant scenarios to the generator. Furthermore, the high Emotion Accuracy (0.478) confirms that retrieving exemplars via Emotion Attributes effectively aligns the generated response with the user's emotional state.

Universality across Architectures The benefits of REG extend to other architectures. (1) For

[1] https://github.com/uuaaaaaaa/REG_Code

Models	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑	BLEU-1 ↑	BLEU-2 ↑
Transformer-based Models						
Transformer	37.65	0.47	2.05	–	18.07	8.34
MIME	37.33	0.41	1.62	0.296	18.60	8.39
KEMP	37.32	0.55	2.31	0.341	18.19	8.15
CASE	35.37	0.74	4.01	0.402	17.90	8.69
CAB	35.06	0.89	2.95	0.405	20.23	9.39
IAMM	34.82	0.88	3.05	0.437	19.51	8.74
ReflectDiffu	**24.56**	0.98	4.35	**0.487**	23.59	**11.25**
REG (Ours)	24.62	**1.15**	**5.12**	0.484	**23.88**	11.22
PLM-based Models						
LEMPEx	26.37	1.41	14.66	–	19.18	8.46
DialoGPT	18.74	2.71	12.01	–	18.69	8.58
BlenderBot	16.71	2.58	16.20	0.470	19.79	9.33
EmpGPT-3	–	3.15	18.63	–	16.38	7.67
Exemplar	16.83	2.71	22.25	–	18.50	9.65
EmpCRL	16.91	4.33	16.32	0.411	20.77	9.85
PECER	16.79	3.69	16.83	–	**21.23**	10.14
T5 + REG (Ours)	**16.67**	**5.79**	**25.41**	**0.478**	19.60	**11.12**
LLM-based Models						
GPT-4o	230.99	1.79	14.72	0.244	6.57	2.68
Llama-3.1-8B + SFT	28.34	0.88	3.96	0.203	22.10	10.45
Llama-3.1-8B + SFT + CoT	24.92	0.92	4.13	0.211	23.38	11.29
Llama-3.1-8B + REG (Ours)	**19.55**	**2.21**	**8.44**	**0.285**	**24.10**	**11.65**

Table 1: Automatic evaluation results. The best results within each category are highlighted in bold.

Transformer-based standard models, REG significantly boosts Dist scores compared to baselines like CASE and MIME. (2) For LLM-based models (Llama-3.1-8B), we observe that while standard SFT and CoT yield high BLEU, they suffer from low diversity (Dist-2: 4.13) and emotional accuracy (Acc: 0.211). Incorporating REG drastically mitigates these issues, raising Acc to 0.285 and Dist-2 to 8.44. This indicates that emotionally aligned exemplars are crucial even for Large Language Models to avoid generic responses.

4.4 Ablation Studies

We conduct ablation studies based on the T5 backbone to quantify the contribution of each component.

Impact of Emotion Attributes As shown in Table 2, Emotion Attributes are fundamental to our framework. Removing them (*w/o* EA.) causes a sharp decline in Emotion Classification Accuracy (0.478 → 0.310). This validates that semantic similarity alone is insufficient for empathy; explicit emotional attributes are necessary to guide the retriever towards exemplars that resonate with the user's feelings.

Effectiveness of Token-level Retrieval We investigate the specific metrics used for retriever optimization: (1) BLEU & F1: Removing either BLEU or F1 leads to consistent performance degradation across all metrics, confirming their role in ensuring precise token-level alignment. (2) ROUGE-L: Interestingly, removing ROUGE-L (*w/o* R-L) leads to a slight increase in diversity (Dist-1/2) but a drop in coherence (BLEU) and Accuracy. This suggests ROUGE-L fosters precision and coherence, potentially at a slight cost to open-ended diversity, effectively balancing the generation. (3) Optimization Strategy: Comparing REG with a standard RAG approach (G_1+RAG), which uses the raw output probability of G_1, we see that our metric-driven supervisor outperforms direct likelihood maximization. This proves that optimizing for discrete quality metrics (like BLEU/F1) provides a robust training signal for the retriever.

augumention

Effectiveness of Retrieval Candidate Augmentation Table 2 shows that removing Retrieval Candidate Augmentation (*w/o* RCA) notably decreases Dist-1/2 scores. This confirms that randomly introducing broader candidates during training prevents the retriever from collapsing into a narrow scope, thereby encouraging the model to reference diverse exemplars and generate richer content.

