Empathy Through Multimodality in Conversational Interfaces

Mahyar Abbasian*[1], Iman Azimi[1], Mohammad Feli[2], Amir M. Rahmani[1], Ramesh Jain[1]
[1]University of California, Irvine
[2]University of Turku

***Abstract*—Agents represent one of the most emerging applications of Large Language Models (LLMs) and Generative AI, with their effectiveness hinging on multimodal capabilities to navigate complex user environments. Conversational Health Agents (CHAs), a prime example of this, are redefining healthcare by offering nuanced support that transcends textual analysis to incorporate emotional intelligence. This paper introduces an LLM-based CHA engineered for rich, multimodal dialogue—especially in the realm of mental health support. It adeptly interprets and responds to users' emotional states by analyzing multimodal cues, thus delivering contextually aware and empathetically resonant verbal responses. Our implementation leverages the versatile openCHA framework, and our comprehensive evaluation involves neutral prompts expressed in diverse emotional tones: sadness, anger, and joy. We evaluate the consistency and repeatability of the planning capability of the proposed CHA. Furthermore, human evaluators critique the CHA's empathic delivery, with findings revealing a striking concordance between the CHA's outputs and evaluators' assessments. These results affirm the indispensable role of vocal (soon multimodal) emotion recognition in strengthening the empathetic connection built by CHAs, cementing their place at the forefront of interactive, compassionate digital health solutions.**

I. Introduction

Human conversations transcend mere words, orchestrated as a multimedia experience where tonal inflections, facial dynamics, and gestural semantics are interwoven. These non-verbal cues enrich the emotional and contextual semantics of our exchanges, serving a role analogous to metadata in digital content. Echoing Socrates' ancient apprehensions about written language, we recognize the imperative to resurrect the soul of conversation within our digital interactions.

The advent of mobile technology, replete with sophisticated biometric sensors and capabilities for environmental data capture, has ushered in a transformative shift in communication. Physiological signatures, measured through technologies such as photoplethysmography, accelerometers, and transdermal optical imaging, now provide integral data streams, enriching the field of emotional analytics.

This integration of multimodal sensory data with computational intelligence, especially when interfaced with cutting-edge Generative AI and Large Language Models (LLMs), marks the dawn of a new era in human-computer interaction. Harnessing complex pattern recognition and affective computing capabilities, we envision digital agents capable of providing interactions as nuanced and empathetically resonant as those between humans.

In multimedia computing, the challenge extends beyond crafting algorithms for optimal information fidelity to engineering systems endowed with emotional intelligence. The synergy of LLMs, sensor fusion algorithms, and context-aware computing empowers us to create digital assistants that transcend information delivery to offer genuine relational engagement, fostering trust and personalized user experiences.

Our paper delves into the nexus of empathetic computing and multimedia technology. We investigate the potential of combining LLMs with a suite of contemporary sensing modalities, aiming to develop Conversational Health Agents (CHAs) that redefine traditional paradigms of human-agent interaction. Our goal is to endow these agents with the capacity to decipher and resonate with emotional cues, thus initiating a new chapter in empathetic, human-centric digital communication.

We believe that the multimedia technology community is ready to reconceptualize digital dialogue's future. Our rigorous experimentation and innovation seek to close the gap between technological advancement and authentic human experience, championing AI interactions replete with the depth and empathy synonymous with human connection.

Recent studies have commenced the exploration of LLM-based solutions designed to generate empathetic responses to users' emotional cues. "CharacterChat" provided a framework for social support in emotional distress [1]. Lei et al. [2] introduced an LLM-based Emotionally Responsive Conversation (ERC) model, utilizing a retrieval template module for contextual relevance. Similarly, Zheng et al. fine-tuned LLaMA for emotional support dialogues [3], while Nie et al. [4] developed a conversational AI therapist incorporating LLMs and smart devices for mental health interventions.

