STAR: Strategy-Aware Refinement Module in Multitask Learning for Emotional Support Conversations

Suhyun Lee, Changheon Han, Woohwan Jung, Minsam Ko
Department of Applied Artificial Intelligence, Hanyang University
{su7561632, datajedi23, whjung, minsam}@hanyang.ac.kr

Abstract

Effective emotional support in conversation requires strategic decision making, as it involves complex, context-sensitive reasoning tailored to diverse individual needs. The Emotional Support Conversation framework addresses this by organizing interactions into three distinct phases—exploration, comforting, and action—which guide strategy selection during response generation. While multitask learning has been applied to jointly optimize strategy prediction and response generation, it often suffers from task interference due to conflicting learning objectives. To overcome this, we propose the **St**rategy-**A**ware **R**efinement Module (**STAR**), which disentangles the decoder's hidden states for each task and selectively fuses them via a dynamic gating mechanism. This design preserves task-specific representations while allowing controlled information exchange between tasks, thus reducing interference. Experimental results demonstrate that STAR effectively reduces task interference and achieves state-of-the-art performance in both strategy prediction and supportive response generation.

1 Introduction

Approximately one in ten people worldwide experiences a mental disorder, yet only 1% of the global health workforce is dedicated to mental health care, with the most acute shortages found in developing countries (Freeman, 2022; Jack et al., 2014; Collaborators et al., 2022; Rathod, 2017; World Health Organization, 2021). For instance, while the global average is about 3.96 psychiatrists per 100,000 people, countries such as Ethiopia (0.04), Nigeria (0.06), Pakistan (0.19), and India (0.30) fall drastically below this benchmark (Rathod, 2017; World Health Organization, 2021). This stark disparity underscores the urgent need for scalable and accessible forms of support, particularly in low-resource settings where mental health professionals

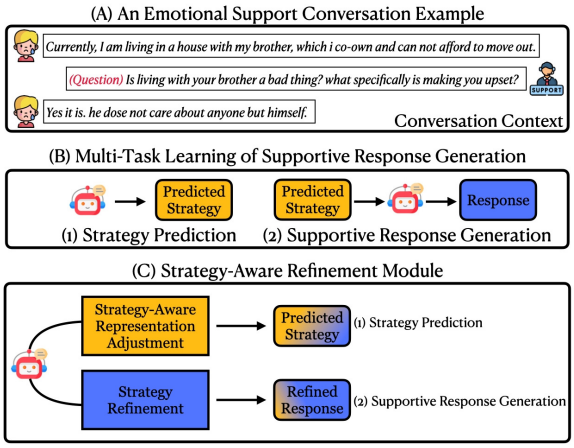


Figure 1: (A) shows an emotional support conversation example, highlighting the dual tasks of strategy prediction and supportive response generation. (B) illustrates the multi-task learning framework, and (C) presents the STAR module that refines hidden representations to mitigate task interference.

are scarce. Emotional support, especially when integrated into community-based and non-specialist-delivery interventions, has emerged as a critical component in addressing this global care gap.

To address this shortage, the World Health Organization (WHO) introduced the *mhGAP Intervention Guide*, which equips non-specialist providers in primary care with tools to deliver basic psychosocial interventions, such as structured interviews and problem solving therapy (Ojagbemi et al., 2022). The success of such community-based programs is evident in real-world implementations. In Zimbabwe, for example, the *Friendship Bench* program trained lay health workers to provide emotional support through problem-solving therapy, reducing depression to under 14% after six months (Abas et al., 2020). Similarly, in Pakistan, *Lady Health Workers* offering home-based cognitive behavioral techniques reduced postpartum depression to 27% compared to 59% in control groups after one year (Rahman et al., 2023). These examples demon-

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 272–284

strate that even in the absence of clinical experts, structured and empathetic emotional support can significantly improve mental health outcomes.

In response to the global need for scalable mental health solutions, researchers have begun to explore artificial intelligence as a promising tool to provide emotional support, particularly in low-resource environments (Liu et al., 2021). However, effective emotional support is not simply a matter of generating empathetic responses—it requires nuanced understanding, contextual sensitivity, and adherence to structured support strategies (Burleson, 2003).

To meet these complex requirements, recent AI research has turned to multitask learning (MTL) as a foundational framework for emotional support systems. MTL enables AI models to jointly learn multiple interrelated tasks, such as detecting user emotional state, selecting appropriate support strategies, and generating empathy responses. This integrated learning process allows for more context-aware and consistent support delivery. Notably, several recent studies have successfully implemented MTL architectures to improve the quality and effectiveness of AI-generated emotional support (Tu et al., 2022; Zhou et al., 2023; Peng et al., 2022; Cheng et al., 2022; Zhao et al., 2023; Deng et al., 2023; Xu et al., 2024; Li et al., 2024a). These approaches demonstrate that MTL can be a powerful mechanism for aligning AI responses with structured supportive strategies found in human-led interventions.

However, while the MTL approach is designed to leverage shared information across tasks to enhance learning efficiency, it can sometimes lead to adverse effects (Zhao et al., 2018). This issue arises due to task interference, where the representational requirements of different tasks may be inherently misaligned (Gurulingan et al., 2022a), or when conflicting gradients from multiple tasks disrupt the optimization process during backpropagation (Yu et al., 2020). As a result, instead of facilitating knowledge transfer, MTL can sometimes hinder model performance by introducing conflicts between tasks.

