Muffin: Mitigating Unhelpfulness in Emotional Support Conversations with Multifaceted AI Feedback

Anonymous ACL submission

Abstract

Emotional support conversation systems are designed to alleviate users' emotional distress and assist them in overcoming their challenges. While previous studies have made progress, their models occasionally generate unhelpful responses, which are intended to be supportive but instead have counterproductive effects. Since unhelpful responses can hinder the effectiveness of emotional support, it is crucial to mitigate them within conversations. Our solution is motivated by two principal considerations: (1) multiple facets of emotional support are expected to be considered when developing emotional support conversation models, and (2) directly reducing the probability of generating unhelpful responses can effectively mitigate their occurrence. Accordingly, we introduce a novel **model-agnostic** framework named Mitigating unhelpfulness with multifaceted AI feedback for emotional support (*Muffin*). It first employs a multifaceted AI feedback module designed to assess the helpfulness model responses across various facets of emotional support. Leveraging contrastive learning, Muffin then reduces the unhelpful responses' likelihoods. To validate the effectiveness of our proposed framework, we apply Muffin to various previous emotional support generation models, including the state-of-the-art. Experimental results demonstrate that Muffin can significantly mitigate unhelpful response generation while enhancing response fluency and relevance.

1 Introduction

Emotional support conversation systems (supporters) are designed to generate responses that can buffer the emotional distress experienced by users (help-seekers) and help users to work through the challenges they are confronting (Liu et al., 2021). Recently, many studies have contributed to this field (Deng et al., 2023; Zhao et al., 2023; Cheng et al., 2022; Tu et al., 2022). Despite great success, their models occasionally generate well-intended responses that produce a counterproductive support effect, i.e., exacerbating the negative emotional states of users or inhibiting effective problem-solving, as shown in Figure 1. In the psychology and communication theories, these failed support attempts are termed "*unhelpful messages*" (Greene and Burleson, 2003; Burleson, 1985).

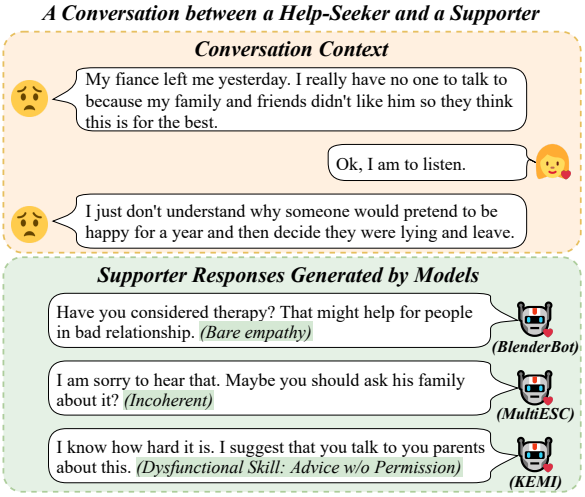

Figure 1: Examples of unhelpful responses generated by recent emotional support conversation models, including BlenderBot (Roller et al., 2021), MultiESC (Cheng et al., 2022), and KEMI (Deng et al., 2023).

The frequency of unhelpful responses generated by some of the previous models is not extremely high, e.g, approximately 30% of responses generated by MultiESC (Cheng et al., 2022) on ESConv benchmark (Liu et al., 2021) are identified as unhelpful under strict evaluation criteria. However, their occurrence can significantly undermine earlier supportive efforts and damage the trust between the help-seeker and the supporter (Llewelyn et al., 1988). Therefore, mitigating models' generation of unhelpful responses is critical. We aim to address this problem with the following two deliberations. ***D1*** - *Consideration of Multiple Facets*: Many pre-

vious studies generate responses that primarily emphasize a single facet of emotional support, e.g., one of empathetic expression (Li et al., 2024), communication skill efficiency (Cheng et al., 2022; Liu et al., 2021), or response coherence (Deng et al., 2023), in each of their models. However, such a singular emphasis on one facet often leads to the oversight of the others, potentially resulting in unhelpfulness (Greene and Burleson, 2003), as exemplified in Figure 1. ***D2*** *- Direct Minimization of Unhelpful Response Probability*: Previous models are typically optimized by minimizing the negative log-likelihood of the golden responses. Moving beyond this optimization objective, we aim to specifically mitigate unhelpful responses by directly targeting and reducing the probability of their generation.

Accordingly, this paper introduces a novel **model-agnostic** framework called Mitigating unhelpfulness with multifaceted AI feedback for emotional support (*Muffin*). For ***D1***, we design a multifaceted AI feedback module within the Muffin framework. This module assesses whether a specific response is unhelpful from multiple facets of emotional support. Leveraging the advanced capabilities of recent large language models (LLMs), we implement this module by instruction-tuning LLaMA, avoiding inefficient and expensive human feedback collection. Then, we continue optimizing an emotional support conversation model with its previous training objective and a new one to implement ***D2***. The additional objective is to minimize the likelihood of unhelpful responses, which is implemented by contrasting unhelpful responses, identified by the feedback module, and the other (non-unhelpful) ones. Through these two steps, we aim to mitigate the unhelpful responses generated by a given emotional support conversation model. Experimental results highlight the effectiveness of our framework, demonstrating that Muffin can enhance the helpfulness of previous emotional support conversation models, including those recognized as state-of-the-art. The main contributions of this work are:

1. We recognize and address a crucial problem in recent emotional support conversation models, i.e., the generation of unhelpful responses, a key concern in effective emotional support.

2. We propose Muffin, a novel model-agnostic framework designed to mitigate unhelpful response generation. It incorporates a multifaceted AI feedback module to distinguish unhelpful generated responses and mitigates responses identified as unhelpful by leveraging contrastive learning.

3. We undertake experiments with the latest emotional support conversation models, including state-of-the-art ones, to demonstrate Muffin's effectiveness in mitigating the models' tendency to produce unhelpful responses.

2 Related Work

2.1 Emotional Support Conversation

In the domain of emotional support conversation generation, prior studies have achieved some success. Specifically, they have each emphasized and incorporated different facets of emotional support, such as empathetic expression, communication skills, and response coherence, into their respective models. Some of them consider a single facet. For example, MultiESC (Cheng et al., 2022) only considers the communication skill efficacy of responses by planning response strategies. KEMI (Deng et al., 2023) incorporates related external knowledge for response generation. Although the response coherence is enhanced, the efficacy of communication skills is ignored. Li et al. (2024) enhance the empathetic expression of LLMs through chain-of-though. Beyond these, MISC (Tu et al., 2022) generates supportive responses considering both commonsense and communication skills, thereby enhancing two facets of emotional support. TranESC (Zhao et al., 2023) incorporates commonsense, communication skills, and emotional elements into response generation. Most all these models are optimized by minimizing the negative log-likelihood of golden responses, instead of the unhelpful response likelihood.

