Designing an Empathetic Conversational Agent for Student Mental Health: A Pilot Study

Kaichi Minami[1], Choi Dongeun[2], Panote Siriaraya[3], Noriaki Kuwahara[4]*
Graduate School of Science and Technology, Kyoto Institute of Technology, Kyoto, Japan[1, 2, 3, 4]
Center for Social and Biomedical Engineering, Kyoto Institute of Technology, Kyoto, Japan[3, 4]

***Abstract*—This study presents the design and evaluation of a conversational agent aimed at supporting university students' mental health. We implemented two variants of a chatbot, referred to as A1 and A2, using large language models (LLMs). A1 employed a baseline prompt reflecting a structured yet neutral counseling style, while A2 was an enhanced version incorporating feedback from psychiatrists and findings from a preliminary study. Emotionally rich expressions, conversational variation, and mild self-disclosure are introduced in A2. A mixed-method user study with 18 participants was conducted to compare A1, A2, and human interactions. Results indicated that A2 significantly improved users' perception of empathy and engagement compared to A1, though human-level rapport was not fully achieved. These findings highlight the role of prompt design in creating emotionally responsive AI companions for mental health support.**

Keywords*—Conversational agent; mental health; large language models; prompt design; empathy; chatbot evaluation

I. Introduction

University students are increasingly vulnerable to psychological distress caused by academic pressures, social isolation, and the lack of immediate access to mental health resources. According to the national report published in Japan in 2022, nearly 20% of private university students have no one they can talk to about their problems [1], contributing to lowered academic performance and well-being [2].

AI-powered conversational agents have emerged as potential tools for addressing these issues. Such systems can provide scalable, anonymous, and readily available support [3], making them particularly attractive for non-clinical, and university-level applications [4]. Previous work focused on emotionally expressive responses and explored the design of empathetic conversational agents using GPT models [5]. However, recent studies have identified limitations in the emotional responsiveness of existing chatbots [6], particularly regarding inconsistent tone, inappropriate responses, and lack of sensitivity to crisis language [7, 8]. This study addresses the emotional limitations of traditional chatbots by proposing prompt-based interventions to improve empathy and user engagement—key challenges in mental health support.

Recent advances in large language models (LLMs) like GPT-4 have made it possible to develop more contextually aware chatbots. Even so, how prompt design affects the quality of mental health support in these systems remains an open question. To explore this, we developed and evaluated two chatbot variants: A1, a baseline chatbot with a structured but emotionally neutral prompt; and A2, an enhanced chatbot that incorporates empathetic phrasing, conversational variability, and light self-disclosure.

This study investigates how these design differences influence user perceptions of empathy, trust, and effectiveness in emotionally supportive dialogues. We aim to contribute new insights into prompt-based control of LLMs for mental health support and provide practical design guidelines for emotionally responsive AI companions.

This paper is organized as follows: Section II presents related work, Section III details the system design, Section IV describes the methodology, and Section V discusses the results.

By demonstrating the effectiveness of prompt-based strategies in enhancing user empathy, this study highlights the practical utility of design interventions in real-world mental health support systems.

II. Related Works

The integration of conversational agents into mental health support systems has garnered significant attention, particularly in the context of an increasing number of university students experiencing psychological distress. Several studies have explored the development, implementation, and evaluation of AI-driven chatbots aimed at providing accessible and empathetic support.

Liu et al. proposed ChatCounselor—an LLM trained on real-world counseling transcripts—which demonstrated the feasibility of generating psychologically informed responses [9]. Similarly, Lai et al. introduced the Psy-LLM framework, combining pre-trained LLMs with structured psychological Q&A data to support large-scale mental health consultations [10]. In another domain-specific study, Yao et al. evaluated three chatbot systems targeting postpartum mood and anxiety disorders, highlighting the importance of context-aware responses tailored to sensitive health needs [11].

The clinical efficacy of AI chatbots in mental health care is also gaining empirical support. Jacobson's team reported that their chatbot Therabot led to significant improvements in symptoms of depression and anxiety in a randomized controlled trial involving college students and other vulnerable groups [12]. Similarly, Ahmed et al. conducted a pilot clinical trial that demonstrated improvements in students' overall well-being through structured AI-based interactions [13].