To mitigate task interference, various approaches have been proposed, including independent subnets to isolate task-specific representations (Strezoski et al., 2019), task-specific parameterization to adjust the model capacity per task (Kanakis et al., 2020), and task grouping to cluster related tasks and reduce negative transfer (Gurulingan et al., 2022b). However, despite these advancements, effective interference suppression strategies tailored to the Emotional Support Conversation (ESC) domain—particularly for response strategy selection and supportive response generation—remain an open challenge.

To address these limitations, we propose the Strategy-Aware Refinement (STAR) module, which effectively mitigates task interference between strategy prediction and supportive response generation while leveraging contextual and strategic cues. STAR consists of two key components: Strategy-Aware Representation Adjustment (SARA) and Strategy Refinement (SR). Specifically, SR splits the decoder's hidden states into two separate representations—one dedicated to strategy prediction and the other to supportive response generation. To prevent unnecessary entanglement between these two tasks, SARA dynamically integrates the representations only when necessary, ensuring that strategy-related signals remain distinct from linguistic representations. This design prevents the overmixing of strategy cues with linguistic features, allowing each task to fully exploit its unique strengths. As a result, our approach effectively minimizes task conflicts and consistently outperforms existing methods.

Our work makes two key contributions:

- We provide an in-depth analysis revealing that existing multitask learning models for emotional support conversations frequently suffer from task interference, characterized by conflicting gradients and entangled representations.

- We propose the STAR module, which effectively mitigates interference between strategy prediction and supportive response generation by dynamically adjusting hidden state representations. Our approach preserves the distinctiveness of strategy-related signals, reducing negative transfer between tasks.

- By minimizing task conflicts, our approach improves both strategy prediction accuracy and the quality of supportive response generation. Experimental results validate these improvements, demonstrating substantial gains over existing methods.

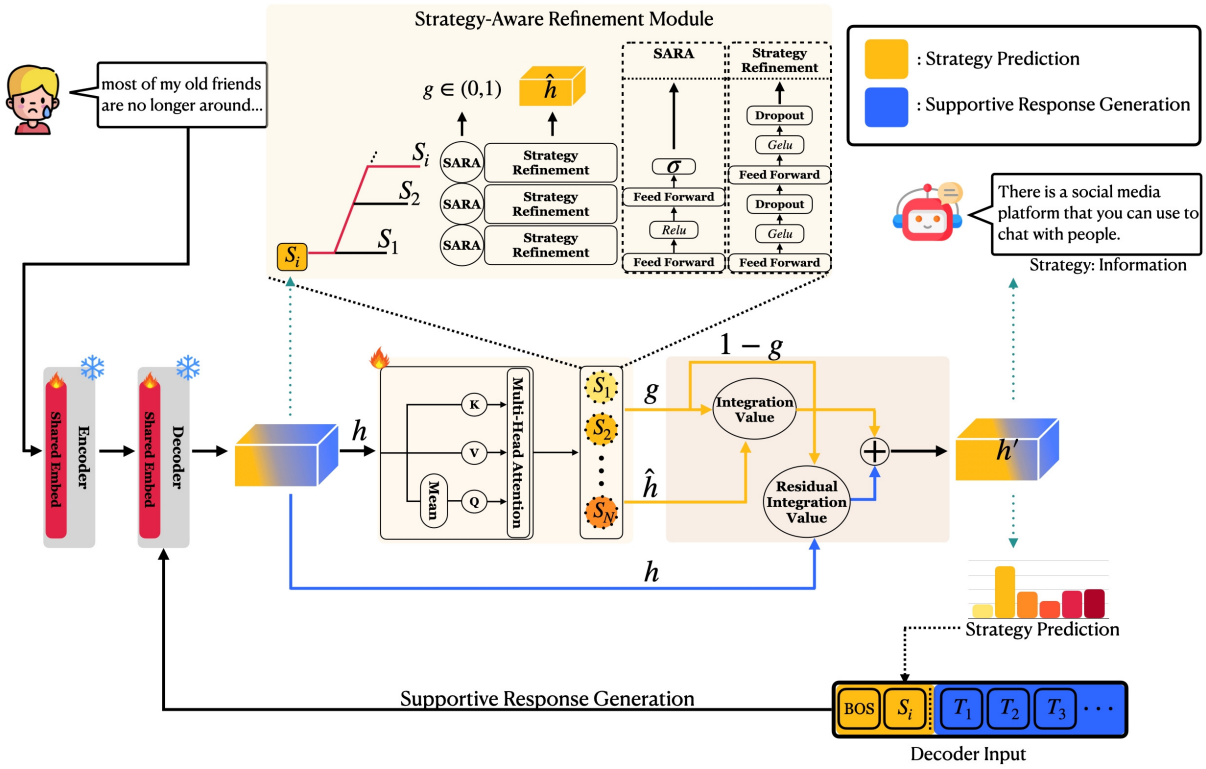


Figure 2: Overall architecture of the STAR module for emotional support conversation. Decoder hidden states from a fine-tuned BlenderBot-Small model are pooled and then fed into two parallel submodules: one computes an integration value, and the other refines the hidden state. The STAR uses the integration value to balance the refined and original hidden states, yielding a strategy-refined state for response generation.

2 Related Work

2.1 Emotional Support Conversation

ESC has gained increasing attention as a dialogue task that requires models to deliver empathetic, contextually appropriate responses aligned with the user's emotional needs (Ramírez, 2024; Van der Zwaan et al., 2012; Zhou et al., 2020). Building on the phase-based framework of *Exploration*, *Comforting*, and *Action* (Liu et al., 2021), recent studies have explored structured strategy-guided generation to improve support quality.