2.2 Contrastive Learning for Text Generation

Contrastive learning was initially employed to learn meaningful representations by contrasting positive and negative samples in the field of natural language processing (Logeswaran and Lee, 2018; Gutmann and Hyvärinen, 2012). Recently, it has been applied to text generation (Jiashuo et al., 2023; Zheng et al., 2023; Liu et al., 2022), achieving impressive performance across various settings. During training, the model is exposed to a range of "hard" negative and positive samples through contrastive learning, enabling the model to distinguish preferred outputs from less desirable ones. Consequently, selecting positive and negative samples

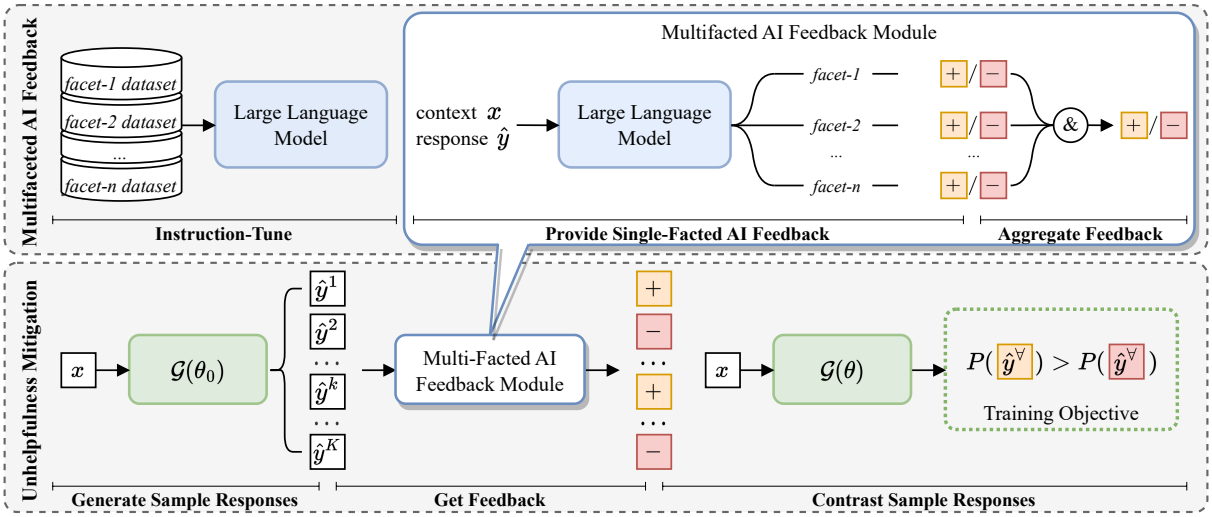

Figure 2: The overview of our proposed model-agnostic framework—Muffin. + and − indicate helpful (non-unhelpful) and unhelpful labels, respectively.

is crucial in this process. In this paper, responses generated by the model that are deemed unhelpful are negative samples, while all other responses are considered positive samples.

3 Preliminaries

3.1 Unhelpfulness of Emotional Support

We draw upon theories in psychology and communication (Greene and Burleson, 2003; Burleson, 1985) and adopt the term "*unhelpful*" to characterize responses that consistently produce negative outcomes in emotional support conversations. Conversely, responses that yield positive outcomes, or at the very least, do not cause negative effects, are termed "*helpful*." These theories also suggest that an unhelpful response often stems from a flaw in merely one specific facet of emotional support. Often, the flaw directly exacerbates the user's negative feelings or hinders effective problem-solving. For instance, a response can be deemed unhelpful if it either neglects the individual's feelings and needs (lacking empathy) or portrays the individual's behavior as problematic (exhibiting a dysfunctional communication skill: confront). In our work, we use this feature to identify unhelpful responses.

3.2 Problem Definition

Our primary goal is to mitigate the generation of unhelpful responses. Rather than training a new model from scratch, we aim to refine a pre-trained emotional support conversation model with the dataset it was originally trained on. This process unfolds as follows. Let $\mathcal{G}(\theta_0)$ represent the model trained on a dataset $\mathcal{D}$, where θ_0 denotes the model parameters. Each instance in $\mathcal{D}$ is denoted as (x, y), with x as the input and y as the expected output. Usually, x is the conversation context, but it contains additional related information in some models. Assume that there are K samples $\{\hat{y}^1, \cdots, \hat{y}^K\}$ with labels $\{\hat{l}^1, \cdots, \hat{l}^K\}$. These samples are the diverse beam search generation results of $\mathcal{G}(x; \theta_0)$. As for the label $\hat{l}^k \in \{0, 1\}$, it represents feedback to indicate whether the sample $\hat{y}^k$ is unhelpful ($\hat{l}^k = 0$) or not unhelpful ($\hat{l}^k = 1$). Our objective is to refine the model's parameters θ such that the likelihood of generating unhelpful samples is reduced relative to helpful ones. In this process, we only modify the training process, ensuring that the model's architecture and the inference mechanism remain untouched. Moreover, our approach is model-agnostic. This implies that $\mathcal{G}(\theta_0)$ can be any deep learning model designed and trained for emotional support conversations.

4 Method

The overall framework of Muffin is outlined in Figure 2. It is composed of two principal components, each specifically designed for deliberations in Section 1: ***D1***: *Consideration of Multiple Facets* and ***D2***: *Direct Mimimization of Unhelpful Response Probality*, respectively. The multifaceted AI feedback module aims to identify whether a response from multiple facets of emotional support is unhelpful. The unhelpfulness mitigation module mitigates the likelihood of unhelpful responses by contrasting helpful and unhelpful responses.

4.1 Multifaceted AI Feedback

We distinguish whether a response is unhelpful from multiple facets. However, collecting feedback from humans is inefficient and costly. In addition, recent large language models (LLMs), such as the GPT series (Ouyang et al., 2022) and LLaMA (Touvron et al., 2023), demonstrate remarkable natural language understanding capabilities. Therefore, we decide to obtain feedback from AI.

Instruction-tuning Prompt engineering provides a simple and straightforward approach to obtaining feedback from LLMs. However, our experiments suggest that it is challenging to manifest the full potential of LLMs for emotional support without investing significant effort in prompt design, which will be detailed later. As an alternative, we elicit the desired capabilities of the LLM via instruction tuning (Wei et al., 2022). Specifically, we design task descriptions and instructions tailored to classification tasks related to different emotional support facets. We employ corresponding datasets for these tasks. The prompt employed is illustrated in Figure 3. Notably, the response class can indicate whether the response is unhelpful regarding this facet. During the training phase, all *texts in italics enclosed within curly braces* are provided. During inference, the model is expected to generate the *response class*, based on the other *italicized inputs within the curly braces*.

```
### Instruction:
{task description and instruction}

### Input:
Conversation Context: {context}
The last supporter statement: {response}
{all possible classes}

### Output:
{response class}
```

Figure 3: The prompt used by the Multifaceted AI Feedback for classifying the supporter's response.

Multifaceted AI feedback module The final feedback for an emotional support response is derived from an aggregation of AI feedback across multiple facets. For the given response and its context, we use the instruction-tuned LLM to provide feedback on all these facets respectively. If feedback from any of these facets suggests the response is unhelpful, the response is accordingly labeled as 0; otherwise, the response is deemed helpful (non-unhelpful) and labeled as 1, as mentioned in Section 3.1.

4.2 Unhelpfulness Mitigation

We mitigate $\mathcal{G}_0$ generating unhelpful responses by contrasting helpful and unhelpful responses generated by $\mathcal{G}_0$ itself, which can be implemented by the following three steps:

Generating sample responses We utilize $\mathcal{G}_0$ to generate responses on its own training dataset $\mathcal{D}$ using diverse beam search (Vijayakumar et al., 2016). Thus, for each instance $(x, y) \in \mathcal{D}$, there are K sample responses $\{\hat{y}^1, \cdots, \hat{y}^K\}$.

Getting feedback These responses can be generated because they have relatively high generation probabilities. However, some of them can be unhelpful responses. Therefore, we adopt the multifaceted AI feedback module to identify whether these responses are unhelpful. Thus, we obtain K labels $\{\hat{l}^1 \cdots, \hat{l}^K\}$, where $\hat{l}^k \in \{0, 1\}$.

