ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447417

GENERATING PERSONA-AWARE EMPATHETIC RESPONSES WITH RETRIEVAL-AUGMENTED PROMPT LEARNING

Zhengjie Huang[1] *Pingsheng Liu*[1] *Gerard de Melo*[2] *Liang He*[1] *Linlin Wang*[1,3*]

[1] East China Normal University [2] Hasso Plattner Institute, University of Potsdam
[3]Shanghai Artificial Intelligence Laboratory

ABSTRACT

Empathetic response generation requires perceiving and understanding the user's emotion to deliver suitable responses. However, existing models generally lack an ability to respond in a persona-specific way, which has been shown to play a vital role in expressing appropriate empathy. To address this problem, we propose a novel Transformer-based architecture that incorporates retrieval-augmented prompt learning to generate persona-aware empathetic responses. Since personalized emotional resonance is subtle and uncontrollable, we employ dense passage retrieval to retrieve exemplary responses that reflect specific persona and context characteristics to cue the generative model on signaling empathy. Extensive experiments confirm the effectiveness of our model for persona-aware empathetic response generation.

Index Terms— Natural Language Processing, Dialogue Generation, Empathetic, Personalized, Exemplar Prompting

1. INTRODUCTION

Empathy, i.e., an individual's ability to perceive and intimately understand the feelings and experiences of others, is a fundamental trait of human daily conversations. Endowing dialogue systems with more empathetic responses can improve user satisfaction and participation in diverse domains and settings [1, 2]. Empathy is a complex construct that consists of two notable aspects, namely affection and cognition [3]. The affective aspect involves the appropriate emotion expression in reaction to the user's experiences [4], such as *sadness*. The cognitive aspect entails perceiving the user's situation [5], which is often expressed in the dialogue act leveraged in the dialogue, such as *suggesting*. Recently, numerous approaches have been introduced to improve the ability of dialogue models to comprehend the feelings of interlocutors [6, 7, 8, 9, 10, 11].

Fig. 1 presents a typical example of dialogue expressing empathy. We believe that despite encountering the same conversational context, different empathetic responses may be provided depending on the agent's persona [12, 13]. However,

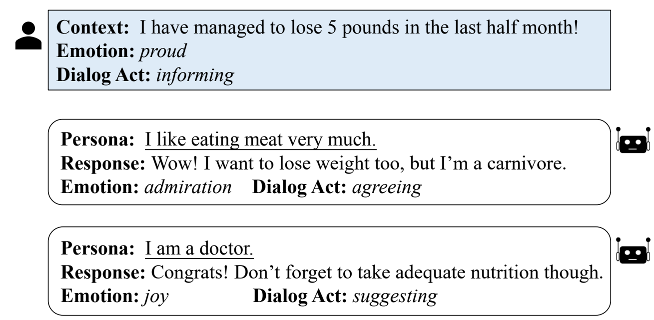

Fig. 1. Example of empathetic response generation by two dialogue agents with different personas. Different personas may entail different styles of expressing sympathy.

most existing work ignores the impact of personality characteristics when generating empathetic dialogue responses. Even in the case of ChatGPT, the responses primarily reflect its character persona in the semantic context rather than displaying distinct emotional expressions. This calls for new research on dialogue models that capture sufficient signals to deliver persona-aware empathetic responses.

Current generative models often struggle to reflect personalized characteristics and perceive a user's real emotional situation, as it may be too subjective and vague for systems to viscerally understand feelings. They often produce utterances that are insufficiently empathetic or entirely appropriate, e.g., merely responding with "*That's awful.*" in response to "*My uncle was screaming for help all last night.*", or even "*How did he die?*". Hence, drawing on the idea of controlled text generation, we retrieve exemplary responses from the training set based on persona information and the dialogue context, which serve as the reference for the dialogue system to perceive empathetic features. It matches the desired embedded representations in a continuous low-dimensional latent space. Although the retrieved exemplars can provide essential cues during the decoding process, they are insufficient to enable high-quality persona-aware empathetic responses [14]. We further incorporate prompt learning so as to elicite extrinsic signals from pre-trained language models, providing supplementary evidence for response generation.