Models	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc ↑	BLEU-1 ↑	BLEU-2 ↑
REG	**16.67**	5.79	25.41	**0.478**	**19.60**	**11.12**
w/o EA.	16.71	5.62	24.72	0.310	17.13	8.55
w/o BLEU	16.96	5.64	23.91	0.473	16.77	7.91
w/o F1	16.83	5.61	22.30	0.471	18.12	10.61
w/o R-L.	16.90	**5.81**	**25.53**	0.469	18.26	9.10
w/o RCA.	16.78	5.51	24.12	0.451	17.25	9.04
G_1+RAG	16.94	3.89	18.11	0.391	17.10	8.61
$REG_{scratch}$	16.93	5.75	25.01	0.463	18.50	8.51

Table 2: Results of ablation study. EA: Emotion Attributes, RCA: Retrieval Candidate Augmentation.

4.5 Analysis

Models	PPL ↓	Dist-1 ↑	Dist-2 ↑	BLEU-1 ↑	BLEU-2 ↑	Acc ↑
$T5^S$	17.14	5.50	21.01	17.04	8.05	0.411
$T5^S$+*REG*	16.67	5.79	25.41	19.60	11.12	0.478
$T5^L$	16.17	5.56	22.37	19.42	9.83	0.437
$T5^L$+*REG*	14.89	6.03	26.02	19.84	11.20	0.481
$T5^{XL}$	15.51	5.78	24.79	19.55	10.91	0.487
$T5^{XL}$+*REG*	13.73	6.51	26.31	19.96	11.28	0.489

Table 3: Results on the size of T5 series models.

Scalability on T5 Series We further verify REG's efficacy across different model sizes (Small, Large, XL). Results in Table 3 demonstrate consistent improvements. REG not only functions effectively on T5-Small but also enhances T5-Large and T5-XL, providing substantial gains in diversity and emotional accuracy while maintaining low perplexity. This scalability underlines the robustness of our framework.

Quality of Retrieved Exemplars To validate the retriever directly, we manually evaluated the emotional alignment of retrieved exemplars for 500 test instances (K=6). As shown in Table 5, our method retrieves a significantly higher number of emotionally aligned exemplars (1786/3000) compared to semantic-only baselines like LEMPEx (1139/3000). This empirical evidence supports our claim that Emotion Attributes effectively guide the retrieval process.

Training Strategy We compared our two-stage training with training REG from scratch ($REG_{scratch}$). The latter shows degraded performance across all metrics (Table 2). This indicates that "warming up" the generator (G_1) is crucial to provide reliable supervision signals for updating the retriever in the second stage.

4.6 Human evaluation

These 200 contexts are randomly selected, and following Sabour et al. (2022), three crowdsourcing workers are asked to choose the better response(Win) by considering the three aspects: **Coherence(Coh.):** response is more coherent in content and more related to the context. **Empathy(Emp.):** response is more empathetic and has a better understanding of user's situations. **Informativeness(Inf.):** response carries more information related to the context. We pay 10 $ per hour for the three workers with undergraduate education.

The results are presented in Table 6. REG consistently outperforms baselines across all metrics. Specifically: (1) Against semantic-retrieval baselines (e.g., LEMPEx, Exemplary-empathy), REG demonstrates significant gains in Empathy and Informativeness, confirming that our attribute-guided retrieval introduces more emotionally relevant and specific content than standard semantic matching. (2) Interestingly, in the comparison with the LLM baseline (Llama-3.1 + SFT), while the Coherence scores are competitive (reflecting the inherent fluency of LLMs), REG achieves a decisive victory in Empathy (58.0% vs. 32.5%). This suggests that while general-purpose LLMs produce fluent text, they often resort to generic "safe" responses. In contrast, REG, by retrieving emotionally aligned exemplars, captures finer emotional nuances and responds with greater empathy and specificity.