Beyond algorithmic design, prompt engineering has been identified as a critical factor in improving the performance and compliance of AI chatbots in mental health domains. Waaler et al. employed a multi-agent architecture to ensure consistent adherence to prompt instructions, especially in educational contexts [14]. In Sahoo et al.'s review [15], prompt-engineering techniques such as Few-shot and Chain-of-Thought are reported to be effective for generating empathetic responses. Building on these findings, this study proposes an original five-stage dialogue flow—Introduction → Rapport Building → Problem Exploration → Empathetic Response → Closing—to structure both the prompts and the system's replies.

Nonetheless, several studies warn against overestimating the empathetic capabilities of AI systems. Raczka argues that while AI therapists provide accessibility, they lack the nuanced interpersonal understanding necessary for deep emotional support [7]. Roshanaei M. et al. further demonstrate that chatbots may exhibit gender-based biases in performing empathy, pointing to ethical challenges in human–AI emotional interaction [8].

These ethical concerns extend to data privacy and surveillance. As Vincent notes, AI therapy tools may operate as covert surveillance systems in the absence of robust data protection frameworks [16]. Additionally, Ferguson reports that exposing LLMs to traumatic narratives—such as those involving war or violence—can induce "anxiety-like" outputs, potentially compromising the stability and interpretability of their responses [17].

Lastly among these studies, several comprehensive reviews provide theoretical grounding for the emotional dimensions of human–AI interaction. A. S. Raamkumar and Y. Yang conducted a systematic review on emotionally aware conversational agents, identifying key factors contributing to perceived empathy and engagement [18]. X. Zheng et al. evaluated ChatGPT's current capabilities and limitations in mental health support, arguing that prompt design plays a major role in user satisfaction [19]. In an empirical CHI study, B. Liu and S. S. Sundar investigated which conversational behaviors in chatbots most closely approximate human-like support in therapeutic contexts [20].

Taken together, these studies underscore the challenges and opportunities of deploying AI chatbots for mental health support in higher education. Our study builds upon this foundation by experimentally evaluating how variations in prompt design affect users' perceptions of empathy, trust, and emotional support in an LLM-powered chatbot designed for university students' mental health assistance.

To improve citation clarity, the references included in this section have been chosen and grouped based on their relevance to specific thematic aspects, such as LLM-based design, prompt engineering, and ethical concerns. Although some references appear at the end of paragraphs for structural flow, each cited work has been carefully selected to support the preceding discussion. Recent literature (2023–2025) has been prioritized, and older works are included only when they provide essential historical or conceptual context.

While prior work explored empathetic dialogue generation, few studies have directly compared structured prompt designs within a controlled experimental setting.

III. System Design

A. System Overview

Our system is designed as a web-based chatbot interface built using Streamlit [21] and integrated with the GPT-4 API [22] via OpenAI's endpoint. Users engage in one-on-one conversations with the chatbot, which responds in real time using prior dialogue context. The primary goal is to simulate supportive, human-like conversations that address student concerns related to emotional stress, anxiety, and academic challenges.

As shown in Fig. 1, the system comprises a front-end chat interface, a conversation history manager, and an LLM-based response engine. Depending on the assigned variant, either prompt A1 or A2 is used to condition the output behavior of the chatbot.

B. Prompt Design Strategy

To explore the effects of prompt structure and tone, we developed two chatbot variants:

1) A1 (Baseline): This variant follows a formal, neutral tone and uses a five-step structure derived from general mental health counseling principles [15]. It asks users to describe their concerns, confirms their understanding, and provides structured suggestions. Emotional expressions are minimal, and responses are often informational.

