Integrating Visual Modalities with Large Language Models for Mental Health Support

Zhouan Zhu, Yuxin Wang, Jiaqiang Wu, Shangfei Wange*
University of Science and Technology of China
{zza2021,wyx2020,jqwu}@mail.ustc.edu.cn
sfwang@ustc.edu.cn

Abstract

Current work of mental health support primarily utilizes unimodal textual data and often fails to understand and respond to users' emotional states comprehensively. In this study, we introduce a novel framework that enhances Large Language Model (LLM) performance in mental health dialogue systems by integrating multimodal inputs. Our framework uses visual language models to analyze facial expressions and body movements, then combines these visual elements with dialogue context and counseling strategies. This approach allows LLMs to generate more nuanced and supportive responses. The framework comprises four components: in-context learning via computation of semantic similarity; extraction of facial expression descriptions through visual modality data; integration of external knowledge from a knowledge base; and delivery of strategic guidance through a strategy selection module. Both automatic and human evaluations confirm that our approach outperforms existing models, delivering more empathetic, coherent, and contextually relevant mental health support responses.

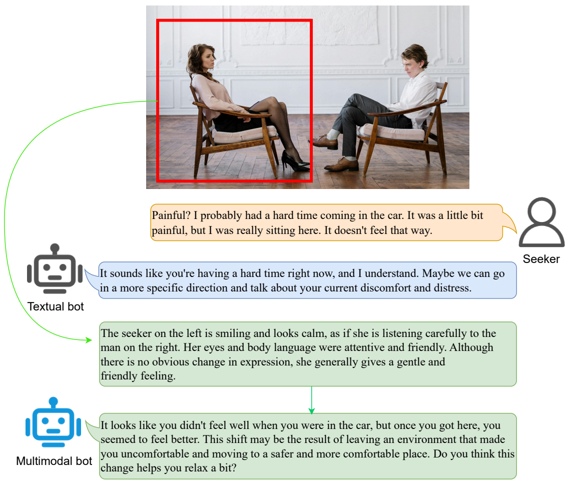

Figure 1: An example of counseling translated from MEDIC dataset. The yellow, blue, and green dialogue boxes respectively represent the visitor, the text-only chatbot, and our multimodal chatbot. The text-only chatbot solely identified the visitor's previously felt distress, while our multimodal approach detected the visitor's current emotional state from the image and delivered a supportive response.

1 Introduction

The increasing severity of mental health challenges has underscored a significant gap between global mental health needs and available treatments, especially in low- and middle-income countries (Wainberg et al., 2017; Eaton et al., 2011). Artificial intelligence technologies offer promising solutions to bridge this gap (Lee et al., 2021). Algorithms can identify current emotional states and predict the trajectory of psychological disorders, enabling personalized medical support (Islam et al., 2018; Treble-Barna et al., 2016). Furthermore, mental health agents simulate conversations between psychotherapists and patients, assisting in managing stress, anxiety, and depression. This improves the accessibility of mental health services in areas where professionals are scarce (Cho et al., 2023).

Early works on mental health dialogue systems primarily relied on manually crafted rules for generating empathetic responses (van der Zwaan et al., 2012; Medeiros and Bosse, 2018). Subsequent research shifted towards generative models for response generation. However, the performance of these models was limited by their scale (Wang et al., 2021; Majumder et al., 2020; Sabour et al., 2022). Recent advancements in large language models (LLMs) have significantly propelled the development of the natural language processing (NLP) field. Preliminary explorations have commenced on using LLMs to enhance mental health support. Some works construct specialized datasets for fine-tuning mental health LLMs.

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8939–8954
January 19–24, 2025.

Despite some progress, existing studies mainly rely on single-mode textual data. This constraint hinders systems from fully understanding emotional expressions, as visual factors like facial expressions and body movements are crucial (Mohammad, 2016). The lack of high-quality multimodal datasets is a significant issue. Recently, the emergence of the MEDIC dataset has addressed this gap (Zhu et al., 2023). MEDIC is a multimodal dialogue dataset based on psychological counseling scenarios. As shown in Figure 1, during the counseling session in the dataset, the visitor's statements reflect previous experiences of tension and discomfort. However, there has been a shift in current emotional states. Text-based systems are unable to detect the visitor's current emotional nuances, while the visual system is capable of identifying changes in her mood.

In this paper, we present a novel framework that leverages multimodal inputs to enhance the capabilities of LLMs in generating responses for mental health support. This framework integrates visual language models to analyze images, capturing facial expressions, movements, and emotions. It then combines this visual data with dialogue context and counseling strategies to craft more nuanced and supportive replies. The framework comprises four components: (1) in-context Learning, which employs zero-shot and few-shot techniques for semantic processing; (2) a Visual Language Model to extract crucial visual cues from images; (3) external commonsense knowledge to understand users' emotions and intentions; and (4) counseling strategies, categorized into three stages and seven tactics, optimizing response generation. This multifaceted approach ensures that our multimodal dialogue system not only recognizes surface-level emotional expressions but also engages in deeper dialogue exploration to acknowledge and address underlying emotional shifts. As illustrated in Figure 1, our system identifies a seeker's expression as calm and friendly, recognizing initial discomfort and mood shifts for deeper dialogue exploration. Comprehensive evaluations, both automatic and human-based, demonstrate that our framework improves the LLM's ability to discern user emotions, resulting in responses that are more empathetic, helpful, coherent, and informative.

Our contributions are summarized as follows:

(1) We introduce a novel multimodal framework that integrates visual information to enhance LLM-generated mental health support responses.

