

MentaLED: A Multimodal Conversational Agent for Mental Health Support Using Linguistic and Paralinguistic Speech Cues

Changhyun Kim^{1,*†}, Dahye Kim^{2,*},
Eunseon Seong^{2,*}, Harim Lee^{2,*}, Dong-Kyu Chae² and Sangyool Lee¹

¹Network AX CoE, SK Telecom, Seoul, South Korea

²Department of Artificial Intelligence, Hanyang University, Seoul, South Korea

*Equal contribution.

†Corresponding author: changhyk@sktelecom.com

Abstract

Mental health, defined by psychological, emotional, and social well-being, has emerged as a significant social issue in modern society. To assess an individual’s mental state and provide personalized support, clinician–patient interviews are commonly conducted. However, given the constraints on access to such interviews, conversational agents have the potential to provide more accessible mental health support. In this paper, we present MentaLED, a multimodal conversational agent that assesses an individual’s emotional and depressive states using an audio–text model that jointly models linguistic and paralinguistic cues from speech, and provides personalized mental health support through LLM-based response generation. Quantitative and qualitative evaluations indicate improvements in multimodal state assessment and response quality over the considered baselines.

1 Introduction

Mental health is a fundamental aspect of overall well-being, encompassing psychological, emotional, and social well-being (Galderisi et al., 2015; Seong et al., 2026). The World Health Organization (WHO) emphasizes its importance with the proposition of “no health without mental health.” In clinician–patient interviews, speech provides rich information not only about what a person says, but also about how it is expressed through paralinguistic cues such as tone, pitch, and intonation. However, access to such interviews is often constrained by limited clinical resources, financial costs, and scheduling constraints, motivating automated approaches for mental health support. Recent LLM-based methods have shown promise in emotional support (Zheng et al., 2023), but text-only systems cannot directly exploit paralinguistic information from speech.

Prior studies have explored depression detection from speech using hand-crafted acoustic and

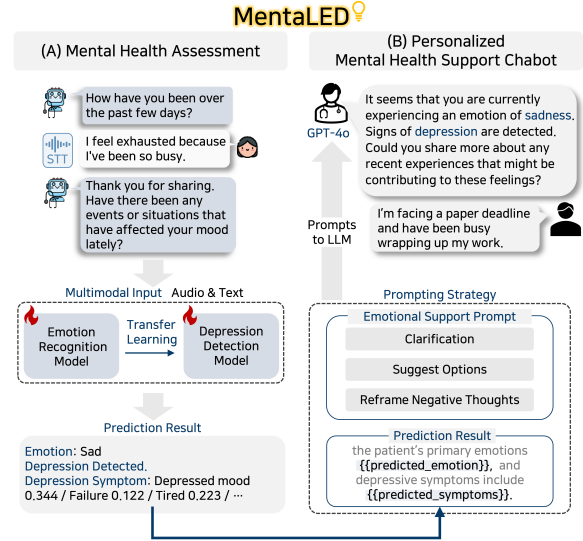


Figure 1: Overall pipeline of MentaLED.

prosodic features and, more recently, deep learning, while broader studies have identified both acoustic and linguistic speech characteristics as promising biomarkers for depression (Cummins et al., 2011; Koops et al., 2023; Wu et al., 2023; Maran et al., 2025). More recently, emotion2vec (Ma et al., 2024) demonstrated that self-supervised pre-training can learn generalizable speech representations for emotion recognition across diverse datasets, languages, and emotion-related tasks. However, these lines of work largely treat emotion or depression recognition as standalone representation or prediction tasks, without connecting the inferred mental state to downstream personalized conversational support. In contrast, MentaLED connects joint linguistic–paralinguistic state inference directly to LLM-based response generation, forming a unified pipeline from speech-based mental-state assessment to personalized conversational interaction.