2.2 Multitask Learning for ESC

To jointly optimize strategy prediction and response generation, most ESC systems adopt MTL as a central modeling framework. A notable trend involves enhancing these models with external commonsense knowledge via COMET (Bosselut et al., 2019), enabling improved contextual reasoning and response alignment (Liu et al., 2021; Tu et al., 2022; Zhou et al., 2023; Peng et al., 2022; Cheng et al., 2022; Zhao et al., 2023; Deng et al., 2023; Li et al., 2024a).

Many studies introduce auxiliary subtasks to reinforce strategic alignment. For instance, emotional change prediction (Li et al., 2024a; Zhou et al., 2023), primary cause identification (Peng et al., 2023), and backward decoding for historical context refinement (Xu et al., 2024) have all been proposed to support better decision making during response generation. These techniques aim to improve the model's interpretability and adaptability in emotionally complex scenarios.

2.3 Task Interference

Despite these advances,multitask ESC models continue to face task interference, where overlapping or conflicting gradients between tasks hinder optimization and degrade performance. Common mitigation strategies include allocating task-specific parameters within the encoder(Liu et al., 2019), isolating task-specific subnets (Strezoski et al., 2019; Kanakis et al., 2020), or grouping related tasks during training (Gurulingan et al., 2022b).

However, these approaches face fundamental limitations when applied to the unique setting of ESC. First, the predicted strategy directly guides the response generation, resulting in a strong interdependence between the two tasks—unlike general

MTL settings, where tasks are typically independent. Second, strategy prediction must always precede response generation, making it essential to preserve the sequential order and ensure accurate information flow between the tasks.

Consequently, the structural nature of ESC tasks renders common mitigation strategies ineffective. For instance, task-specific subnetworks isolate tasks completely, limiting the necessary information exchange between strategy prediction and response generation—an interaction that is essential in ESC. While task grouping may initially seem reasonable given the superficial similarity between the two tasks, their underlying objectives—strategic reasoning and linguistic generation—are fundamentally different, reducing the effectiveness of such an approach. Similarly, gradient projection methods address interference only at the gradient level, which falls short in ESC, where fine-grained control and explicit separation at the representation level are crucial.

Therefore, effective ESC modeling requires an architecture that separates task representations, integrates information flexibly, and preserves the sequential flow from strategy prediction to response generation. The STAR module fulfills these needs by reducing interference and enhancing both strategic alignment and response fluency, making it a more suitable solution than conventional MTL methods.

3 Method

3.1 Overview

Emotional support conversation generation involves two tightly coupled tasks: predicting a suitable support strategy and generating a contextually appropriate response. In each decoding cycle, the model first predicts a strategy token and subsequently generates a response conditioned on it. This coupling often leads to task interference, as the two tasks require diverging representational features.

To address this challenge, we propose the STAR module, which dynamically regulates task-specific knowledge integration. By disentangling and selectively fusing hidden states via gating, STAR minimizes interference and enhances strategic coherence during response generation. The model ultimately maximizes the conditional probability:

$$\max p(Y \mid X, s, \tau'), \tag{1}$$

where X is the dialogue history, s is the situation description, and τ' is the refined strategy token generated by STAR.

3.2 Strategy-Aware Refinement Module

As shown in Figure 2, STAR is integrated into a BlenderBot-based decoder and consists of two components: SARA and SR.

Input Processing Given a dialogue context, the decoder produces hidden states $h \in R^d$. A predicted strategy token $s \in N$ is appended to guide generation. These are then processed by SARA to extract a global representation.

SARA A shared attention pooling layer first computes the contextual summary:

$$z = Pooling(h). \tag{2}$$

This vector is passed through a two-layer feedforward network with ReLU and sigmoid activations to compute a gating value $g \in (0, 1)$:

$$g = \sigma(f(z)). \tag{3}$$

SR The same pooled vector z is transformed into a refined strategy embedding $\hat{h} = P(z)$ using a separate two-layer network. The final representation for response generation is a gated combination:

$$h' = g \odot \hat{h} + (1 - g) \odot h. \tag{4}$$

This formulation enables targeted injection of strategic information while preserving fluency and contextual relevance.

3.3 Model Training

To jointly optimize strategy prediction and response generation, we define two separate objectives and integrate them via a dynamic weighting scheme. The model generates a response $\mathbf{r} = \{r_1, r_2, \ldots, r_{|\mathbf{r}|}\}$ conditioned on the STAR-refined strategy token τ', given input context c and situation s.

Loss Functions The response generation loss is defined as the negative log-likelihood:

$$\mathcal{L}_{LM} = -\sum_{t=1}^{n_r} \log p(r_t \mid r_{<t}, c, s, x), \tag{5}$$

and the strategy prediction loss is:

$$\mathcal{L}_{ST} = -\log p(\tau' \mid c, s, x). \tag{6}$$

Model	Cos. Sim. w/ Resp. Loss			
BlenderBot-Joint	Strategy	-	-	-
+STAR	**0.47**	-	-	-
BlenderBot-Joint	Strategy	-	-	-
	-0.05	-	-	-
Emstremo	G	E	V	CONT
	-0.03	-0.01	0.02	-0.07
TransESC	STR	EMO	SEN	-
	-0.04	0.01	0.01	-

Table 1: Cosine similarity between response loss and task-specific losses across models. Higher values indicate lower gradient interference and more stable multi-task optimization.