4.7 Case Study

Table 4 presents two cases from five models, showing that REG tends to generate more accurate, empathetic, and profound responses, demonstrating a better understanding of user situations. This performance mainly stems from three advantages:

(1) Emotion Attributes. EA helps retrieve emotionally similar exemplars. In Case 1, with a terri-

Context	I was driving and it began to rain so hard that I could see nothing in front of me.	My best friend is getting married. I'm so excited for her!
Emotion	**Terrified**	**joyful**
Transformer	I am sorry to hear that. I hope you have a good time!	
KEMP	This is so bad. I am so sorry to hear that.	Oh, that's so good news!
CASE	Oh! I hope you will be sure.	Wow, I think it so excited too.
EmpCRL	Really? That is so dangerous. Are you safe now?	I am happy to hear this, congratulations!
PECER	I am so sorry to here that and did you go?	Oh, are you going to the wedding?
Llama-3.1-8B (SFT)	That sounds absolutely terrifying! It is very dangerous to drive when visibility is low. I hope you managed to pull over or get home safely.	That is wonderful news! It creates such a happy atmosphere when a close friend gets married. You must be looking forward to the celebration!
LEMPEx	Oh so bad! Did you get wet?	So happy! Congratulations!
Exemplars (by LEMPEx)	*Speaker:*S I got in my car to go to a baseball game. I was really excited to see some of my favorite players. When I got there, heavy rain started to come down. Eventually they canceled the game. *Listener:* That has been happening a lot lately. Everyone in the US seems to be getting wet.	*Speaker:* I'm getting married this fall! *Listener:* Wow, congrats on my behalf.
REG (Ours)	So sorry to hear that. That's so terrible and did you sucessfully arrive?	Such good news! When will the wedding hold?
Exemplars (by REG)	**Exemplar 1.** *Speaker:* When a storm came through my town a few springs ago, I felt prepared. I knew when to leave and where to go. *Listener:* That must have been very frightening. Did you have somewhere safe that you went to? **Exemplar 2.** *Speaker:* We have had so much rain the past few weeks and my house ended up flooding. I lost everything! *Listener:* I am really sorry. Do you have house insurance?	*Speaker:* I felt contentment when my brother was going to get married. I just felt glad for him. *Listener:* Congrats for your brother! When is the wedding? *Speaker:* My parents have been married for many years. I cannot wait to help them celebrate their anniversary this year! *Listener:* That's sweet. How long have they been married?"
Ground-Truth	Oh no, what happened to your wipers?	When is she getting married?

Table 4: Case study of the generated responses by REG and the baselines.

Methods	**RAG**	**Fixed**	**LEMPEx**	**Ours**
Aligned Count	257	1098	1139	**1786**
Total Count	3000	3000	3000	3000
Accuracy	8.6%	36.6%	38.0%	**59.5%**

Table 5: Evaluation of retrieved exemplar alignment quality.

fied context, REG finds emotionally similar examples like *a storm came through my town; had so much rain; my house ended up flooding.*, leading to a better understanding of the situation. In Case 2, especially via Interpretation and Exploration attributes, REG retrieves more profound exemplars, such as *When is the wedding?*, which guides the model towards deeper engagement and potentially greater user satisfaction.

(2) Token-level Retrieval. REG employs token-level retrieval to refine the retrieval process, yielding more accurate and relevant exemplars. In Case 1, TLR retrieves examples with less noise, contributing to more empathetic responses. In Case 2, it mines examples that better match the context, further enhancing empathetic dialogue generation and user experience.

(3) Retrieval Candidate Augmentation. RCA is used to retrieve more diverse exemplars. In Case 1, this leads to more diverse input, enriching the model's output variation. In Case 2, REG also retrieves more diverse context-similar examples, resulting in more empathetic and diverse responses.