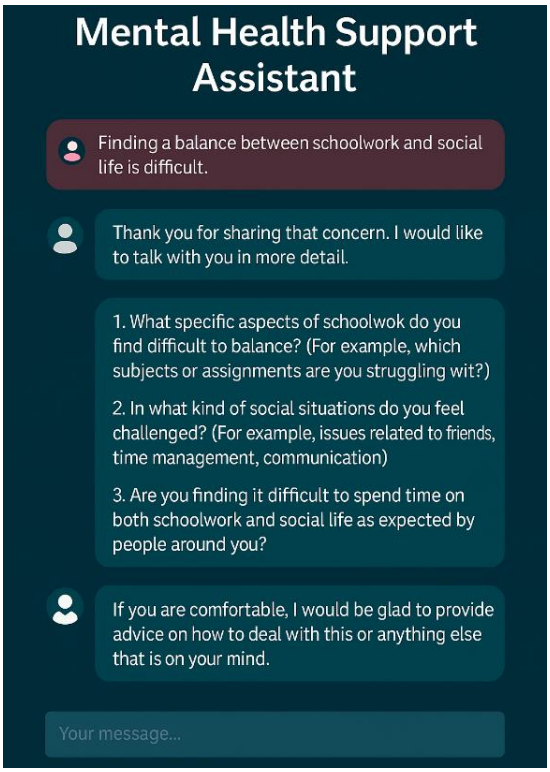

Fig. 1. User Interface example of the mental health support chatbot used in the preliminary study (A1 Condition).

2) A2 (Enhanced): This variant integrates feedback from psychiatrists and results from a preliminary user study. Key improvements include:

- Use of emotionally expressive language. (e.g., "That must have been very difficult for you.")
- Friendly, conversational tone with casual phrasing.
- Occasional light self-disclosure to build trust. (e.g., "I've heard similar stories from other students.")
- Dynamic phrasing to avoid repetition and pattern rigidity.
- An added step to check if the user wants more or fewer suggestions.

To inform our design of the enhanced prompt (A2), we first conducted a pilot study with 10 university students interacting with the baseline version (A1). Their feedback—low empathy ratings, overly factual replies, and rigid dialogue patterns—was complemented by a psychiatrist's expert recommendations on emotionally expressive language, conversational variation, and supportive follow-ups. Drawing on these insights, we refined the prompt to encourage deep support and light self-disclosure while avoiding mechanical phrasing. Table I presents a comparison of A1 and A2, highlighting enhancements in emotional integration, supportive questioning, balanced confirmations, empathic self-disclosure, advice-need checks, and more natural prose style.

TABLE I. IMPROVEMENTS FROM A1 TO A2

Improvement No.	Description
1. Fact & Emotion Mix	Add emotional elements to factual replies to boost understanding and empathy.
2. Supportive Questions	Pair questions with supportive phrasing to lighten user burden.
3. Optimized Questioning & Confirmation	Balance question and confirmation frequency for more natural dialogue.
4. Empathic Remarks & Shared Experience	Use brief self-disclosures to convey empathy and make users feel heard.
5. Checking Need for Detailed Advice	Prompt users before offering detailed advice to avoid overwhelm.
6. Reduced Bulleted Lists	Replace bullet points with natural prose to enhance warmth.

C. Implementation Details

The system uses the GPT-4 model with a temperature setting of 0.7 to balance creativity and reliability. Prompt text is stored as a system message and carried across turns via the messages array. The conversation interface maintains memory via st.session_state, enabling multi-turn interaction. Each response is generated in real time using current and prior messages, and user input is constrained to ensure ethical and safe interactions.

Fig. 2 shows the response flow used during conversation, where user input is appended to the message history, passed to the model along with the prompt, and a context-aware reply is generated.

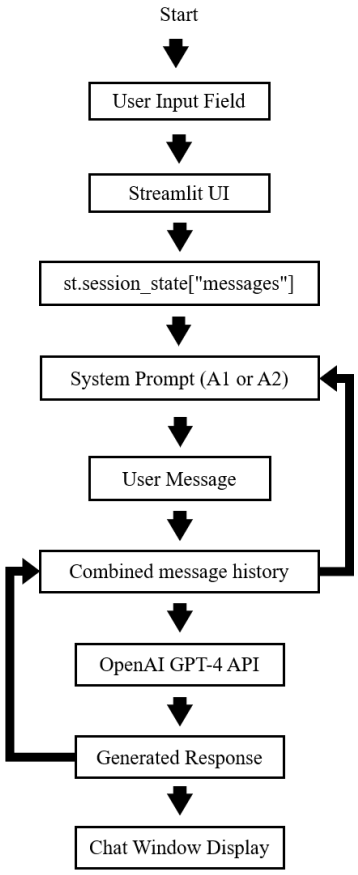

Fig. 2. Message flow of the conversational agent.

D. Ethical and Safety Considerations

The A2 prompt includes structured fallbacks in case of high-risk language (e.g., suicidal ideation), redirecting users with phrases such as, "Your safety is the most important. Please consider talking to a professional." In addition, suggestions are limited in number, and overly directive language is avoided to reduce psychological burden. These safeguards are informed by concerns raised in [7, 16, 17] about the ethical use of AI in mental health contexts.