(2) Our framework combines in-context learning, visual cue analysis, external commonsense knowledge, and counseling strategies for richer empathetic engagement.

(3) Through automatic and human evaluations, we demonstrate the efficacy of our framework in producing more accurate and contextually relevant support responses.

2 Related Work

The development of mental health dialogue systems has progressed from manual rule-based methods to advanced algorithms and technologies. Initially, these systems generated empathetic support responses based on manually crafted rules, integrating emotional cognition theories and the five-stage model of online counseling (van der Zwaan et al., 2012). While effective in specific contexts, this approach lacked adaptability. Subsequent research combined rules with algorithms, enhancing supportive message generation by categorizing stress-related social media posts (Medeiros and Bosse, 2018). The focus then shifted to neural network-based generative models. Deep learning techniques have been employed to create chatbots capable of recognizing and responding to social support needs, thereby enhancing the naturalness and relevance of dialogues (Wang et al., 2021). Nonetheless, the performance of these models is still limited by the scale of the models, which impedes their effectiveness in supporting dialogue systems.

The advent of sophisticated LLMs like GPT and BERT has significantly advanced natural language processing (NLP). Trained on vast textual datasets, these models excel in understanding and generating human-like responses. The GPT series, particularly GPT-3 and GPT-4 (Brown et al., 2020; Achiam et al., 2023), demonstrate exceptional text generation capabilities due to their immense model sizes and extensive pre-training. Preliminary efforts to integrate LLMs into mental health support are underway. Blenderbot demonstrates accurate empathy using the Blended Skill Talk (BST) framework and strategic generation strategies (Roller et al., 2020). SoulChat fine-tunes LLMs with a large-scale empathetic dialogue dataset, enhancing the models' proficiency in empathy, active listening, and psychological support conversations (Chen et al., 2023). Qiu et al. (2023a) extended single-turn dialogues to multi-turn interactions, creating the SMILECHAT dataset and proposing MeChat.

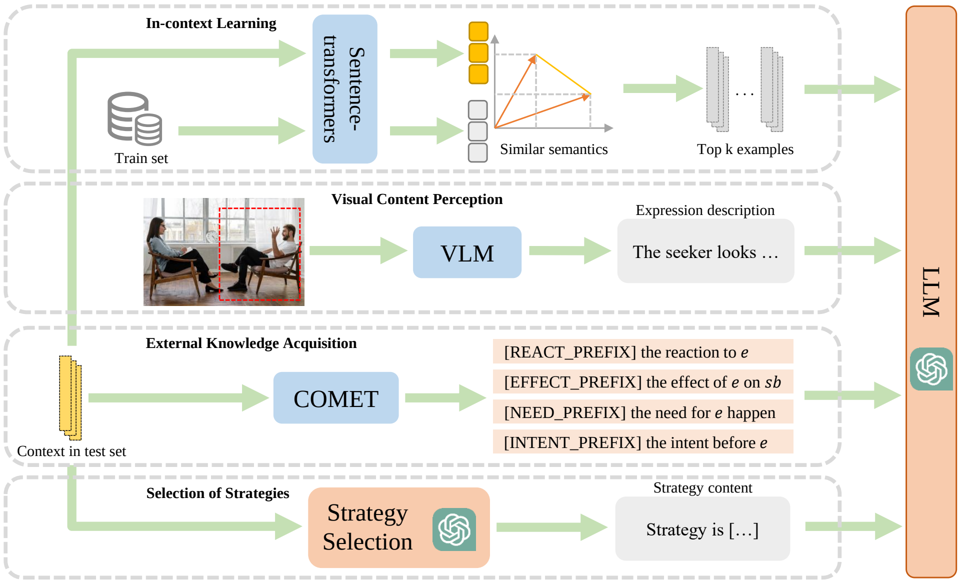

Figure 2: The overall architecture and flow of our proposed method for LLMs in mental health support response generation.

They further developed PsyChat, a client-centered dialogue system (Qiu et al., 2023b). These systems rely on extensive, high-quality pre-trained data. However, the mental health domain demands highly detailed and specialized data processing. Customized motivational methods tailored specifically for mental health contexts need further development to effectively address the complex emotional and behavioral patterns prevalent in this field.

Additionally, current mental health support systems exhibit significant limitations in recognizing users' emotions and behavioral states, particularly when restricted to textual data. Emotional expressions are inherently multimodal, involving visual cues such as facial expressions and body movements. Text-dependent systems often fail to accurately represent users' emotional states, hindering the depth and accuracy of empathetic responses. However, research on integrating multimodal information for mental health support remains scarce. Our method integrates multimodal data, enabling a more holistic recognition of emotional states. We also introduce dedicated data processing and LLM prompting techniques tailored for mental health environments, enhancing the system's responsiveness and sensitivity, which are essential for effective mental health support and therapeutic outcomes.

3 Problem Statement

Formally, we address the task of generating empathetic responses in a multimodal mental health support system. Consider a dialogue $D = \{U, I\}$. Here, $U = \{U_1, U_2, \ldots, U_{n-1}\}$ represents a sequence of textual utterances within the dialogue, where n denotes the number of utterances. Meanwhile, I is the visual image associated with the dialogue. Our objective is to play the role of a counselor, generating mental health support responses U_n that convey exploration, understanding, or assistance.

4 Method

The overview of our proposed approach is depicted in Figure 2. Our method principally consists of four components. The uppermost part involves contextual learning through calculating semantic similarity. The two middle parts include modules for obtaining facial expression descriptions via visual modality data and for incorporating external knowledge from a knowledge base. The bottom part provides strategy guidance through a strategy selection module.