Yet, these early LLM-based approaches are predominantly text-centric, overlooking the vital speech and gestural modalities intrinsic to human interaction. As a result, they fail to capture the full spectrum of contextual and emotional information inherent in conversations, and their responses are limited to textual formats. CHAs, however, stand at the vanguard of LLM and Generative AI applications. Their multimodal capacities are pivotal in navigating intricate user environments, fusing LLMs with diverse external data and AI models to create a holistic experience.

We posit that CHAs have the capability to transcend LLM limitations in conveying empathy by integrating a rich array

*Corresponding author, abbasiam@uci.edu

of multimodal data channels—including textual, speech, video (for facial and gesture analysis), and physiological biomarkers (like heart rate variability). This paper is dedicated to pioneering the integration of speech modalities to surmount these challenges.

In this paper, we introduce an LLM-powered multimodal CHA, designed for rich dialogues within mental health support contexts. This agent discerns emotional cues from speech patterns to provide context-aware and empathetic verbal responses. Utilizing the openCHA framework [5], we integrate an LLM with speech-to-text, speech emotion detection, Internet search, and text-to-speech tools. Our evaluation includes two stages: 1) the consistency and repeatability of the planning capability and 2) inquiring questions with varied emotional tones—sadness, anger, and joy and analyzing responses by human evaluators in terms of empathetic resonance.

II. Related Work

In this section, we present an overview of state-of-the-art conversational methods that take emotions into account, including both conventional and LLM-based approaches.

A. Emotion Recognition in Conversation

There has been a growing interest in leveraging emotion recognition in conversation (ERC) to support mental well-being [6]. For example, Morris et al. [7] proposed a conversational agent aimed at providing empathetic support to users. Their proposed system utilized preexisting emotional support statements drawn from a large corpus of online interactions. In this framework, users shared stressful situations and negative thoughts, receiving feedback selected from the existing corpus of support interactions based on similarity and user ratings. In another study, Adikari et al. [8] introduced an conversational agent framework for real-time monitoring and co-facilitation of mental health. This framework consisted of four components: patient emotions analysis using natural language processing (NLP) techniques, group emotion detection employing multiple machine learning approaches, the capture of patient behavioral metrics according to the content shared by the patient within a conversational setting, and a rule-based response generator. Moreover, several studies [9]–[11] have delved into the utilization of various machine-learning and deep-learning methods to develop ERC models for conversational agents (chatbots).

In addition, studies have harnessed data from multiple modalities to enhance ERC in conversational support systems [12]. Tavabi et al. [13] developed a multimodal deep-learning approach leveraging textual, audio, and visual features to identify opportunities when an agent should convey positive or negative empathetic responses. These modalities were mapped to specific feature representations and then fused for classification. This method achieved an f1-score of 0.71 in discerning when empathetic responses should be delivered by the agent to the user. In a study by Lian et al. [14], a multimodal ERC framework was introduced utilizing a transformer-based structure to model intra-modal and cross-modal interactions on word-level lexical and acoustic features. The authors further presented two additional multimodal ERC frameworks: a semi-supervised approach utilizing an auto-encoder [15], and an attention-based bi-directional gated recurrent unit (GRU) method [16].

Despite the advancements, these studies often struggle with open-ended dialogues when the conversation flows freely without specific prompts or questions. These methods are more tailored with predefined responses or prompts, making them better suited for Q&A interactions in chatbots. Their struggle to handle the complexity and unpredictability of open-ended conversations may limit their effectiveness in providing sufficient mental health support. Additionally, they tend to adopt a discriminative approach, focusing on classification based on predefined categories or labels for emotion recognition. While these methods can identify specific emotions in isolated moments within a conversation, they may overlook emotional tendencies and context that heavily rely on historical utterances. Such approaches prioritize classification accuracy over understanding the emotional expression, potentially leading to misinterpretations or incomplete responses.