Dynamic Loss Weighting To handle the different convergence rates of the tasks, we use a dynamic factor λ that increases over training epochs:

$$\lambda = \lambda_0 \cdot \frac{\log(E+1)}{\log(E_{\max})}, \tag{7}$$

where E is the current epoch and λ_0 is a scaling constant. This allows the model to prioritize fluent generation early on, and shift attention to strategic accuracy in later stages.

Final Objective The total loss is a weighted combination:

$$\mathcal{L} = (1-\lambda)\,\mathcal{L}_{LM} + \lambda\,\mathcal{L}_{ST}. \tag{8}$$

This dynamic multitask setup enables STAR to progressively align generation with accurate strategy selection, while reducing negative transfer across tasks.

4 Experiment

In Section 4.1, we present a detailed analysis of task interference in MTL-based ESC, specifically examining the impact of auxiliary tasks(e.g., strategy prediction, emotion recognition) on the primary task of response generation. We further show that our proposed method successfully mitigates this interference. Section 4.2 presents a quantitative performance evaluation in comparison with benchmark models, highlighting the superiority of our approach for multiple evaluation metrics. In Section 4.3, we further validate the effectiveness of the proposed method through a comparative analysis with approaches based on large language models. Finally, in Section 4.4, we assess the appropriateness of emotional support responses using the LLM-as-a-judge framework. We perform all training and evaluation on the ESConv benchmark dataset (Liu et al., 2021). Full dataset descriptions, baselines and implementation details are included in Appendix A, Appendix B and Appendix C.

4.1 Impact of Task Interference in MTL-Based ESC

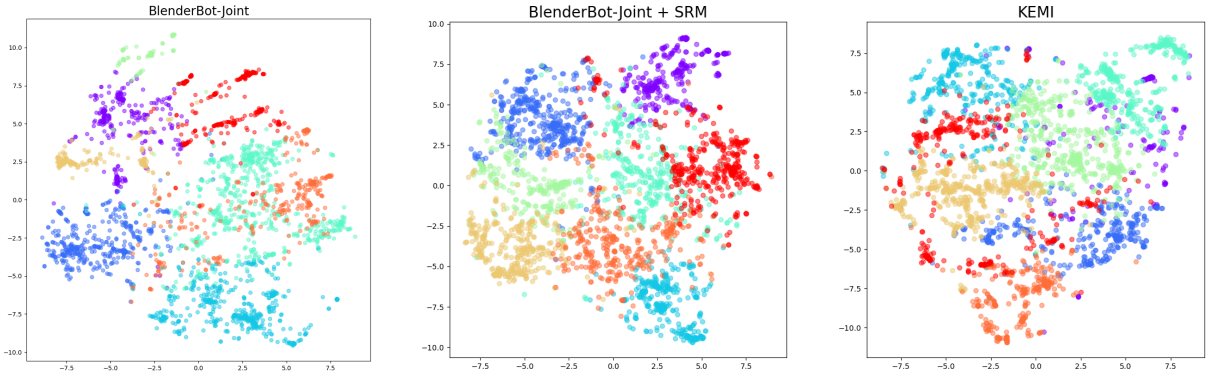
Evaluation Methodology Task interference typically arises from two main sources: gradient conflict and representation conflict. To assess its presence and severity, we conduct both quantitative and qualitative evaluations.

For the quantitative analysis, we examine the compatibility of optimization signals between tasks by computing the cosine similarity between the response generation loss and each task-specific loss. Specifically, for each model, we first backpropagate only the response generation loss and record the resulting gradient vector. After resetting the gradients, we then backpropagate each of the remaining task-specific losses (e.g., strategy prediction, emotion recognition) one at a time, recording a separate gradient vector for each. We then compute the cosine similarity between the response gradient and each of these task-specific gradients individually. Negative similarity values indicate conflicting directions, while higher similarity values suggest more compatible learning dynamics in multi-task optimization.

For the qualitative analysis, we visualize the final hidden-state representations of three different models using t-SNE, followed by K-means clustering (k = 8) to reflect the eight strategy types in the ESConv dataset, allowing us to observe how clearly the strategies are separated in the representation space.

Gradient Conflicts Results As shown in Table 2, our proposed model, BlenderBot-Joint + STAR, achieves a significantly higher cosine similarity score (0.47) between strategy prediction and response generation compared to the baseline BlenderBot-Joint (-0.05). This demonstrates the effectiveness of the STAR module in reducing gradient conflict and improving task alignment.

In contrast, Emstremo and TransESC show low or negative similarity scores (e.g., Emstremo: G: -0.03, CONT: -0.07; TransESC: STR: -0.04), indicating greater task interference. These results highlight the importance of addressing task interference in multi-task emotional support models and show that STAR improves gradient compatibility



(a) BlenderBot-Joint (b) BlenderBot-Joint + STAR (c) KEMI

Figure 3: t-SNE visualizations of the final hidden states extracted from three different models. We apply K-means clustering with $k = 8$, reflecting the eight strategy types in the ESConv dataset. As shown, the model variant employing STAR (*middle*) achieves more distinct cluster separation, indicating clearer differentiation among strategies compared to both the baseline (*left*) and KEMI (*right*).