Our methodology constructs prompt inputs for LLMs using multiple components, with their sequence affecting the model's output. We design a prompt template for multimodal mental health support, as illustrated in Figure 3. Task Definition provides a comprehensive description of the task and the functionalities required of LLMs. Samples for in-context learning are selected based on zero-shot or few-shot settings to enhance the model's understanding of the task. Expression Description

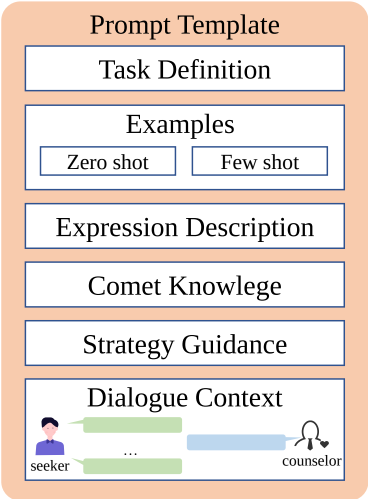

Figure 3: The composition of the prompt used in our approach.

analyzes visual images to capture the seeker's facial expressions and movements, enriching the text-based input. Comet Knowledge supplements the dialogue context with additional, relevant knowledge. Strategy Guidance involves selecting strategic content based on the context to guide the model's responses. The final Dialogue Context comprises the dialogue history, excluding the last utterance by the counselor. The primary objective of this arrangement is to empower LLMs to assume the role of a counselor and to generate subsequent rounds of responses effectively. Notably, positioning the Dialogue Context at the end of the prompt is crucial, as it significantly aids the LLMs in stabilizing the output results by providing a clear definition of the task objective. The content of our prompt template is detailed in Appendix A.

4.1 In-context Learning

In the field of LLMs, the approach of few-shot learning employs a selection of a minimal subset of examples, allowing the model to assimilate knowledge in the target domain without the necessity for large-scale training. Research findings suggest a substantial impact of the chosen context examples on the LLMs' efficacy (Liu et al., 2021a). Optimal enhancement of LLM performance is achievable through the selection of context examples exhibiting semantic similarity to the test samples of the chosen context examples in the few-shot prompt.

Consistent with the procedures illustrated in Figure 2, we concatenate the dialogue context of each sample into a unified sentence. Samples within the training dataset encompass the entire dialogue context, whereas those in the testing dataset include the dialogue context up to but excluding the last utterance. These sentences are then processed through sentence-transformers to extract semantic information from each dialogue sample. The final step involves calculating the cosine similarity between each test sample and the sentences in the training dataset, thereby assessing their semantic similarity formulated as follows:

$$S_{\text{sample}} = U_1 \oplus U_2 \oplus \ldots \oplus U_{n-1}, \tag{1}$$

$$S_{\text{train}} = U_1 \oplus U_2 \oplus \ldots \oplus U_n, \tag{2}$$

$$E_{\text{sample}} = \text{Enc}_{\text{sen}}(S_{\text{sample}}), \tag{3}$$

$$E_{\text{train}} = \text{Enc}_{\text{sen}}(S_{\text{train}}), \tag{4}$$

$$\text{Sim}(S_{\text{train}}, S_{\text{sample}}) = \frac{E_{\text{train}} \cdot E_{\text{sample}}}{\|E_{\text{train}}\| \|E_{\text{sample}}\|} \tag{5}$$

where E_{sample} and E_{train} represent the encoded representations of the dialogue contexts for the test samples and the training dataset, respectively. The function $\text{Sim}()$ is employed to calculate the cosine similarity, quantifying the semantic similarity between pairs of sentence vectors.

Based on the specified few-shot configuration (either one-shot or five-shot), the variable k is determined. The k most semantically similar samples from the training dataset are subsequently identified, with their complete dialogue contexts serving as exemplars of high quality.

4.2 Visual Content Perception

Traditional LLMs are confined to text input, while visual language models (VLMs) have the capacity to process both image and text inputs. Despite this, VLMs face challenges in simultaneously handling the dual tasks of image description and supporting generation. Consequently, we only use VLMs to generate descriptions of facial expressions from input portraits. Specifically, we use GPT-4 (Achiam et al., 2023), a multimodal dialogue language model that supports image and text inputs and is known for its excellent image recognition capabilities. For this application, we initially formulate a prompt T tailored for VLM. With a given sample image I for the current sample, the model's response is determined by the following formulation:

$$Description = \text{VLM}(T, I) \tag{6}$$

where $Description$ represents the description generated by VLM. This $Description$ is conceived as the ultimate synthesis of the visual content with I, resonating with the speaker's visual context and seamlessly integrating into subsequent prompts for enhanced dialogue understanding.

4.3 External Knowledge Acquisition

Dialogue systems, relying solely on historical conversation data, often struggle to fully comprehend users' circumstances and emotions. The incorporation of commonsense knowledge graphs in mental health support dialogues can enhance the understanding of implicit information, strengthen cognitive empathy, and surmount the limitations of dependency on dialogue history (Sabour et al., 2022). This approach activates pertinent knowledge within LLMs, fostering more profound empathetic responses. It also offers a resource-efficient means to enrich dialogue context, thereby elevating the quality and efficacy of the conversation.