5 Conclusion and Future Work

In this paper, we propose the **REG** framework, a novel retrieval-augmented model designed to enhance empathetic dialogue generation. A key contribution is the introduction of four Emotion Attributes which guide the retrieval process to identify exemplars that are not only semantically similar but also emotionally aligned with the user's context, thereby fostering a deeper understanding of their situation. To address issues of retrieval noise and promote diversity, we employ a Token-level Retrieval strategy, utilizing generator metrics to refine exemplar selection, and incorporate Retrieval Candidate Augmentation to broaden the pool of reference examples. For future work, we aim to

Comparisons	Aspects	Win	Lose	κ
REG vs. CASE	Coh.	**53.7**‡	32.1	0.56
	Emp.	**56.3**‡	31.0	0.57
	Inf.	**56.0**‡	31.8	0.55
REG vs. LEMPEx	Coh.	**52.1**†	35.5	0.50
	Emp.	**53.8**‡	35.1	0.51
	Inf.	**52.6**‡	35.2	0.53
REG vs. Exemplary-empathy	Coh.	**50.8**‡	45.3	0.51
	Emp.	**49.8**‡	40.5	0.49
	Inf.	**52.7**‡	43.3	0.47
REG vs. PECER	Coh.	**51.3**‡	45.3	0.51
	Emp.	**50.6**‡	43.5	0.49
	Inf.	**51.9**‡	43.2	0.47
REG vs. Llama-3.1 (SFT)	Coh.	**46.5**	44.2	0.48
	Emp.	**58.0**‡	32.5	0.52
	Inf.	**54.3**‡	35.1	0.50

Table 6: Human evaluation results (Win/Lose %). Ties are omitted for brevity. Kappa (κ) scores indicate moderate inter-annotator agreement. † and ‡ denote significance at $p < 0.1$ and $p < 0.05$.

further explore methods for retrieving even more precise and contextually relevant exemplars to continue improving the depth and quality of empathetic dialogue generation.

Limitations

The primary limitation of our study lies in the mismatch between automatic evaluation metrics and human judgment. While automated metrics primarily assess the quality of generated responses and the accuracy of emotion prediction, they fall short in capturing the nuanced aspects of empathy. This highlights the need for a more comprehensive and standardized approach to evaluate empathetic dialogue generation in the future.

Ethics Statement

Our experiments are conducted using the widely adopted EmpatheticDialogues dataset, which has been carefully filtered during its creation to eliminate any sensitive or personally identifiable information. We ensure that the dataset contains no private data. Furthermore, all human evaluations are carried out anonymously to safeguard the privacy of the participants. Throughout our study, we strictly adhere to ethical guidelines in both dataset usage and human evaluation, ensuring that no harm, bias, or breaches of privacy occur.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376188, 62272340, 62276187, 62376192, 62166022).

References

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Pecer: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10631–10635. IEEE.

Changyu Chen, Yanran Li, Chen Wei, Jianwei Cui, Bin Wang, and Rui Yan. 2024. Empathetic response generation with relation-aware commonsense knowledge. In *WSDM*, pages 87–95.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-CORE: Emotion correlation enhanced empathetic dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10568–10586, Singapore. Association for Computational Linguistics.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. 2024. The llama 3 herd of models.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. *Advances in neural information processing systems*, 31.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

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 *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill 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, San Diego, California. Association for Computational Linguistics.

Qintong Li, Pijian Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2020. Knowledge bridging for empathetic dialogue generation. In *AAAI Conference on Artificial Intelligence*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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)*, volume 1, pages 3469–3483. Association for Computational Linguistics.

Yunlong Ma, Bo Wang, Yihong Tang, Zifei Yu, Chenyun Xue, Gaoke Zhang, and Yuexian Hou. 2026. Queryaligner: Customizing user query to match llms preferences for better intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32446–32454.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2021. 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102 2:246–68.

