Enhancing Emotional Support Conversations: A Framework for Dynamic Knowledge Filtering and Persona Extraction

Jiawang Hao[1,2,3*] **and Fang Kong**[1,2†]
[1]Laboratory for Natural Language Processing
[2]Soochow University School of Computer Science and Technology, Suzhou, China
[3]TAL Education Group, Beijing, China
20225227107@stu.suda.edu.cn, kongfang@suda.edu.cn

Abstract

With the growing need for accessible emotional support, conversational agents are being used more frequently to provide empathetic and meaningful interactions. However, many existing dialogue models struggle to interpret user context accurately due to irrelevant or misclassified knowledge, limiting their effectiveness in real-world scenarios. To address this, we propose a new framework that dynamically filters relevant commonsense knowledge and extracts personalized information to improve empathetic dialogue generation. We evaluate our framework on the ESConv dataset using extensive automatic and human experiments. The results show that our approach outperforms other models in metrics, demonstrating better coherence, emotional understanding, and response relevance.

1 Introduction

Empathy refers to the ability to perceive others' emotions, consider their perspectives, and respond appropriately. With the rapid advancement in the field of dialogue systems, the question of how to imbue machines with empathy has gained significant attention(Cameron et al., 2019; Daley et al., 2020; Denecke et al., 2020). At the same time, a growing number of people are experiencing mental health issues and seeking support. The cost of professional mental health care and counseling is high, and access is often limited (Olfson, 2016; Cullen et al., 2020; Vindegaard and Benros, 2020). This highlights the importance of using conversational agents and chatbots to automate these tasks (Denecke et al., 2020; Kraus et al., 2021).

To address this issue, Liu et al. (2021) introduced the Emotional Support Conversation (ESC) task, using neural network models to reduce users' emotional distress, improve their mood, and ultimately help resolve their problems. Leveraging external knowledge bases, MISC (Tu et al., 2022) innovatively integrates COMET (Bosselut et al., 2019) to enhance the model's capability in emotional reasoning. PAL (Cheng et al., 2023) addresses the need for personalized user responses by incorporating persona information into dialogue generation, which enhances the diversity of the replies.

However, not all introduced knowledge contributes to dialogue generation: some knowledge even interferes with dialogue generation. For example, in Figure 1, the intent knowledge "*have fun*" generated based on the user's current statement is clearly irrelevant. To clarify the contextual relevance of each piece of commonsense knowledge, we follow Gao et al. (2022) and categorize knowledge into four types: **RPA** (directly relevant), **RPP** (related to the past and present), **RPF** (related to the future), and **IRR** (irrelevant). Based on the

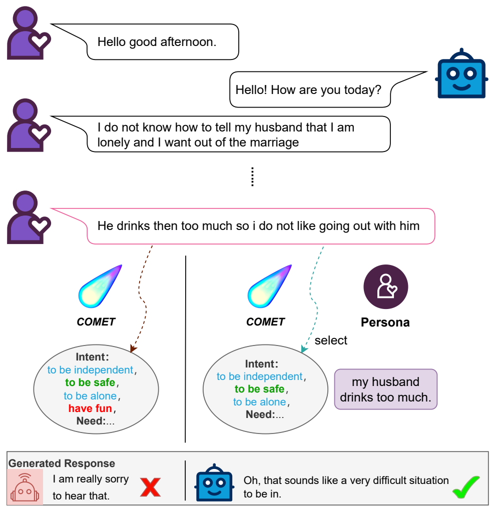

Figure 1: This example is drawn from the ESConv dataset. Red, blue, and green represent irrelevant knowledge, knowledge related to historical context, and knowledge strongly related to the current text, respectively.