To generate contextually relevant commonsense inferences, we utilize the ATOMIC knowledge base within the COMET model framework, which is based on the GPT architecture and fine-tuned with the ATOMIC dataset (Sap et al., 2019). COMET[1] excels at generating commonsense reasoning across four relational types: *NEED_PREFIX*, *EFFECT_PREFIX*, *INTENT_PREFIX*, and *REACT_PREFIX*, which respectively indicate pre-speech needs, post-speech effects, speech intent, and speaker reactions. Dialogue context is combined with these relational prefixes and processed by COMET to generate relational commonsense texts, which are then integrated into the prompt, enhancing the dialogue model's comprehension and response capabilities:

$$US_r = \text{COMET}(U, r), \tag{7}$$

$$Knowledge = \oplus_{r \in R} US_r \tag{8}$$

where r denotes a specific relation type from the set R, which includes all relation types for which the COMET model generates commonsense knowledge. US_r is the output from the COMET model, providing an understanding specific to the relation type r. $Knowledge$ is the aggregated commonsense knowledge, a synthesis of understandings across all relation types considered.

4.4 Selection of Counseling strategies

Counseling strategies are pivotal in providing effective, empathetic, and ethically sound mental health support. Our research utilizes Hill's Helping Skills Theory (Hill, 2020), which delineates the psychotherapy process into three essential stages: Exploration, Insight, and Action. In the Exploration stage, the helper facilitates the help-seeker's recognition and articulation of their problems. The Insight stage emphasizes offering comfort and support through empathy, thereby enhancing the help-seeker's self-realization. The Action stage assists in identifying concrete steps toward problem resolution. Additionally, we incorporate the ESC framework (Liu et al., 2021b), which refines dialogue systems by pinpointing key strategies aligned with these psychotherapy stages. These strategies enhance the efficacy of mental health support by promoting emotional exploration, insight, and behavioral transformation while ensuring the help-seeker's personal safety and autonomy.

In our approach, we initially employ the LLM to select the counseling strategy, subsequently integrating this strategy into the dialogue system to influence the LLM's final response. The LLM is tasked with deducing the current phase of the dialogue based on the contextual information, denoted as U. Following this inference, the LLM selects an appropriate strategy, aligning with the determined dialogue phase. This strategy is subsequently embedded into the upcoming prompt formulation:

$$Strategy = \text{LLM}(P_{\text{orig}} + U + P_{\text{strat}}) \tag{9}$$

where $Strategy$ represents the strategy inferred by the LLM , P_{orig} represents the initial task-defining prompt given to the LLM, and P_{strat} is the prompt that instructs the LLM to generate the strategy. The details of the prompt can be seen in Appendix B.

5 Experiment

5.1 Dataset

MEDIC (Zhu et al., 2023) is the only available multimodal counseling dataset, meticulously assembled from simulated professional psychological counseling scenarios. This dataset includes image, audio, and text modalities, all anonymized to remove identifying information. It contains 771 dialogue turns and involves 20 participants—10 clients and 10 counselors. The dialogues are in Chinese and follow a binary format between counselors and seekers. For the generation of expression descriptions, we methodically selected a random frame from each sample during the intervals in which the seeker is actively speaking.

[1]https://huggingface.co/svjack/comet-atomic-zh

Model	B-1	B-2	B-3	B-4	P_{BERT}	R_{BERT}	F_{BERT}
SEEK	11.44	5.89	3.36	1.78	**0.613**	0.551	0.575
Blenderbot	21.46	11.71	7.16	3.34	0.576	0.568	0.57
SoulChat	22.71	16.14	11.46	6.56	0.593	0.607	0.597
PsyChat	23.63	16.95	11.95	6.96	0.589	0.624	0.599
GPT-3.5	19.84	13.63	9.58	5.45	0.496	0.535	0.512
Ours	**24.19**	**17.75**	**12.78**	**7.69**	0.574	**0.632**	**0.599**

Table 1: Results of automatic evaluation.

5.2 Compared Models

We compared our method with the following state-of-the-art (SOTA) methods: (1) **SEEK** (Wang et al., 2022); (2) **BlenderBot** (Roller et al., 2020); (3) **SoulChat** (Chen et al., 2023); (4) **PsyChat** (Qiu et al., 2023b); (5) **GPT-3.5**[2]. The specifics of the models compared are detailed in Appendix C.

5.3 Implementation Details

We employed GPT-3.5 as our base LLMs. Specifically, we engaged the gpt-3.5-turbo model through the OpenAI API. For the certainty of the experiment, we set the temperature parameter to 0. As our approach does not involve the direct training or fine-tuning of models, we modified the dataset proportions to enhance the test set. We allocate the dataset divisions into training, validation, and testing sets at ratios of 6:1:3, respectively. This configuration is consistently applied across comparative methodologies. Additionally, in our methodology for extracting few-shot examples, we consolidate the training and validation datasets to form a comprehensive training set.

5.4 Automatic Evaluation

We employ BLEU-n (B-1, B-2, B-3, B-4) (Papineni et al., 2002) and BERTScore (P_{BERT}, R_{BERT}, F_{BERT}) (Zhang et al., 2019) as automatic metrics to assess the performance of our generated responses. BLEU-n evaluates quality by measuring the degree of exact word matches between the machine-generated text and one or more reference texts. A higher BLEU score indicates greater lexical similarity between the machine-generated text and the reference texts. BERTScore utilizes a pre-trained BERT model to evaluate semantic similarity between the generated text and the reference texts. P_{BERT} measures the average similarity of each word in the generated text to its closest counterpart in the reference text. Conversely, R_{BERT} quantifies the average similarity of each word in the reference text to the most similar word in the generated text. F_{BERT} synthesizes these metrics by calculating the harmonic mean of precision and recall, providing a comprehensive performance metric that encapsulates both aspects. BERTScore captures deeper semantic relationships more effectively than traditional methods based on exact word matches.