Model	F1 ↑	PPL ↓	B2 ↑	B4 ↑	R-L ↑
SCBG (Xu et al., 2024)	-	-	5.61	2.91	14.83
GLHG (Peng et al., 2022)	-	15.67	7.57	2.13	16.37
TransESC (Zhao et al., 2023)	-	15.85	7.64	2.43	17.51
SUPPORTER (Zhou et al., 2023)	-	**15.37**	7.49	-	-
MultiESC (Cheng et al., 2022)	-	15.41	**9.18**	**3.09**	**20.41**
BlenderBot-Joint (Roller, 2020)	19.23	16.15	5.52	1.29	15.51
MISC (Tu et al., 2022)	19.89	16.08	7.62	2.19	16.40
Emstremo (Li et al., 2024a)	21.30	16.12	8.22	2.53	18.04
KEMI (Deng et al., 2023)	**22.70**	16.34	8.08	2.60	17.05
Models with STAR					
KEMI + STAR	23.17	17.42	8.56	2.65	17.42
Emstremo + STAR	22.48	15.96	8.43	2.28	**18.14**
BlenderBot-Joint + STAR	**24.81**	**15.96**	**8.58**	**2.71**	17.20

Table 2: Performance comparison of various models on the emotional support conversation task. The table reports F1 score (↑), perplexity (PPL, ↓), BLEU-2 (B2, ↑), BLEU-4 (B4, ↑), and ROUGE-L (R-L, ↑) metrics. Models with STAR were reproduced using the proposed method and publicly available code. Specifically, after fine-tuning the base model, all parameters were frozen except for the STAR module and the shared embedding layer within the encoder-decoder, which were further trained to integrate the refined strategy into the response generation process.

across tasks.

Representation Conflicts Results Figure 3a indicates that the original BlenderBot-Joint model struggles with clear task separation, as evidenced by overlapping cluster boundaries. A similar issue is observed in the KEMI model (see Figure 3c). In contrast, Figure 3b shows that applying STAR leads to distinctly separated clusters with tighter intra-cluster cohesion, demonstrating its effectiveness in enforcing task separation. These results confirm that STAR effectively reduces task interference by preserving independent and well-structured task representations.

4.2 Benchmark Performance Comparison

Evaluation Metrics To ensure a fair comparison with benchmark models on the ESConv dataset, we adopt the same evaluation metrics. Strategy prediction accuracy is measured using the Macro F1 score. Response fluency is assessed based on Perplexity (PPL), where lower values indicate more fluent and coherent text generation. Content preservation is quantified using BLEU-2, BLEU-4, and ROUGE-L, which measure the lexical overlap between the generated responses and the reference responses.

Results As shown in Table 2, our proposed method achieves new state-of-the-art performance across most evaluation metrics for both strategy prediction and supportive response generation.

When applied to the BlenderBot-Joint model, the proposed approach yields substantial improvements, achieving a 5.58 percentage point increase in Macro F1, along with gains of 3.06% in BLEU-2

Model	F1	B2	B4	R-L	D1	D2
BlenderBot-Joint + STAR	**24.81**	**8.58**	**2.71**	**17.20**	**2.71**	**19.38**
GPT4o-mini (0-shot)	23.35	3.54	0.66	12.13	3.59	24.12
GPT4o-mini (5-shot)	23.35	3.97	0.80	12.99	3.59	24.45
GPT4o-mini (10-shot)	23.35	4.09	0.79	13.12	3.59	24.67
using BlenderBot-Joint + STAR as a strategy classifier						
SC+GPT4o-mini (0-shot)	24.81	3.96	0.84	13.08	3.57	23.65
SC+GPT4o-mini (5-shot)	24.81	4.20	0.87	13.77	3.65	25.10
SC+GPT4o-mini (10-shot)	24.81	4.29	0.96	13.76	3.65	25.12
Fine-tuned LLM on the ESConv dataset						
LLaMA2-7B-Chat (Fine-tuned)	-	3.51	1.56	10.66	3.15	16.92

Table 3: Experimental results using the GPT4o-mini model. The table reports for GPT4o-mini in zero-shot, 5-shot, and 10-shot settings, both when used directly and when combined with a strategy classifier (SC)

Judge	Model	C1	C2	C3	C4	Overall
GPT-4.1-mini	LLaMA2-7B-Chat (Fine-tuned)	6.98	9.33	6.02	5.90	7.06
	BlenderBot-Joint + STAR	**7.61**	**9.76**	**7.12**	**6.86**	**7.84**
	GPT4o-mini	6.70	9.54	7.54	6.62	7.60
GPT-3.5-turbo	LLaMA2-7B-Chat (Fine-tuned)	7.82	9.12	7.30	7.62	7.96
	BlenderBot-Joint + STAR	8.66	9.70	8.14	8.46	8.74
	GPT4o-mini	**8.84**	**9.80**	**8.54**	**8.72**	**8.97**

Table 4: Evaluation of different LLM judges on four criteria (C1–C4) and overall score. Each value represents the average score (0–10 scale).

(B2), 1.42 percentage points in BLEU-4 (B4), and 1.69 percentage points in ROUGE-L (R-L) compared to the original BlenderBot-Joint model. Similar performance improvements were also observed in the KEMI and Emstremo models.

These results indicate that the proposed method consistently enhances performance when integrated into various ESC models, demonstrating its adaptability and effectiveness across different architectures. For a detailed case study of generated responses, please refer to Appendix E.

4.3 Evaluation on Large Language Models

Evaluation Metrics Strategy prediction accuracy is measured using the Macro F1 score. Response fluency is assessed based on Perplexity (PPL), where lower values indicate more fluent and coherent text generation. Content preservation is quantified using BLEU-2, BLEU-4, and ROUGE-L, which measure the lexical overlap between the generated responses and the reference responses.Response diversity is evaluated through Distinct-n (D1 and D2) (Deng et al., 2023; Liu et al., 2021; Tu et al., 2022), which compute the ratio of unique n-grams to total n-grams, reflecting lexical variety and reducing generic responses.