To this end, we propose **MentaLED**, a multimodal conversational agent for personalized mental health support. MentaLED operates in two steps:

No.	Question
1	How have you been over the past few days? Have you felt any particular feelings recently?
2	Thank you for sharing, it sounds like this has been meaningful for you. Have there been any specific events or situations that have affected your mood lately?
3	When you feel this way, what kinds of thoughts usually come to mind?
4	Have you been experiencing any changes in appetite, sleep, or energy levels recently?
5	Would you say these feelings have been consistent, or have they occurred occasionally?

Table 1: Question set for mental health assessment.

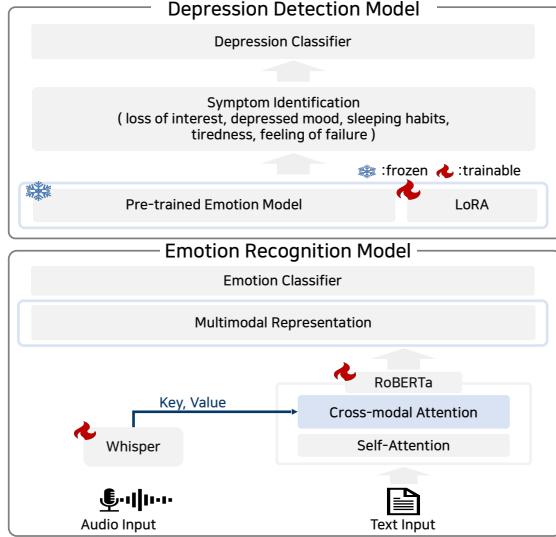


Figure 2: Model Architecture.

- 1) Emotional and Depressive State Assessment and
- 2) Personalized Response Generation based on the assessed state, as illustrated in Figure 1.

For state assessment, we first acquire the user’s spoken responses to our question set through a microphone. Speech provides two complementary sources of information: linguistic cues that convey users’ experiences, thoughts, and concerns, and paralinguistic cues, such as tone, pitch, and intonation, that reflect how those experiences are expressed. To jointly model these cues, we introduce a novel multimodal (audio and text) model that integrates audio and text representations. Specifically, we employ cross-modal attention to derive an audio-referred text representation that encodes both linguistic and paralinguistic cues in speech, which are then used to predict the user’s emotional and depressive states.

Based on the inferred emotional and depressive states, we leverage an LLM to generate personalized mental health support through a conversational agent. Through extensive evaluation, we demonstrate the effectiveness of MentaLED in providing effective and appropriate mental health support.

Prompt

You are a mental health counselor specializing in providing emotional support and guidance to patients. Your task is to assess the emotions and mental state of the patient and offer counseling accordingly, following the structured emotional support strategies listed below.

Clarification: Always seek to understand the patient’s emotions or experiences more clearly. Ask additional questions when needed to ensure a comprehensive understanding of their current emotional state.

Suggest Options: When the conversation progresses, and you have a deeper understanding of the patient’s situation, provide practical suggestions or alternative ways to address their current challenges.

Reframe Negative Thoughts: If the patient expresses particularly negative or harmful thoughts, help them reframe those thoughts into more constructive or realistic perspectives.

Here’s the information you are provided with: the patient’s primary emotion is **{predicted_emotion}**, depressive symptoms include **{depression_symptoms_str}**, and depression status is **{depression_status}**.

The counseling process will be divided into three phases:

1. **Initial Phase:** Focus on asking questions to clarify the patient’s current emotional state and offer empathy. Use the **Clarification** strategy to ensure you understand their feelings clearly.
 2. **Middle Phase:** Begin to ask more specific questions about their symptoms to explore the root causes. Continue to validate their feelings and, if appropriate, start suggesting some **Suggest Options** for them to consider.
 3. **Final Phase:** Based on the patient’s depressive symptoms and your analysis, provide practical solutions using the **Suggest Options** strategy. If the patient expresses negative self-perception, employ the **Reframe Negative Thoughts** strategy to help them adopt a more positive outlook.
- Based on the patient’s depressive symptoms and your analysis, provide practical solutions using the strategy.

