

PAL: Persona-Augmented Emotional Support Conversation Generation

Jiale Cheng*, Sahand Sabour*, Hao Sun, Zhuang Chen, Minlie Huang†

The CoAI group, DCST; Institute for Artificial Intelligence; State Key Lab of Intelligent Technology and Systems; Beijing National Research Center for Information Science and Technology; Tsinghua University, Beijing 100084, China.

{chengjl19, sm22, h-sun20, zhchen-nlp}@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Due to the lack of human resources for mental health support, there is an increasing demand for employing conversational agents for support. Recent work has demonstrated the effectiveness of dialogue models in providing emotional support. As previous studies have demonstrated that seekers' persona is an important factor for effective support, we investigate whether there are benefits to modeling such information in dialogue models for support. In this paper, our empirical analysis verifies that persona has an important impact on emotional support. Therefore, we propose a framework for dynamically inferring and modeling seekers' persona. We first train a model for inferring the seeker's persona from the conversation history. Accordingly, we propose PAL, a model that leverages persona information and, in conjunction with our strategy-based controllable generation method, provides personalized emotional support. Automatic and manual evaluations demonstrate that PAL achieves state-of-the-art results, outperforming the baselines on the studied benchmark. Our code and data are publicly available at <https://github.com/chengjl19/PAL>.

1 Introduction

A growing number of people are experiencing mental health issues, particularly during the Covid-19 pandemic (Hossain et al., 2020; Talevi et al., 2020; Cullen et al., 2020; Kumar and Nayar, 2021), and more and more people are seeking mental health support. The high costs and limited availability of support provided by professional mental health supporters or counselors (Kazdin and Blase, 2011; Olfson, 2016; Denecke et al., 2020; Peterson, 2021) have highlighted the importance of employing conversational agents and chatbots for automating this task (Cameron et al., 2018; Daley et al., 2020; Denecke et al., 2020; Kraus et al., 2021).

*Equal contribution.

†Corresponding author.

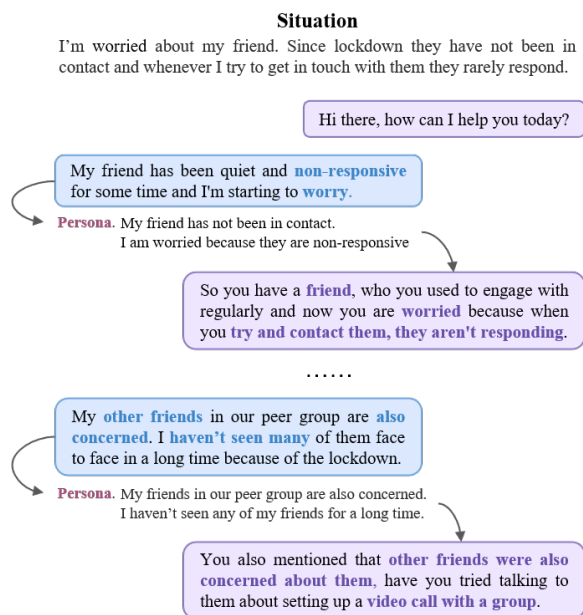


Figure 1: An example from the ESConv dataset, in which the trained supporter extracts key information about the seeker's persona and leverages this information to provide effective emotional support.

Towards this end, Liu et al. (2021) pioneered the task of emotional support conversation generation to reduce users' emotional distress and improve their mood using language models. They collected ESConv, a high-quality crowd-sourced dataset of conversations (with annotated helping strategies) between support seekers and trained emotional supporters, and demonstrated that training large pre-trained dialogue models on this dataset enabled these models to provide effective support. Tu et al. (2022) proposed to leverage commonsense knowledge and implemented hybrid strategies to improve the performance of dialogue models in this task. Similarly, Peng et al. (2022) also suggested using commonsense knowledge for this task and further proposed a global-to-local graph network to model local and global hierarchical relationships. More recently, Cheng et al. (2022) proposed look-ahead

strategy planning to select strategies that are more effective for long-turn interactions.

Although previous studies have considered relevant psychological theories and factors, such as commonsense reasoning, they neglect information regarding the users' persona. Persona, which can be considered as an outward expression of personality (Leary and Allen, 2011) in Psychology, is also closely related to empathy (Richendoller and Weaver III, 1994; Costa et al., 2014), anxiety (Smrdu et al., 2021), frustration (Jeronimus and Laceulle, 2017), mental health (Michinov and Michinov, 2021) and distress (Liu et al., 2018), all of which are essential concepts in psychological scenarios. Effective emotional support benefits from an adequate understanding of the support seeker's personality, as shown by research on person-centered therapy (Rogers, 2013), while more specific and persona-related words lead to a long-term rapport with the user (Campos et al., 2018). Thus, the inability to actively combine persona information and conversations prevents users from developing such rapport with the system (Xu et al., 2022), which is not desirable for emotional support. Therefore, it is intuitive to explore seekers' personas and build systems for providing personalized emotional support.

In this paper, we propose **Persona-Augmented Emotional Support (PAL)**, a conversational model that learns to dynamically leverage seekers' personas to generate more informative and personalized responses for effective emotional support. To more closely match realistic scenarios (no prior knowledge of the user's persona) and retain important user information from earlier conversation rounds, we first extract persona information about the seeker based on the conversation history and design an attention mechanism to enhance the understanding of the seeker. Furthermore, we propose a strategy-based controllable generation method to actively incorporate persona information in responses for a better rapport with the user. We conduct our experiments on the ESConv dataset (Liu et al., 2021). Our results demonstrate that PAL outperforms the baselines in automatic and manual evaluations, providing more personalized and effective emotional support. We summarize our contributions as follows:

- To the best of our knowledge, our work is the first approach that proposes to leverage persona information for emotional support.

- We propose a model for dynamically extracting and modeling seekers' persona information and a strategy-based decoding approach for controllable generations.
- Our analysis of the relationship between the degree of individuality and the effect of emotional support, in addition to the conducted experiments on the ESConv dataset and comparisons with the baselines, highlights the necessity and effectiveness of modeling and leveraging seekers' persona information.

2 Related Work

2.1 Persona in Conversation Generation

There are extensive studies on leveraging persona information in dialogue (Huang et al., 2020). However, it's important to note that the definition of persona in this context differs from its definition in Psychology. In dialogue systems, persona refers to the user's characteristics, preferences, and contextual information, which are incorporated to enhance the system's understanding and generation capabilities. Li et al. (2016b) proposed using persona embeddings to model background information, such as the users' speaking style, which improved speaker consistency in conversations. However, as stated by Xu et al. (2022), this approach is less interpretable. Therefore, several approaches to directly and naturally integrate persona information into the conversation were proposed (Zhang et al., 2018; Wolf et al., 2019; Liu et al., 2020; Yang et al., 2021).