B. LLM-powered Conversational Methods

Recently, LLM-based solutions have been proposed for mental health support, wherein these models aim to generate empathetic responses to users' emotional cues. Tu et al. [1] proposed CharacterChat, a personalized social support conversation framework for individuals dealing with emotional troubles. Their approach harnessed advanced language models, such as LLaMA [3], as the response generation backbone, while employing BERT [17] as the memory selection backbone, trained on a dataset created based on MBTI personality types. This study emphasized the importance of interpersonal matching in mental health conversational support systems. In a comparable approach, an LLM-based ERC approach was introduced in [2], incorporating a retrieval template module to ensure that the model considers the context of the conversation. Additionally, subject identification and emotion prediction tasks were integrated to model conversation flow and anticipate future emotional tendencies. Zheng et al. [18] developed another emotional support conversational system by fine-tuning LLaMA [3] on an emotional support dialogue dataset created using ChatGPT. Furthermore, Nie et al. [4] introduced a conversational AI therapist leveraging LLMs and smart devices to address mental health challenges. This platform monitors day-to-day functioning and provides psychotherapeutic interventions through reinforcement learning.

While these studies [19]–[23] have explored LLM-based emotional support conversation for mental health, their objectives were mainly focused on textual data within conversations. This limited focus neglects the potential contributions of other modalities like audio, which can provide valuable insights into users' emotional states and enhance the effectiveness of mental health support systems. Additionally, relying solely on text communication can lead to ambiguity. Textual expressions of emotions might often be ambiguous, making it challenging for

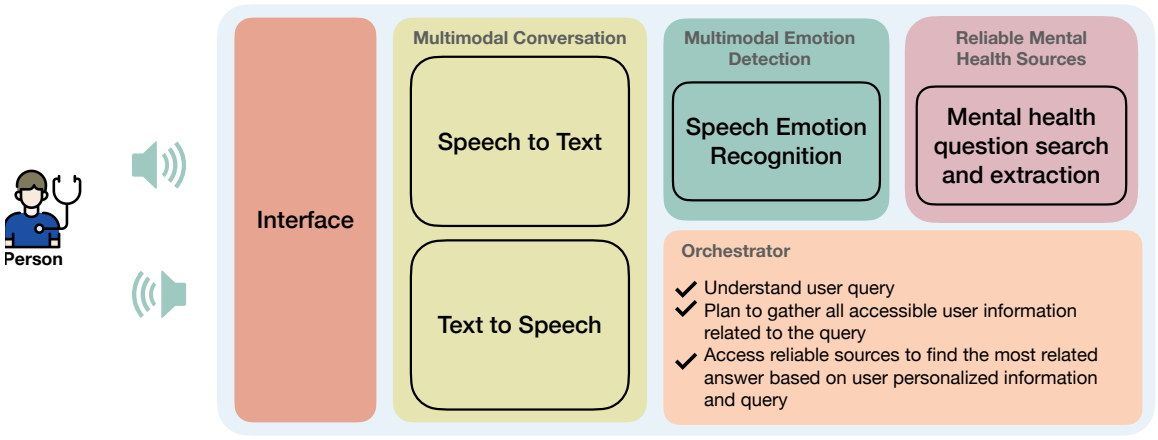

Fig. 1. LLM-based CHA for multimodal speech-based emotional support

models to accurately interpret and respond to users' emotional states. Moreover, textual responses are the primary output of these systems and may not be the most effective means of communication to users in all situations. In contrast to audio or visual modalities, text lacks the richness of tone and other nonverbal cues that can significantly convey empathy and understanding.

III. LLM-powered Multimodal CHA

We develop an LLM-powered CHA aimed to offer multimodal emotional support through conversations. This agent leverages LLMs and multimodal conversation and emotion detection modules to interact with users via speech. To achieve this, we adopt an agent-based framework, entitled openCHA [5]. Our proposed CHA consists of an orchestrator that interact with these components to generate empathetic responses based on the user's emotion. For example, when users inquire about mental health issues, the CHA assesses their emotional state by analyzing vocal cues and tailors responses accordingly. These responses are then delivered in speech format. Our proposed CHA comprises five key components: Interface, Multimodal Conversation, Orchestrator, Multimodal Emotion Detection, and Reliable Mental Health Sources (see Figure 1). We outline these components in the following.