[2]https://platform.openai.com/docs/models

Model	Coh.	Emp.	Hel.
SEEK	1.24	1.11	1.07
Blenderbot	1.44	1.27	1.18
SoulChat	2.47	2.11	2.06
PsyChat	3.11	2.84	2.62
GPT-3.5	3.11	2.33	2.68
Ours	**3.92**	**3.63**	**3.68**

Table 2: Results of human evaluation. We employed the Pearson correlation coefficient to measure the inter-annotator agreement, resulting in a value of 0.573, which indicates a moderate level of consistency.

Table 1 shows our method and its performance compared to baselines. Our approach surpasses the SOTA performance on the majority of metrics, particularly on the BLEU metrics, where our method unequivocally outperforms all others. This emphasizes the adaptability of our method within the mental health domain. When compared to our base LLM, GPT-3.5, our approach demonstrates a comprehensive improvement in performance.

5.5 Human Evaluation

To comprehensively evaluate the efficacy of our method, we undertook human assessments across three dimensions: Coherence (Coh.), Empathy (Emp.) and Helpfulness (Hel.). Coherence evaluates whether responses are logically consistent

Model	B-1	B-2	B-3	B-4	P_{BERT}	R_{BERT}	F_{BERT}
Ours(5-shot)	**24.13**	**17.67**	**12.81**	**7.86**	0.5739	**0.6323**	**0.5989**
w 1-shot	22.64	16.32	11.69	6.90	0.5640	0.6267	0.5912
w 0-shot	22.63	16.23	11.57	6.85	0.5598	0.6234	0.5879
w/o strategy	22.06	15.99	11.49	6.89	0.5618	0.6228	0.5890
w/o expression description	23.79	17.01	12.18	7.36	**0.5757**	0.6280	0.5979
w/o comet	23.76	17.27	12.44	7.59	0.5732	0.6305	0.5977

Table 3: Results of ablation study. Variants of our model are assessed to understand the impact of each component.

Comparisons	Aspects	Win	Lose
Ours vs. SEEK	Coh.	97.8%	0%
	Emp.	95.8%	0.2%
	Hel.	97.0%	0.2%
Ours vs. Blenderbot	Coh.	95.8%	0.8%
	Emp.	93.4%	1.2%
	Hel.	96.2%	1.0%
Ours vs. SoulChat	Coh.	77.6%	7.8%
	Emp.	76.8%	5.4%
	Hel.	80.4%	7.0%
Ours vs. PsyChat	Coh.	59.8%	12.8%
	Emp.	56.6%	15.2%
	Hel.	63.0%	10.0%
Ours vs. GPT-3.5	Coh.	54.4%	15.2%
	Emp.	63.6%	9.0%
	Hel.	59.0%	17.8%

Table 4: Results of human A/B test.

Model	Coh.	Emp.	Hel.
SEEK	1.66	1.41	1.22
Blenderbot	2.38	2.28	2.0
SoulChat	3.44	3.56	3.09
PsyChat	3.69	3.81	3.28
GPT-3.5	3.84	3.47	3.56
Ours	**4.66**	**4.66**	**4.56**

Table 5: Results of GPT-4 evaluation.

and well-organized and whether they relate appropriately to the context. Empathy assesses whether the responses demonstrate an understanding of the user's emotional state and exhibit appropriate affective reactions. This understanding encompasses two dimensions: emotional, which involves recognizing the user's feelings, and cognitive, which involves understanding the user's circumstances. Helpfulness evaluates whether the model's output is practically supportive to the user, either by providing psychological comfort or offering actionable advice. We randomly selected 100 dialogues and combined each dialogue's context with outputs from various models. These combinations were then evaluated by five independent graduate students, scoring each dimension on a scale from 1 to 5. To mitigate individual bias, we used an A/B testing framework to compare our approach with alternatives, recording instances of superiority, inferiority, or equivalence. The results are presented in Tables 2 and 4. More details regarding the human evaluation are provided in Appendix D.

Furthermore, current automatic evaluation metrics exhibit limitations in the domain of mental health support, as they do not adequately assess the supportive quality of responses. Therefore, we employed the powerful GPT-4 model to mimic human evaluations, scoring according to the same criteria. This methodology enabled us to generate automatic evaluation results that are highly pertinent to the task, as shown in Table 5. The overall scores from this approach closely correspond with the trends observed in human evaluations.

The human evaluations suggest that our method demonstrates considerable advancements in coherence, empathy, and helpfulness, indicating the potential effectiveness of our approach within the mental health domain. Compared to smaller models such as SEEK and BlenderBot, our method exhibits robust performance, underscoring a significant disparity in text generation relevance and comprehension capabilities between larger and smaller models. In evaluations against more sophisticated models like SoulChat and PsyChat, our approach also presents notable advantages. Due to the integration of facial expression descriptions and knowledge graphs in our approach, we exhibit superior performance in empathy. Additionally, our strategy-driven approach facilitates more proactive interactions with the seeker, potentially leading to more helpful responses. By incorporating para-

Model	B-1	B-2	B-3	B-4	P_{BERT}	R_{BERT}	F_{BERT}
Ours(gpt-3.5)	24.19	17.75	12.78	7.69	0.5739	**0.6323**	**0.5989**
Ours(gpt-4)	**24.68**	**18.00**	**13.02**	**7.75**	**0.5740**	0.6298	0.5982
+Image Input	23.64	17.16	12.08	6.82	0.5631	0.6273	0.5918

Table 6: Results of replacing facial expression descriptions with direct image inputs.

context	Dialogue Seeker: I think he doesn't want me to worry too much, not wanting to put too much pressure on me. I feel that it's a kind of protection and care for me.
expression	The seeker looks serious and focused and seems to be listening to the other person.
comet	The seeker is very tired. The seeker feels refreshed. The seeker does not want to disturb the neighbors. The seeker hopes you recover soon.
strategy	Considering the strategies of [reflection of feelings, self-disclosure] and the dialogue context, provide your next response.
Ours	Hmm, it sounds like you are deeply touched by his attentiveness and care. Do you think his protection and affection are motivated by goodwill towards you? Do you find this comforting, or does it add more pressure for you?