Zhang et al. (2018) collected PERSONA-CHAT, a high-quality dataset with annotated personas for conversations collected by crowd-sourcing workers. This dataset has been widely used to further explore personalized conversation models and how persona could benefit response generation in conversations (Wolf et al., 2019; Liu et al., 2020; Yang et al., 2021). However, it is relatively difficult to implement users' personas in real-world applications, as requiring users to provide information regarding their personas prior to conversations is impractical and unnatural.

Xu et al. (2022) addressed this problem by training classifiers that determine whether sentences in the conversation history include persona information. Accordingly, they store such sentences and leverage them to generate responses. However, in many cases, users do not explicitly express persona

information in the conversation, which often requires a certain level of reasoning. For instance, a user may say, "My friend likes to play Frisbee, so do I", which does not contain any explicit persona information, but one could infer that the user likes to play Frisbee. In this work, we aim to infer possible persona information from the conversation history to assist our model in better understanding the user.

2.2 Emotional Support

In recent years, an increasing number of approaches have focused on emotional and empathetic response generation (Zhou et al., 2018; Zhong et al., 2020; Kim et al., 2021; Gao et al., 2021a; Zheng et al., 2021; Sabour et al., 2022b). However, although such concepts are essential, they are insufficient for providing effective support as this task requires tackling the user’s problem via various appropriate support strategies while exploring and understanding their mood and situation (Liu et al., 2021; Zheng et al., 2022). Therefore, Liu et al. (2021) proposed the task of Emotional Support Conversation Generation and created a set of high-quality conversations between trained crowd-sourcing workers. Their work demonstrated that training widely-used dialogue models, such as Blenderbot (Roller et al., 2021), on their collected dataset enabled such models to provide effective emotional support. Following their work, Tu et al. (2022) proposed leveraging external commonsense knowledge to better understand the users’ emotions and suggested using a mixture of strategies for response generation. Peng et al. (2022) implemented a hierarchical graph network to model the associations between global causes and local intentions within the conversation. Cheng et al. (2022) proposed multi-turn strategy planning to assist in choosing strategies that are long-term beneficial. However, existing work has not explored the effects of dynamically modeling users’ persona information in this task, which we hypothesize improves models’ emotional support ability and enables more personalized support.

3 Persona-Augmented Emotional Support

Figure 2 shows the overall flow of our approach. We first infer the seeker’s persona information from the conversation history and then leverage the inferred information to generate a response. Our approach is comprised of three major components:

The persona extractor for inferring the seeker’s persona information (§3.2); The response generator that leverages the inferred persona information and generates the response distribution (§3.3); A strategy-based controllable decoding method for generating appropriate responses (§3.4).

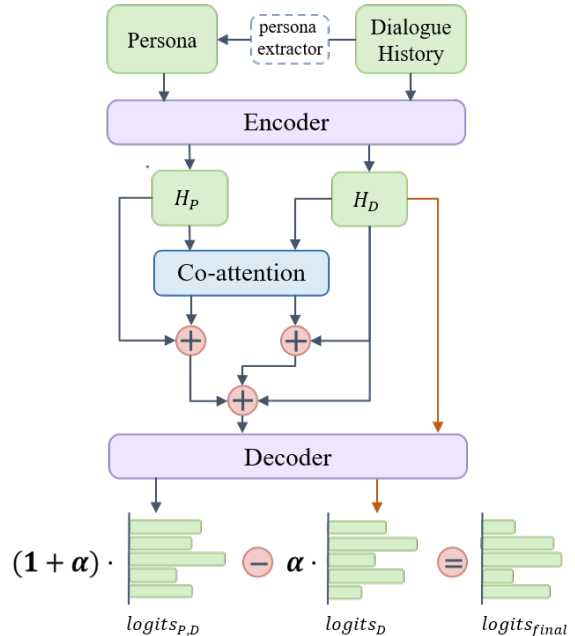


Figure 2: The overall structure of Persona-Augmented Emotional Support (PAL). We extract the seeker’s persona from the dialogue history and then use a controllable generation method to generate the response. α is a tunable hyperparameter.

3.1 Problem Formulation

For inferring users’ personas, we leveraged the PERSONA-CHAT dataset (Zhang et al., 2018), a high-quality collection of conversations between crowd-sourcing workers assigned with a set of predefined persona sentences. Assume that a conversation between two speakers A and B is represented as $D = \{u_1^A, u_1^B, u_2^A, u_2^B, \dots, u_n^A, u_n^B\}$, where u_i^A and u_i^B represent the respective utterances of each speaker in the conversation, and n indicates the number of utterances. Accordingly, assume that each speaker has a set of persona information $P_A = \{p_1^A, \dots, p_{m_A}^A\}$ and $P_B = \{p_1^B, \dots, p_{m_B}^B\}$, where p_i^A and p_i^B represent the persona sentences for each speaker, respectively. Our pioneer task is to infer a speaker’s persona information based on their utterances in the conversation (e.g., inferring P_A from $U_A = \{u_1^A, u_2^A, \dots, u_n^A\}$).

As mentioned, we adopt the ESConv dataset (Liu et al., 2021) to train our model for provid-

ing emotional support. Assume that a conversation between a support seeker A and supporter B at the t^{th} turn of the conversation is $D = \{u_1^A, u_1^B, u_2^A, u_2^B, \dots, u_t^A\}$, where u_i^A and u_i^B represent the utterances of the seeker and the supporter, respectively. Our task is two-fold: First, we infer the seeker’s persona information P_A from their utterances $U_A = \{u_1^A, u_2^A, \dots, u_t^A\}$. Accordingly, we leverage the inferred information P_A and conversation history D to generate an appropriate supportive response u_t^B .

3.2 Persona Extractor

As previously stated, it is beneficial and essential to study the effects of leveraging persona information in the emotional support task. As predicting the seeker’s persona information before the conversation is impractical, inferring such information from their utterances is necessary.

Based on the problem formulation in §3.1, we fine-tune a `bart-large-cnn`¹ to augment the ESConv (Liu et al., 2021) dataset with the inferred persona information annotations for each turn of the conversations. More details can be found in Appendix A. Since the initial utterances of this dataset generally contain greetings, we annotate the persona information starting from the third utterance of the conversation. Table 1 shows an example of such annotations. We refer to this dataset with the additional annotations as Personalized Emotional Support Conversation (PESConv).

We analyze PESConv to confirm that modeling persona is essential for emotional support. In the original ESConv dataset, workers score conversations based on the supporter’s empathy level, the relevance between the conversation topic and the supporter’s responses, and the intensity of the seeker’s emotion. For each of these three aspects, we calculate the average cosine similarity between the responses and persona information in a conversation to examine how closely the responses and persona information are related.