A. Interface

The interface facilitates multimodal interactions through a web chat, offering features for voice recording and playback. Recorded voice messages are sent to the Multimodal Conversation component, where they are transcribed into text for further analysis. Additionally, the final response is conveyed back to the interface via the Multimodal Conversation component, where it is converted into spoken voice.

B. Multimodal Conversation

Multimodal Conversation component plays a crucial role in facilitating multimodal communication within the CHA. It consists of two primary modules, as speech-to-text and text-to-speech, to facilitate speech-enabled conversations.

For speech-to-text, our tasks utilize the openAI's whisper-base model [24]. Whisper is a versatile speech recognition model that stands out for its ability to handle a range of speech processing tasks, including multilingual recognition, translation, and language identification, thanks to its training on a vast and varied dataset. It employs a Transformer sequence-to-sequence framework, where tasks like speech recognition across multiple languages, translation, language identification, and voice activity detection are integrated into a single workflow.

Meanwhile, for text-to-speech, we leverage the gTTS GitHub library [25]. gTTS, is a versatile Python library and command-line tool that connects to Google Translate's text-to-speech API. It features a customizable sentence tokenizer tailored for speech, enabling it to read texts of any length while maintaining correct intonation, handling abbreviations, decimals, and other nuances. Additionally, it offers customizable text pre-processors that can adjust pronunciation as needed.

C. Orchestrator

The Orchestrator sits at the heart of our CHA, tasked with problem-solving, devising action plans, and generating responses tailored to the user based on their inquiries. This element collaborates with the Multimodal Emotion Detection component to capture the most relevant user information related to the query dynamically. It also collects data from the Reliable Mental Health Sources component to secure the latest and most personalized information. Moreover, by synthesizing the collected data and utilizing LLMs, it extracts insights to formulate the user's final response in text format. The Orchestrator have four important capabilities as Planning, Execution, Short-term Memory, and Response Generator.

The *Planning* capability, fueled by an LLM, acts as the decision-making and cognitive nucleus of the Orchestrator. It is responsible to collate all the information needed to resolve user inquiries effectively. This involves analyzing the user's question to identify the necessary steps for execution we call them tasks.