Rui Zhou Xuejiao Zhang Zikun Wang Pan Gao, Donghong Han. 2023. Cab: Empathetic dialogue generation with cognition, affection and behavior. *arXiv preprint arXiv:2302.01935*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Ashish Sharma, Inna Lin, Adam Miner, David Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*. ACM.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468.

Yihong Tang, Bo Wang, Dongming Zhao, Jinxiaojia Jinxiaojia, Zhangjijun Zhangjijun, Ruifang He, and Yuexian Hou. 2024. Morpheus: Modeling role from personalized dialogue history by exploring and utilizing latent space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676.

Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31.

Xu Wang, Bo Wang, Yihong Tang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. Ecc: Synergizing emotion, cause and commonsense for empathetic dialogue generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5475–5485.

Jeanne Watson, Rhonda Goldman, and Margaret Warner. 2002. Clientcentered and experiential psychotherapy in the 21st century: Advances in theory, research, and practice.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 2024. Dynamic demonstration retrieval and cognitive understanding for emotional support conversation. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2024. An iterative associative memory model for empathetic response generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3081–3092, Bangkok, Thailand. Association for Computational Linguistics.

Jiahao Yuan, Zixiang Di, Zhiqing Cui, Guisong Yang, and Usman Naseem. 2025. ReflectDiffu: Reflect between emotion-intent contagion and mimicry for empathetic response generation via a RL-diffusion framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25435–25449, Vienna, Austria. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Y Zhao, B Wang, and Y Wang. Explicit vs. implicit: Investigating social bias in large language models through self-reflection (2025).

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237, Toronto, Canada. Association for Computational Linguistics.

Ying Zhu, Bo Wang, Dongming Zhao, Kun Huang, Zhuoxuan Jiang, Ruifang He, and Yuexian Hou. 2023. Grafting fine-tuning and reinforcement learning for empathetic emotion elicitation in dialog generation. In *ECAI 2023*, pages 3148–3155. IOS Press.

A Appendix

A.1 VADER

VADER outperforms individual human raters (F1 Classification Accuracy = 0.96 and 0.84, respectively) and the Pearson Product Moment correlation coefficient shows that VADER ($r = 0.881$) performs as well as individual human raters ($r = 0.888$) at matching ground truth. This indicates that VADER's performance is on par with, or even superior to, individual human annotators.

A.2 The Three Classifiers

The EmpathyMentalHealth dataset (Sharma et al., 2021) is leveraged to create synthetic labels for the data in EmpatheticDialogues dataset. The EmpathyMentalHealth includes context with annotated emotional presence, interpretation, exploration labels. Labels for all the dimensions are provided on a scale of 1/2/3, indicating low/mid/high levels. The pretrained T5 model is incorporated as the backbone of the classification models, which has only the encoder part. A linear layer is added on top of the encoders with softmax activation for the three class classification.

The classification models are trained on EmpathyMentalHealth dataset, using triplets consisting of a context and a corresponding label. To prepare the input, the context is fed into a T5 encoder model. The classification label is derived from the output vector associated with the initial <s> token in the final layer. For evaluation, LEMPEx use 20% of the annotated samples from the EmpathyMentalHealth dataset as a validation set. The performance on this validation split for classifying emotional presence, interpretation, and exploration is reported in Table 7. Among these, the exploration category shows the highest predictive performance, achieving nearly a 94% weighted F1 (W-F1) score. In contrast, emotional presence and interpretation are more challenging to predict, with W-F1 scores around 83% and 84%, respectively.

Based on the best-performing checkpoints on the validation set, we generate synthetic labels for samples in the EmpatheticDialogues dataset. For each instance, we input the context into the trained model, and assign the predicted class as the synthetic gold label.

Dimension	**ACC**	**W-F1**
Emotional Presence	82.99	82.89
Interpretation	84.47	83.70
Exploration	94.04	93.92

Table 7: Accuracy (Acc) and weighted-F1 (W-F1) scores of 3-way classification for emotional presence, interpretation, and exploration prediction in validation set of EmpathyMentalHealth.