*Corresponding author. Email: llwang@cs.ecnu.edu.cn

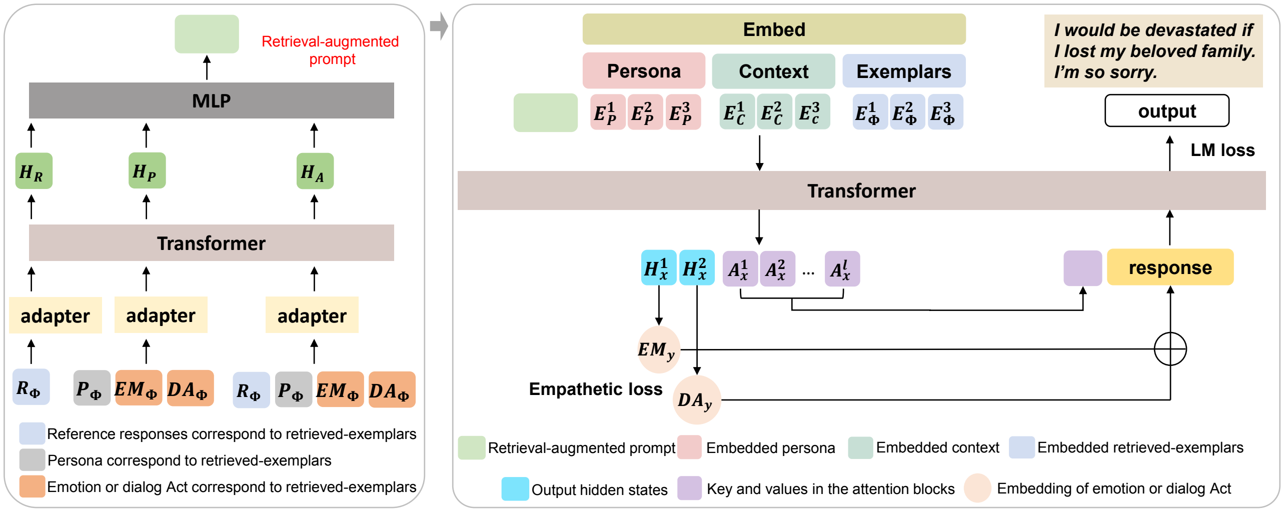

Fig. 2. Overall architecture of the proposed model.

In this work, we introduce the **P**ersona-aware **E**mpathetic **R**esponse **G**eneration **M**achine (PERGM). Our key contributions are: (1) We propose a novel Transformer-based architecture for persona-aware empathetic response generation, whereas most existing work ignores persona characteristics. (2) Our model incorporates retrieval-augmented prompt learning into a generative architecture, leveraging exemplars as guiding signals. (3) Automatic and manual evaluations demonstrate that PERGM can generate more informative and empathetic responses than prior work.

2. METHODOLOGY

Given a conversation $C = [u_1, u_2, ..., u_m]$ involving a participating user and a dialogue agent, each of the m utterances is labeled with emotion (EM) and dialogue act (DA) tags, and the agent's persona $P = [p_1, p_2, ..., p_n]$ is described by n sentences. We aim to generate a response u_{m+1} that is empathetic to the user's situation and reflects the agent's persona entries. As depicted in Fig. 2, our model first retrieves exemplars, and subsequently incorporates retrieval-augmented prompt learning into a pre-trained GPT-based decoder, seeking to generate persona-aware empathetic responses.