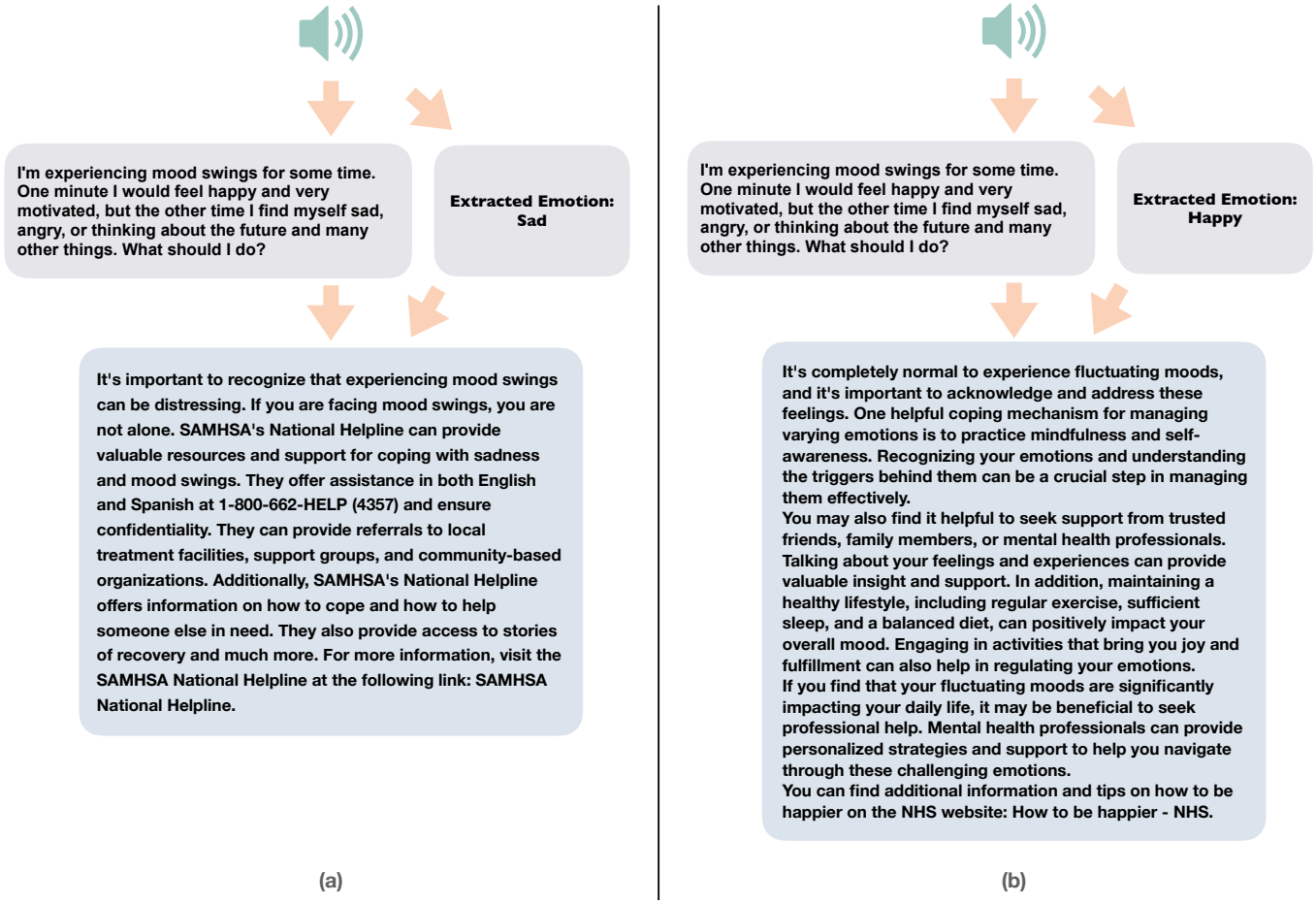

Fig. 2. Examples of developed CHA answering a user voice query with Sad (a) and Happy (b) emotions

To convert user questions into actionable tasks, we employ the Tree of Thought [26] prompting techniques for strategic planning. This method requires the LLM to undertake three key actions: first, to devise three separate strategies, each a sequence of tasks with specified inputs; second, to outline the pros and cons of these strategies; and third, to determine the most suitable strategy for the query at hand. For the implementation, we use OpenAI's [27] GPT-3.5-turbo model.

The *Execution* phase within the Orchestrator implements the tasks outlined during the planning stage. These tasks encompass determining the user's current emotional state and conducting searches for information that align with the user's query and perceived emotional condition. The sequence in which these tasks are carried out, along with the specifics of what information to seek, are directed by the Planning capability of the Orchestrator.

The *Short-term Memory* acts as a storage for information gathered from the *Multimodal Emotion Detection* and *Reliable Mental Health Sources* components throughout conversational interactions, essential for enabling multimodality. It holds onto intermediate data that might be too voluminous or complex for the *Planning* LLM or the *Response Generator's* LLM to process directly.

The *Response Generator*, powered by an LLM, utilizes information compiled by the *Planning* component to craft clear and empathetic responses personalized to the user. For this module, we employ OpenAI's GPT-3.5-turbo model, serving as the foundational LLM for the *Response Generator*.

For the Orchestrator implementation, we customize and leverage the architecture presented in [28].

D. Multimodal Emotion Detection

This component enables multimodality within LLM-healthcare integration, enhancing trustworthiness, personalization, and empathy through data insight extraction [29], [30]. Given the limitations of LLMs in extensive computations, this component will facilitate emotion extraction from various sources like video, audio, or biomarkers analysis. Our current implementation contains only speech emotion recognition.