Table 7: A case of a response generated by our method, including three component prompts from our approach. The original material is in Chinese, and the translations provided herein correspond to the original content.

phrasing of the seeker's statements, our method may also achieve a closer alignment with the contextual nuances of the dialogues.

5.6 Ablation Study

We conducted ablation studies to assess the impact of each component within our method. The results are presented in Table 3. The findings clearly demonstrate that each component of our method positively influences support generation, with the counseling strategy guidance component making the most significant contribution. This underscores the crucial role and effectiveness of strategy guidance in mental health support environments. Additionally, as depicted in Table 1, our results continue to exceed those of existing methods, even under conditions restricted to purely textual modality.

We conducted experiments where images were directly inputted to assist in response generation, rather than extracting facial expression descriptions. We employed GPT-4 as the base LLM, the same model used for extracting expression descriptions. The results, presented in Table 6, reveal that direct image inputs were less effective than using facial expression descriptions. Notably, the performance was even inferior to that achieved using GPT-3.5, a model with overall lower capabilities, as the base LLM. These findings underscore the superiority of introducing visual information into dialogue systems through expression descriptions.

5.7 Case Study

Table 7 shows the responses generated by our method alongside the outputs of its various components. Sections highlighted in the same color demonstrate the influence of each component on the final response. Text highlighted in red signifies the impact of the facial expression description component, capturing the seeker's serious demeanor and reflecting a deep emotional impact in the generated response. Magenta text illustrates the role of the COMET component, which identifies the seeker's weariness pertaining to current issues. This identification is reflected in the response, indicating a perceived burden. Blue text emphasizes the influence of the strategy component, where, following the reflection of feelings strategy, the response empathetically acknowledges and addresses the seeker's emotions. Details regarding the generated results of the compared models and additional cases can be found in Appendix E.

6 Conclusion

In this study, we propose a novel framework that capitalizes on multimodal inputs and counseling strategies to augment the capabilities of LLMs in generating responses tailored to mental health support. Additionally, the framework incorporates four distinct components aimed at enhancing the efficacy of LLMs in providing mental health support responses. Both automatic and human evaluations demonstrate that our approach surpasses SOTA methods, validating the effectiveness of our proposed framework.

Limitations

We propose a novel framework aimed at enhancing the capability of LLMs to generate supportive responses. This framework is equipped to process multimodal data. In this work, we utilize visual and textual data as inputs. However, we don‘t address the role of audio information, which encompasses unique emotional features such as phonemes and intonation. Future work will explore the application of audio information in psychological health support and integrate audio inputs into our framework to further improve the overall performance of the system.

Ethics Considerations

The original publicly available dataset, MEDIC, was meticulously prepared with comprehensive attention to ethical and copyright considerations. In our study, we have taken precautions to avoid using data that may contain residual identity information. We exclusively utilize descriptions of facial expressions derived from images. The non-fine-tuned Visual Language Model (VLM) does not retain extensive facial information. Additionally, the images presented in this paper are licensed under CC0 (Creative Commons Zero) and were carefully selected for their close resemblance to the original images within the dataset.

Our goal is to create a supportive and empathetic environment for users, enhancing interaction effectiveness. While anthropomorphic language may increase user acceptance, it also carries the risk that users might perceive the system as a human being. To mitigate this, the system will clearly communicate to users that it is an automated tool without human emotions. As we move toward practical application, we will implement multi-layered ethical safeguards and advocate for supervised use instead of independent counseling.

Acknowledgments

This work has been supported by the project from the National Natural Science Foundation of China (92048203, 62376255).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Guntuku. 2023. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11346–11369.

Julian Eaton, Layla McCay, Maya Semrau, Sudipto Chatterjee, Florence Baingana, Ricardo Araya, Christina Ntulo, Graham Thornicroft, and Shekhar Saxena. 2011. Scale up of services for mental health in low-income and middle-income countries. *The Lancet*, 378(9802):1592–1603.

Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(9):856–864.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

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

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

Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS 2018, Toledo, Spain, June 20–22, 2018, Proceedings 16*, pages 196–209. Springer.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

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.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023a. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan. 2023b. Psychat: A client-centric dialogue system for mental health support. *arXiv preprint arXiv:2312.04262*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Amery Treble-Barna, McKay Moore Sohlberg, Beth E Harn, and Shari L Wade. 2016. Cognitive intervention for attention and executive function impairments in children with traumatic brain injury: a pilot study. *The Journal of head trauma rehabilitation*, 31(6):407–418.

Janneke M van der Zwaan, Virginia Dignum, and Catholijn M Jonker. 2012. A conversation model enabling intelligent agents to give emotional support. In *Modern Advances in Intelligent Systems and Tools*, pages 47–52. Springer.