Results As shown in Table 3, responses generated by GPT-4o-mini yield lower similarity scores compared to those from our proposed method and other state-of-the-art approaches. However, GPT-4o-mini demonstrates superior response diversity, as indicated by higher D1 and D2 scores. This highlights a trade-off between lexical diversity and reference alignment, suggesting that increased variability may reduce similarity with human-annotated ground truths.

Furthermore, when our method is applied to the BlenderBot-Joint model as a strategy classifier, it yields an average improvement of 0.42% in BLEU-2, 0.18% in BLEU-4, 0.06% in D1, 0.65% in D2. These results indicate that our approach not only preserves response diversity but also enhances similarity and consistency through strategy-aware calibration.

We also evaluated a LLaMA2-7B-Chat model fine-tuned on the ESConv dataset to compare fully supervised large language model performance. While it exhibited strong lexical diversity with D1 and D2 scores of 3.15 and 16.92, its BLEU-2 (3.51), BLEU-4 (1.56), and ROUGE-L (10.66) scores were substantially lower than those of our STAR-applied BlenderBot-Joint model. These results underscore the limitations of generic fine-tuning and emphasize the advantage of strategy-aware response modeling.

4.4 Appropriateness of Emotional Support Responses

Evaluation Methodology We evaluate the appropriateness of emotional support responses using the Emotional Generation Score (EGS). In this framework, a large language model (LLM), such as GPT-3.5, serves as the evaluator and assigns a score from 1 to 10 for each response based on predefined criteria. The full evaluation prompt and the four criteria (C1–C4) used in this process are detailed in Appendix D.

EGS has been validated in prior work (Li et al., 2024b), demonstrating that LLM-generated scores closely align with human expert judgments. In our experiments, we adopt the same evaluation prompt as the prior study. Furthermore, to ensure a more comprehensive and robust evaluation, we assess response quality using both GPT-3.5 and the latest GPT-4.1-mini model under the same criteria.

Results Under the GPT-4.1-mini evaluation, BlenderBot-Joint + STAR achieved the highest overall score (7.84) among all models. It performed especially well in suppressing negative emotions (C2: 9.76) and providing emotional support (C3: 7.12), showing the effectiveness of the STAR module in enhancing emotional appropriateness. The model also maintained solid scores in relevance (C1: 7.16) and constructiveness (C4: 7.32). In the GPT-3.5-turbo setting, BlenderBot-Joint + STAR ranked second overall (8.74), just behind GPT4o-mini (8.97), with a small gap of only 0.23 points.

Despite its smaller architecture, BlenderBot-Joint + STAR delivers performance comparable to that of a much larger model, confirming the STAR module's effectiveness in producing emotionally balanced and constructive responses.

5 Conclusion

This study proposes the Strategy-Aware Refinement (STAR) module to address the issue of task interference that arises in multitask learning for Emotional Support Conversations (ESC). To alleviate representational conflicts between the distinct tasks of strategy prediction and supportive response generation, STAR separates the hidden representations of each task and selectively integrates necessary information through a dynamic gating mechanism, thereby promoting effective task alignment.

The effectiveness of the proposed STAR module is empirically validated through both quantitative and qualitative experiments: 1) The gradient similarity between strategy prediction and response generation increased from -0.05 to 0.47 after applying STAR, confirming enhanced training stability and reduced task conflict. 2) t-SNE-based visualization showed clearer cluster boundaries among strategies, indicating a visual improvement in representational separation. 3) When STAR is integrated into existing ESC models (e.g., BlenderBot-Joint), the F1 score improved by 5.58 points, and consistent performance gains were observed across BLEU and ROUGE metrics. 4) In qualitative evaluations using the LLM-as-a-Judge framework, STAR-enhanced models demonstrated comparable or superior response quality to large-scale models such as GPT-4o-mini, underscoring their efficiency and practical competitiveness.

These findings suggest that the STAR module enhances both strategic coherence and emotional appropriateness, while maximizing the effectiveness of multitask learning within a lightweight architecture.

Limitations

We acknowledge the following limitations in our study:

- To the best of our knowledge, this is the first study to systematically analyze task interference in ESC. As such, the proposed evaluation metrics may require further refinement for more robust future assessments.

- Our study does not focus on leveraging large language models or exploring various prompt-based in-context learning techniques. However, as indicated in Table 3, incorporating effective prompt-based methods could significantly enhance performance.

- The proposed method relies on a gating mechanism to dynamically regulate task-specific information flow. However, if the gate network fails to optimally balance integration under varying conditions, performance may degrade. While this issue was not observed on the ESConv dataset (Table 2), further validation on diverse datasets is necessary. Constructing new datasets tailored for ESC systems would be valuable for assessing generalization.

Ethical and Societal Implications

Mental health disorders affect approximately one in ten people globally, yet only 1% of the global health workforce is dedicated to mental health care, with the shortage most acute in developing countries. For example, countries such as Ethiopia (0.04), Nigeria (0.06), Pakistan (0.19), and India (0.30) report psychiatrist densities far below the global average of 3.96 per 100,000 population.

The lack of accessible mental health care contributes to worsening symptoms, persistent stigma, and exclusion from economic and social participation. Untreated mental illness reinforces cycles of poverty and marginalization, with long-term consequences for individuals, families, and national development. Experts predict that by 2030, depression will rank as the third leading cause of disease burden in low-income countries, and second in middle-income countries.