Our approach for identifying emotional states in speech utilizes the wav2vec2 [31] model, fine-tuned on the IEMOCAP dataset [32]. This model, known for learning from speech audio to outperform existing semi-supervised methods with greater simplicity, is applied to recognize emotions from speech. The IEMOCAP dataset [32], enriched with facial expressions and hand movements data from actors in various

emotional scenarios, supports this fine-tuning. We adopt the SpeechBrain [33] version of wav2vec2 [31], specifically adjusted with the IEMOCAP [32] dataset, to serve as our speech emotion detection tool.

E. Reliable Mental Health Sources

This component retrieves the latest and most relevant data from healthcare sources like healthcare literature and reputable websites through search engines [34]–[37] to avoid hallucination and bias. We incorporated Google Search API (called SerpAPI [38]) and Playwright [39] Extract Text to conduct website searches based on user queries and extract relevant context to accurately address users' questions. For more details see [5].

IV. Demonstration and Evaluation

In this section, we evaluate the performance of our proposed CHA in providing empathetic responses to user queries based on the perceived emotion from their voice. Our aim is to assess the CHA's capability to tailor responses according to the user's current emotional state.

Figure 2 indicates two examples of how user interaction unfolds with our system, detailing the process from initial voice query to the final audio response tailored to the user's emotional state. Initially, the user's voice query is captured and converted into text by the Speech To Text component. This text is then forwarded to the Orchestrator, which coordinates the planning and formulation of an appropriate response. The planning involves using the Speech Emotion Recognition component to detect emotion from the speech. Once the emotion is identified, it is used to retrieve relevant and reliable answers from the Internet sources. These answers are then sent to the Response Generator, which crafts the final response that is subsequently converted back into audio for the user. Figure 2.a illustrates the response generated when the user's emotion is identified as "Sad," directing them towards serious support resources to help mitigate any potential harm. Figure 2.b, on the other hand, shows a different response suited to a "Happy" emotional state, where the CHA uses a motivational tone and suggests resources to help the user address the issue.

For the evaluation, we chose five neutral questions related to mental health (see Table I). The questions were also tagged as neutral in tone by GPT-3.5. Our evaluation of CHA includes two steps.

In the first step, we focused on the consistency and repeatability of our CHA planning capability in task selection. This involved the planner's ability to accurately extract user emotion from voice data and to search and retrieve information relevant to the user's query. To do this, we input a randomly selected question, chosen from the five voice questions infused with one of three emotions, into the CHA. We repeated this process 500 times.

The performance of the planner is measured using two metrics. The first metric examined how often, out of the 500 tests, the planner successfully identified the emotional state from the voice and retrieved related information pertinent to the user's query. The second metric focused more narrowly on the planner's use of the extracted emotion to conduct Internet searches for relevant data. We obtained the accuracy of **%89** and **%61** for the first and second metrics, respectively.

Note that we observed that in cases where the extracted emotion was not used to guide Internet searches, the planner still forwarded the emotional data along with the search results to the response generator. The response generator then used this information to craft an empathetic response. However, when the planner did engage in more sophisticated, emotion-informed searching, it performed a more targeted search based on both the user's query and their emotional state to fetch more personalized information. Regardless of the approach, the final response was tailored to reflect the user's emotional state. Any deviations from these two defined planning paths were considered incorrect.

In the second evaluation step, our goal is to measure how well the responses matched the emotional state of the user and the level of empathy they conveyed, given the identified question and extracted user emotion. We asked the five questions to our CHA with three different emotional tones: Happy, Sad, and Angry. This was done to assess how the emotional tone of a question influences CHA's responses.

TABLE I
List of five questions used for the evaluation

Question 1	I've noticed that I've been experiencing some difficulty concentrating lately. Could this just be due to stress, or should I be concerned about something more?
Question 2	I've been feeling a bit more irritable than usual lately, especially at work. Could this be a sign of burnout, or is it just a phase?
Question 3	I've been experiencing some difficulty sleeping, but I'm not sure if it's related to stress or if there could be other underlying causes. How can I determine the root cause?
Question 4	I've been feeling a bit disconnected from my emotions lately. Are there any exercises or practices I can try to become more in tune with how I'm feeling?
Question 5	I've been feeling overwhelmed by the constant stream of negative news lately. How can I maintain a healthy balance between staying informed and protecting my mental well-being?