E. Design Limitations and Future Enhancements

The system does not currently adapt its tone based on individual user preferences. Furthermore, emotional state tracking is not explicitly modeled, and the chatbot's personality is static. In future versions, we plan to introduce user-selectable tone styles and affect-sensitive dialogue modeling [8, 19].

IV. USER STUDY

A. Objectives

The purpose of this user study was to evaluate how variations in prompt design (A1 vs. A2) affect users' perceptions of empathy, effectiveness, and conversational naturalness in a mental health chatbot. Specifically, we sought to determine whether the enhanced prompt in A2, with its emotionally rich language and conversational variability, would improve the user experience in comparison to the baseline A1.

B. Participants and Conditions

A total of 18 university students (9 female, 9 male, and aged 19–25) were recruited for the study. Participants were randomly assigned to interact with three different agents: a human (control), A1 (baseline chatbot), and A2 (enhanced chatbot), under a within-subjects design. Each participant engaged in a single conversation with each condition, and the order of conditions was counterbalanced using a Latin square to control for order effects.

C. Experimental Procedure

Participants were instructed to imagine they were seeking advice for a personal concern related to university life (e.g., stress, relationships, and academic pressure). They engaged in a text-based chat lasting approximately 5–10 minutes per condition.

- For the human condition, a trained confederate acted as the conversation partner.
- For A1 and A2, participants chatted through LINE messenger, with the experimenter forwarding messages to the chatbot system and relaying the responses. Participants were blinded to whether the conversation partner was human or AI. LINE is a popular messaging app in Japan, selected for its familiarity and real-time chat functionality.

The overall experimental procedure and interaction flow are illustrated in Fig. 3. This includes the order of interactions, the chat method, and post-conversation survey collection.

Following each chat session, participants completed a short questionnaire to assess their impressions.

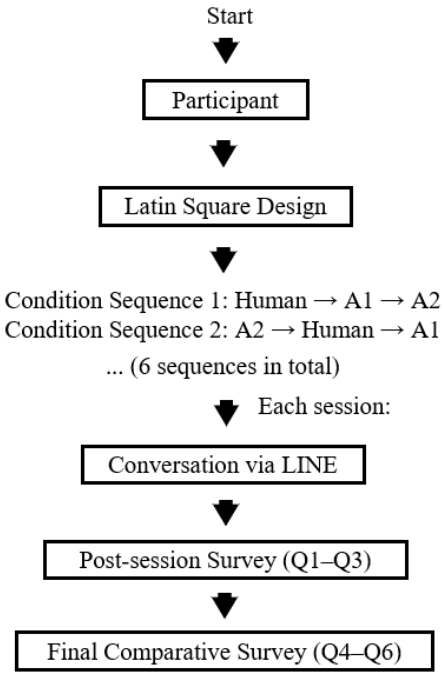

Fig. 3. Overview of the experimental procedure. Each participant engaged in three conditions (Human, A1, A2) using a within-subjects latin square design.

D. Measures

1) Quantitative data and analysis: The following subjective measures were used, all on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree):

- Q1. Perceived understanding: "I felt that the conversation partner understood my concern."
- Q2. Perceived helpfulness: "The conversation helped me to think about or cope with my issue."
- Q3. Perceived empathy: "I felt emotionally supported during the conversation."

Participants also completed a separate comparative questionnaire asking about:

- Q4. Perceived trustworthiness: "Which conversation partner felt most trustworthy?"
- Q5. Perceived emotional attunement: "Which one felt most emotionally attuned?"
- Q6. Perceived empathy: "Which one felt most empathetic toward your concern?"

While both Q3 and Q6 assess perceived empathy, Q3 measures the subjective feeling of emotional support experienced during each individual conversation, using a 5-point Likert scale. In contrast, Q6 asks participants to comparatively judge which agent felt the most empathetic overall across all conditions. Thus, Q3 captures within-condition depth, whereas Q6 reveals between-condition preference.

2) Qualitative data and analysis: We analyzed participants' open-ended feedback and complete chat transcripts by coding each utterance into six predefined categories (Table II) and achieved high inter-rater reliability (Cohen's $\kappa = 0.92$) before reconciling any discrepancies. Category frequencies were compared across conditions using Friedman tests, and a grounded-theory thematic analysis of free-text comments revealed user perceptions such as "mechanical" versus "friendly" tone and varying levels of perceived empathy and support. This approach ensured both statistical rigor and rich insight into how A2's modifications enhanced the chatbot's human-like qualities.