* Part of this work was completed during an internship at TAL Education Group, Beijing, China.
† Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3193–3202
January 19–24, 2025. ©2025 Association for Computational Linguistics

four categories, we can classify the knowledge in Figure 1. Clearly, "*have fun*" falls under IRR, as it is irrelevant. "*To be safe*" is closely related to the current context and belongs to RPA. Meanwhile, "*to be independent*" and "*to be alone*" are strongly connected to the previous user intention "*out of the marriage*", and only weakly related to the current utterance, thus classified as RPP. Since our task does not have access to future text, we exclude the RPF category. By following these classifications, we can fully utilize the most relevant knowledge while excluding irrelevant information. The remaining commonsense knowledge closely aligns with the context, which enhances the consistency of generated responses but may reduce their diversity. So we incorporate the user's persona information as supplementary input into our generation process, providing rich background information and contextual clues to increase the diversity of responses. For example, in Figure 1, the persona information summarizes the user's current situation: "*My husband drank too much.*". Combined with the filtered user intention (as shown on the right in Figure 1), we can easily generate an appropriate response, as shown in Figure 1.

In this paper, we propose a method for dynamic knowledge filtering and persona extraction to generate empathetic support dialogues. Specifically, we use the *ComFact* dataset (Gao et al., 2022) to train a knowledge filter that selects relevant commonsense knowledge from COMET. Next, we encode the context, commonsense knowledge, and persona information separately. Through a bidirectional cross-attention mechanism, we integrate context with commonsense knowledge and context with persona information to generate the final response. We evaluate our method on the ESConv dataset. The results show that our approach outperforms current baseline models in both automatic and human evaluations.

Our contributions are as follows:

- We propose a method for dynamic knowledge filtering and persona extraction to generate high-quality responses.
- Both automatic and human evaluations indicate that our model demonstrates a deep understanding of the user's personality and situation compared to other SOTA methods.
- In the ablation study, we explore the impact of different methods for selecting commonsense knowledge on dialogue generation.

2 Related Works

2.1 Empathetic Dialogue Systems

Empathic dialogue systems are designed to recognize and understand user emotions and generate more meaningful interactions (Rashkin et al., 2019; Liu et al., 2021; Sabour et al., 2022). Most previous research focuses on detecting emotions to address the emotional aspect of empathy (Majumder et al., 2020; Roller, 2020; Li et al., 2022). However, there is growing awareness of the need to address the cognitive aspect of empathy. Knowledge graphs, especially those containing commonsense and emotional information, have been shown to produce better empathetic responses (Hwang et al., 2021; Gao et al., 2022). Sabour et al. (2022); Tu et al. (2022); Zhao et al. (2023); Cheng et al. (2023) have explored incorporating knowledge graphs to improve understanding and performance of empathy, but fully understanding the psychological state of a dialogue partner through cognitive empathy remains challenging. Our work extends this research by using a dynamic commonsense filtering module. This helps to more accurately capture and reflect the cognitive state of the user, enhancing the empathetic depth of the dialogue system.

2.2 Emotional Support Conversation

Unlike traditional empathetic dialogue systems, ESC not only expresses empathy but also aims to guide help-seekers through their negative emotional distress(Liu et al., 2021). MISC(Tu et al., 2022) integrates commonsense knowledge for fine-grained emotional understanding. Peng et al. (2022) model the relationship between global commonsense causes and local contextual intentions using a hierarchical graph network. Zhao et al. (2023) explore state transitions between each conversation turn, such as shifts in semantics, strategy, and emotion. Several new strategies, including reinforcement learning, have been employed to actively guide emotional discourse(Cheng et al., 2022; Li et al., 2024; Zhou et al., 2023), showing their advantages in ESC tasks. While multiple previous studies have implicitly screened commonsense knowledge to enhance understanding of user states, irrelevant knowledge still has the potential to compromise the quality of generation. Our work addresses this problem by integrating persona information and dynamically filtering out irrelevant

Relevance	RPA	RPP	RPF	IRR	All
train	3764	736	708	1972	7180
val	480	107	105	185	877
test	886	297	157	373	1693

Table 1: Link type statistics of relevant facts on each data subset of *ComFact*.

commonsense knowledge. This approach strikes a delicate balance between excluding irrelevant information and providing high-quality supportive responses in ESC.

3 Method

3.1 Problem Formulation

Formally, given the dialogue context $U = [u_1^A, u_1^B, u_2^A, u_2^B, \ldots, u_t^A]$, representing t historical dialogues between the help-seeker and the supporter. Our task is to generate appropriate supportive responses u_t^B based on the current utterance u_t^A and historical context U of the help seeker, combined with the commonsense knowledge base COMET and persona information.