2.1. Exemplar Retrieval

Given a query (i.e., the agent's persona P and the dialogue context C), and a set of candidate responses d_Φ from the training set, we adopt dense passage retrieval (DPR) [15] to retrieve exemplary responses as in Fig. 3, learning a probability distribution $p(d_\Phi|P, C)$ according to the similarity:

$$p(d_\Phi|P,C) = \exp(E_{\mathrm{PC}}(P,C) \cdot E_{\mathrm{R}}(d_\Phi)^T), \qquad (1)$$

where $E_{\mathrm{PC}}(\cdot) \in \mathbb{R}^{1\times H}$ and $E_{\mathrm{R}}(\cdot) \in \mathbb{R}^{m\times H}$ indicate the encoding for persona-context pairs and candidate exemplars

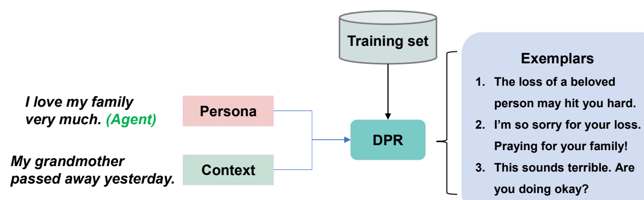

Fig. 3. Hard prompt invokation.

with BERT, H indicates the hidden state size of the encoder, m refers to the number of candidate responses. Finally, we select top-q responses as retrieved exemplars $\hat{d_\Phi}$ with Eq. 1.

2.2. Retrieval-augmented Prompting

Given the retrieved exemplars $\hat{d_\Phi}$, which include the reference responses R_Φ, persona P_Φ, emotion id E_Φ, dialogue act id D_Φ, we generate the prompt by using a disentangled-attention decoder with three independent parameter attention adapters. In order to enable the prompt to focus on the interplay between personality and emotions, and model how they influence responses, three adapters are devised to process inputs $[R_\Phi]$, $[P_\Phi; E_\Phi; D_\Phi]$, and $[R_\Phi; P_\Phi; E_\Phi; D_\Phi]$ separately:

$$\begin{aligned}
\mathbf{H}_R &= \mathrm{Transformer}(R_\Phi) \\
\mathbf{H}_P &= \mathrm{Transformer}([P_\Phi; E_\Phi; D_\Phi]). \qquad (2) \\
\mathbf{H}_A &= \mathrm{Transformer}([R_\Phi; P_\Phi; E_\Phi; D_\Phi]).
\end{aligned}$$

where $\mathbf{H}_R$ captures the impact of semantic context, $\mathbf{H}_P$ models the effect of personality–emotion interactions, and $\mathbf{H}_A$ explores the influence of the above factors. We GPT-2 as our Transformer model. An MLP combines the three adapters' outputs to obtain prompt O:

$$O = \mathrm{MLP}([\mathbf{H}_R; \mathbf{H}_P; \mathbf{H}_A]). \qquad (3)$$

2.3. Empathetic Response Decoding

We use the Transformer-based GPT-2 as the backbone for empathetic response generation, as shown in Fig. 2. Our model takes the persona P, dialogue context C, retrieved exemplars $\hat{d_\Phi}$ and the prompt O as input. Subsequently, we feed the concatenated sequence $x = [O; P; C; \hat{d_\Phi}]$ to the autoregressive GPT-2 model, obtaining the output hidden states, keys, and values in the attention blocks:

$$\mathbf{H}_x, \mathbf{A}_x = \text{Transformer}(x). \tag{4}$$

Next, we adopt the hidden state of the last position in the sequence, $\mathbf{h}_x = \mathbf{H}_x[-1] \in \mathbb{R}^d$ to predict the emotion (EM) id of the target response, denoted by $E_{\tilde{y}}$:

$$\mathbf{h}_E = \mathbf{F}_E(\mathbf{h}_x) \in \mathbb{R}^d,\ \ E_{\tilde{y}} = \text{softmax}(\mathbf{M}_E \mathbf{h}_E), \tag{5}$$

where $\mathbf{F}_E$ is a non-linear layer, d indicates the hidden state size of the Transformer, $\mathbf{M}_E \in \mathbb{R}^{10\times d}$ represents the EM embedding matrix. The hidden state at the penultimate position of the sequence, $\mathbf{h}_{x'} = \mathbf{H}_x[-2] \in \mathbb{R}^d$ is used to predict the dialogue act (DA) id of the target response, denoted by $D_{\tilde{y}}$:

$$\mathbf{h}_D = \mathbf{F}_D(\mathbf{h}_{x'}) \in \mathbb{R}^d,\ \ D_{\tilde{y}} = \text{softmax}(\mathbf{M}_D \mathbf{h}_D), \tag{6}$$

where $\mathbf{F}_D$ is a non-linear layer, $\mathbf{M}_D \in \mathbb{R}^{9\times d}$ represents the DA embedding matrix. Then we add the EM embedding and DA embedding to acquire the fused embedding $\mathbf{e}_r$, which controls the empathy expression of the response:

$$\mathbf{e}_r = \mathbf{M}_E[E_{\tilde{y}}] + \mathbf{M}_D[D_{\tilde{y}}]. \tag{7}$$