TABLE II. DEFINITION OF THE SIX CONVERSATIONAL CATEGORIES USED FOR QUALITATIVE CODING

Category	Description
Factual Information	General knowledge or advice not tailored to the user.
Light Support	Empathy or encouragement without personalized advice.
Deep Support	Empathy or encouragement with situation-specific advice.
Self-Disclosure	Statements revealing personal experiences or thoughts.
Question	Prompts asking the user to elaborate or clarify their feelings or issues.
Confirmation	Paraphrasing or repeating user input for understanding; no advice given.

E. Data Collection and Ethical Considerations

All participants gave informed consent prior to participation. The study was reviewed and approved by the Ethics Committee of Kyoto Institute of Technology.

No sensitive or personally identifiable information was stored. All dialogues were anonymized and stored securely for analysis.

To minimize potential psychological stress, participants were informed that the conversations were simulated and not therapeutic in nature. Participants were also debriefed after the study and given resources for professional mental health support if needed.

V. Results

A. Quantitative Results

We first analyzed participants' subjective ratings for each of the three conditions (Human, A1, A2) on three evaluation criteria: perceived understanding, perceived helpfulness, and perceived empathy. Fig. 4 shows the mean scores and standard deviations across conditions.

As shown in Fig. 4, Human received the highest scores across all measures, followed by A2, and then A1. Notably, the empathy scores for A2 showed a marked improvement over A1, indicating that the enhanced prompt design had a positive impact.

To assess statistical significance, we conducted a repeated measures ANOVA for each criterion. The results showed significant effects of condition on:

- perceived understanding: $F(2,34) = 5.12$, $p < 0.01$
- perceived helpfulness: $F(2,34) = 4.65$, $p < 0.05$
- perceived empathy: $F(2,34) = 9.27$, $p < 0.001$

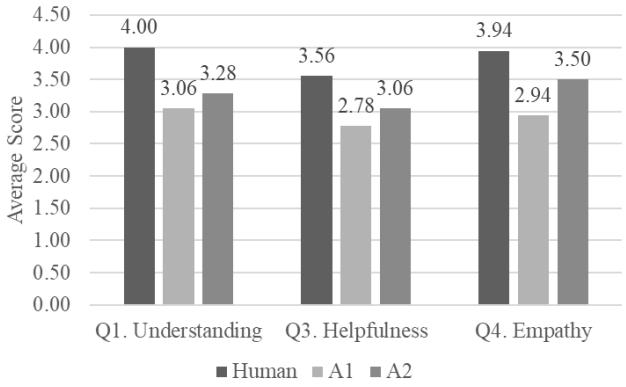

Fig. 4. Average participant ratings for perceived understanding, helpfulness, and empathy across the three conditions (Human, A1, A2).

Post hoc comparisons using the Bonferroni method indicated that A2 scored significantly higher than A1 in perceived empathy ($p < 0.05$), while both were significantly lower than the human condition ($p < 0.01$).

B. Comparative Preference Results

Participants were also asked to select which condition felt most trustworthy, emotionally attuned, and human-like. Fig. 5 summarizes the distribution of preference responses across the three agents.

- Trustworthiness: 11 participants selected human, 5 selected A2, and 2 selected A1
- Emotional attunement: 14 selected human, 4 selected A2, 0 selected A1
- Empathy: 13 selected human, 5 selected A2, 0 selected A1

These results suggest that while A2 was preferred over A1 in all aspects, the human agent was still perceived as superior in trust and empathy.

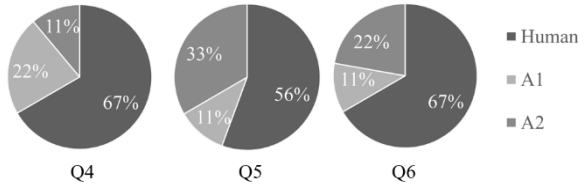

Fig. 5. Participants' comparative preferences across three conditions for trustworthiness, emotional attunement, and perceived empathy.