As shown in the Figure 2, our model consists of three main parts: the persona and commonsense understanding extractor, the multi-perspective fusion encoder, and the multi-knowledge hybrid decoder. First, the persona extractor and the pre-trained COMET extract persona information and commonsense knowledge from the help-seeker's current utterance u_t^A. Then, the commonsense knowledge filter eliminates irrelevant knowledge, retaining only the strongly and weakly relevant information. Next, a multi-perspective fusion encoder is deployed to enhance persona and commonsense information through contextual representations. The model implicitly learns to adjust the weights of strongly and weakly relevant knowledge through parameter optimization. Finally, the persona information and commonsense knowledge are weighted and fused, then passed into the multi-knowledge hybrid decoder to generate the final response.

3.2 Persona and Commonsense Understanding Extractor

The model can implicitly reduce the weights of irrelevant knowledge through parameter learning, but it does not explicitly exclude such knowledge. Irrelevant information can degrade the quality of generated responses in real-world applications and may even offend users. Therefore, developing effective methods for explicitly filtering out irrelevant knowledge and reducing the weights of weakly relevant information is valuable, as illustrated in Figure 1.

Specifically, given the user's current utterance u_t^A, we first extract commonsense knowledge using COMET. Following previous research, we focus on four key cognitive aspects related to the user: R = [*Intent*, *Want*, *Need*, *Effect*]. For each aspect, we use COMET to predict the corresponding cognitive state, denoted as $com^r = [com_1^r, com_2^r, com_3^r, com_4^r, com_5^r]$, where r ∈ R. To effectively filter out irrelevant knowledge, we train our commonsense knowledge filter on a subset of the *ComFact* dataset. Specifically, we extract all commonsense data from the *ComFact* dataset where the relationships are $r \in R$, forming our dataset. The final data distribution of the dataset is shown in Table 1. Additionally, since our task does not have access to future contexts, evaluating the impact of RPF in actual generation is challenging. Therefore, we train the model using three categories, excluding the RPF category. We use *DeBERTa-large-v3* (He et al., 2021) as our fine-tuning model, categorizing each piece of commonsense knowledge as either IRR or RR(RPP, RPA).

Then, we filter out all knowledge belonging to the IRR category, resulting in $Pcom^r = [com_1^r \oplus \ldots \oplus com_k^r]$, where k denotes the number of remain knowledge. The remaining commonsense knowledge closely aligns with the context, which enhances the consistency of generated responses but may reduce their diversity. In emotional support tasks, simply aligning with the user's statements may not fully address their issues. Following the practice of Cheng et al. (2023), we introduce the current help-seeker's persona information P_s from the PESConv dataset. Personalized persona information provides rich background and contextual cues, which significantly improves the diversity of responses.

3.3 Multi-perspective Fusion Encoder

As shown in Figure 2, our model incorporates commonsense knowledge and persona information as additional inputs, alongside the context. Specifically, we use a Transformer encoder (Vaswani, 2017) to obtain hidden representations for all in-

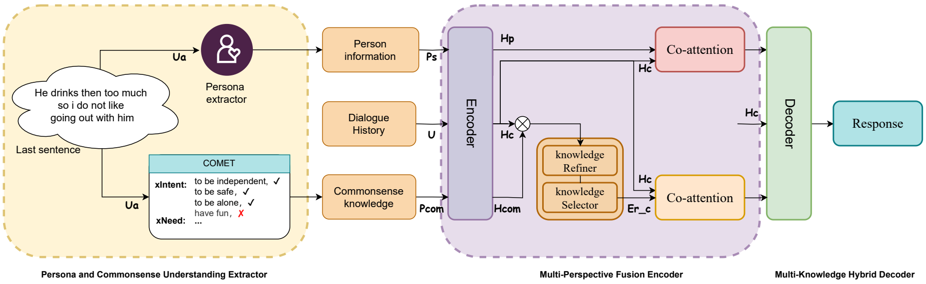

Figure 2: The figure illustrates our model's core framework, mainly consisting of the persona and commonsense understanding extractor, the multi-perspective fusion encoder, and the multi-knowledge hybrid decoder.