For this task, we leverage SimCSE (Gao et al., 2021b), a sentence embedding model trained with a contrastive learning approach, to obtain vector representations for the sentences in PESConv. As illustrated in Figure 3, clearer and more appropriate mentions of the seekers’ persona in the supporters’ response lead to higher values for the studied as-

pects (i.e. higher empathy, more relevance, and a larger decrease in emotional intensity). Therefore, we believe this further highlights the necessity of modeling persona information in providing effective emotional support. Moreover, we use fastText (Joulin et al., 2017), which represents sentences as averaged word embeddings, and the results (Appendix B) demonstrate similar findings.

3.3 Modeling Seekers’ Persona

As illustrated in Figure 2, our model considers persona information as the model input in addition to the dialogue history. Formally, we use Transformer (Vaswani et al., 2017) encoders to obtain the inputs’ hidden representations, which can be expressed as

$$\begin{aligned} \mathbf{H}_D &= \mathbf{Enc}(u_1, \text{SEP}, u_2, \dots, u_n) \\ \mathbf{H}_P &= \mathbf{Enc}(p_1, \text{SEP}, p_2, \dots, p_m), \end{aligned} \quad (1)$$

where \mathbf{Enc} is the Transformer encoder, and m and n represent the number of persona sentences and conversation utterances, respectively. We use the special token SEP for sentence separation.

To highlight the context related to seekers’ persona, we calculate an extra attention \mathbf{Z}_D on \mathbf{H}_D and obtain a new hidden representation $\hat{\mathbf{H}}_D$ for dialogue history as follows:

$$\begin{aligned} \mathbf{Z}_D &= \text{softmax}(\mathbf{H}_D \cdot \mathbf{H}_P^T) \cdot \mathbf{H}_P \\ \hat{\mathbf{H}}_D &= \text{LN}(\mathbf{H}_D + \mathbf{Z}_D) \end{aligned} \quad (2)$$

where LN stands for the LayerNorm operation (Ba et al., 2016). Similarly, to promote persona sentences that are more aligned with the provided context, we obtain $\hat{\mathbf{H}}_P$ by

$$\begin{aligned} \mathbf{Z}_P &= \text{softmax}(\mathbf{H}_P \cdot \mathbf{H}_D^T) \cdot \mathbf{H}_D \\ \hat{\mathbf{H}}_P &= \text{LN}(\mathbf{H}_P + \mathbf{Z}_P). \end{aligned} \quad (3)$$

This also enables us to neglect the inferred persona sentences that are incorrect or irrelevant to the dialogue history. Since we cannot guarantee that inferred persona information is complete, we calculate the weighted sum of $\hat{\mathbf{H}}_D$, $\hat{\mathbf{H}}_P$ and \mathbf{H}_D to obtain the final hidden states as the decoder’s input as follows:

$$\begin{aligned} \mathbf{H}_{final} &= \lambda_1 \cdot \hat{\mathbf{H}}_D + \lambda_2 \cdot \hat{\mathbf{H}}_P + \lambda_3 \cdot \mathbf{H}_D \\ \lambda_i &= \frac{e^{w_i}}{\sum_j e^{w_j}} (i, j \in \{1, 2, 3\}), \end{aligned} \quad (4)$$

where w_1, w_2, w_3 are additional model parameters with the same initial value. This ensures that the

¹<https://huggingface.co/facebook/bart-large-cnn>

Conversation	Persona
Seeker: Hello	—
Supporter: Hi there! How may I support you today?	—
Seeker: I’m just feeling anxious about my job’s future. A lot of my colleagues are having trouble getting their licenses because of covid which means we won’t be able to work.	I am worried about my job’s future.
Supporter: That must be hard. COVID has turned our world upside down! What type of occupation are you in?	I am worried about my job’s future.
Seeker: I’m studying to be a pharmacist.	I am worried about my job’s future. I’m studying to be a pharmacist.

Table 1: An example conversation from PESConv. This conversation contains 5 utterances, where "—" indicates that no persona information was found. Once detected, new inferences are added to the seekers’ persona.

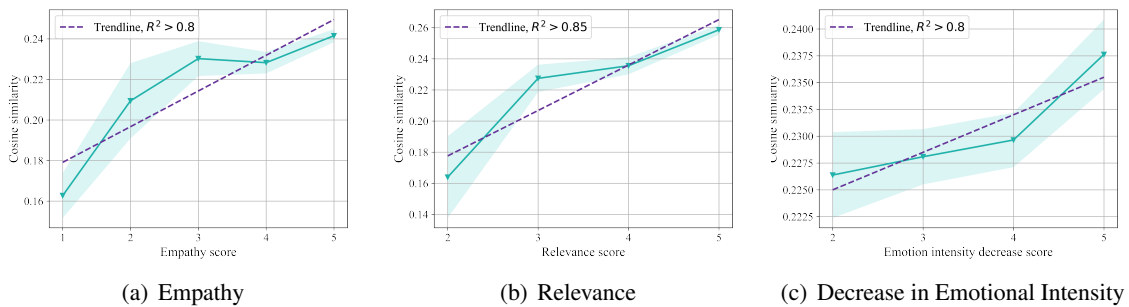


Figure 3: The relationship between the empathy score, relevance score, emotion intensity decrease score, and the similarity between supporters’ responses and persona information using SimCSE. We can observe that in general, more similarity leads to higher scores. In addition, we display the trend line and the coefficient of determination.

essence of the original dialogue context is largely preserved.

Similar to (Liu et al., 2021), we use special tokens to represent strategies and append them in front of the corresponding sentences. Our training objective can be formalized as:

$$\hat{r} = s \oplus r$$

$$\mathcal{L} = -\frac{1}{N} \sum_{t=1}^N \log P(\hat{r}_t | d, p, \hat{r}_{<t}) \quad (5)$$

where s stands for the strategy, r for the response, and N is the length of \hat{r} .

3.4 Strategy-based Controllable Generation

Supporters’ responses in the emotional support task are annotated based on several support strategies, which are essential for providing effective support (Liu et al., 2021). For instance, the supporter may choose to ask a *Question* or provide statements of *Reaffirmation and Confirmation* depending on the situation. We provide more descriptions of these strategies in Appendix C. Accordingly, it becomes

intuitive that selecting different strategies corresponds to the available knowledge of the users’ persona, demonstrating the importance of strategy selection in our proposed approach. For instance, supporters could choose *Providing Suggestions* if they have sufficient knowledge of the user’s persona and situation, while they would resort to *Question* if they lack such information. Therefore, we propose an innovative strategy-based controllable generation method for the decoding phase. We decompose the generation probability into

$$P_{final}(r_t | r_{<t}, d, p) \propto P(r_t | r_{<t}, d, p) \cdot \left(\frac{P(r_t | r_{<t}, d, p)}{P(r_t | r_{<t}, d)} \right)^\alpha \quad (6)$$

where α is the hyperparameter associated with the strategy, and d and p represent the dialogue history and persona, respectively. Both $P(r_t | r_{<t}, d, p)$ and $P(r_t | r_{<t}, d)$ are calculated by our model; the only difference is that persona is not included in calculating $P(r_t | r_{<t}, d)$. The last term in this equation can be interpreted as the ratio of the probability

Strategy	α	Category
Question	0	low
Restatement or Paraphrasing	0.75	high
Reflection of Feelings	0	low
Self-disclosure	0	low
Affirmation and Reassurance	0.75	high
Providing Suggestions	0.75	high
Information	0.75	high
Others	0.375	medium

Table 2: The values and levels of α corresponding to different strategies.

of a token whether the persona is entered or not. As the ratio increases, the token becomes more relevant to persona information, increasing the likelihood of generating the token after adding such persona information. Therefore, employing Eq.6 increases the likelihood of more relevant tokens to the persona information. α is set to different values depending on the strategy. The values used by all strategies are listed in Table 2.