Contrasting sample responses We expect that the model $\mathcal{G}$ can sign a higher likelihood to the helpful responses than the unhelpful ones. Therefore, we contrast them using the following loss:

$$\mathcal{L}_{cl} = \frac{1}{2K} \sum_{i} \sum_{j \neq i} \max(0, - (\hat{l}^i - \hat{l}^j) \times (\mathrm{P}(\hat{y}^i|x) - \mathrm{P}(\hat{y}^j|x) + \lambda)), \quad (1)$$

where λ is the margin hyperparameter. Moreover, $\mathrm{P}(\hat{y}^i|x)$ is the length-normalized log-probability of the response $\hat{y}^i$, and it is be computed by:

$$\mathrm{P}(\hat{y}^i|x) = \sum_{t=1}^{|\hat{y}^i|} \frac{\log \mathcal{G}(\hat{y}^i_t|x, \hat{y}^i_{<t}; \theta)}{|\hat{y}^i|^\alpha}, \quad (2)$$

where α is the length penalty hyperparameter. In addition to the above loss, we also consider the negative log-likelihood loss to prevent the model's generation from deviating too much from the ground truth. The loss can be formulated as:

$$\mathcal{L}_{gen} = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log \mathcal{G}(y_t|x, y_{<t}; \theta). \quad (3)$$

The final loss is the combination of the above two losses:

$$\mathcal{L} = \beta_{cl}\mathcal{L}_{cl} + \beta_{gen}\mathcal{L}_{gen}, \quad (4)$$

where β_{cl} and β_{gen} are weight hyperparameters.

5 Experiments

5.1 Experimental Setups

Base models ($\mathcal{G}(\theta_0)$) Our proposed method, i.e., Muffin, is a model-agnostic approach designed to mitigate the unhelpfulness of an existing emotional support conversation model. To examine its effectiveness, we experiment with five recent models: BlenderBot (Vanilla) (Roller et al., 2021), BlenderBot-Joint (Joint) (Liu et al., 2021), MultiESC (Cheng et al., 2022), TransESC (Zhao et al., 2023), and KEMI (Deng et al., 2023). We obtain each model's parameters θ using its official implementation and the default hyperparameters.

When training Muffin$_{\mathcal{G}(\theta_0)}$, we try to use the same training hyperparameters as the base model $\mathcal{G}(\theta_0)$, including batch size and random seed. However, we use a smaller learning rate, i.e., 3×10^{-5}, to help the model converge more efficiently. We set the epoch number as 1 since the training loss converged within one epoch. The margin hyperparamter λ is 0.01, the length penalty hyperparameter α is 1, and the weight hyperparamters β_{cl} and β_{gen} are both 1. In the response generation phase, as described in Section 4.2, we set the number of sample responses K to 10. Consequently, both the beam size and the number of beam groups are configured to be 10, while all other generation hyperparameters are the same as its base model.

ESConv dataset ($\mathcal{D}$) The ESConv dataset (Liu et al., 2021) is used to train the aforementioned base models. ESConv is a benchmark for emotional support conversations, comprising approximately 1K conversations with 13K utterances. All base models, with the exception of TransESC, follow the original division of ESConv for training, validation, and testing, using an 8:1:1 ratio. TransESC employs a random split while maintaining the same ratio. Notably, each model adopts different data preprocessing methods. We adhere to each base model's specific data division and pre-processing.

5.2 Facets of Emotional Support

We consider three essential facets of emotional support: empathetic expression, skill efficiency, and response coherence, which the base models incorporate into their models. Here, we would like to describe the unhelpfulness of each facet and detail the corresponding classification dataset.

Empathetic expression Empathetic expressions signify the supporter's interest in and comprehension of the help-seeker's perspective. Conversely, their absence can impede conversation engagement and obstruct establishing a trust-based relationship between the supporter and the help-seeker (Morse et al., 1992). While empathy encompasses various aspects (Paiva et al., 2017; Batson, 2009; Bohart and Greenberg, 1997), we adopt the comprehensive framework proposed by (Sharma et al., 2020). It identifies three empathy communication mechanisms, i.e., emotional reaction, interpretations, and explorations, each assessed across three levels: no communication, weak communication, and strong communication. Our work considers responses exhibiting no empathy across all mechanisms unhelpful. For example, "sleeplessness can result in upsetness," which inappropriately offers mere information, is considered unhelpful in empathetic expression, especially when responding to a statement like "I am upset".

For training and testing LLMs in classifying unhelpful responses regarding empathetic expression, we also utilize the dataset compiled by Sharma et al. (2020). It consists of 3K context-response pairs. Each response within this dataset is assessed based on three previously mentioned empathy communication mechanisms. Responses consistently labeled as "no communication" across all these mechanisms are identified as unhelpful.

Skill efficiency The applications of effective strategies can help supporters convey appropriate and impactful support messages (Greene and Burleson, 2003), deepening the understanding of the help-seeker's state and facilitating problem solution (Hill, 2009). However, some dysfunctional skills lead to opposite effects (Barsky and Coleman, 2001; Burleson and Samter, 1985). For example, while *advice* is typically seen as beneficial, *advice without permission* can be less effective than employing no specific skills at all. This study assesses skill efficiency using motivational interviewing skill codes (Moyers et al., 2003), which include three general categories: MI adherent, MI non-adherent, and others. Responses classified as MI non-adherent category are deemed unhelpful.

We utilize the Motivational Interviewing (MI) dataset proposed by Welivita and Pu (2022) to classify unhelpful responses in the context of skill efficiency. This dataset comprises 17K context-response pairs, where each response is annotated with one of three MI codes: MI adherent, MI non-adherent, or others. Responses labeled as "MI non-

adherent" are considered unhelpful.

Response coherence While response coherence is a fundamental expectation in almost all conversational systems (Huang et al., 2020), it holds particular significance in emotional support conversation systems. Incoherent responses can confuse and impede effective communication. We categorize responses as coherent or incoherent, with the latter labeled unhelpful.

We have synthesized a dataset, derived from the base model's training set $\mathcal{D}$, specifically for the purpose of response coherence detection. Specifically, we randomly selected 4K context-response pairs from $\mathcal{D}$. For each pair, the original response is categorized as coherent. To introduce incoherence, we employ two methods: firstly, by selecting a response from a different conversation, and secondly, by modifying keywords or important information in the original response to create a subtly incoherent variant. These methods result in a total of approximately 12K context-response pairs, which are then utilized for classifying unhelpful responses in terms of their response coherence.

5.3 Multifaceted AI Feedback

Instruction-tuning settings We instruction-tune a 7B LLaMA (Touvron et al., 2023) for the multifaceted AI feedback module and use a low-rank adaptor (LoRA) (Hu et al., 2022) for efficiency. Specifically, we freeze LLaMA's weight and inject trainable rank decomposition matrices into query, key, value, and output layers. The learning rate is 3×10^{-4}, and the training epoch is 12. We merge the datasets described in Section 5.2, formatting each context-response pair according to the instruction template presented in Figure 3. Consequently, this merged dataset encompasses a total of 22K instances, which are utilized for instruction-tuning. The dataset is partitioned into training, validation, and test sets following an 8:1:1 split ratio.

Module performance The instruction-tuned LLaMA model is subsequently employed within the Multifaceted AI Feedback Module. The module's effectiveness is evaluated using the instruction-tuning test set, with the results depicted in Figure 4. We report both accuracy and F1 scores for data instances throughout the entire testing set and across distinct facets. Additionally, we extend our evaluation to include the module's performance when utilizing three other LLMs: GPT-3.5, GPT-4, and the original (vanilla) LLaMA model.

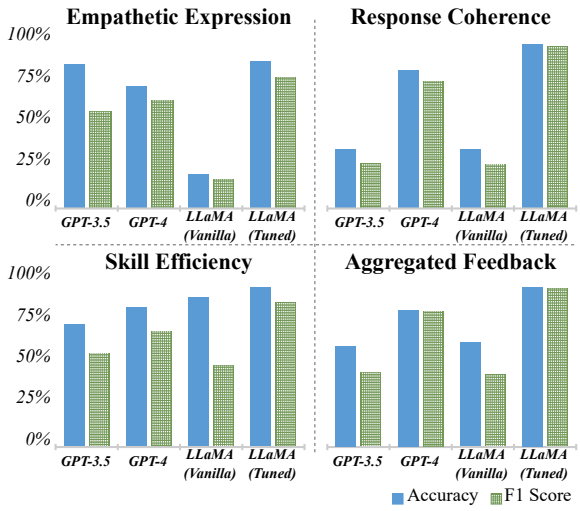

Figure 4: Comparison of performance among various Language Model Models (LLMs) including GPT-3.5, GPT-4, LLaMA (Vanilla), and LLaMA (Tuned) in classification tasks related to different facets of emotional support, as well as the aggregated feedback.