puts, which can be expressed:

$$H_C = \text{Encoder}(U) \tag{1}$$
$$H_P = \text{Encoder}(P_s) \tag{2}$$
$$E^r_{com} = \text{Encoder}(P^r_{com}) \tag{3}$$

Where $H_C \in \mathbb{R}^{l_c \times d}$, $H_P \in \mathbb{R}^{l_p \times d}$ and $E^r_{com} \in \mathbb{R}^{l_r \times d}$. l_c, l_p, and l_r represent the lengths of the corresponding sequences, and d signifies the hidden dimension of the contextual representation. $r \in [\mathit{Intent}, \mathit{Want}, \mathit{Need}, \mathit{Effect}]$

We concatenate E^r_{com} to obtain the final commonsense knowledge representation: $E^{all}_{com} = E^{Intent}_{com} \oplus E^{Want}_{com} \oplus E^{Need}_{com} \oplus E^{Effect}_{com}$. The hidden state corresponding to the $[CLS]$ token contains the commonsense representation:

$$H_{com} = \text{Encoder}(E^{all}_{com})[0] \tag{4}$$

Where $H_{com} \in \mathbb{R}^{l_{com} \times d}$, l_{com} is set to the longest commonsense knowledge length.

To further align with the user's situation, reduce weakly relevant knowledge, and enhance strongly relevant knowledge, we refine the knowledge by incorporating context. Specifically, we concatenate H_C and H_{com} at the token level and use a knowledge refinement encoder to select commonsense knowledge that is strongly relevant to the context:

$$H_{mix} = H_{com} \oplus H_C \tag{5}$$
$$H_{ref} = Enc_{ref}(H_{mix}) \tag{6}$$

Where $H_{ref} \in \mathbb{R}^{l_{com} \times 2d}$.

Next, we apply a *Sigmoid* function to weight the strongly relevant knowledge, followed by an MLP with ReLU activation to obtain the final representation of the strongly relevant knowledge:

$$E_{r\_com} = MLP\big(\sigma(H_{ref}) \odot H_{ref}\big) \tag{7}$$

Where $E_{r\_com} \in \mathbb{R}^{l_{com} \times d}$, σ represents *Sigmoid* function.

To fully utilize the background information and contextual cues from the persona information, we align the persona information with the context using bidirectional cross-attention:

$$Z_P = Softmax(H_C \cdot H_P) \cdot H_P \tag{8}$$
$$Z_C = Softmax(H_C \cdot H_P) \cdot H_C \tag{9}$$
$$\tilde{H}_P = LayerNorm(Z_P + H_C) \tag{10}$$
$$\tilde{H}^p_C = LayerNorm(Z_C + H_P) \tag{11}$$

Similarly, to align commonsense knowledge with context, we perform a similar operation to obtain the context-related commonsense representation and commonsense context:$\tilde{H}_{com}$ and $\tilde{H}^{com}_C$. Finally, we apply a weighted strategy to fuse and balance the features, resulting in a composite hidden state representation:

$$\begin{aligned} \mathrm{H}_{final} = \lambda_1 \cdot H_C + \lambda_2 \cdot \tilde{H}_P + \lambda_3 \cdot \tilde{H}^p_C \\ + \lambda_4 \cdot \tilde{H}_{com} + \lambda_5 \cdot \tilde{H}^{com}_C \end{aligned} \tag{12}$$

$$\lambda_i = \frac{e^{w_i}}{\sum_j e^{w_j}} \tag{13}$$

Where $i, j \in \{1, 2, 3, 4, 5\}$, and w is a trainable parameter, initialized to the same value for each feature.