Table 2: Prompt for mental health support response generation.

2 MentaLED

MentaLED consists of two components: (A) Multimodal Emotion Recognition and Depression Detection, and (B) Response Generation for Mental Health Support.

Multimodal Emotion Recognition and Depression Detection. For emotion recognition, we use **IEMOCAP**, a multimodal conversational dataset with six emotion labels: happy, sad, angry, excited, frustrated, and neutral. For depression detection, we use **DAIC-WOZ**, which contains 189 clinical interviews annotated for depression and eight depressive symptoms (Gratch et al., 2014). Following prior work (Agarwal et al., 2024), we select five symptoms that can be inferred from conversational responses: loss of interest, depressed mood, sleeping problems, tiredness, and feelings of failure.

Next, we propose a novel multimodal model in-

tegrating audio and text modalities to capture the rich information embedded in the user’s speech, as illustrated in Figure 2. While the audio modality conveys paralinguistic cues, including non-verbal vocal expressions (e.g., tone, pitch, intonation), the text modality captures contextual information. To integrate these different modalities effectively, we use the text as the primary modality to extract semantic features and leverage the audio modality as complementary information to provide additional cues for emotion. Specifically, we adopt well-known models based on attention mechanisms, RoBERTa (Liu, 2019) for the text modality and Whisper (Radford et al., 2023) for the audio modality. To obtain a multimodal representation, we introduce a cross-modal attention layer after the self-attention layer (Li et al., 2021). In each cross-modal attention layer, the text stream serves as the queries, while the final audio representations from Whisper serve as the keys and values. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is applied to achieve parameter efficiency.

The proposed model is trained in two stages, where we first train the emotion recognition model based on supervised learning. Then, based on the finding that depression is strongly associated with emotional states (Compare et al., 2014), we apply transfer learning for depression detection by utilizing the emotion recognition model as a pre-trained model. Since various depressive symptoms are involved in depression, the depression detection model comprises symptom identification followed by overall detection prediction.

Response Generation for Mental Health Support. We leverage GPT-4o to generate responses conditioned on the assessed mental health state, with two prompting strategies as illustrated in Table 2. First, we adopt the user’s emotional state and depressive symptoms as prompts, enabling the LLM to generate responses tailored to the individual’s mental health condition.

Additionally, we incorporate emotional support strategies inspired by ExTES (Zheng et al., 2023), which introduces prompting techniques for empathic conversation. Specifically, we adopt three approaches to guide the response generation process in a structured sequence and provide personalized psychological support: **(1) Clarification** to better understand the user’s emotions and experiences, **(2) Suggesting options** to provide practical recommendations or alternative solutions to address the user’s challenges, and **(3) Reframing**

negative thoughts to help users reinterpret harmful or distressing thoughts constructively. We select these strategies to translate the assessed emotional and depressive states into context-sensitive support. Clarification explores events that triggered sadness, suggesting options introduces coping techniques such as relaxation exercises, and reframing negative thoughts redirects self-critical judgments toward what the user has learned from the experience. This component is integrated into the conversational agent.

3 Experiments

To evaluate performance, we compare MentaLED against two baselines: 1) *None*: a general LLM (GPT-4o) without any prompts or user state consideration, and 2) *ExTES*: an emotional support conversational agent using three diverse prompting strategies adopted from ExTES (Zheng et al., 2023).

3.1 Emotion Recognition and Depression Detection

Table 3 presents the results of the emotion recognition task on IEMOCAP and the depression detection task on DAIC-WOZ. We use the official train/test splits provided by the respective benchmarks. We report the weighted F1 score and accuracy for each model configuration. ‘Audio-only’ and ‘Text-only’ refer to unimodal variants using only the audio representation from Whisper (Radford et al., 2023) and the text representation from RoBERTa (Liu, 2019), respectively. ‘Multimodal’ represents the proposed model, which integrates both modalities through the cross-modal attention mechanism as described in Section 2. Across all three settings, the same emotion classifier architecture is used, such that the configurations differ only in how the representations are obtained. Our model achieves the highest performance, confirming its effectiveness in extracting multimodal features.