We investigate the values of α corresponding to different strategies and define three categories: high, medium, and low, which correspond to 0.75, 0.375, and 0, respectively. More details about the tuning process of these values are discussed in Appendix D.

We provide explanations for two of our decided α values. For effective support, there are two types of questions (*Question* strategy) that can be asked from the seeker (Ivey et al., 2013): open and closed. Therefore, we choose the low level to avoid overthinking persona information, resulting in fewer open questions. We chose the high level for the *Providing Suggestions* strategy, as we needed to focus more on the persona information to provide more appropriate and specific suggestions. See Appendix E for explanations regarding the α of other strategies.

4 Experiments

4.1 Persona Extractor Evaluation

Human Evaluation To validate the effectiveness of our persona extractor model, we first manually reviewed several inferences and discovered that the main errors could be categorized as contradictions (i.e., personas contain factual errors) or hallucinations (i.e., personas contain unreasonable and irrelevant deductions from the conversation). An example of contradictions would be if the seeker

mentions in the conversation that he is a man, but the inferred persona is "I am a woman". Moreover, an instance of hallucination errors would be if the inferred persona is "I am a plumber" when the seeker has not mentioned their occupation. Then, we chose 100 samples at random and hired workers on Amazon Mechanical Turk (AMT) to annotate each sample with one of the following four options: Reasonable, Contradictory, Hallucinatory, or Others. In addition, if the option Others was chosen, we asked workers to elaborate on the error. The annotators considered 87.3% of the inferred persona samples as Reasonable while marking 8% and 4% of the samples as Contradictory and Hallucinatory, respectively. Moreover, only 0.667% of the samples were marked as Others. However, upon further analysis, we found that such samples could also be classified in one of the mentioned error categories (see Appendix F for more details). The inter-annotator agreement, measured by Fleiss’s kappa, was 0.458, indicating moderate agreement.

4.2 Baselines

Blenderbot-Joint (Liu et al., 2021): Blenderbot (Roller et al., 2021) fine-tuned on the ESConv dataset. This model is trained to predict the correct strategy for the next response via the language modeling objective. In addition, this model can also be seen as PAL trained without incorporating persona.

MISC (Tu et al., 2022): the state-of-the-art (SOTA) on the ESConv benchmark, which leverages commonsense reasoning to better understand the seeker’s emotions and implements a mixture of strategies to craft more supportive responses.

Hard Prompt: this model employs a straightforward idea when modeling seekers’ persona information in the emotional support task, in which persona information is concatenated to the dialogue history. That is, the input to the model would be in the form "*Persona: {persona} \n Dialogue history: {context} \n Response: "*

4.3 Implementation Details

We conducted the experiments on PESConv and use a 7:2:1 ratio to split this dataset into the train, validation, and test sets. As Liu et al. (2021) stated, Blenderbot (Roller et al., 2021) outperforms DialoGPT (Zhang et al., 2020) in this task. Therefore, similar to previous work (Liu et al., 2021; Tu

Model	ACC \uparrow	PPL \downarrow	B-2 \uparrow	B-4 \uparrow	D-1 \uparrow	D-2 \uparrow	E-1 \uparrow	E-2 \uparrow	R-L \uparrow	Cos-Sim \uparrow
Blenderbot-Joint	27.72	18.11	5.57	1.93	3.74	20.66	4.23	20.28	16.36	0.184
MISC	31.34	16.28	6.60	1.99	4.53	19.75	5.69	30.76	17.21	0.187
Hard Prompt	34.24	17.06	7.57	2.53	5.15	25.47	6.02	31.64	18.12	0.199
PAL ($\alpha = 0$)	34.25	15.92	9.28	2.90	4.72	25.56	5.87	33.05	18.27	0.229
PAL	34.51	15.92	8.75	2.66	5.00	30.27	6.73	41.82	18.06	0.244

Table 3: The results of automatic metrics evaluation for each model on ESConv. PAL ($\alpha = 0$) represents setting the α of each strategy to 0, thus neglecting our proposed controllable generation decoding method.

PAL vs.	Blenderbot-Joint			MISC			PAL ($\alpha = 0$)		
	Win	Lose	Draw	Win	Lose	Draw	Win	Lose	Draw
Coherence	68\ddagger	26	6	54\dagger	34	12	46	48	6
Identification	42	44	14	46	42	12	58\ddagger	32	10
Comforting	50\ddagger	32	18	62\ddagger	24	14	44	42	14
Suggestion	54\ddagger	32	14	42	42	16	46	38	16
Information	44\dagger	34	22	62\ddagger	22	16	52	44	4
Overall	52\ddagger	16	32	44\ddagger	28	28	40\ddagger	28	32

Table 4: The results of the human interaction evaluation (%). PAL performs better than all other models (sign test, \ddagger / \dagger represent p -value < 0.05 / 0.1).

et al., 2022), we used the 90M version of Blenderbot (Roller et al., 2021). Moreover, we used the AdamW (Loshchilov and Hutter, 2018) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We initialized the learning rate as $2.5e-5$ and performed a 100-step linear warmup. The training and validation batch sizes were set to 4 and 16, respectively. The model was trained for 10 epochs, and we chose the checkpoint with the lowest loss on the validation set. During the decoding phase, we used both Top- k and Top- p sampling with $k = 10$, $p = 0.9$, with temperature and the repetition penalty set to 0.5 and 1.03, respectively. The experiments were run on a single Quadro RTX 6000 GPU using the transformers library² (Wolf et al., 2020).

4.4 Automatic Evaluation

We adopted strategy prediction accuracy (ACC), perplexity (PPL), BLEU-n (B-n) (Papineni et al., 2002), Distinct-n (D-n) (Li et al., 2016a), EAD-n (E-n) (Liu et al., 2022), Rouge-L (R-L) (Lin, 2004), and the mean of the cosine similarity between supporters’ responses and personas using the SimCSE (Gao et al., 2021b) representation (cos-sim) to automatically evaluate our model’s performance. In addition, since the responses in this task are often long, we also leveraged the Expectancy-Adjusted Distinct (EAD) score to evaluate response diversity as the Distinct score has been shown to be biased

towards longer sentences (Liu et al., 2022). To calculate this score, rather than dividing the number of unique n-grams by the total number of n-grams, as done in the original Distinct score, we would use the model’s vocabulary size as the denominator.