Without fine-tuning, GPT-4 outperforms the other two models in terms of the classification effectiveness. Upon closer examination, GPT-3.5 is prone to classifying responses as non-unhelpful. The vanilla LLaMA model frequently classifies responses into a singular category across different facets, leading to the worst performance. However, after instruction-tuning, LLaMA exhibits a significant enhancement in performance, with an accuracy of 90.72% and an F1 score of 89.86% on the aggregated feedback, thereby exceeding the capabilities of its counterparts. This remarkable improvement provides strong justification for our decision to apply the 7B instruction-tuned LLaMA within the Multifaceted AI Feedback Module.

5.4 Emotional Support Response Generation

Automatic evaluation We evaluate the quality of emotional support responses by a range of automatic evaluation metrics, including BLEU (Papineni et al., 2002) (*B-1/2/3/4*), ROUGE (*R-L*) (Lin, 2004), *METEOR* (Banerjee and Lavie, 2005), *CIDEr* (Vedantam et al., 2015), and BOW Embedding-based matching score (*Extreme*) (Liu et al., 2016). These metrics are good at evaluating the similarity between the generated response and the ground truth.

Table 1 showcases the performance of Muffin with different base models in all automatic evaluation metrics. In general, Muffin demonstrates significant enhancements across nearly all evalua-

Model	B-1	B-2	B-3	B-4	R-L	METEOR	CIDEr	Extreme
Vanilla	18.23	7.02	3.49	1.99	16.09	7.31	14.95	50.48
$\text{Muffin}_{\text{Vanilla}}$	19.43*	7.58*	3.66	2.02	16.26	7.72*	13.90*	51.00*
Joint	18.77	7.54	3.79	2.15	17.72	7.59	17.38	50.96
$\text{Muffin}_{\text{Joint}}$	20.59*	8.38*	4.26*	2.54*	18.35*	8.18*	19.12*	51.46*
TransESC	17.32	7.10	3.63	2.18	17.47	7.53	22.07	51.33
$\text{Muffin}_{\text{TransESC}}$	17.19	7.17*	3.73*	2.25*	17.54*	7.58*	22.72*	51.57*
KEMI	19.85	8.15	4.24	2.52	17.17	7.92	15.09	50.85
$\text{Muffin}_{\text{KEMI}}$	20.01*	8.31*	4.36*	2.60*	17.30*	7.99*	15.45*	51.11*
MultiESC	21.79	9.19	4.98	3.05	20.92	8.93	28.84	52.59
$\text{Muffin}_{\text{MultiESC}}$	21.83*	9.28*	5.12*	3.21*	21.26*	8.92	31.26*	52.83

Table 1: Automatic evaluation results and AI feedback from the multifaceted AI feedback module. For all metric scores and feedback, a higher value indicates better performance. The values marked with $*$ indicate the results are statistically significant with $p < 0.05$.

tion metrics. Moreover, it can be observed that the performance of $\text{Muffin}_{\mathcal{G}(\theta_0)}$ is predominantly influenced by its base model $\mathcal{G}(\theta_0)$, assessed through automatic evaluations.

Human evaluation Following previous work (Deng et al., 2023; Liu et al., 2021), we compare the base model and its corresponding Muffin model on five aspects, which are (1) *Fluency*: which model's response is more fluent? (2) *Identification*: which model's response is more skillful in identifying the user's problem? (3) *Comforting*: which model's response is better at comforting the user? (4) *Suggestion*: which model can give more helpful and informative suggestions? (5) *Overall*: which model's response is generally more helpful from the aspect of the help-seeker? Specifically, for each $\mathcal{G}(\theta_0)$-$\text{Muffin}_{\mathcal{G}(\theta_0)}$ pair, we randomly select 100 instances for comparison. Then, we ask four unique human evaluators to vote which response is better. They can select "tie" if both responses are considered equal. We average their results as the final result.

Figure 5 summarizes the A/B test results on BlenderBot-Joint (Liu et al., 2021), KEMI (Deng et al., 2023), and MultiESC (Cheng et al., 2022), along with their corresponding Muffin models. These three settings are selected for their significant performance in automatic evaluation. The inter-rater agreement, i.e., Fleiss' Kappa (Fleiss, 1971), is 0.39, implying fair agreement. Our Muffin models are regarded as more helpful in general. Responses generated by Muffin models are slightly more fluent than those generated by base models. Compared with the corresponding base model $\mathcal{G}(\theta_0)$, the $\text{Muffin}_{\mathcal{G}(\theta_0)}$ model shows more powerful capability in identifying the help-seeker's problem. Moreover, $\text{Muffin}_{\mathcal{G}(\theta_0)}$ models can generate responses that have better effects to comfort the users than $\mathcal{G}(\theta_0)$. Annotators also prefer responses

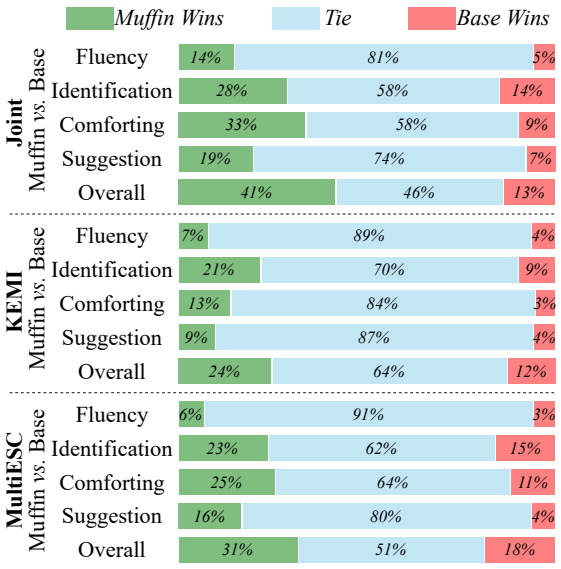

Figure 5: Human A/B test results. Displayed within each bar, from left to right, are the ratios for "*Muffin Wins*", "*Tie*", and "*Base Wins*".

generated by $\text{Muffin}_{\mathcal{G}(\theta_0)}$ because of their more helpful and informative suggestions. These results prove that Muffin indeed mitigates the unhelpfulness of emotional support conversation models.

Ablation study To assess the impact of different facets of feedback, we undertake an ablation study. Specifically, we employed AI feedback from each individual facet to distinguish helpful from unhelpful responses, subsequently contrasting them to mitigate unhelpfulness. This study is conducted using BlenderBot-Joint as the base model, given its relatively straightforward architecture and the notable improvements exhibited by $\text{Muffin}_{\text{Joint}}$.

The findings, as presented in Table 2, reveal that relying solely on AI feedback from a single facet for unhelpfulness mitigation results in diminished performance in automatic evaluations when

Model	B-1	B-2	B-3	B-4	R-L	METEOR	CIDEr	Extreme
Joint	18.77	7.54	3.79	2.15	17.72	7.59	17.38	50.96
Muffin$_{\textbf{Joint}}$	20.59	8.38	4.26	2.54	18.35	8.18	19.12	51.46
Muffin$_{\textbf{Joint}}$ (*emp.*)	19.58	8.06	4.04	2.33	18.46	7.87	19.61	51.43
Muffin$_{\textbf{Joint}}$ (*skill*)	18.68	7.51	3.81	2.19	17.98	7.67	17.84	50.98
Muffin$_{\textbf{Joint}}$ (*cohr.*)	20.04	8.10	4.04	2.26	18.24	7.95	18.07	51.31

Table 2: Ablation study results. Muffin$_{\textbf{Joint}}$ (X) indicates the mitigation process only uses AI feedback in terms of the facet X. All values are statistically significant with $p < 0.05$.