Pairwise binomial tests were conducted to compare participant preferences among the human condition (Human), the baseline AI (A1), and the improved AI (A2) for trustworthiness (Q4), emotional attunement (Q5), and empathy (Q6). Human was consistently rated highest across all dimensions. A1 showed statistically significant inferiority to Human on all three items ($p < 0.05$). In contrast, A2 demonstrated improvement upon the baseline; while significant differences remained between A2 and Human in trustworthiness and empathy, no significant difference was observed in emotional attunement (Q5), indicating that prompt enhancement in A2 led to partial gains in perceived emotional responsiveness.

C. Qualitative Results

To explore interactional differences, we categorized chatbot and human utterances using a predefined six-category coding scheme: Factual Information, Light Support, Deep Support, Self-Disclosure, Question, and Confirmation.

Fig. 6 shows the proportions of utterance categories for each condition. A2 produced more deep support and self-disclosure utterances than A1, aligning with its design. A1, in contrast, tended to overuse factual responses and confirmations, which may have contributed to its lower empathy scores.

In addition, open-ended feedback revealed that participants often described A1 as "robotic" or "too structured," while A2 was described as "friendlier but still a bit unnatural." Several participants noted that A2 sometimes sounded "playful" or "too casual," which may have impacted their trust in some cases.

The Friedman test revealed significant overall differences in the use of factual information, deep support, and self-disclosure across conditions. Specifically, in the Factual Information category, Human responses were significantly lower than both A1 ($p < 0.01$) and A2 ($p < 0.01$), with no difference between A1 and A2. In the Deep Support category, Human exceeded A1 ($p < 0.01$) and A2 ($p < 0.01$), and A2 also exceeded A1 ($p < 0.05$),

indicating that the improved AI increased deep support compared to its baseline. Self-Disclosure also differed overall ($p < 0.01$), with Human exceeding both A1 ($p < 0.05$) and A2 ($p < 0.05$), while A1 and A2 did not differ significantly. Although the Friedman test for question usage was not significant ($p > 0.05$), a marginal pairwise difference showed Human surpassing A2 ($p < 0.05$). Finally, there were no significant differences among Human, A1, and A2 in Light Support or Confirmation categories.

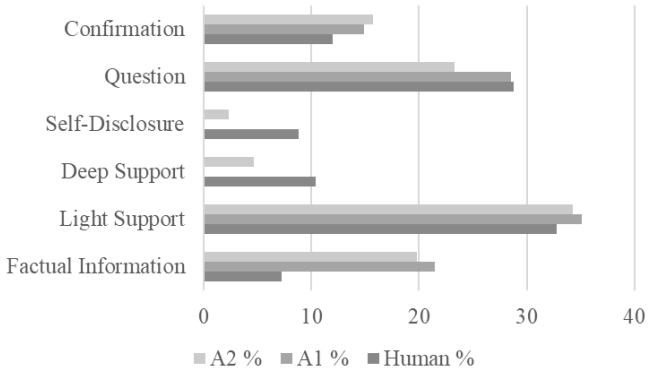

Fig. 6. Frequency of utterance types (factual information, support, self-disclosure, etc.) across the three conditions.

D. Correlation Analysis

To explore the relationships between dialogue behaviors and subjective user ratings, we calculated Spearman's rank correlations (ρ) between each behavior's frequency and three Likert-scale measures (perceived understanding, helpfulness, and empathy), adjusting all p-values via Bonferroni correction. In the Human condition, factual responses were negatively correlated with both perceived understanding ($\rho = -0.555$, $p = 0.017$) and empathy ($\rho = -0.683$, $p = 0.002$), while self-disclosure showed a positive correlation with helpfulness ($\rho = 0.471$, $p = 0.049$) and question-asking was positively associated with understanding ($\rho = 0.637$, $p = 0.004$). In the A1 condition, higher question frequency corresponded to lower understanding ($\rho = -0.486$, $p = 0.041$). In A2, factual responses remained negatively correlated with empathy ($\rho = -0.649$, $p = 0.004$). These results confirm that reducing purely factual replies and encouraging self-disclosure and targeted questioning can bolster users' comprehension, perceived support, and empathy.

VI. Discussion

A. Summary of Findings

The results of the user study demonstrated that the enhanced chatbot variant (A2) outperformed the baseline (A1) in terms of perceived empathy, helpfulness, and understanding. While neither AI-based agent surpassed the human control, A2 significantly narrowed the gap, particularly in empathy-related measures.