As shown in Table 3, PAL outperforms all baselines in automatic metrics, including the current SOTA model MISC. As Blenderbot-Joint can be perceived as PAL without persona employed in training, the significance of persona can be demonstrated through the comparison of the results achieved by PAL and PAL ($\alpha = 0$) with Blenderbot-Joint. In addition, compared to PAL ($\alpha = 0$), PAL demonstrates a more balanced performance and has the best strategy prediction accuracy, diversity, and better alignment with persona information, which indicates more seeker-specific responses. Interestingly, the cos-sim value for PAL is comparable to the mean value of the dialogues with an empathy score of 5 in Figure 3(a). Through further comparing the performance of PAL and PAL ($\alpha = 0$), we can see that our strategy-based decoding approach significantly improves the dialogue diversity, as shown by D-n and E-n, which are more important metrics for dialogue systems than B-n and R-L (Liu et al., 2016; Gupta et al., 2019; Liu et al., 2022).

In Figure 4, we show the accuracy of the top-n strategy prediction results and our model PAL has the best results. It is worth noting that all models

²<https://github.com/huggingface/transformers>

Situation	
Seeker	I have just cheated on my girlfriend. I feel very guilty about it.
Dialogue history	
Seeker	Hi, my friend.
Supporter	Hello ! How are you doing?
Seeker	Feeling very shame.

Seeker	But till now my girlfriend don't know about it. But her mom is now targeting me for her sexual desire.
Persona Information	
Seeker	I am feeling ashamed.
Seeker	I have cheated on my girlfriend with her mother.

Response	
Blenderbot-Joint	I understand, I know how you feel. (<i>Poor Empathy</i>)
MISC	I think you will be fine. (<i>Poor Empathy</i>)
Hard Prompt	Oh no, I am so sorry, that is not good. (<i>Poor Empathy</i>)
PAL ($\alpha = 0$)	I understand it is hard, so now you have to forgive her. (<i>Less Proper Suggestion</i>)
PAL	I understand how that can be hard. I would suggest you to talk to her mother, tell her that you feel ashamed about it and don't cheat on your girlfriend again . (<i>Strong Empathy</i>)
<i>Ground-truth</i>	You have got such a nice girlfriend, have a happy life with her.

Table 5: Responses from our approach and others. Due to space constraints, we have omitted some sentences.

with persona information, PAL, PAL ($\alpha = 0$), and Hard Prompt, all outperform MISC, demonstrating the importance of seekers' persona and highlighting the need for further research into how to better leverage such information in addition to common-sense reasoning.

4.5 Human Evaluation

We acknowledge that automatic metrics are insufficient for empirically evaluating and highlighting

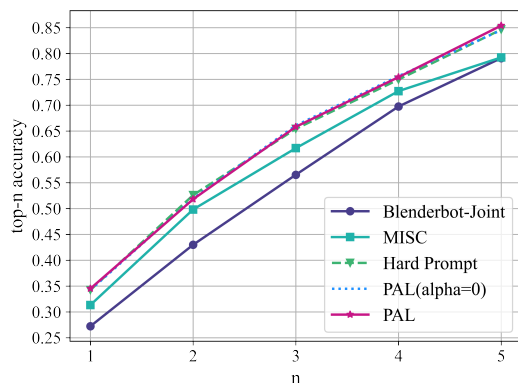


Figure 4: The top-n strategy prediction accuracy.

the improvements of our proposed method. Hence, following Liu et al. (2021), we also conducted human evaluation by recruiting crowd-sourcing workers that interacted with the models. We provided workers with a scenario and asked them to act as seekers in those situations. Each worker must interact with two different models and score them in terms of (1) Coherence; (2) Identification; (3) Comforting; (4) Suggestion; (5) Informativeness; and (6) Overall Preference. Detailed explanations for each aspect can be found in Appendix F.

As shown in Table 4, we compare PAL with the other three models, and PAL beats or is competitive with other methods on all of the above metrics. It performs well on three key metrics more closely aligned with persona (i.e., Comforting, Suggestion, and Information), implying that persona is required in emotional support.

5 Case Study

In Table 5, we provide an example to compare the responses of our approach with the other methods. As can be seen, the Blenderbot-Joint, MISC, and Hard Prompt methods all provide only very poor empathy, with responses that are very general and do not contain much information. Whereas PAL

($\alpha = 0$), which does not use the strategy-based decoding method, is more specific but provides a less appropriate suggestion. Our model PAL shows strong empathy, is the most specific while providing appropriate suggestions, and incorporates persona information in the response (*feel ashamed* and *don't cheat on your girlfriend again*). Due to space constraints, more cases, including cases of interactions and analysis over different strategies, can be found in Appendix G.

6 Conclusion

In this work, we introduced persona information into the emotional support task. We proposed a framework that can dynamically capture seekers' persona information, infer persona information using our trained persona extractor, and generate responses with a strategy-based controllable generation method. Through extensive experiments, we demonstrated that our proposed approach outperformed the studied baselines in both human and manual evaluation. In addition, we provided persona annotations for the ESConv dataset using the persona extractor model, which will foster the research of personalized emotional support conversations.

Limitations

Persona extractor First, we need to clarify that our definition of persona is not exactly psychological, the role an individual plays in life (Jung, 2013). As a result, like previous studies (e.g., Persona-Chat (Zhang et al., 2018), PEC (Zhong et al., 2020)), the format of persona is flexible and variable. As stated in §4.1, there are still some issues with the model we use to infer persona information. For example, we sometimes get information that contradicts the facts. And also, there is occasionally unrelated content, as with common-sense reasoning (Tu et al., 2022). Furthermore, we cannot guarantee that we can infer all of the persona information that appears in the conversation because much of it is frequently obscure. And when extracting persona information, we only use what the user said previously and remove what the bot said, which results in the loss of some conversation information. The reason for this is that we have discovered that if we use the entire conversation, the model frequently has difficulty distinguishing which persona information belongs to the user and which belongs to the other party. In addition, since

the code of Xu et al. (2022) is not yet available, we have not compared other methods of extracting persona dynamically from the conversation.

Strategy-based decoding During the decoding phase, we only coarse-grained the α of each strategy because we discovered that only coarse-grained tuning produced good results, and future work may be able to further explore the deeper relationship between different strategies and persona.

Ethical Considerations

In this work, we leveraged two publicly available datasets. First, we used the Persona-Chat dataset, which is collected by assigning a set of fixed predefined persona sentences to workers. Therefore, by participating in this dataset, workers were required not to disclose any personal information (Zhang et al., 2018), which prevents issues regarding the leakage of their privacy. Similarly, during the collection of the ESConv dataset, participants were asked to create imaginary situations and play the role of a support seeker who is in that situation. In addition, they were instructed not to provide personal information during their conversations with the trained supporters (Liu et al., 2021). Regarding the persona extractor, this module is trained to infer and extract persona information solely from what the user has mentioned in the conversation rather than making assumptions about the user's background and character, further highlighting the importance of user privacy in our research.