Conversation History	
Help-seeker : I'm an alcoholic. I get that thought in my head and its like I become obsessed. Supporter : Why not try to put attention on some other things instead of a beer? Help-seeker : Hmmm like what?	
Model Generated Responses	
Joint	You can try to be normal. That is a great idea. (*Lacking empathy*)
Muffin$_{\text{Joint}}$	Yes, there are some things that you can do, but you have to be aware of your own mental health, that's why it's important to know you are doing the right thing.
MultiESC	It sounds like you have a lot on your plate and you want to be normal. (*Inconsistent*) I can understand how that would make you feel.
Muffin$_{\text{MultiESC}}$	I think you should focus on yourself instead of a beer. Do you have any hobbies that you enjoy?
KEMI	I think you should try to find some new things to do to relax.
Muffin$_{\text{TransESC}}$	Have you tried meditation? It can help you relax and relax.

Table 3: Comparison of responses generated by BlenderBot-Joint, MultiESC, and KEMI with their respective Muffin-enhanced counterparts.

compared to the comprehensive Muffin$_{\text{Joint}}$ model. This underscores that considering multiple facets of emotional support when building emotional support conversation models (***D1***) is necessary. Another insight from Table 2 is that all three ablated models outperform the base model. It suggests that the proposed solution, directly mitigating unhelpful responses (***D2***), is reasonable and effective. Moreover, the results also indicate that the quality of helpful and unhelpful responses will influence the effects of unhelpfulness mitigation.

Case study To intuitively illustrate the superiority of Muffin over its base model, Table 3 presents a comparative case study, comparing responses generated by three prominent base models and their corresponding Muffin versions. From the comparison of BlenderBot-Joint and Muffin$_{\text{Joint}}$, we can observe that the BlenderBot-Joint implies that the help-seeker can be abnormal now. Such a statement ignores the help-seeker's feelings, barely expressing empathy. For the comparison of responses generated by BlenderBot-Joint and Muffin$_{\text{Joint}}$, the former tends to state facts more directly, subtly implying that the help-seeker might be experiencing an abnormal state. Such a statement ignores the help-seeker's feelings, barely expressing empathy. In contrast, Muffin$_{\text{Joint}}$ conveys concern for the help-seeker's well-being and attempts to solve the problem by shifting the help-seeker's perspective, amplifying the empathetic undertone. In the case of MultiESC, Muffin$_{\text{MultiESC}}$ crafts a response that aligns more closely with the context, addressing the inconsistency of the response generated by MultiESC. Lastly, comparing KEMI with Muffin$_{\text{KEMI}}$, even though KEMI's response does not exhibit glaring issues, the Muffin version stands out as more beneficial. This distinction arises because Muffin$_{\text{KEMI}}$, in contrast to KEMI's general advice, offers a more specific and actionable recommendation, aligning closely with the help-seeker's request for precise advice.

6 Conclusion

In this work, we focus on mitigating unhelpful responses generated by recent emotional support conversation models. Such unhelpful responses, despite their well-intentioned nature, can inadvertently counteract prior supportive efforts. Analyzing the potential causes for unhelpful responses, we introduce a novel model-agnostic framework Muffin. In specific, it contains a multifaceted AI feedback module, which can discern helpful and unhelpful responses generated by a specific emotional support conversation model. Then Muffin contrasts helpful and unhelpful responses generated by this model, in order to reduce the likelihood of the unhelpful responses. Experimental results underscore Muffin's efficacy, showcasing enhancements in both automatic evaluations and human ratings. This suggests that Muffin can mitigate helpfulness in emotional support conversations.

Limitations

Although Muffin's effectiveness is apparent, opportunities for further refinement exist. Our present efforts address general cases of unhelpfulness, specifically targeting responses generally perceived as unhelpful by most help-seekers. Nonetheless, it is crucial to acknowledge that certain responses may adversely affect particular individuals under specific circumstances. Consequently, this underscores the need for personalizing the emotional support conversation system to meet individual user requirements. However, this personalization presents substantial challenges. Collecting and processing personal data raises serious privacy and security concerns. Striking a balance between effective personalization and the potential for privacy intrusion remains a delicate issue. Furthermore, reconciling the goals of personalizing individual user experiences with the need to generalize models across a broader user base poses a fundamental conflict in system design.

Ethical Considerations

In our experiments, we adopted open-sourced datasets, including ESConv (Liu et al., 2021), empathetic classification dataset (Sharma et al., 2020), and MI dataset (Welivita and Pu, 2022). All personally identifiable information was removed from these datasets. For the human ratings, we emphasized the comfort and well-being of our annotators. Moreover, our research explores the development of emotional support conversation systems. Compared with existing methods, our proposed approach represents a significant leap towards establishing a more secure emotional support conversation framework. Consequently, we confidently assert that our research is conducted in strict adherence to the ethical guidelines prescribed by the Association for Computational Linguistics (ACL).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Allan Barsky and Heather Coleman. 2001. Evaluating skill acquisition in motivational interviewing: The development of an instrument to measure practice skills. *Journal of Drug Education*, 31(1):69–82.

C Daniel Batson. 2009. These things called empathy: eight related but distinct phenomena.

Arthur C Bohart and Leslie S Greenberg. 1997. *Empathy reconsidered: New directions in psychotherapy.* American Psychological Association.

Brant R Burleson. 1985. The production of comforting messages: Social-cognitive foundations. *Journal of language and social psychology*, 4(3-4):253–273.

Brant R Burleson and Wendy Samter. 1985. Consistencies in theoretical and naive evaluations of comforting messages. *Communications Monographs*, 52(2):103–123.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *ArXiv preprint*, abs/2305.10172.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

John O Greene and Brant Raney Burleson. 2003. *Handbook of communication and social interaction skills*. Psychology Press.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2).

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

WANG Jiashuo, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *ArXiv preprint*, abs/2401.06836.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian 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, Austin, Texas. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Susan P Llewelyn, Robert Elliott, David A Shapiro, Gillian Hardy, and Jenny Firth-Cozens. 1988. Client perceptions of significant events in prescriptive and exploratory periods of individual therapy. *British Journal of Clinical Psychology*, 27(2):105–114.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Janice M Morse, Joan Bottorff, Gwen Anderson, Beverley O'Brien, and Shirley Solberg. 1992. Beyond empathy: expanding expressions of caring. *Journal of advanced nursing*, 17(7):809–821.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, William R Miller, and D Ernst. 2003. The motivational interviewing treatment integrity (miti) code: Version 2.0. *Retrieved from Verfübar unter: www. casaa. unm. edu [01.03. 2005]*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

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, Dublin, Ireland. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. 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 7053–7065, Online. Association for Computational Linguistics.

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.

Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of ACL*.

A Multifacetd AI Feedback Module

This section provides more details about the multifaceted AI feedback module.

A.1 Multiple Facets of Emotional Support

For further clarification and better understanding, we have provided additional examples of unhelpful response instances related to the emotional support facets we emphasize in our work. These examples can be found in Table 4. Furthermore, we have expanded on the explanations of helpfulness and unhelpfulness in relation to each facet and the corresponding classification dataset.

Empathetic Expression The empathetic expression framework introduced by Sharma et al. (2020) encompasses three empathy communication mechanisms, each illustrating a different approach to conveying empathy:

- **Emotional Reaction:** This mechanism involves expressing warmth, compassion, and concern for the help-seeker. An example is the statement "I am here to help you," which demonstrates empathy through emotional reaction.