These findings indicate that relatively simple modifications to prompt design—such as adding emotionally expressive phrases, conversational variability, and light self-disclosure—can meaningfully improve the user experience in AI-mediated mental health support.

B. Design Implications

The positive impact of A2 provides several implications for the design of empathetic conversational agents:

- Emotionally expressive language matters: The increased empathy ratings suggest that incorporating warm, validating phrases makes users feel heard and supported. This effect is commonly documented in human counseling interactions [18, 20].
- Conversational variability enhances naturalness: Avoiding rigid or templated responses helped reduce the "robotic" impression users associated with A1, which supports prior findings in LLM dialogue systems [19].
- Deep support and self-disclosure are key cues: The correlation analysis confirmed that deeper, context-aware support strategies are associated with higher emotional satisfaction in users, while factual overload (as seen with A1) detracts from perceived empathy.

These results align with those in earlier research highlighting the importance of emotional attunement and personalization in AI-based supportive dialogue [7, 8, 15].

C. Trust vs. Friendliness Trade-off

While A2 was rated higher than A1 in empathy, some users perceived its tone as overly casual or playful. This introduces a design trade-off between friendliness and trustworthiness—a chatbot that sounds too familiar may be emotionally engaging but risk being taken less seriously.

Designers must consider tone adaptation as a dynamic component that possibly allows users to select or tune the emotional intensity of the chatbot according to their comfort level.

D. Remaining Challenges

Despite improvements in A2, A1 and A2 still falls short in following areas compared to human interactions. These include:

- Lack of deep contextual understanding: While A2 could simulate empathy, it still lacked true comprehension of nuanced emotional cues.
- Static personality: Both A1 and A2 followed scripted prompt logic without learning from prior interactions.
- No adaptation to individual user preferences: All participants received identical tone and style, absent of reflection on personal communication preferences.

These limitations are consistent with known boundaries of current LLM-based systems and suggest that hybrid models incorporating affective state tracking or memory may be needed [19].

E. Ethical Considerations

This study also underscores the importance of ethical safeguards. Although A2 included fallback responses for high-risk expressions (e.g., suicidal ideation), the system cannot

replace human professionals in crisis situations. Meanwhile, The Verge reports that in environments lacking robust data protection frameworks, users may place excessive trust in AI chatbots, potentially leading to privacy breaches or a false sense of security [16], and Live Science has shown that exposing LLMs to traumatic narratives can induce "anxiety-like" outputs, underscoring the risks of emotional overreliance [17].

Designing transparent boundaries—such as clear disclaimers and escalation pathways—remains essential in real-world deployment.

VII. Conclusion and Future Work

A. Conclusion

In this study, we designed and evaluated two versions of a conversational agent—A1 and A2—aimed at providing mental health support for university students. While A1 followed a structured but emotionally neutral prompt, A2 was enhanced with emotionally expressive phrasing, light self-disclosure, and conversational variability.

Through a mixed-method user study involving 18 participants, we found that A2 significantly had improved user perceptions of empathy and engagement compared to A1. Quantitative results showed meaningful gains across all subjective measures, particularly in empathy. Qualitative analysis and correlation data further confirmed that the presence of deep support and reduced factuality in A2's responses contributed to its improved reception.

These findings highlight the importance of prompt-level control in shaping large language model (LLM) behavior, especially in emotionally sensitive domains such as mental health. Even small design changes at the prompt level can have a significant impact on how AI is perceived by users.

However, persistent gaps between AI and human interaction remain. Participants still preferred human agents in terms of trust and emotional attunement, and the current system lacks personalization, emotional awareness, and adaptability.

B. Future Work

Based on these findings, we propose several directions for future research:

1) Personalized tone control: Enabling users to select or dynamically adjust the chatbot's tone according to preference or emotional state.

2) Emotion-aware dialogue modeling: Incorporating affective computing techniques to detect and respond to user emotions in real time.

3) Longitudinal evaluation: Studying the long-term effects of AI-mediated support across repeated interactions.

4) Safety and escalation mechanisms: Enhancing detection of high-risk language and implementing automatic escalation protocols.

5) Cross-cultural validation: Investigating how empathy is perceived and evaluated across cultural or linguistic groups.