The embedding of each token $\tilde{y}_t$ in the response is:

$$\mathbf{e}_{\tilde{y}_t} = \mathbf{e}_r + \text{Embedding}(\tilde{y}_t). \tag{8}$$

Finally, we can predict the next token $\tilde{y}_{t+1}$ by:

$$\mathbf{s}_t = \text{Transformer}(\mathbf{e}_{\tilde{y}_t}, \mathbf{A}_x), \tag{9}$$

$$\tilde{y}_{t+1} \sim \mathbb{P}(y_{t+1}|\tilde{y}_{\leq t}; x) = \text{softmax}(\mathbf{M}_W \mathbf{s}_t). \tag{10}$$

where $\mathbf{s}_t$ refers to the output hidden state corresponding to $\tilde{y}_t$, and $\mathbf{M}_W \in \mathbb{R}^{|\mathcal{V}|\times d}$ represents the word embedding matrix.

2.4. Loss Function

The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{EM}} + \mathcal{L}_{\text{DA}}, \tag{11}$$

where $\mathcal{L}_{\text{NLL}}$ is the negative log likelihood loss of the target response, $\mathcal{L}_{\text{EM}}$ and $\mathcal{L}_{\text{DA}}$ are the empathy factor (i.e., Emotion, Dialog Act) prediction losses. Specifically,

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{l_y}\sum_{t=1}^{l_y} log\mathbb{P}(y_t^*|y_{<t}^*; x, E_y^*, D_y^*), \tag{12}$$

$$\mathcal{L}_{\text{EM}} = -log\mathbb{P}(E_y^*|x),\ \ \mathcal{L}_{\text{DA}} = -log\mathbb{P}(D_y^*|x). \tag{13}$$

	Happy			Offmychest		
	train	valid	test	train	valid	test
#Conversation.	69K	9K	11K	55K	7K	6K
#Utterance.	161K	21K	26K	130K	17K	14K
#Speaker.	66K	14K	15K	59K	13K	14K
#Average.Persona	61.0	62.7	60.5	54.3	61.2	63.4

Table 1. Detailed Statistics of the PEC-AE Dataset.

3. EXPERIMENTS

3.1. Setup

Dataset: PEC [12] is a large-scale persona-grounded empathetic conversation dataset crawled from Reddit. It consists of two different domains, **Happy**, which primarily has posts with positive sentiment, and **Offmychest**, where negative posts prevail. To promote higher quality empathy research, a more fine-grained annotation with Dialogue Act (DA) and Emotion (EM) labels [10] has been created on top of the PEC dataset. For experiments, we use the processed PEC-AE dataset, with detailed statistics given in Table 1.

Baselines: Several representative Transformer-based models are compared with our model: (1) **GPT** [16], a vanilla model which generates responses based only on the dialogue context. (2) **EmpTransfo** [8], a transfer learning based scheme together with a multi-task objective for empathetic response generation. (3) **CoMAE** [10], a framework hierarchically modeling the empathy factors for empathetic response generation, achieving state-of-the-art performance on PEC-AE. (4) **ChatGPT**, a representative model specifically trained to interact with users in a genuinely conversational manner.

Implementation Details: All the implementations are based on PyTorch and the HuggingFace Transformers library. For training the DPR, two *BERT-base* [17] models are employed to encode query and candidates respectively, the initial learning rate is set to 1×10^{-5} with a linear decay of step size 0.1 per epoch. For the decoder, we employ *GPT2-small* [16] with 12 layers, 12 heads, and hidden state size of 768 as our backbone. The parameter count of PERGM is 249M. We adopt AdamW optimization with an initial learning rate of 6.25×10^{-5} and linear decay of step size 0.1 per epoch. The full model converges in 3 epochs with batch size 16, and we adopt an early-stopping criterion.