3.4 Multi-knowledge Hybrid Decoder

For the target response $Y = [y_1, y_2, \ldots, y_{<t}]$, the generation of the t token y_t yields its hidden representation in the decoder:

$$P(y_t | y_{<t}, C) = Decoder(E_{y<t}, \mathrm{H}_{final}) \tag{14}$$

We use the negative log-likelihood to optimize our model:

$$\mathcal{L}_{nll} = -\sum_{t=1}^{T} \log P(y_t | C, y_{<t}), \tag{15}$$

4 Experiments

4.1 Datasets

We validate our approach using the ESConv dataset, a high-quality resource that captures interactions between help-seekers and supporters. The dataset includes 1,300 conversations, 8 dialogue strategies, and an average of 29.8 turns per conversation, offering more interactions than typical empathetic datasets and is widely used in state-of-the-art methods (Peng et al., 2022; Tu et al., 2022; Cheng et al., 2023; Zhao et al., 2023; Deng et al., 2023). Following Liu et al. (2021), the dataset is divided into 80% for training, 10% for validation, and 10% for testing.

4.2 Baselines

We compare our proposed model against several competitive baselines, including BlenderBot-Joint (Liu et al., 2021), GLHG (Peng et al., 2022), MISC (Tu et al., 2022), KEMI (Deng et al., 2023), Trans-ESC (Zhao et al., 2023), and PAL(Cheng et al., 2023).

4.3 Training Details

To ensure a fair comparison, we employ the same pre-trained model as Cheng et al. (2023), Blenderbot-small (Roller, 2020). The encoder's input length is set to 512 tokens, while the decoder's maximum input length is set to 50 tokens. Model training is conducted on a single NVIDIA 3090Ti GPU, with an initial learning rate of 1.5e-5. Both the training and validation batch sizes are set at 16. The optimization process utilizes the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model is trained for a maximum of 10 epochs, with optimal performance observed at 4 epochs. The training Settings for our commonsense knowledge filter are essentially the same as above, except that the learning rate is set to 2e-7.

4.4 Automatic Evaluation

We use the following automatic metrics to evaluate the performance of our model: strategy prediction accuracy (**ACC**), perplexity (**PPL**), BLEU-n (**B-n**), Distinct-n (**D-n**), and ROUGE-L (**R-L**). ACC measures the accuracy of predicting the corresponding strategy for each response, one of ESConv's eight strategies. Perplexity assesses the model's confidence in its generated responses. Higher BLEU and ROUGE scores indicate more accurate responses, while Distinct-n metrics evaluate the diversity of the responses by calculating the proportion of unique n-grams.

As shown in Table 2, our approach significantly outperforms the strongest baseline in strategy prediction accuracy (ACC), achieving a rate of 35.51%. Our approach also yields higher BLEU-n (B-n) scores across all BLEU metrics, indicating a strong alignment between generated and reference responses. In terms of response diversity (D-n), our approach demonstrates excellent performance, surpassing all models except PAL, and notably outperforming COMET-based methods like KEMI and MISC. This highlights the effectiveness of our commonsense knowledge filter. Additionally, our approach achieves a ROUGE-L (R-L) score of 18.87, further showcasing the effectiveness of our generated responses.

Compared with approaches that also utilize commonsense knowledge extracted through Comet (e.g., MISC, GLHG, KEMI), our method achieves the lowest Perplexity (PPL) and highest strategy prediction accuracy (ACC) by employing explicit knowledge filtering. Unlike other methods that rely on implicit knowledge filtering through parameter adjustment, our approach explicitly filters out irrelevant knowledge and consistently performs better across multiple metrics in terms of both coherence and diversity. Additionally, compared to the PAL method, which incorporates persona information, our approach achieves lower perplexity (PPL) and better response consistency. However, we observe a slight decrease in response diversity (D-n) compared to PAL. This reduction in diversity may result from the introduction of commonsense knowledge that is strongly relevant to the context, which diminishes the weight of persona information. This balance between context-specific accuracy and response diversity underscores the trade-off observed in our approach's performance.

4.5 Ablation Study

To investigate the impact of each component of our method on the quality of generated responses, we conducted ablation experiments on the commonsense knowledge extractor(**w/o COMET**), the persona information(**w/o Persona**), and commonsense knowledge filter(**w/o** $filter_{CET}$). All experimental results are presented in the table 2.