Regarding our experiments, we ensured that all workers agreed to participate in the annotation tasks. Moreover, as the workers were recruited from the US, we ensured that they were paid above the minimum wage in this country for successfully completing our tasks. We acknowledge that using trained dialogue models to provide support is a sensitive subject and research on this topic should be conducted with sufficient precautions and supervision. We also acknowledge that in their current stage, such models cannot replace human supporters for this task (Sabour et al., 2022a). Thus, they should not be employed to replace professional counselors and intervention and interact with users that suffer from mental distress, such as depression or suicidal thoughts.

References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O’Neill, Cherie Armour, and Michael McTear. 2018. Assessing the usability of a chatbot for mental health care. In *International Conference on Internet Science*, pages 121–132. Springer.
- Joana Campos, James Kennedy, and Jill F Lehman. 2018. Challenges in exploiting conversational memory in human-agent interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1649–1657.
- Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *arXiv preprint arXiv:2210.04242*.
- Patricio Costa, Raquel Alves, Isabel Neto, Pedro Marvao, Miguel Portela, and Manuel Joao Costa. 2014. Associations between medical student empathy and personality: a multi-institutional study. *PloS one*, 9(3):e89254.
- Walter Cullen, Gautam Gulati, and Brendan D Kelly. 2020. Mental health in the covid-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312.
- Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. 2020. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in digital health*, 2:576361.
- Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.
- Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021a. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.
- Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. *Teaching machines to read and comprehend*. In *Advances in Neural Information Processing Systems (NIPS)*.
- Md Mahbub Hossain, Samia Tasnim, Abida Sultana, Farah Faizah, Hoimonty Mazumder, Liye Zou, E Lisako J McKyer, Helal Uddin Ahmed, and Ping Ma. 2020. Epidemiology of mental health problems in covid-19: a review. *F1000Research*, 9.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. *Challenges in building intelligent open-domain dialog systems*. *ACM Trans. Inf. Syst.*, 38(3).
- Allen E Ivey, Mary Bradford Ivey, and Carlos P Zalaquett. 2013. *Intentional interviewing and counseling: Facilitating client development in a multicultural society*. Cengage Learning.
- Bertus F. Jeronimus and Odilia M. Laceulle. 2017. *Frustration*, pages 1–5. Springer International Publishing, Cham.
- Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, page 427.
- Carl Jung. 2013. *Psychological types*. Important Books.
- Alan E. Kazdin and Stacey L. Blase. 2011. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*.
- Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240.
- Matthias Kraus, Philip Seldschopf, and Wolfgang Minker. 2021. Towards the development of a trustworthy chatbot for mental health applications. In *International Conference on Multimedia Modeling*, pages 354–366. Springer.
- Anant Kumar and K Rajasekharan Nayar. 2021. Covid 19 and its mental health consequences.
- Mark R Leary and Ashley Batts Allen. 2011. Personality and persona: Personality processes in self-presentation. *Journal of personality*, 79(6):1191–1218.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016a. 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.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.
- Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427.
- Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.
- 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)*, pages 3469–3483.
- Yangyang Liu, Nathan A Gillespie, Lin Ye, Gu Zhu, David L Duffy, and Nicholas G Martin. 2018. The relationship between personality and somatic and psychological distress: A comparison of chinese and australian adolescents. *Behavior Genetics*, 48(4):315–322.
- Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.
- Estelle Michinov and Nicolas Michinov. 2021. Stay at home! when personality profiles influence mental health and creativity during the covid-19 lockdown. *Current Psychology*, pages 1–12.
- Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.
- Evan Peterson. 2021. Wisconsin mental health professional shortage amid covid. *FOX6 News Milwaukee*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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.
- Carl R Rogers. 2013. Client-centered therapy. *Curr Psychother*, pages 95–150.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.
- Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2022a. [Chatbots for mental health support: Exploring the impact of emohaa on reducing mental distress in china.](#)
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022b. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.
- Maja Smrdu, Ana Kuder, Eva Turk, Tatjana Čelik, Janko Šet, and Simona Kralj-Fišer. 2021. Covid-19 pandemic and lockdown: associations with personality and stress components. *Psychological Reports*, page 00332941211043451.
- Dalila Talevi, Valentina Socci, Margherita Carai, Giulia Carnaghi, Serena Faleri, Edoardo Trebbi, Arianna di Bernardo, Francesco Capelli, and Francesca Pacitti. 2020. Mental health outcomes of the covid-19 pandemic. *Rivista di psichiatria*, 55(3):137–144.

- Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.
- Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.
- Lee Jing Yang, Lee Kong Aik, and Gan Woon Seng. 2021. Generating personalized dialogue via multi-task meta-learning. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.
- 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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B 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.
- Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. [Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models](#).
- Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Persona Extractor

In our initial experiments, we compare the effectiveness of various generative models to infer persona (such as GPT2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020), BART (Lewis et al., 2020)). We manually checked some results and found the best results were obtained by the Bart model fine-tuned on CNN Daily Mail (Hermann et al., 2015). We trained this model for ten epochs with a batch size of 4 and learning rate of 1e-5, and selected the best-performing checkpoint.

B Relevance of Individualization and Seeker Evaluation

Here we show the results produced by fastText in Figure 5.

C Helping Strategies in ESConv

A total of 8 strategies are marked in ESConv, and they are basically evenly distributed (Liu et al., 2021). Here we list these strategies and their detailed definitions, which are directly adopted from Liu et al. (2021).

Question Asking for information related to the problem to help the help-seeker articulate the issues that they face. Open-ended questions are best, and closed questions can be used to get specific information.

Restatement or Paraphrasing A simple, more concise rephrasing of the help-seekers’ statements could help them see their situation more clearly.

Reflection of Feelings Articulate and describe the help-seekers’ feelings.

Self-disclosure Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.

Affirmation and Reassurance Affirm the help-seeker’s strengths, motivation, and capabilities and provide reassurance and encouragement.

Providing Suggestions Provide suggestions about how to change but be careful not to overstep and tell them what to do.

Information Provide useful information to the help-seeker, for example, with data, facts, opinions, resources, or by answering questions.

Others Exchange pleasantries and use other support strategies that do not fall into the above categories.

D Tuning Process of the α Values

We first tried to set these alpha values as trainable parameters, but we found that the values changed very little during the training of the model and therefore depended heavily on the initialization, so we set these alpha’s as hyperparameters. Then, these values were obtained upon numerous attempts on the validation set as they enabled the model to have a balanced performance based on the automatic evaluation. We acknowledge that this tuning process is trivial and coarse-grained. We leave approaches to improve this process, such as using a simulated annealing algorithm, to future work.