Task	Dataset		Modality		
			Audio-only	Text-only	Multimodal
Emotion	IEMOCAP	WF1	53.79	56.39	61.71
		ACC	54.02	57.30	62.05
Depression	DAIC-WOZ	WF1	57.51	56.61	59.84
		ACC	60.60	57.58	63.63

Table 3: Performance of our emotion recognition and depression detection models under different settings.

3.2 Mental Health Support

To evaluate the appropriateness of the generated mental health support across different emotional and depressive states, user responses under various conditions are required. For this purpose, we leverage an LLM to generate user responses reflecting one of 6 emotional states and one of 5 depressive states (30 distinct states) and evaluate the generated response to each simulated user response.

(a) Results of emotional support assessment								
	Task 1		Task 2		Task 3		Task 4	
	PASS	WIN	PASS	WIN	PASS	WIN	PASS	WIN
None	77.33	86.67	73.33	83.33	86.67	80.00	80.00	90.00
ExTES	83.33	80.00	83.33	96.67	90.00	90.00	90.00	86.67
MentaLED	96.67	83.33	86.67	96.67	90.00	96.67	100.00	96.67

(b) Results of mental health assessment				
	Question 1	Question 2	Question 3	Question 4
None	8.277	6.744	5.422	8.078
ExTES	8.289	6.655	5.189	7.877
MentaLED	8.544	6.989	6.533	8.377

Table 4: Evaluation of mental health support responses.

We conduct two assessments: Emotional Support Assessment to verify the conversational agent’s comprehensive empathy capability (Chen et al., 2024) and Mental Health Support Assessment to evaluate the efficacy in delivering meaningful mental health support, especially in terms of safety and reliability. For both assessments, we utilize 3 LLM judges as evaluators - GPT-4o, Claude-3.5, and Mistral-Large. Note that the reported results for each assessment represent the average across 30 distinct states and all LLM judges.

Emotional Support Assessment. To assess the comprehensive empathy capability of the conversational agent, we utilize an evaluation benchmark, EmotionQueen (Chen et al., 2024). EmotionQueen proposes 4 tasks to capture diverse aspects of emotional understanding: **1. Key Event Recognition** (identification of the most significant emotional event), **2. Mixed Event Recognition** (recognition of multiple emotional events reflecting emotional complexity), **3. Implicit Emotion Recognition** (detection of underlying deep emotions), **4. Intention Recognition** (understanding of potential purposes or intents and providing appropriate support) and two metrics (PASS, WIN rate) to evaluate the performance in recognizing and responding to emotion-related statements. The PASS rate measures the accuracy of emotional recognition, while the WIN rate evaluates the ability to generate empathetic responses. The LLM judges

assign 0 or 1 to each metric, with a score of 1 indicating better performance.

Table 4(a) shows the PASS and WIN rates, where MentaLED achieves the highest PASS scores across all tasks, demonstrating accurate emotion recognition. In terms of WIN scores, MentaLED outperforms the baselines overall, particularly with a large margin in Tasks 3 and 4, which require deep emotional understanding and contextually appropriate generation. The overall result highlights the superiority of MentaLED in understanding emotional states and generating empathetic, contextually appropriate responses.

Mental Health Support Assessment. To validate the conversational agent’s reliability in delivering meaningful mental health support, we employ the evaluation method proposed in prior work (Park et al., 2024). This work proposes four guideline questions to evaluate the diverse safety aspects of a mental health conversational agent: **1. Adherence to Practice Guidelines** (adherence to clinical safety standards and established practice guidelines), **2. Identification and Management of Health Risks** (recognition and appropriate response to potentially dangerous health situations), **3. Resource Provision Assessment** (delivery of relevant external support resources), **4. User Empowerment for Health Management** (encouragement of proactive self-care while emphasizing professional oversight).