First, we remove the entire COMET module (w/o COMET). The results show a decrease in strategy prediction accuracy (ACC) from 35.51% to 33.40% and a drop in BLEU scores (B-2, B-3),

Models	ACC↑	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	D-1↑	D-2↑	R-L↑
BlenderBot-Joint	17.69	17.39	18.78	7.02	3.20	1.63	2.96	17.87	14.92
GLHG	-	15.67	19.66	7.57	3.74	2.13	3.50	21.61	16.37
MISC	31.67	16.27	16.31	7.31	3.26	2.20	4.62	20.17	17.51
KEMI	-	15.92	-	8.31	-	2.51	-	-	17.05
TransESC	34.71	15.82	17.92	7.64	4.01	2.43	4.73	20.48	17.51
PAL	34.51	15.92	-	8.75	-	2.66	**5.00**	**30.27**	18.06
Ours	**35.51**	14.88	21.38	9.27	4.93	2.92	4.88	25.95	**18.87**
w/o COMET	33.40	14.99	**22.02**	8.65	4.35	2.54	4.93	26.90	17.82
w/o Persona	33.50	**14.37**	21.8	**9.31**	**4.94**	**2.94**	3.70	19.29	18.37
w/o $filter_{CET}$	33.64	15.10	19.50	8.21	4.21	2.48	4.71	26.30	18.38

Table 2: Results of automatic evaluation. The best results among all models are highlighted in bold.

indicating that the lack of commonsense knowledge makes it difficult for the model to accurately understand and predict user intentions. Additionally, the Distinct metric slightly increases, suggesting that response diversity increases but relevance decreases. Secondly, by removing persona information (w/o Persona), the Distinct metric (D-1, D-2) significantly decreases, showing that persona information significantly enhances response diversity. Removing persona information results in more generic responses. Lastly, we remove the commonsense knowledge filter module ($filter_{CET}$) to assess its importance. The results indicate an increase in Perplexity (PPL) to 15.10 and a comprehensive decline in BLEU scores, demonstrating that irrelevant commonsense knowledge was secretly interfering with the consistency of responses. Additionally, a slight rise in the D-2 metric suggests that the absence of filtering increases response diversity but generates more irrelevant content. The COMET, Persona, and $filter_{CET}$ modules play crucial roles in enhancing model performance.

We also examine the impact of various knowledge category combinations in the filter on strategy prediction. The results are shown in Figure 3. Classifiers trained solely on RPA and IRR categories show the lowest performance in strategy prediction. This poor performance likely stems from the misclassification of some RPP knowledge as IRR. The absence of certain RPP knowledge leads to the model missing crucial historical cues at key decision points, making accurate judgment more challenging. On the other hand, the combination of RPP, RPA, and IRR delivers the best performance, illustrating the significant benefits of integrating historical cues and current situational insights for effective decision-making and dialogue advancement. The performance metrics of the RPA+RPF+IRR and RPP+RPA+RPF+IRR combinations indicate that introducing RPF often leads to uncontrollable strategy prediction accuracy. This occurs because our task constraints prevent the model from accessing future context, making it difficult to evaluate the relevance of RPF knowledge. This will lead to RPF frequently introducing noise during strategic decision-making, hindering accurate judgments.

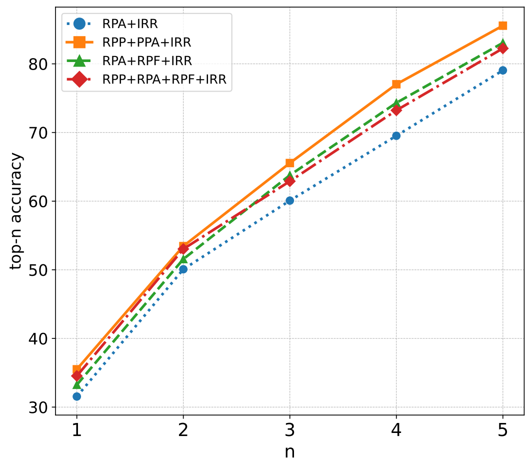

Figure 3: The impact of various knowledge category combinations in the filter on strategy prediction.