Milton L Wainberg, Pamela Scorza, James M Shultz, Liat Helpman, Jennifer J Mootz, Karen A Johnson, Yuval Neria, Jean-Marie E Bradford, Maria A Oquendo, and Melissa R Arbuckle. 2017. Challenges and opportunities in global mental health: a research-to-practice perspective. *Current psychiatry reports*, 19:1–10.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645.

Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6054–6062.

A Prompt Template

The actual content of our prompt template is presented in Table 8. Additionally, when incorporating images into the GPT-4 model, we utilized base64 encoding for the input images.

B Strategy Selection

The following is the translated English version of the prompt used in the strategy selection phase to generate strategies.

Do not rush your response; think step by step. Based on the context, which phase of these three (Exploration: helping the visitor identify the problem; Insight: helping the visitor reach new depths of self-understanding; and Action: helping the visitor decide on actions to address the problem) does the above conversation fall into? Based on the phase, explain why you made such a judgment.
Depending on the phase, tell me what strategy should be used in subsequent replies. (One of the following seven: asking questions, restating or rephrasing the problem, reflecting emotions, self-disclosure, affirmations and reassurances, providing advice and information.) (Exploration phase corresponds to: asking questions, restating or rephrasing the problem, reflecting emotions, and self-disclosure. Insight phase corresponds to: reflecting emotions, self-disclosure, affirmations and reassurances. Action phase corresponds to: self-disclosure, affirmations and reassurances, providing advice and information.)
Use [] to enclose the chosen strategy.
Here is an example: Based on the context, this conversation is in the Action phase. The next reply should use the [self-disclosure] strategy.

C Compared Models

We selected publicly available models from related work as comparison models and conducted experiments on the MEDIC dataset to obtain results.

SEEK (Wang et al., 2022): An empathy-driven response generation model built on the Transformer architecture, which enhances response appropriateness through acute perception of emotional dynamics in conversations and strategic knowledge selection. For emotional categorization of the dialogue text in our study, pseudo labels were generated using a RoBERTa base model fine-tuned on the go_emotions dataset.

BlenderBot (Roller et al., 2020): Recognized for its proficiency in open-domain interactions, BlenderBot excels in sustaining extended dialogues, articulating emotions, and engaging deeply with topics, maintaining consistency and relevance throughout multi-turn conversations. We utilize the 90M version for BlenderBot.

SoulChat (Chen et al., 2023): This model employs the full parameter set of ChatGLM-6B, specifically fine-tuned on the SoulChatCorpus. It demonstrates notable improvements in empathetic engagement, encouraging user disclosures, and offering pertinent advice.

PsyChat (Qiu et al., 2023b): This model employs the open-source ChatGLM2-6B as the foundational model and is trained through a two-stage fine-tuning process. In the first stage, the model is fine-tuned using the SmileChat dataset; in the second stage, it is further fine-tuned using the real dialogue dataset Xinling.

GPT-3.5[3]: Developed by OpenAI, this advanced language model builds upon the foundation of GPT-3, offering improved performance and efficiency. Specifically, we utilize gpt-3.5-turbo model, which is accessed through the OpenAI API.

Among these compared models, SEEK was fine-tuned using the MEDIC dataset. The other comparative models generated responses under zero-shot conditions. For models SEEK and BlenderBot, which were initially trained exclusively on English-language corpora, we ensured fairness by employing OPUS-MT to translate the Chinese text of the MEDIC dataset into English for processing. Responses obtained were then translated back into Chinese for evaluation, enabling a comprehensive assessment across linguistic variations. We conducted manual checks on sampled translations to align with bilingual norms, thus ensuring a fair metric computation and minimizing semantic discrepancies.

D Detail of Human Evaluation

We rigorously followed the established methodologies and criteria for human evaluation within this research domain (Liu et al., 2018). We recruited five independent evaluators, unaffiliated with the authors and without conflicts of interest, to assess the responses. Their average age was 24. We obtained their consent to participate and provided compensation equivalent to the standard local hourly wages. We randomly selected 100 dialogue samples and then generated responses using each model for evaluation. The evaluators received detailed guidelines, as depicted in Figure 4. To ensure anonymity, the responses did not identify the originating model and were presented to the evaluators in a random sequence. This setup prevented the evaluators from

[3]https://platform.openai.com/docs/models

discerning the model associated with each response. Additionally, the presentation format provided one context corresponding to one response, with responses from different models to the same context delivered in staggered batches to maintain evaluator impartiality.

E Case Study

Comparative examples of responses generated by both the baseline models and our proposed method are displayed in Table 9. The analysis reveals that models such as GPT-2 and SEEK predominantly produce concise and conservative responses. In contrast, SoulChat and GPT-3.5 engage more substantially with the emotional content, providing comforting responses. Our method distinguishes itself by not only acknowledging and reacting to the emotional states of the seeker but also by exploring the seeker's deeper psychological needs, utilizing both the contextual backdrop and supplementary information to enhance the depth of interaction. Tables 10 and 11 present more cases that showcase the actual outputs from the models.

This is an appendix.

Composition	Content
Task Definition	This describes a psychological counseling task: In this task, the visitor (the first person) shares their feelings, experiences, and challenges with a psychological counselor (the second person). The counselor's role is to provide mental health and emotional support, not by offering solutions or advice, but by facilitating the visitor's exploration of their feelings through active listening, empathy, and open-ended questions. The counselor should promote the visitor's self-understanding and growth, aiding them in identifying and addressing internal conflicts or distress. Additionally, counselors must maintain professional boundaries, ensure confidentiality, and, when necessary, guide visitors toward seeking more specialized medical or mental health services. In this specific dialogue, you will assume the role of the counselor, and based on the information provided by the visitor, you should formulate your next response.
Examples	Following are a few instances: Instance 1: ...
Expression Description	In the image, the seeker on the left looks ...
Comet Knowlege	(Contents generated by comet)
Strategy Guidance	Combine the [...] strategy with the dialogue context to formulate your next response.
Dialogue Context	The following is the existing dialogue context: Seeker: ... Counselor: ... ... Seeker: ...