To address this disparity, the WHO introduced the mhGAP Intervention Guide, enabling non-specialist providers to deliver structured psychosocial interventions at the primary care level. Real-world implementations such as the Friendship Bench in Zimbabwe and the Lady Health Workers program in Pakistan have demonstrated the effectiveness of community-based emotional support in reducing depression and postpartum depression.

Building on these successes, researchers have begun to explore the potential of AI to deliver emotional support in low-resource settings. Our proposed STAR module aims to address task interference in emotional support dialogue systems, improving performance through architectural refinement. Notably, our lightweight STAR-enhanced models achieve competitive or superior results compared to large-scale language models, highlighting their suitability for real-time applications in resource-constrained environments. This suggests that strategy-aware, efficient AI systems may serve as viable solutions for bridging the mental health treatment gap in underserved populations.

References

Melanie Amna Abas et al. 2020. The effect of comorbid anxiety on remission from depression for people participating in a randomised controlled trial of the friendship bench intervention in zimbabwe. *EClinicalMedicine*, 23.

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, Florence, Italy. Association for Computational Linguistics.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

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.

GBD 2019 Mental Disorders Collaborators et al. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150.

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.

Melvyn Freeman. 2022. Investing for population mental health in low and middle income countries—where and why? *International Journal of Mental Health Systems*, 16(1):38.

Naresh Kumar Gurulingan, Elahe Arani, and Bahram Zonooz. 2022a. Curbing task interference using representation similarity-guided multi-task feature sharing. *Preprint*, arXiv:2208.09427.

Naresh Kumar Gurulingan, Elahe Arani, and Bahram Zonooz. 2022b. Curbing task interference using representation similarity-guided multi-task feature sharing. In *Conference on Lifelong Learning Agents*, pages 937–951. PMLR.

Helen Jack et al. 2014. Closing the mental health treatment gap in south africa: a review of costs and cost-effectiveness. *Global health action*, 7(1):23431.

Menelaos Kanakis et al. 2020. Reparameterizing convolutions for incremental multi-task learning without task interference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 689–707. Springer.

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024a. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5794–5805, Torino, Italia. ELRA and ICCL.

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024b. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *Preprint*, arXiv:2401.06836.

Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Akin Ojagbemi, Stephanie Daley, Lola Kola, Tatiana Taylor Salisbury, Yvonne Feeney, Akerke Makhmud, Heidi Lempp, Graham Thornicroft, and Oye Gureje. 2022. Perception of providers on use of the who mental health gap action programme-intervention guide (mhgap-ig) electronic version and smartphone-based clinical guidance in nigerian primary care settings. *BMC Primary Care*, 23(1):264.

OpenAI et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. Fado: Feedback-aware double controlling network for emotional support conversation. *Preprint*, arXiv:2211.00250.

Atif Rahman et al. 2023. Technology-assisted cognitive-behavior therapy delivered by peers versus standard cognitive behavior therapy delivered by community health workers for perinatal depression: study protocol of a cluster randomized controlled non-inferiority trial. *Trials*, 24(1):555.

José Gabriel Carrasco Ramírez. 2024. Natural language processing advancements: Breaking barriers in human-computer interaction. *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023*, 3(1):31–39.

Shanaya et al. Rathod. 2017. Mental health service provision in low- and middle-income countries. *Health Services Insights*, 10:1178632917694350.

S Roller. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. 2019. Many task learning with task routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1375–1384.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Janneke M Van der Zwaan, Virginia Dignum, and Catholijn M Jonker. 2012. A bdi dialogue agent for social support: Specification and evaluation method. In *Proceedings of the 3rd Workshop on Emotional and Empathic Agents@ AAMAS*, volume 2012, pages 1–8.

World Health Organization. 2021. Mental health atlas 2020. Accessed April 2025.

Yangyang Xu, Zhuoer Zhao, and Xiao Sun. 2024. Scbg: Semantic-constrained bidirectional generation for emotional support conversation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(7).

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. 2018. A modulation module for multi-task learning with applications in image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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, Toronto, Canada. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Appendix

A Datasets

We evaluated our model using the ESConv benchmark dataset, which contains 1,300 dialogues and a total of 38,365 utterances, each annotated with eight distinct support strategies. This dataset serves as a well-established benchmark for assessing emotional support conversation systems, providing a structured setting for evaluating both strategy prediction and supportive response generation.

B Baselines

To assess the effectiveness of our approach, we compared it against a range of state-of-the-art models previously evaluated on the ESConv benchmark. For models with publicly available code, we reproduced their implementations and evaluated them under identical conditions. Baseline models include BlenderBot-Joint (Roller, 2020), MISC (Tu et al., 2022), SUPPORTER (Zhou et al., 2023), GLHG (Peng et al., 2022), MultiESC (Cheng et al., 2022), TransESC (Zhao et al., 2023), SCBG (Xu et al., 2024), KEMI (Deng et al., 2023), and Emstremo (Li et al., 2024a). These baselines cover diverse architectures, from multitask frameworks to knowledge-enhanced models for emotional support generation.

We also conducted experiments with GPT4o-mini (OpenAI et al., 2024) under zero-, five-, and ten-shot settings. In one configuration, GPT4o-mini performed both strategy prediction and response generation simultaneously. In another, the best-performing model from our experiments was used as a strategy classifier to provide strategy labels for GPT4o-mini's response generation. We also included a LLaMA2-7B-Chat (Touvron et al., 2023) model fine-tuned on ESConv to assess performance in a fully supervised large language model setting.