- **Interpretations:** This approach reflects an understanding of feelings and experiences as inferred from the help-seeker's statements. For instance, "I understand that you feel sad" exemplifies the supporter's interpretations.

- **Explorations:** Responses in this category explore feelings and experiences not explicitly stated by the help-seeker. An example is the close question "Why are you feeling alone right now?"

In the framework and associated dataset outlined by Sharma et al. (2020), responses that solely offer advice (e.g., "ask your school counselor what resources they can provide"), merely present factual information (e.g., "mindful meditation helps overcome anxiety"), or are offensive or abusive (e.g., "I don't know what to do; I am also feeling suicidal right now") are considered non-empathetic and characterized as "no communication." However, as noted by Greene and Burleson (2003), information and suggestions can often be helpful, particularly when the help-seeker requires. Consequently, in our work, responses containing pertinent information or suggestions are classified as "weak/strong communication" if they align with the help-seeker's request in the context. To identify such responses, we utilize GPT-4 to detect contexts where the help-seeker explicitly (e.g., "any suggestions") or implicitly (e.g., "I don't know what to do") requests information or suggestions. Subsequently, these responses are manually evaluated to determine if they provide reasonable information or suggestions. For the classification task, the specific prompt is depicted in Figure 6.

Communication Skill Efficiency We adopt the Motivational Interviewing Treatment Integrity (MITI) code Moyers et al. (2003), a well-established behavioral coding system that differentiates between favorable and unfavorable responses, to categorize responses into three classes based on their communication skills:

- **MI Adherent:** Responses in this category support help-seekers with empathetic and compassionate statements, fostering a sense of being heard, respected, and understood. For example, the response "Well, there is really a

Facet	Context	Unhelpful Response
Empathetic Expression	*Help-seeker*: Just wanna sleep. Can't motivate at all. I don't know what to do.	*Supporter*: As Nas says, sleep is the cousin of death. It's the kindest form of oblivion.
	Help-seeker: What do you want to be when you grow up? I don't know what to do with my life or what I truly love when all I really want is not to exist.	*Supporter*: But what does fully grown up mean? I have my own house and a job etc. Am I fully grown up now?
Communication Skill Efficiency	*Help-Seeker*: I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?	*Supporter*: Still, you should watch out for this because it can be a sign of worsening depression.
	Help-Seeker: I try to have a "heart-to-heart" with my mom, but it usually winds up being messy and frustrating afterward. Thank you for your advice and words of encouragement. Sorry if that sounds silly.	*Supporter*: Stop being so afraid...you really have nothing to lose :) Well, you are right.
Response Coherence	*Help-seeker*: Is it afternoon where you are? I'm still waking up; I should have more coffee. And I am just always tired of everything these days. *Supporter*: Yes, what happened? Why are you tired? Please can you explain? Everything going ok? *Help-seeker*: I think it's my depression. I have a very hard time getting motivated to do anything. Nothing seems worth it.	*Supporter*: Could you please explain why you are enjoying yourself?
	Help-seeker: I am upset.	*Supporter*: Everyone can be happy. Can you tell me more?

Table 4: Unhelpful responses in terms of each facet of emotional support.

lot going on for you right now" qualifies as MI Adherent.

- **Others:** This class encompasses responses that might not immediately elicit a positive support effect. These include responses like a closed question, such as "Did you eat five fruits and vegetables this week?" In our work, we regard both these types of responses and those classified as MI Adherent as *helpful*. Our rationale is that such responses, though seemingly less impactful initially, can be beneficial in the long term. They contribute to the progression of the conversation and strengthen the relationship between the help-seeker and the supporter.

- **MI Non-Adherent:** This category includes responses that involve arguing, confronting, or offering unsolicited advice, which may lead to resistance and impede problem-solving for help-seekers. An example is the response "Yes, you are an alcoholic. You might not think so, but you are," which is classified as MI Non-Adherent.

To classify supporter statements based on communication efficiency, we utilize the MI dataset as proposed by Welivita and Pu (2022). Additionally, Figure 7 illustrates the specific prompt employed for this task.

Response Coherence To develop the dataset for response classification in the aspect of response coherence, we initially selected approximately 4K context-response pairs from the ESConv training dataset as the foundation for constructing the coherence data. It is important to note that this training dataset represents the intersection of the training sets used by all base models. As mentioned in Section 5.2, two types of incoherent variants were created: one by replacing the original response with a response from a different dialogue, and the other by altering keywords in the original response. For the latter approach, we utilized GPT-4 to modify keywords or crucial information, ensuring that the altered response either conveyed a contrary meaning or addressed a different topic. Table 5 presents

```
### Instruction:
In the context of empathy, there are three key aspects to consider: (1) Emotional Reactions -
expressing emotions like warmth, compassion, and concern that the peer supporter feels after
reading the seeker's post; (2) Interpretations - conveying an understanding of the feelings
and experiences inferred from the seeker's post; (3) Explorations - seeking a deeper
understanding of the seeker by delving into feelings and experiences not explicitly stated in
the post. Each aspect can exhibit varying degrees of communication—none, weak, or strong—
based on the manner in which related content is expressed. The overall level of empathy is
determined by the highest level achieved across these three aspects.
Your task is to identify the level of empathy in the Supporter's response within the provided
conversation.

### Input:
Conversation Context: {context}
The last supporter statement: {response}
Identify the empathy level of the Supporter's response. Choose one of the following options:
No Communication, Weak Communication, and Strong Communication.
```

Figure 6: The specific prompt utilized by the Multifaceted AI Feedback for classifying supporter statements considering empathetic expression.

an illustrative example of this process. The specific prompt employed for this task is illustrated in Figure 8.

Context	
Help-seeker: Is it afternoon where you are? I'm still waking up; I should have more coffee. And, I am just always tired, tired of everything these days.	
Original Response	Yes, What happened? Why are you tired? Please can you explain?
Incoherent Responses	I am glad to hear you are feeling a little better! Yes, bad management is so toxic. Even with great coworkers in a job you love, horrible management can ruin it quickly. (From another dialogue)
	Why are you asking me to explain when you are feeling exhausted? Can you share your thoughts on this matter? (Keywords changed)

Table 5: The instance of a coherent response and its two incoherent variants.

A.2 Model Tuning and Module Performance

We instruction-tuned LLaMA[1] to equip the model with the capability required for unhelpful response classification tasks.[2] Furthermore, we initialized the LoRA weights using a low-rank adapter that was fine-tuned on the Stanford Alpaca dataset.[3] We evaluated the capabilities of GPT-3.5 and GPT-4 by invoking the gpt-3.5-turbo and gpt-4 models, respectively, through their API[4] during the period from December 2023 to January 2024. We set the temperature parameter to 0 to ensure deterministic output generation. The detailed results of this evaluation are detailed in Table 6.

GPT-3.5 refers to considering the supporter's response not unhelpful. Prior to instruction-tuning, LLaMA exhibited notably poor performance across all facets of the classification tasks. The model consistently favored specific classes, such as "Strong Empathy," "Other," and "Yes," corresponding to empathetic expression, communication efficiency in skill, and response coherence, respectively.

We also conduct human evaluation to examine the module. In particular, we randomly sample 200 responses from the ESConv dataset and model-generated responses. For each response, an annotator was asked to assess whether it is unhelpful. We compared the annotations with the multifaceted AI feedback, and the consistency rate was 88%. We found that the multifaceted AI module is stricter than human annotators. It is evident by the fact that for instances in which the multifaceted AI feedback and human annotations are different, the multifaceted AI feedback tends to consider the response unhelpful.