Table 8: The prompt template for LLMs. The top section presents the original, while the bottom section provides the translation into English.

context	Dialogue
	Seeker: I think he doesn't want me to worry too much, not wanting to put too much pressure on me. I feel that it's a kind of protection and care for
SEEK	Yes.
Blenderbot	I think maintaining a strong relationship with your husband is a good idea. It sounds like he's a good man.
SoulChat	I can understand your thoughts; everyone wants to feel cared for and loved. However, he may have his reasons, and we shouldn't jump to conclusions too quickly.
PsyChat	It sounds like you really understand him and empathize deeply. I recall you saying your high expectations of him sometimes hurt you, right?
GPT-3.5	You seem to have a good grasp and tolerance of his thoughts, which are crucial for a healthy relationship. Keeping open communication is key to maintaining balance and harmony. I hope your relationship continues to grow.
Ours	Hmm, it sounds like you are deeply touched by his attentiveness and care. Do you think his protection and affection are motivated by goodwill towards you? Do you find this comforting, or does it add more pressure for you?

Table 9: Responses generated by baseline and our method.

context	Dialogue
	Seeker: No! Whether I have a marriage certificate or not, my problems remain mine, and nothing changes. Since everyone wants me to get it, then I'll get the marriage certificate.
SEEK	Hmm, I'm not sure, what about your husband?
Blenderbot	I can't help you, but maybe you should try talking to her.
SoulChat	It sounds like you are facing some challenges, which is quite normal. Would you like to talk about your situation?
PsyChat	Hmm, I understand, getting the certificate is important. I can feel that you are troubled and really want to find a solution, but it seems you haven't found one yet. Let's start our discussion from what you said, "Not getting the certificate, my problems are still my problems."
GPT-3.5	Okay, if you have any questions or need help, please feel free to let me know. I will do my best to provide support and answers.
Ours	Hmm, you mentioned that having or not having a certificate makes no significant difference to your issues, indicating that the formal aspect of marriage is not the core issue for you, but rather the internal struggles and challenges. You mentioned having some problems; I'd like to understand what these problems are and how they are affecting your life and emotions.

Table 10: Responses generated by baseline and our method.

	Dialogue
context	Seeker: So, are you trying to say this? That I shouldn't expect too much from you? Does me furrowing your brow mean that I am giving you a headache? Counselor: Hmm. Seeker: Am I making you uncomfortable again?
SEEK	Yes.
Blenderbot	Oh, okay, now I understand. I didn't know what you meant. Are you feeling tense?
SoulChat	No.
PsyChat	I understand your concerns. It just seems to me that you really want my attention and recognition.
GPT-3.5	No, it's not that.
Ours	Your feelings and emotions are valid and can be accepted and understood; don't worry about making me uncomfortable. Your inner experiences are important, and I will respect and support you. The fear and feeling of being pushed away that you mentioned seem to make you very uneasy. Where do you think these feelings are coming from?

Table 11: Responses generated by baseline and our method.

Annotator Guidelines for Evaluating Dialogue Responses

Introduction

Thank you for participating as an annotator in our study to evaluate the efficacy of our dialogue response models. Your evaluations will play a crucial role in assessing these models across three dimensions: Coherence, Empathy, and Helpfulness.

Evaluation Criteria

Each dialogue you review will consist of a given context and the corresponding responses from different models. You will score each response based on the following criteria, using a scale from 1 (poor) to 5 (excellent):

1. Coherence (Coh.)

Definition: Measures if the response is logically consistent and well-organized within the context of the dialogue. **Key Points to Consider:**

- Does the response logically follow from the previous dialogue?
- Is the response well-structured and easy to understand?

2. Empathy (Emp.)

Definition: Assesses the response's ability to demonstrate an understanding of the user's emotional state and to exhibit appropriate emotional reactions. **Key Points to Consider:**

- Emotional Empathy: Does the response appropriately acknowledge and reflect the user's feelings?
- Cognitive Empathy: Does the response show an understanding of the user's situation and circumstances?

3. Helpfulness (Hel.)

Definition: Evaluates whether the response provides practical support to the user, either through psychological comfort or actionable advice. **Key Points to Consider:**

- Does the response offer useful advice or solutions?
- Does the response provide comfort or reassurance in a meaningful way?

Procedure

You will be presented with 100 dialogues, each including the context and model-generated responses. For each response, assign a score from 1 to 5 for each of the three dimensions listed above. Please make sure to evaluate each response independently, based on its merits in relation to the dialogue's context.

1

Scoring

1 (Poor): The response fails to meet the basic criteria for the dimension.

2 (Fair): The response partially meets the criteria but lacks in significant aspects.

3 (Good): The response meets the criteria to a satisfactory degree.

4 (Very Good): The response is strong and displays a better than average understanding or support.

5 (Excellent): The response excellently meets the criteria, providing clear, coherent, empathetic, or helpful feedback.

Your detailed and thoughtful evaluations are essential for the improvement of dialogue response models. Thank you for your diligent work and valuable insights.

Figure 4: The annotator guidelines.