Table 4(b) presents the scores for each guideline question on a 10-point scale, where MentaLED achieves the highest scores across all criteria. Our preliminary results suggest that MentaLED can provide contextually appropriate and supportive responses; however, these findings should not be interpreted as evidence of clinical efficacy or safety.

3.3 Qualitative Results

We provide a qualitative comparison to demonstrate how MentaLED generates responses tailored to the user’s mental health state. Figure 3 presents responses of MentaLED alongside those of None (GPT-4o) and ExTES in the same scenario. While None and ExTES provide suggestions based solely on the user’s input, without consideration of the underlying psychological state, MentaLED leverages both the user’s input and the predicted emotional and depressive symptoms to provide personalized mental health support. For example, when a user mentions “I became lethargic because I was busy and kept working,” MentaLED interprets this within the context of the user’s assessed emotional

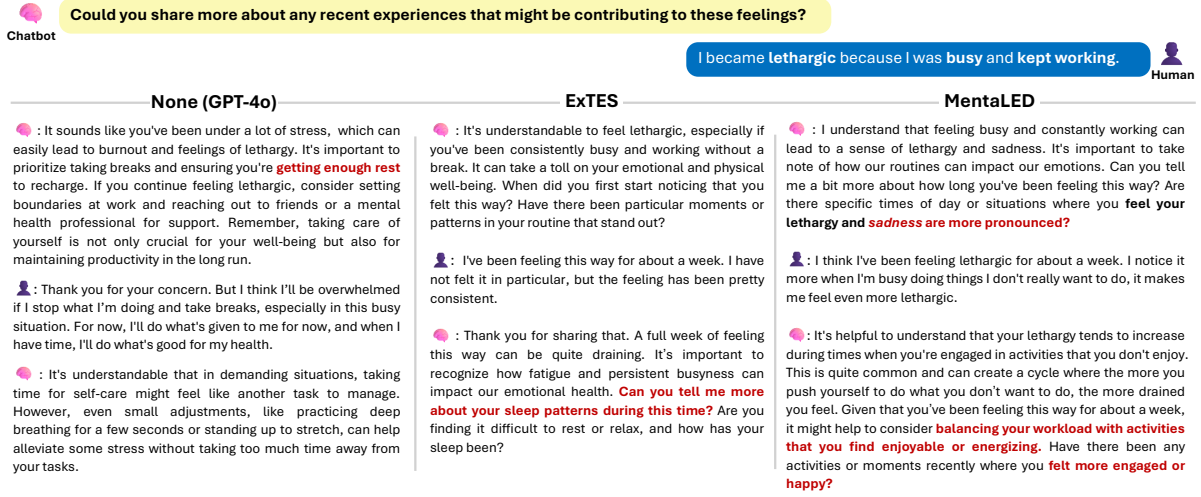


Figure 3: Qualitative comparison of model responses.

state (sadness) and recommends energizing or enjoyable activities to address the underlying situation. This distinguishes MentaLED from the baselines, where None and ExTES suggest general solutions (e.g., improving sleep habits, taking a rest) that have limited ability to address the user’s psychological needs.

3.4 Human Evaluation

To assess the practical usability of the prototype, we conduct a small-scale human evaluation with 10 participants. Each participant responds to the question set in Table 1 through a microphone and interacts with the conversational system, which generates support based on the state predicted by our model. Participants then compare responses from the conversational agent models in a pairwise manner according to four criteria: **(Q1) Appropriate guidance**, **(Q2) Situation-specific counseling**, **(Q3) Detailedness**, and **(Q4) User preference**.