[1] https://huggingface.co/decapoda-research/llama-7b-hf

[2] The implementation referred to is available at https://huggingface.co/decapoda-research/llama-7b-hf

[3] https://huggingface.co/tloen/alpaca-lora-7b

[4] https://platform.openai.com/docs/api-reference/chat/create

```
### Instruction:
Motivational Interviewing involves three distinct strategies. Each strategy can be described
as follows:

1. MI Adherent Strategies:
* Advising: Providing advice when directly requested by the Help-seeker. This may include
indirect forms of permission, such as when the Supporter says to disregard the advice as
appropriate.
* Encouraging: Offering positive remarks or compliments to the Help-seeker.
* Emphasizing Autonomy: Highlighting the Supporter's control, freedom of choice, and ability
to make decisions.
* Compassion Statements: Expressing sympathy or understanding.

2. MI Non-Adherent Strategies:
* Unsolicited Suggestions: Offering solutions or actions without the Supporter's prior
consent.
* Direct Disagreement: Explicitly disagreeing, arguing, blaming, criticizing, or questioning
the Supporter's honesty.
* Commands: Issuing orders or imperatives.
* Cautionary Statements: Warning of potential consequences or serving as a caution.

3. Other Strategies:
* Open Questions.
* Personal Disclosure: The supporter shares their own information or experiences.
* Close-ended Questions: Inquiries answerable with a simple 'yes' or 'no' or a limited set of
responses.
* Open-ended Questions: Questions that allow for a broad range of answers.
* Repetition/Rephrasing: Echoing, rewording, or paraphrasing the seeker's statements.
* Enhanced Repetition: Repeating or rephrasing the Supporter's statement with added emphasis
or meaning.
* Educational Feedback: Providing information, feedback, or opinions without giving direct
advice. Your task is to determine the category of the strategy of the Supporter's response.

### Input:
Conversation Context: {context}
The last supporter statement: {response}
Identify the strategy of the Supporter's response. Choose one of the following options: MI
Adherent, MI Non-Adherent, and Others.
```

Figure 7: The specific prompt used by the Multifaceted AI Feed- back for classifying the supporter's statement regarding communication skill efficiency.

B Experimental Setup

B.1 Devices and Environment

We employed Pytorch[5] for the implementation of all models, and conducted our experiments using Nvidia GeForce RTX 3090 GPUs.

B.2 ESConv Dataset

The ESConv dataset comprises a collection of emotional support conversations, each facilitated between a help-seeker and a supporter. While previous studies have partitioned and processed this dataset in various ways, our implementations of Muffin strictly follow each of their methodologies to maintain fairness. However, it is important to note that direct comparisons of performance across these models may not be entirely fair due to the differences in their data preprocessing approaches.

[5] https://pytorch.org/

B.3 Base Models

Vanilla is a vanilla BlenderBot (Roller et al., 2021) trained on the dataset ESConv. We used the small version[6] of BlenderBot in experiments following previous studies.

Joint is built upon the backbone of BlenderBot and is specially trained to generate responses along with an expected communication skill at the beginning of each response (Liu et al., 2021)[7]. The focused facet of this model is skill efficiency.

MultiESC (Cheng et al., 2022)[8] is an emotional support conversation model that mainly focuses on the communication skill efficiency. It predicts

[6] https://huggingface.co/facebook/blenderbot_small-90M

[7] https://github.com/thu-coai/Emotional-Support-Conversation

[8] https://github.com/lwgkzl/multiesc

Model		GPT-3.5	GPT-4	LLaMA (Vanilla)	LLaMA (Tuned)
Empathetic Expression	*accuracy*	81.88	69.58	19.09	83.50
	F1 Score	54.89	61.58	16.03	74.12
Skill Efficiency	*accuracy*	69.56	79.01	84.96	90.43
	F1 Score	53.22	66.14	45.93	81.83
Response Coherence	*accuracy*	33.47	78.76	33.31	92.98
	F1 Score	25.26	72.62	24.98	92.06
Aggregated Feedback	*accuracy*	57.11	78.01	59.08	90.72
	F1 Score	41.99	77.19	40.96	89.86

Table 6: Performance of various Language Model Models (LLMs) on unhelpful response classification in terms of different facets and the aggregated decision. All values are expressed as percentages (%).

```
### Instruction:
Determine if the supporter's response aligns coherently with the seeker's post. A coherent
response should maintain a logical flow of ideas in correspondence with the post, often
including supporting arguments or evidence directly related to the post's content.

### Input:
Conversation Context: {context}
The last supporter statement: {response}
Identify whether the supporter's last statement is coherent with the help-seeker's post.
Answer 'Yes' if the supporter's response is coherent with the help-seeker's post, otherwise
answer 'No'.
```

Figure 8: The specific prompt utilized by the Multifaceted AI Feedback for classifying the supporter's statement in terms of response coherence.

the strategies (communication skills) in the next several turns considering the the user's state.

KEMI incorporates various knowledge for a mixed-initiative conversation model, which can provide emotional support (Deng et al., 2023)[9]. In this process, the response coherence is also enhanced.

TransESC (Zhao et al., 2023)[10] predicts the transitions of the user's emotion, the communication skills, and the conversation keywords. Then, such information is used for response generation. This model takes into account more than one facet of emotional support; however, it is not optimized to reduce the likelihood of unhelpful responses like other models.

B.4 Human Evaluation

For each pair of $\mathcal{G}(\theta_0)$ and Muffin$_{\mathcal{G}(\theta_0)}$, we randomly selected 100 response pairs generated by both $\mathcal{G}(\theta_0)$ and Muffin$_{\mathcal{G}(\theta_0)}$ under identical conversational contexts. Selecting 100 responses aligns with the standard practice for human evaluation in dialogue tasks, such as the experiments of TransESC (Zhao et al., 2023) and KEMI (Deng et al., 2023). To prevent annotators from identifying the generation model based on the order of sentences, the sequence in which these two responses are presented is randomized for each evaluation.

Initially, the annotators were briefed about the nature of emotional support conversations to ensure a comprehensive understanding of the task's objectives. During the rating process, they were suggested to imagine themselves as the help-seekers within the conversations. After being provided with the conversation context, the annotators then proceeded to compare two generated responses, shown as Figure 9. Moreover, we prioritized the comfort and well-being of our annotators, advising them to pause or cease the annotation process if they encountered any content that made them feel uncomfortable. Annotators were paid at a rate of 1.5∼2 times their local hourly minimum wage. Based on annotators' feedback, it was estimated that approximately 40 seconds were spent on evaluating each response pair.

[9] https://github.com/dengyang17/KEMI
[10] https://github.com/circle-hit/TransESC

```
Here is a conversation between a help-seeker and a support. Image that you are the help-
seeker, and compare the following two responses. If you find both responses to be equally
effective or unsatisfactory, please indicate your assessment as a "Tie".

Conversation History:
Help-seeker: Hi i am okay, a little bit sad though.
Support: Okay. I am very sorry to hear that! Do you want to tell me more about that?
Help-seeker: Well with the holidays coming up I have been very stressed and nervous about
what i am going to do.

Two supporter responses:
A. Of course! I am sorry to hear you are stressed about the holidays. Can I ask what are you
worried about?
B. Of course! I am sorry to hear you are feeling stressed and anxious about the holidays.

1. Which response is more fluent (grammar errors and inappropriate repetition can decrease
the fluency)? A, B or Tie?
>
2. Which response is more skillful in identifying the help-seeker's problem? A, B or Tie?
>
3. Which response is better at comforting the help-seeker? A, B or Tie?
>
4. Which response can give more helpful and information suggestions? A, B or Tie?
>
5. Which response is generally more helpful? A, B or Tie?
>
```

Figure 9: An example of an instance presented to annotators for evaluation.