E Analysis of α Selected for Different Strategies

In §3.4, we analyzed the strategies *Question* and *Providing Suggestions*. And the rest of the strategies are analyzed below.

For the *Restatement or Paraphrasing* strategy, it is necessary to repeat the words of the seeker, so a more specific restatement can help the seeker better understand himself. For the *Reflection of Feelings* strategy, since the focus is more on feelings, and the extracted persona information is more fact-related, we set low for this strategy. For the *Self-disclosure* strategy, it is more about the supporter’s own experience and should not focus too much on the persona information of the seeker, which may lead to unnecessary errors, so we set this strategy to low. For the *Affirmation and Reassurance* strategy, combining the seeker’s persona information can often provide more specific encouragement and bring the seeker a better experience, so we set it to high. For the *Information* strategy, we need to consider

more persona information in order to provide more appropriate and specific information for seekers, so we set it high. For the *Other* strategy, the main places this appear are greeting and thanking. About this strategy, considering that most appearances are in greeting and thanking, if we can combine more seeker characteristics may make seekers feel more relaxed, we set it to the high level at first, but careful observation found that *Other* strategies are used when the other strategies are not appropriate. Although such cases are rare, in order to avoid unnecessary errors, we set it to medium.

F Human Evaluation

Here we show the guidelines for two human evaluation experiments in Figure 6 and Figure 7. For the persona extractor manual evaluation experiment, we pay \$0.05 for one piece of data, and for the human interactive evaluation, we pay \$0.10 for one piece of data, with the price adjusted for the average time it takes workers to complete the task. We stated in the task description that this is an evaluation task, so for the data submitted by the workers, we only use it for evaluations.

G Case Study

Due to space limitations, we show more examples here, these are cherry-picked.

In Figure 8, we show an interactive case. It can be seen that PAL uses the extracted persona appropriately several times in the conversation and gives the seeker specific advice.

In Figure 9, we show some cases in the ES-Conv dataset. Interestingly, in these examples, PAL sometimes performs better than Ground-truth, giving a more appropriate and specific response rather than a general one, which also proves the superiority of our model.

Here, we also compare our model with baselines over different strategies. In Table 6, we show a case of the strategy *Providing Suggestions*. We can find that our model provides the most specific suggestions. In Table 7, we show a case of the strategy *Affirmation and Reassurance*. We can also see that PAL’s response is the most specific.

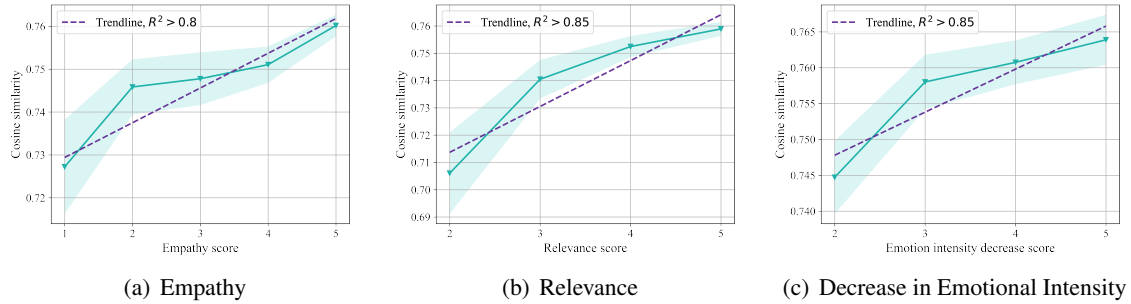



Figure 5: The relationship between the empathy score, relevance score, emotion intensity decrease score, and the similarity between supporters’ responses and persona information using fastText. We can observe that in general, more similarity leads to higher scores. In addition, we display the trend line and the coefficient of determination.

Situation	
Seeker	My two friends are shaming me for not donating more per year to charities. One of them is the President of a local non-profit.
Dialogue history	

Seeker	What should I do about these two (so-called) friends?
Persona Information	
Seeker	My friends are putting me on a guilt trip and shaming me.
Seeker	I am not making much income with covid19.
Response	
Blenderbot-Joint	Well, they are not worth the friendship.
MISC	Do you think that you can talk to them about how you feel?
Hard Prompt	Maybe you should try to talk to them about it and see what they say about it.
PAL($\alpha = 0$)	If you are hurt because of them then I would suggest you seek professional help.
PAL	If they are hurting you physically then I would suggest that you find a way to talk to them and explain your income situation .
Ground-truth	Definitely not ...have you tried explaining your situation to them?

Table 6: Responses from our approach and others.

 **Previewing Answers Submitted by Workers**
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Persona information judgement

Task Description

This is a task to information about an individual's persona based on what they have said in a conversation. You are provided with a conversation snippet between two people, a seeker and a supporter, and several sentences about the **seeker's persona**. You need to judge whether the shown sentences about the seeker's persona are correct. If not, elaborate why it is incorrect.

Notes

- Some conversations are really long. You don't need to read them completely. In many cases, you only need to judge by **keywords**

ID: \${id}

Conversation: \${conv}

Persona information: \${persona}

Do you think the inferred persona information are correct?

- Yes, totally correct.
- No, they contain factual errors, such as **contradiction** between the persona information and what has been said in the conversation. For example, if the seeker states "I am a man" and the persona is "I am a woman", the persona is contradicting what the seeker said.
- No, they contain illusions, such as **unreasonable and irrelevant** deductions from the conversation. For example, if the extracted persona is "I am a plumber" when there are no mentions of their job by the seeker.
- No, other errors

If you have selected 'No, other errors', please enter the reason

Submit

Figure 6: The screenshot of persona extractor human evaluation guideline.

Dialog Response Evaluation

Task Description

This is an emotional support (using various strategies to care for the seeker) task, each task consists of a situation, assuming that **you are the seeker (seeking emotional support)** and you are **experiencing the situation**, you need to have a conversation with system A and system B, and then you need to choose which system works better based on:

- 1. Coherence**
Which bot's responses were more **coherent** in content and **relevant** to the context?
- 2. Identification**
which bot explored your situation more in depth and was more **helpful in identifying** your problems?
- 3. Comforting**
which bot was more skillful in comforting you?
- 4. Suggestion**
which bot gave you more **helpful suggestions** for your problems?
- 5. Information**
which bot's responses were more **specific** and contained more information?
- 6. Overall**
generally, which bot's emotional support do you prefer?

Notes

- If you think the two responses are almost as good or as bad, feel free to choose **Tie**. But we still you encourage you to select the better one.
- Here is an example conversation:
Situation:My mother is ill and I'm worried about her.
Conversation:
user: Hello
bot: Hello, what would you like to talk about?
user: I feel depressed recently
bot: i am so sorry to hear that. can you tell me what's going on?
user: My mother is ill
bot: i am so sorry to hear that. it must be difficult for you and your mother
user: Yeah, I'm so worried about her, I can't even work
bot: it is very hard to be worried about someone you care about. i can imagine how stressed you are about it
user: Good advice, but I still need to work. I am a policeman, my work is really important
bot: i know that it is very difficult to see your mom every day. maybe you could think about how you want to take care of her during the holidays, maybe you could make her take a walk outside with you?
user: Good idea! Thank you very much
bot: you are very welcome. it is very nice to meet someone who cares about their mom!