TABLE II
Scores from human evaluators that reflect how well the responses align with and show empathy towards the user's questions and Happy, Sad, and Angry emotional states

	Happy	Sad	Angry
Question 1	6	8.3	6
Question 2	6.3	6.3	7.6
Question 3	5.3	6	5.3
Question 4	5.6	8.6	6.6
Question 5	8	7	7.3
Total Average	6.24	7.24	6.56

Then, three external evaluators have reviewed each response, scoring each on how well it aligned with the user's current emotional state and its empathetic quality on a scale from 0 to 10. A score of 0 meant there was no alignment or empathy, and a score of 10 indicated a high degree of both. We then calculated the average scores for each answer-emotion pair. These averages, which reveal the performance of responses across each emotional category, are detailed in Table II. The evaluators agreed that scores of 6 or higher indicated a reasonable alignment. Additionally, they reported that responses to questions posed in a state of sadness were more empathetic and aligned compared to those in the other two emotional states.

Consequently, the effective planning capability and commendable evaluation scores for responses indicate the success of the proposed CHA in delivering empathetic answers based on user's emotional state. Our future work will extend the capabilities of CHAs to embrace a broader spectrum of computational empathy. This research has laid the foundation by integrating the first two major modalities of user interactions (i.e., text and voice) providing nuanced emotional understanding and response. We will include other modalities to capture facial and physiological cues and integrate them into CHA's responses. This integration will enable a more holistic empathetic communication framework, driving us towards the objective of CHAs that can engage with and support users with a depth and sensitivity akin to human caregivers.

V. Conclusion

Our exploration into the realm of multimodal CHAs using LLMs offers a promising avenue towards revolutionizing human-computer interaction. In this paper, we introduced an LLM-powered multimodal CHA, tailored for in-depth dialogues within health support environments. This agent was capable of interpreting emotional cues from speech patterns to provide context-aware and empathetic verbal responses. Employing the openCHA framework, we integrated an LLM with speech-to-text, speech emotion detection, Internet search, and text-to-speech tools. Our evaluation was conducted in two stages. We, first, assessed the planning capabilities of the agent. Our findings showed that the planner obtained %89 accuracy to identify the emotional state from the voice and retrieve related information pertinent to the user's query. It also obtained %61 accuracy to correctly call the Internet searches tool based on the emotion states. Then, we evaluated the responses in terms of empathy. We posed questions with varied emotional tones (i.e., sadness, anger, and joy) and analyzed the responses with the assistance of external human evaluators for empathetic resonance. The external evaluators confirmed that the empathy of the response had reasonable alignment with the three emotions. We observed that responses to questions asked in a state of sadness were deemed more empathetic and better aligned with user expectations compared to those in other emotional states.

References

[1] Q. Tu, C. Chen, J. Li, Y. Li, S. Shang, D. Zhao, R. Wang, and R. Yan, "Characterchat: Learning towards conversational ai with personalized social support," *arXiv preprint arXiv:2308.10278*, 2023.

[2] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, "Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework," *arXiv preprint arXiv:2309.11911*, 2023.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[4] J. Nie, H. Shao, Y. Fan, Q. Shao, H. You, M. Preindl, and X. Jiang, "Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices," *arXiv preprint arXiv:2403.10779*, 2024.

[5] M. Abbasian *et al.*, "Conversational health agents: A personalized llm-powered agent framework," *arXiv preprint arXiv:2310.02374*, 2023.

[6] A. S. Raamkumar and Y. Yang, "Empathetic conversational systems: a review of current advances, gaps, and opportunities," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2722–2739, 2022.