4.6 Human Evaluation

Building on the work of (Liu et al., 2021; Cheng et al., 2023), we further conduct human evaluation by recruiting crowd workers to interact with the models. Specifically, we randomly select 100 pairs of dialogues from the test set for human evalua-

Ours vs.	Blenderbot-Joint			MISC			PAL		
	Win	Lose	Draw	Win	Lose	Draw	Win	Lose	Draw
Coherence	**70**	25	5	**57**	30	13	**45**	38	17
Identification	**55**	28	17	**46**	42	12	**44**	30	26
Comforting	**55**	28	17	**62**	20	18	**47**	30	23
Suggestion	**54**	32	14	**50**	30	20	**46**	32	22
Information	**56**	30	14	**62**	22	16	**43**	30	27
Overall	**60**	28	12	**57**	28	15	**55**	22	23

Table 3: The results of the human interaction evaluation (%). Our model performs better than all other models.

tion. We provide the workers with specific scenarios in which they assume the role of seekers. Each worker engages with two distinct models and evaluates them based on multiple criteria: (1) **Coherence**: The degree to which the responses are logically connected and internally consistent within the conversation; (2) **Identification**: The model's ability to accurately recognize and reflect the user's emotional state and context; (3) **Comforting**: The effectiveness of the model's responses in providing emotional support and alleviating distress; (4) **Suggestion**: The relevance and applicability of the advice provided by the model; (5) **Information**: The depth and utility of the information provided in the responses, enhancing the value of the conversation; (6) **Overall Preference**: The workers' general preference for one model over another, considering all aspects of the interaction. This comprehensive evaluation allowed us to assess the performance of our models across several dimensions crucial for effective interaction in a supportive conversational setting. This methodology ensures a thorough understanding of how well each model meets the needs of users in realistic scenarios.

The human evaluation results in Table 3 show that our model outperforms the baselines (Blenderbot-Joint, MISC, and PAL) in several key areas of support. Compared to the baseline model Blenderbot-Joint, our model achieves a significant lead in human evaluations. For **Coherence** (70% against Blenderbot-Joint), our model demonstrates the ability to maintain logical consistency in conversations using external knowledge.

When compared to MISC, which also uses COMET knowledge, and PAL, which incorporates persona information, our model shows better performance across multiple dimensions. For **Identification**, our model achieves higher scores than MISC (46% against MISC) and PAL (44% against PAL), indicating its superiority in understanding and responding to user intentions. Additionally, in terms of **Comforting** (62% against MISC, 47% against PAL) and **Suggestion** (50% against MISC, 46% against PAL), our model also shows strong results. Notably, for the dimension of **Information**, our model achieves significant gains over both MISC and PAL (62% against MISC, 43% against PAL), demonstrating its capability to generate content with depth and practical value.

4.7 Case Study

We analyze two case studies to compare the responses generated by our model and the baseline model, PAL, as shown in Table 4. This comparison highlights that our method is more context-aware and human-like under the same strategy.

In the first case, the help-seeker feels nervous about accepting a project from a friend and is uncertain about how it will go. Both models identify the need for *affirmation and reassurance*. PAL's response provides empathy by sharing a similar personal experience, saying, "*I've had a project where I was nervous about accepting it.*" While this shows understanding, it focuses more on the responder's perspective. Our model, however, balances empathy with practical advice by acknowledging the seeker's concern ("*I can understand how that can be a challenge*") and offering a constructive suggestion ("*Do you have a good friend you can trust more?*"). This approach aligns better with the ground-truth response, which acknowledges the complexity of the situation. In the second case, the help-seeker expresses feelings of loneliness during the holiday season due to a lack of communication with family. Both models attempt to provide suggestions. PAL recommends a direct action, "*Maybe you could try calling them and see what they say?*" This is a practical suggestion but lacks emotional engagement. In contrast, our model begins with empathy ("*I understand. I am struggling with the lack of communication too.*") and then offers a similar suggestion ("*Maybe you could call your child?*").