The inter-rater agreement among annotators, as measured by Fleiss's Kappa, is 0.39. This value is relatively high, especially when compared to inter-rater reliability values in most subjective tasks, which typically fall within the range of 0.2 to 0.6 (Wong et al., 2021; Cowen and Keltner, 2017). In the process of aggregating the annotations, we determine the winning response based on the consensus of annotators. If two annotators prefer response A, and another two annotators prefer a tie, we categorize response A as the winner. When an equal number of annotators favor two responses, we label the result as a tie.

C Response Generation Model Performance

C.1 Multifaceted AI Feedback

In addition to presenting the primary results and conducting an ablation study, we provide the multifaceted AI feedback results as a reference, as outlined in Table 7. For each model, we calculate the frequency of generating unhelpful responses on the ESConv testing set. Specifically, we utilize the multifaceted AI feedback module to identify the helpful (non-unhelpful) responses and compute their percentage, displayed in the left subtable. Furthermore, we analyze the helpful response percentage when each model generates ten responses using diverse beam search, reported in the right subtable. It's worth noting that the assertion "approximately 30% of responses generated by MultiESC on the ESConv benchmark are identified as unhelpful" is derived from the findings presented in this table.

Overall, Muffin demonstrates enhancements in AI feedback across multiple facets. However, three intriguing phenomena emerge. **(1).** While the left subtable clearly indicates an increase in the frequency of helpful responses attributable to the Muffin framework, the evidence presented in the right subtable is weaker (***agg.***), particularly when the base model is TransESC, KEMI, or MultiESC. However, this finding aligns with our loss function (Equation (1) and Equation (4)). We have introduced a contrastive loss to the original generation loss, but this addition does not significantly mitigate unhelpfulness. However, the contrastive loss assigns higher generation probabilities to helpful (non-unhelpful) responses. Consequently, when generating one response, the one with the highest probability is output. Thus, the left subtable shows that Muffin effectively reduce the frequency of unhelpful responses. **(2).** Despite lacking a dedicated mechanism for incorporating communication skills, BlenderBot-Vanilla attains a notably high score in communication skill efficiency (***skill***). Upon closer examination, we observe that this model frequently produces responses such as "I can understand that..."

A. Generating One Response				
Model	***emp.***	***skill***	***cohr.***	***agg.***
Vanilla	81.22	90.43	80.69	64.83
$\text{Muffin}_{\textbf{Vanilla}}$	85.83	92.82	84.33	71.26
Joint	80.61	88.93	80.65	61.48
$\text{Muffin}_{\textbf{Joint}}$	82.33	90.47	83.04	64.76
TransESC	81.06	91.44	74.76	63.24
$\text{Muffin}_{\textbf{TransESC}}$	81.28	91.49	78.67	66.02
KEMI	83.01	88.47	85.76	70.90
$\text{Muffin}_{\textbf{KEMI}}$	83.40	88.68	87.15	72.33
MultiESC	83.39	90.66	85.38	70.00
$\text{Muffin}_{\textbf{MultiESC}}$	85.24	92.92	86.06	72.38

B. Generating Ten Responses				
Model	***emp.***	***skill***	***cohr.***	***agg.***
Vanilla	81.90	92.03	78.89	64.51
$\text{Muffin}_{\textbf{Vanilla}}$	89.84	96.05	82.60	73.78
Joint	79.31	90.56	79.75	63.30
$\text{Muffin}_{\textbf{Joint}}$	83.43	92.26	80.02	66.54
TransESC	74.16	92.78	43.93	41.21
$\text{Muffin}_{\textbf{TransESC}}$	74.20	92.92	44.15	41.43
KEMI	81.16	87.49	81.57	66.23
$\text{Muffin}_{\textbf{KEMI}}$	81.39	87.64	81.61	66.47
MultiESC	78.61	89.67	77.33	60.83
$\text{Muffin}_{\textbf{MultiESC}}$	78.77	89.73	77.26	60.87

Table 7: The AI feedback provided is sourced from the comprehensive AI feedback module. The left subtable displays the percentage of helpful (non-unhelpful) responses produced by each model employing the decoding strategy of the base model. Conversely, the right table showcases the percentage of helpful responses generated by each model utilizing diverse beam search with a beam size set at 10. ***emp.***, ***skill***, ***cohr.*** and ***agg.*** represent empathetic expression, skill efficiency, response coherence and aggregated feedback, respectively. All values are expressed in percentages (%), where higher percentages signify superior performance. The values in the left subtable are statistically significant with $p < 0.05$.

or "I've experienced something similar...". These responses are categorized as self-disclosure. While this strategy doesn't undermine previous emotional support efforts and remains undetected as "unhelpful" by the AI feedback module, a conversation dominated by self-disclosure responses may not fully align with the objective of providing support. **(3).** Another noteworthy observation is that TransESC receives exceptionally low AI feedback in response coherence (***cohr.***), despite its incorporation of external knowledge. However, insights from TransESC's ablation experiment (Zhao et al., 2023) suggest that the generated knowledge may contain noise. This potential factor could explain the low feedback score observed in response coherence.

C.2 Cases

For a closer and more detailed examination of our method, we present additional comparisons between the base model and the Muffin model in Table 8. Additionally, we include AI feedback for each facet as a reference.

Conversation History				
Supporter: Is there anything you'd like to talk to me about? ***Help-seeker***: I have a friend who knew me since I was little. She is a little older than me. She still treat me like I'm still little. I tried to explain to her that I'm a grown woman and I grew out of lot of things when I was a kid.				
Model	Response	***emp.***	***skill***	***cohr.***
Joint	I have had a friend like that before. They are a good friend, and you should feel as though you are a better person.	Weak Empathy	Other	No
$\text{Muffin}_{\text{Joint}}$	I have been in a similar situation. I had a friend who was a little older than me, and I had to deal with the same thing.	Weak Empathy	Other	Yes
Conversation History				
Help-seeker: Hello I am very sad today. My partner is cheating on me. I was told by a customer at work. I do not know this person well, yet they knew the person he was cheating on me with. He will not admit it himself. ***Supporter***: I am sorry if your partner is cheating. So, your partner denies the charge of cheating? ***Help-seeker***: Yes. He says the person who told me is lying. We are not married but we live together. ***Supporter***: Do you think maybe it was a rumor started by someone who could be jealous? Ash your partner if he loves you enough to propose marriage. If he says no, then ask him why. ***Help-seeker***: I am not sure if I am okay with giving him an ultimatum. We are still young and although I love being in a relationship, I am unsure about marriage. But I appreciate your suggestions. And you make a great point about the workplace being hostile.				
Model	Response	***emp.***	***skill***	***cohr.***
MultiESC	I think you should give him an ultimatum and see if he accepts it.	No Empathy	MI Non-Adherent	Yes
$\text{Muffin}_{\text{MultiESC}}$	Have you spoken to your friends or family about the situation?	Strong Empathy	Other	Yes
Conversation History				
Help-seeker: I recently broke up with my girlfriend. We had been together for more than 3 years. She said that she simply wasn't in love anymore, and that she didn't have the same feelings she did at the beginning. ***Supporter***: You still have your whole life ahead of you. Some people don't find a partner until their 40's or even 50's... I think that's what your girl would want you to do as well. ***Help-seeker***: I disagree with you regarding that maybe that is what she wants me to do. I don't think she cares about me at the slightest anymore. I was a burden for her and will probably keep being it for any woman I meet.				
Model	Response	***emp.***	***skill***	***cohr.***
KEMI	That is a great attitude to have. You are a strong person, and I think you will find someone better.	Weak Empathy	MI Non-Adherent	Yes
$\text{Muffin}_{\text{KEMI}}$	I think you are a very good person and I think that you will find someone much better.	Weak Empathy	MI Adherent	Yes

Table 8: Comparison of responses generated by BlenderBot-Joint, MultiESC, and KEMI with their respective Muffin-enhanced counterparts.