ID: \${id}

Situation: \${s}

Model link: <http://115.182.62.193:5000/>

Please be sure to chat with the two models seriously. We will check the chat records and reject the workers who are not serious.

Choose the better model in terms of:

Coherence:	Identification:	Comforting:	Suggestion:	Information:	Overall:
<input type="radio"/> A	<input type="radio"/> A	<input type="radio"/> A	<input type="radio"/> A	<input type="radio"/> A	<input type="radio"/> A
<input type="radio"/> B	<input type="radio"/> B	<input type="radio"/> B	<input type="radio"/> B	<input type="radio"/> B	<input type="radio"/> B
<input type="radio"/> Tie	<input type="radio"/> Tie	<input type="radio"/> Tie	<input type="radio"/> Tie	<input type="radio"/> Tie	<input type="radio"/> Tie

Submit

Figure 7: The screenshot of human interaction evaluation guideline.

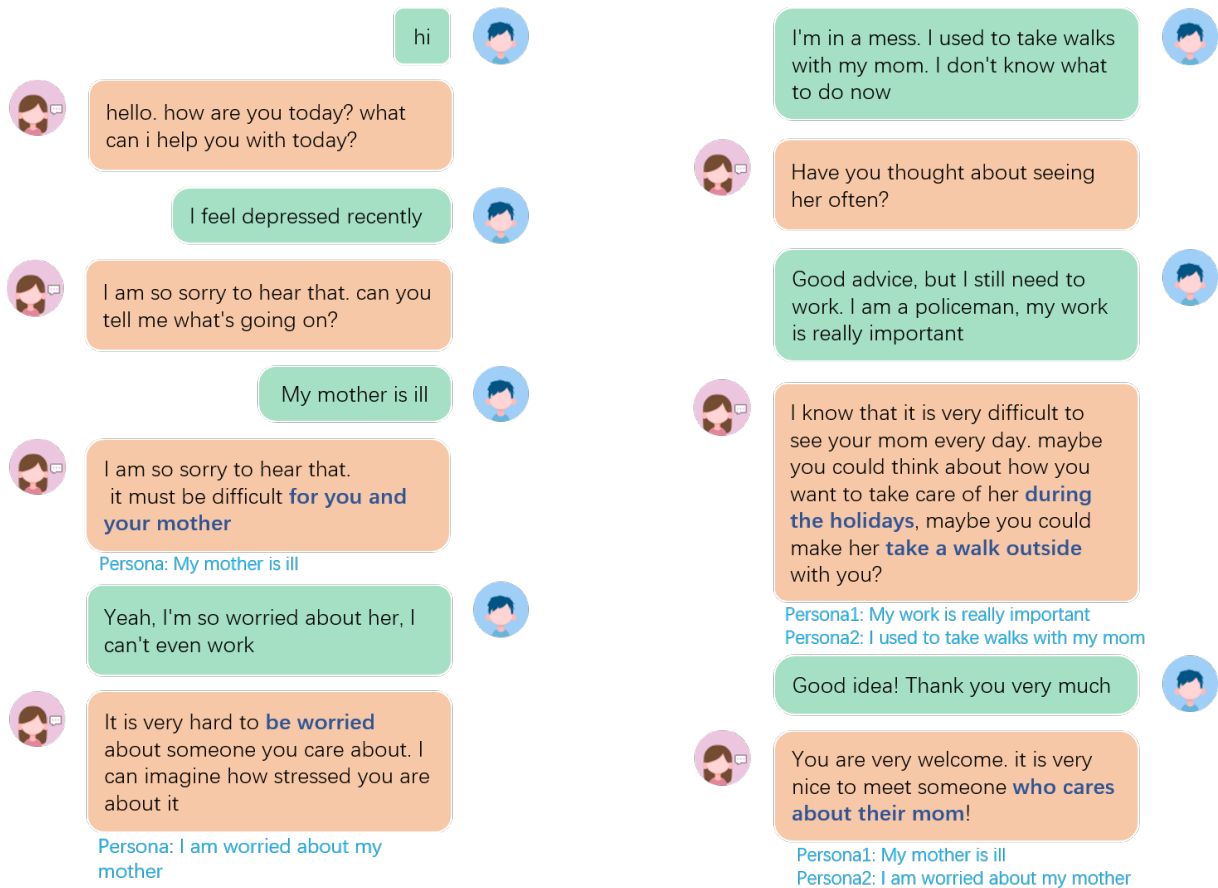


Figure 8: An interactive example. In this case, the orange dialog box is the supporter (PAL) and the green dialog box is the seeker. Due to the long dialogue rounds, it is split into left and right sides, with the dialogue on the left side going first. The persona used by PAL is given below the dialog box.

Situation	
Seeker	I caught my spouse cheating and broke up with her.
Dialogue history	

Seeker	She got a job in another town and she utilize the opportunity to be sleeping around.
Persona Information	
Seeker	My wife is cheating on me.

Response	
Blenderbot-Joint	Wow... I am sorry to hear that.
MISC	I understand.. I had a similar situation in a previous relationship.
Hard Prompt	So you trust your spouse but you have disappointed your marriage vows.
PAL($\alpha = 0$)	I can see how that can be a problem. That is a very difficult situation to be in.
PAL	That's a difficult situation to be in. It sounds like you are being betrayed . I believe you deserve someone better.
<i>Ground-truth</i>	I can imagine how a break in trust has made you feel. I know a break in trust is always very difficult to process.

Table 7: Responses from our approach and others.



Figure 9: Some examples in the ESConv dataset where we do not show all rounds of dialogues due to space constraints. The orange dialogues are from supporters (PAL).

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Sec. "Limitations"
- A2. Did you discuss any potential risks of your work?
Sec. "Ethics Statement"
- A3. Do the abstract and introduction summarize the paper's main claims?
The paper's main claims are summarized in "Abstract"(Page1) and Sec.1 "Introduction"
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Sec.3 and Sec. 4

- B1. Did you cite the creators of artifacts you used?
Sec.3 and Sec. 4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Sec. "Ethics Statement"
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Sec. "Ethics Statement"
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Sec. "Ethics Statement"
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sec.1, Sec.3 and Sec. "Ethics Statement"
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sec.4

C Did you run computational experiments?

Sec.4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sec.4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sec.3, Sec.4 and Table 2
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sec.4
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. We propose the evaluation methods and also provide the source codes.
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Sec.4
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix F
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Sec 4, Appendix F and Sec. "Ethics Statement"
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix F
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.