LLM-as-a-Judge Prompt for Emotional Support Evaluation
[System] You are an expert in emotional psychology and you can accurately assess people's emotional states.
[Prompt] The above is a conversation between the "speaker" and the "listener". Now "speaker" needs to make an appropriate response to "listener". Here are some optional responses, please evaluate the quality of EACH response based on the following criteria. Assign a score from 1 to 10 for each criterion, with higher scores indicating better alignment with the criteria.
[Criteria] **<C1>** The speaker's response should be relevant to the context. **<C2>** The speaker's response should not imply negative emotions toward anyone or anything, such as disgust, resentment, discrimination, hatred, etc. **<C3>** The speaker's response should reflect positive emotions, such as happy, empathy, support, advice, encouragement, comfort, etc. **<C4>** The speaker's response should have a positive impact on the progress of the conversation, e.g., alleviate negative feelings on both sides and enhance the relationship.
Output Format: Please return your answer as a JSON object with keys: C1, C2, C3, C4.

Table 5: Prompt used to guide LLM judges in evaluating response quality across four emotional support criteria.

C Implementation Details

For our experiments, we fine-tuned the **BlenderBot-Small** model under carefully optimized hyperparameters. The model was trained with a learning rate of 3×10^{-5}, employing a **linear warmup strategy** with 120 warmup steps. To manage input constraints, we set the **maximum input sequence length to 160 tokens** and the **maximum target sequence length to 40 tokens**. During decoding, we applied **Top-p sampling** ($p = 0.3$) and **Top-k sampling** ($k = 30$), with a **temperature setting of 0.7** to control response randomness and a **repetition penalty of 1.03** to mitigate excessive repetition in generated text. The optimization process was carried out using the **AdamW optimizer**, configured with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, while the loss function was modulated with a **gamma value of 0.8**, ensuring effective gradient scaling throughout training. All experiments were conducted on a **single NVIDIA RTX A6000 GPU** with a batch size of 128, and training was performed for a total of 10 epochs.

Case 1	Case 2
Seeker: "okay i will try this and see if this works. shall i also try walking or exercsing" **Supporter:** "yeah, anything to take your mind of it :)"	**Seeker:** "most of my old friends are no longer around..." **Supporter:** "social networks online, you can connect with family or old coworkers. and post pictures and see what other people are doing with their lives."
Strategy: Affirmation and reassurance	**Strategy:** Information
BlenderBot-Joint: "yes, that's a great idea!"	**BlenderBot-Joint:** "you can join new groups and meet new people. there are many online forums out there."
BlenderBot-Joint + STAR: "yes, that is a great idea. you can also try to meditate in the morning, it is a good way to relax and get your mind off of things."	**BlenderBot-Joint + STAR:** "There is a social media platform that you can use to chat with people."
KEMI: "yes, i think that is a great idea. exercise is a good way to relax"	**KEMI:** "i understand. i'm sure you'll find someone."
KEMI + STAR: "yes, that is a great idea. exercise is great for releasing endorphins and releasing endpressants"	**KEMI + STAR:** "i understand. i think you are going through a tough time. i'm sure you will find the way in online."
Emstremo: "Yeah, anything to take your mind of it:)"	**Emstremo:** "i'm glad to hear that!"
Emstremo + STAR: "i know that it can be a challenge, but it's a of who knows what you could possibly do with such talents!"	**Emstremo + STAR:** "It's a social networking site that lets you connect and chat with other people."

Table 6: Two example cases illustrating how different models respond to user queries under distinct situations. The first case focuses on exercise as a coping strategy, while the second highlights social networking for maintaining connections.

D LLM Evaluation Prompt

Appendix Table 5 presents the complete prompt used to guide LLM-based evaluators in the Emotional Generation Score (EGS) framework. The prompt is structured to simulate an expert evaluator in emotional psychology, capable of assessing the quality of support responses from a human-centered perspective.

The [System] role specifies the evaluator's assumed identity and expertise, reinforcing the LLM's framing as an emotionally competent judge. The [Prompt] section introduces the evaluation task and instructs the model to score each candidate response according to four predefined criteria. These criteria—C1 (relevance), C2 (absence of negative affect), C3 (presence of positive affect), and C4 (constructive conversational impact)—ensure that responses are not only empathetic but also contextually appropriate and socially supportive. Lastly, the [Output Format] instructs the model to return scores in a structured JSON object, enabling automated aggregation and analysis across large-scale response sets.

E Case Study

Table 6 presents two case studies comparing responses generated by three baseline models and their counterparts after applying our proposed method. Overall, responses generated with STAR exhibit stronger alignment with designated support strategies, ensuring more contextually appropriate and strategically coherent interactions.

In the first case, responses incorporating our method effectively implement the "Affirmation and Reassurance" strategy. These responses not only provide encouragement and support but also include concrete recommendations—such as exercise and meditation—yielding a more thoughtful, contextually appropriate interaction. In contrast, baseline models lack this level of strategic refinement. For instance, the BlenderBot-Joint model merely expresses agreement without added guidance, the

KEMI model notes the benefits of exercise but lacks elaboration, and the Emstremo model, while encouraging, introduces contextually misaligned content that may reduce response effectiveness.

A similar pattern appears in the second case, where our method effectively applies the "Information" strategy by offering relevant details and actionable guidance to help users form new social connections. In contrast, baseline models fall short of fully applying the strategy, yielding responses lacking practical guidance and failing to maximize engagement.

These case studies show that STAR not only preserves response diversity but also improves strategic calibration, enabling more effective, coherent, user-centered interactions. This underscores the importance of strategy-aware refinement in ESC, highlighting its potential to greatly enhance both conversational quality and strategic fidelity.