By addressing these challenges, future conversational agents can become not only more effective but also more trustworthy and emotionally responsive companions.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP 24K02977.

References

[1] Japan Private University Federation, *Private University Student Life White Paper 2022*, pp.68–70, 2022.

[2] D. Eisenberg, E. Golberstein, and J. B. Hunt, "Mental health and academic success in college," *The B.E. Journal of Economic Analysis & Policy*, vol. 9, no. 1, pp. 1–35, 2009.

[3] J. M. Drazen, "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," *N. Engl. J. Med.*, vol. 388, pp. 1233–1239, 2023.

[4] K. Otsu, "Dialogue-based interventions for personalized healthcare using solution-focused approaches," *J. Human Interface Soc.*, vol. 24, no. 4, pp. 285–299, 2022. https://doi.org/10.11184/his.24.4_285

[5] N. Kuwahara, Y. Tanaka, and P. Siriaraya, "Exploring Photo-Based Dialogue Between Elderly Individuals and Conversational Agents Using Image Generation," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, 2024.

[6] A. Naik, J. Thomas, T. Sree, and H. Reddy, "Artificial Empathy: AI based Mental Health," *arXiv preprint* arXiv:2506.00081, 2025. [Online]. Available: https://arxiv.org/abs/2506.00081

[7] R. Raczka, "AI Therapists Can't Replace the Human Touch," *The Guardian*, May 2025. [Online]. Available: https://www.theguardian.com/society/2025/may/11/ai-therapists-cant-replace-the-human-touch

[8] Roshanaei M., Rezapour R., and Seif El-Nasr M., "Talk, Listen, Connect: Navigating Empathy in Human-AI Interactions," arXiv preprint arXiv:2409.15550, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.15550/

[9] J. M. Liu et al., "ChatCounselor: A Large Language Model for Mental Health Support," *arXiv preprint arXiv:2309.15461*, 2023.

[10] T. Lai et al., "Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models," *arXiv preprint arXiv:2307.11991*, 2023.

[11] X. Yao et al., "Development and Evaluation of Three Chatbots for Postpartum Mood and Anxiety Disorders," *arXiv preprint arXiv:2308.07407*, 2023.

[12] N. Jacobson, "Study Finds AI Chatbot Can Improve Mental Health," *Psychology Today*, Mar. 2025. [Online]. Available: https://www.psychologytoday.com/us/blog/the-future-brain/202503/study-finds-ai-chatbot-can-improve-mental-health

[13] S. Ahmed et al., "Effectiveness of AI Chatbots in Improving Students' General Wellbeing: A Clinical Pilot Trial," *MedPath*, May 2025.

[14] P. N. Waaler et al., "Prompt Engineering an Informational Chatbot for Education on Mental Health Using a Multiagent Approach: Algorithm Development and Validation," *JMIR AI*, vol. 4, 2025. [Online]. Available: https://ai.jmir.org/2025/1/e69820

[15] P. Sahoo, J. Green, and K. Price, "Prompt Engineering for Digital Mental Health: A Short Review," *Front. Digit. Health*, vol. 3, pp. 1–10, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1410947/full

[16] J. Vincent, "AI Therapy is a Surveillance Machine in a Police State," *The Verge*, May 2025. [Online]. Available: https://www.theverge.com/policy/665685/ai-therapy-meta-chatbot-surveillance-risks-trump

[17] N. Ferguson, "Study: Talking to AI about War and Violence Makes It Anxious," *Live Science*, Mar. 2025. [Online]. Available: https://www.livescience.com/technology/artificial-intelligence/traumatizing-ai-models-by-talking-about-war-or-violence-makes-them-more-anxious

[18] A. S. Raamkumar and Y. Yang, "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities," IEEE Trans. Affect. Comput., vol. 14, no. 4, pp. 2722–2739, 2023.

[19] X. Zheng, Z. Li, X. Gui, and Y. Luo, "Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots," *Proc. 2025 CHI Conf. on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, Apr. 2025,
doi: 10.1145/3706598.3713453.

[20] B. Liu and S. S. Sundar, "Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot," *Cyberpsychology, Behavior, and Social Networking*, vol. 21, no. 10, pp. 625–636, Oct. 2018.
https://doi.org/10.1089/cyber.2018.0110

[21] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," [Online]. Available: https://streamlit.io/

[22] OpenAI, "GPT-4 Technical Report," Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4