Evaluation Metrics: We use both automatic metrics and human assessments for evaluation. As automatic metrics, we utilize **ROUGE-L** [18] to measure the similarity between the generated and reference responses. **Dist-2**, which refers to the ratio of distinct bi-grams in responses, is used to measure diversity. We also report the emotion prediction accuracy (**EA**) for the generated responses. For human evaluation, we perform pairwise comparisons between the gold responses and the model-generated responses. Three human annotators are employed to rate 100 samples from testing data on three aspects critical for practice use: (1) **Empathy**: How empathetic is the response to the user? (2) **Relevance**: How coherent

Models		Dist-2	ROUGE-L	EA	Human ratings		
					Emp.	Rel.	Flu.
Happy	GPT	19.11	11.32	10.9	2.91	3.27	4.52
	EmpTransfo	19.74	17.31	55.1	3.22	3.45	4.30
	CoMAE	20.62	17.84	53.2	3.31	3.47	4.43
	ChatGPT	7.16	11.25	/	**4.01**	**3.86**	**4.76**
	PERGM	**21.83**	**17.85**	**56.4**	3.54	3.78	4.48
Offmychest	GPT	26.04	13.07	8.9	2.69	3.02	4.63
	EmpTransfo	27.91	16.66	32.1	3.16	3.28	4.57
	CoMAE	27.69	16.81	29.3	3.12	3.25	4.58
	ChatGPT	11.13	10.73	/	**3.94**	**3.74**	**4.82**
	PERGM	**28.56**	**17.70**	**33.8**	3.43	3.64	4.7

Table 2. Results of overall evaluation.

Models		Dist-2	ROUGE-L	EA	Human ratings		
					Emp.	Rel.	Flu.
Happy	**PERGM**	**21.83**	17.85	**56.4**	**3.54**	**3.78**	4.48
	w/o Persona	19.45	17.69	55.4	3.48	3.39	4.50
	w/o Exemplars	19.22	17.77	47.2	3.39	3.46	**4.51**
	w/o Prompts	20.71	**18.12**	55.8	3.50	3.65	4.49
Offmychest	**PERGM**	**28.56**	**17.70**	**33.8**	**3.43**	**3.64**	**4.67**
	w/o Persona	27.27	17.07	33.2	3.30	3.24	4.57
	w/o Exemplars	27.69	17.04	26.8	3.19	3.27	4.60
	w/o Prompts	27.93	17.68	33.6	3.36	3.42	4.62

Table 3. Ablation study of different components.

is the response with the persona and context? (3) **Fluency**: How fluent and human-like is the response? The rating scale ranges from 1 to 5 (best score). Scores across 100 samples and three annotators are averaged to obtain the final rating.

3.2. Results

Table 2 shows the evaluation results on Test-Happy and Test-Offmychest domains. As observed, our PERGM yields the best performance on almost all the objective metrics. For ChatGPT, despite achieving the highest human ratings, it exhibited subpar diversity and significantly deviated from the reference response. For instance, on Test-Offmychest, PERGM obtains the highest Dist-2 score and ROUGE-L score compared to the strongest baseline CoMAE, which indicates that our generated responses are more informative and have better fidelity. Furthermore, the highest emotion prediction accuracy also demonstrates that our model can more easily detect the empathetic information for the target responses. Regarding the human ratings, responses from our model obtain the best scores in terms of Empathy and Relevance except for ChatGPT. Although EmpTransfo and CoMAE achieve a comparable empathy score, they still fare substantially worse than our model. This suggests that our model can leverage persona information and exemplars to model empathetic response generation.