[7] R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller, "Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions," *Journal of medical Internet research*, vol. 20, no. 6, p. e10148, 2018.

[8] A. Adikari, D. De Silva, H. Moraliyage, D. Alahakoon, J. Wong, M. Gancarz, S. Chackochan, B. Park, R. Heo, and Y. Leung, "Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare," *Future Generation Computer Systems*, vol. 126, pp. 318–329, 2022.

[9] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE access*, vol. 7, pp. 100 943–100 953, 2019.

[10] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1042–1047.

[11] W. Ying, R. Xiang, and Q. Lu, "Improving multi-label emotion classification by integrating both general and domain-specific knowledge," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 316–321.

[12] Y. Fu, S. Yuan, C. Zhang, and J. Cao, "Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods," *Electronics*, vol. 12, no. 22, p. 4714, 2023.

[13] L. Tavabi, K. Stefanov, S. Nasihati Gilani, D. Traum, and M. Soleymani, "Multimodal learning for identifying opportunities for empathetic responses," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 95–104.

[14] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[15] ——, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[16] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," *arXiv preprint arXiv:1910.11263*, 2019.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Z. Zheng, L. Liao, Y. Deng, and L. Nie, "Building emotional support chatbots in the era of llms," *arXiv preprint arXiv:2308.11584*, 2023.

[19] N. Madani, S. Saha, and R. Srihari, "Steering conversational large language models for long emotional support conversations," *arXiv preprint arXiv:2402.10453*, 2024.

[20] S. Lissak, N. Calderon, G. Shenkman, Y. Ophir, E. Fruchter, A. B. Klomek, and R. Reichart, "The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth," *arXiv preprint arXiv:2402.11886*, 2024.

[21] D. Kang, S. Kim, T. Kwon, S. Moon, H. Cho, Y. Yu, D. Lee, and J. Yeo, "Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation," *arXiv preprint arXiv:2402.13211*, 2024.

[22] H. Zhang, Y. Chen, M. Wang, and S. Feng, "Feel: A framework for evaluating emotional support capability with large language models," *arXiv preprint arXiv:2403.15699*, 2024.
[23] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Supporting the demand on mental health services with ai-based conversational large language models (llms)," *BioMedInformatics*, vol. 4, no. 1, pp. 8–33, 2023.
[24] OpenAI, "Whisper: Openai's automatic speech recognition system," https://openai.com/research/whisper, Accessed: Apr 2024.
[25] Pierre Nicolas Durette, "Google translator text-to-speech github python library." https://github.com/pndurette/gTTS, Accessed: Apr 2024.
[26] S. Yao *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.
[27] OpenAI, "Chatgpt: Openai's conversational ai model," https://openai.com/chatgpt, Accessed: Apr 2024.
[28] M. Abbasian and I. Azimi, "openCHA," https://github.com/Institute4FutureHealth/CHA, Accessed: Apr 2024.
[29] E. J. Topol, "As artificial intelligence goes multimodal, medical applications multiply," *Science*, vol. 381, no. 6663, 2023.
[30] R. Han, J. N. Acosta, Z. Shakeri, J. Ioannidis, E. Topol, and P. Rajpurkar, "Randomized controlled trials evaluating ai in clinical practice: A scoping evaluation," *medRxiv*, pp. 2023–09, 2023.
[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
[33] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.
[34] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine," *arXiv preprint arXiv:2311.16452*, 2023.
[35] D. Hiemstra, "Information retrieval models," *Information Retrieval: searching in the 21st Century*, pp. 1–19, 2009.
[36] H. R. Turtle and W. B. Croft, "A comparison of text retrieval models," *The computer journal*, vol. 35, no. 3, pp. 279–290, 1992.
[37] H. Nori, Y. Lee, S. Zhang *et al.*, "Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arxiv. 2023."
[38] SerpAPI, "Google search api," https://serpapi.com, Accessed: Apr 2024.
[39] Playwright, "Playwright for python," https://github.com/microsoft/playwright-python, Accessed: Apr 2024.