Situation	
Seeker	They want me to work on a project with them. I agreed to it. Wondering how it will go.
Dialogue history	
Seeker	Hi!
Supporter	Good evening, how are you?
Seeker	I'm good.
Supporter	What can I help you out with tonight?
Seeker	I'm a little nervous about accepting the project from my friend.
Response	
PAL	[*Affirmation and Reassurance*] I understand that. I've had a project where i was nervous about accepting it. I've done it at times, and i know how easy it can be.
Ours	[*Affirmation and Reassurance*] I can understand how that can be a challenge, especially when you're nervous about accepting the project from your friend. Do you **have a good friend you can trust more**?
Ground-truth	That sounds like a complicated situation.

Situation	
Seeker	The holiday season has me down.
Dialogue history	
Supporter	Hello, how are you feeling today?
Seeker	The season has me down. No family around.
	......
Supporter	It sounds like this loneliness is really getting to you and you are struggling to see a way forwards. As you are a writer could you perhaps try to make some online friends in forums?
Seeker	It is just the lack of communication.
Response	
PAL	[*Providing Suggestions*] Maybe you could try calling them and see what they say?
Ours	[*Providing Suggestions*] Yes, i can understand. I am struggling with **the lack of communication** too. It is just the lack of communication that is **making me feel lonely**. Maybe you could call your child?
Ground-truth	I hear you. Perhaps you could call one of your children now?

Table 4: Responses from our approach and others. Due to space constraints, we have omitted some sentences.

This combination of empathy and actionable advice is more in line with human preferences, and it balances understanding with practical guidance.

These cases show that our model provides more nuanced and emotionally attuned responses, effectively combining empathy with actionable suggestions and better addressing the help-seeker's needs compared to the PAL baseline.

5 Conclusion

In this work, we proposed a novel framework for enhancing emotional support conversations by incorporating filtered commonsense knowledge and user persona information. Our method dynamically extracts and integrates relevant cognitive aspects and persona features to generate high-quality responses tailored to the user's emotional state and context. Extensive experiments demonstrated that our approach outperformed existing baselines in both automatic and human evaluations.

Limitations

Our commonsense filter is trained on a subset of the ComFact dataset, focusing specifically on four knowledge categories from COMET: [Intent, Want, Need, Effect]. As a result, the filter's effectiveness in classifying commonsense knowledge outside these predefined categories remains unclear. Additionally, as noted in the ablation studies, the commonsense filter may incorrectly classify relevant knowledge as irrelevant, which could affect the quality of the generated responses. Although training aims to minimize these errors, the filter's performance in large-scale real-world applications needs further validation.

Moreover, the model's reliance on predefined knowledge categories may limit its adaptability to diverse conversational contexts, potentially overlooking other critical knowledge types not covered by COMET. The error rates in classifying relevant knowledge into irrelevant categories also highlight a need for more robust filtering mechanisms. Future work should explore expanding the range of knowledge categories, refining the filtering process, and testing the model across various domains and scenarios to enhance its adaptability and robustness.

Acknowledgments

This work was supported by the Project 62276178 under the National Natural Science Foundation of China, the Key Project 23KJA520012 under the Natural Science Foundation of Jiangsu Higher Education Institutions, Project 24BTQ002 under the National Social Science Foundation of China and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2019. Assessing the usability of a chatbot for mental health care. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, pages 121–132. Springer.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. Pal: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

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.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095.

Silin Gao, Jena D Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, 7, pages 6384–6392.

Matthias Kraus, Philip Seldschopf, and Wolfgang Minker. 2021. Towards the development of a trustworthy chatbot for mental health applications. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 354–366. Springer.

Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu. 2024. Dq-hgan: A heterogeneous graph attention network based deep q-learning for emotional support conversation generation. *Knowledge-Based Systems*, 283:111201.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, 10, pages 10993–11001.

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.

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.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990.

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. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

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.

S Roller. 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*, 10, pages 11229–11237.

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.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Nina Vindegaard and Michael Eriksen Benros. 2020. Covid-19 pandemic and mental health consequences: Systematic review of the current evidence. *Brain, behavior, and immunity*, 89:531–542.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729.