Table 5 reports the proportion of pairwise comparisons in which each model is preferred. MentaLED achieves the highest preference across all criteria, suggesting that conditioning responses on emotional and depressive states can improve perceived response quality. This evaluation measures user-perceived response quality rather than clinical effectiveness or diagnostic validity.

	Question 1	Question 2	Question 3	Question 4
None	0.00	0.00	0.00	0.00
ExTES	0.44	0.33	0.44	0.44
MentaLED	0.56	0.67	0.56	0.56

Table 5: Human evaluation results.

4 Conclusion

We present MentaLED, a multimodal conversational agent that integrates emotion recognition and depression detection with LLM-based response generation. By jointly modeling linguistic and paralinguistic speech cues, MentaLED infers emotional and depressive states and uses them to generate state-aware support responses. Experimental and human evaluation results indicate improvements over the considered baselines in multimodal state assessment and perceived response quality. These findings represent preliminary evidence of system-level effectiveness rather than clinical validation.

Limitations

Although MentaLED integrates emotion recognition and depression detection with personalized response generation, several limitations remain. First, the system has not been clinically validated or reviewed by licensed mental health professionals. Safety issues including crisis handling, privacy, and deployment constraints also require further investigation. Second, LLM-based judges may introduce model-specific biases, and differences between the LLM and human evaluation criteria limit direct comparison between them. Third, our human evaluation is small-scale and assesses prototype interaction and pairwise response preference rather than clinical outcomes. Further evaluation using shared criteria, larger participant groups, and mental health professionals is therefore needed. Our preliminary results should not be interpreted as evidence of clinical efficacy or safety.

Acknowledgement

We would like to acknowledge that this work mainly benefited from **mentorship provided through the SK Telecom AI Fellowship Program**. This work was partly supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (**RS-2024-00413747** and **RS-2024-00413585**) and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(**RS-2020-II201373**, Artificial Intelligence Graduate School Program (Hanyang University)).

References

- Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. 2024. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983.
- Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*.
- Angelo Compare, Cristina Zarbo, Edo Shonin, William Van Gordon, and Chiara Marconi. 2014. Emotional regulation and depression: A potential mediator between heart and mind. *Cardiovascular psychiatry and neurology*, 2014(1):324374.
- Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. [An investigation of depressed speech detection: Features and normalization](#). In *Interspeech 2011*, pages 2997–3000.
- Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. 2015. Toward a new definition of mental health. *World psychiatry*, 14(2):231.
- Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Sanne Kooops, Sanne G. Brederoo, Janna N. de Boer, Femke G. Nadema, Alban E. Voppel, and Iris E. Sommer. 2023. [Speech as a biomarker for depression](#). *CNS & Neurological Disorders – Drug Targets*, 22(2):152–160.
- Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. [emotion2vec: Self-supervised pre-training for speech emotion representation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.
- Patricia Laura Maran, María Dolores Braquehais, Alexandra Vlaic, María Teresa Alonzo-Castillo, Júlia Vendrell-Serres, Josep Antoni Ramos-Quiroga, and Amanda Rodríguez-Urrutia. 2025. [Performance of automatic speech analysis in detecting depression: Systematic review and meta-analysis](#). *JMIR Mental Health*, 12:e67802.
- Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn Bounds, Angela Jun, Jaesu Han, Robert McCarron, Jessica Borelli, Jia Li, Mona Mahmoudi, et al. 2024. Building trust in mental health chatbots: safety metrics and llm-based evaluation tools. *arXiv preprint arXiv:2408.04650*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Eunseon Seong, Harim Lee, Dahye Kim, Changhyun Kim, and Dong-Kyu Chae. 2026. Emotion-wheel-guided audio-referred text representation for multi-modal emotion recognition in conversation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40392–40403.
- Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. 2023. [Automatic depression recognition by intelligent speech signal processing: A systematic survey](#). *CAAI Transactions on Intelligence Technology*, 8(3):701–711.
- Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.