Ablation Study. We disentangled the contributions of each part of PERGM in Table 3. All the metrics become worse after we eliminate the persona information of interlocutors, demonstrating the effectiveness of persona in model generation quality. The removal of the retrieved exemplars causes a 7% accuracy drop in EA, a relative 3.8% ROUGE-L drop and 3.1% in Dist-2 on Test-Offmychest, indicating that the retrieved exemplars can greatly facilitate emotional understanding. When removing the prompts, almost all the metrics also deteriorate slightly. This demonstrates that the necessary prompts are beneficial to empathetic response generation.

Analysis. We further analyzed the influence of exemplars for empathetic response generation in Table 4. The retrieved exemplars are highly semantically similar to the references, corresponding to the dialogue context and the responder's persona. The exemplars can assist the decoder in generating relevant words in the response. Specifically, the generated response is semantically close to the retrieved exemplar, and both express *joy* at finding the lost kitten. On the contrary, the generation from *w/o Exemplar* is generic and does not exhibit the persona character of "*I like cute animals*". We also show the impact of number of retrieved exemplars in Fig. 4.

Context:	My first adopted fur baby was lost for four years. We were reunited yesterday afternoon.
Persona:	I like cute animals.
Reference Response:	So happy for you! How did you find the cute little creature?
Retrieved Exemplar:	I'm so happy that your kitten is doing okey now! Could we see a picture of it? I'm a big sucker for cute things.
Ours:	Yay! So happy you got the kitten back, and I really like cute things.
w/o Exemplar:	That's nice. Your cat is very important to you.

Table 4. Case study on leveraging exemplary responses.

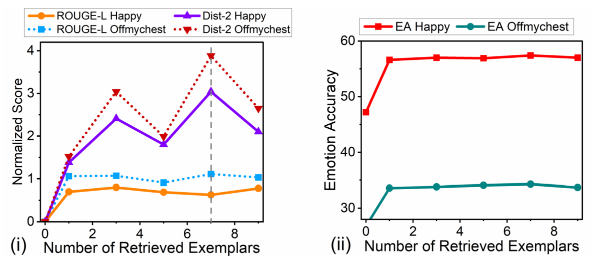

Fig. 4. Impact of adjusting the number of adopted exemplars.

4. CONCLUSION

This paper proposes a novel Transformer-based architecture that incorporates retrieval-augmented prompt learning to generate persona-aware empathetic response. By retrieving empathetically suitable exemplars based on persona and context, it elicits essential linguistic cues from a pre-trained language model. Experiments show our model can deliver more informative and empathetic responses than existing methods.

5. ACKNOWLEDGMENTS

This work was supported by the National Innovation 2030 Major S&T Project of China (No. 2022ZD0160705). We also thank the financial support from Shanghai Chinafortune Co., Ltd and ECNU Multifunctional Platform for Innovation(001).

6. REFERENCES

[1] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang, "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*, 2021, pp. 3469–3483.

[2] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang, "Cass: Towards building a social-support chatbot for online health community," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–31, 2021.

[3] Mark H Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach.," *Journal of personality and social psychology*, vol. 44, no. 1, pp. 113, 1983.

[4] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat, "Empathy: A review of the concept," *Emotion review*, vol. 8, no. 2, pp. 144–153, 2016.

[5] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy, "Therapist empathy and client outcome: An updated meta-analysis.," *Psychotherapy*, vol. 55, no. 4, pp. 399, 2018.

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

[7] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Mime: Mimicking emotions for empathetic response generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 8968–8979.

[8] Rohola Zandie and Mohammad H Mahoor, "Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems," in *The Thirty-Third International Flairs Conference*, 2020.

[9] Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria, "Exemplars-guided empathetic response generation controlled by the elements of human communication," *IEEE Access*, vol. 10, pp. 77176–77190, 2022.

[10] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang, "CoMAE: A multi-factor hierarchical framework for empathetic response generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Aug. 2021, pp. 813–824, Association for Computational Linguistics.

[11] Sahand Sabour, Chujie Zheng, and Minlie Huang, "Cem: Commonsense-aware empathetic response generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11229–11237, Jun. 2022.

[12] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao, "Towards persona-based empathetic conversational models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6556–6566.

[13] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2204–2213.

[14] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[15] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[18] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.