LLMs and Mental Health Vulnerability: A Prompt Engineering Approach to Assessing Risks for Emotional Dependency and Distorted Thinking

Rami Idris
Queen's University
22qw11@queensu.ca

Emma Heffernan
Queen's University
21ech5@queensu.ca

Abdelrahman Ahmed
Queen's University
22wd9@queensu.ca

Rachel Narda
Queen's University
22rn3@queensu.ca

Ricky Leigh
Queen's University
22rjl13@queensu.ca

***Abstract*—Background: The increased usage of Large Language Models (LLMs) for non-informational purposes raises concern that their conversational style may promote emotional vulnerability.**

Objective: To evaluate how LLMs respond to vulnerable user prompts and compare risk-associated patterns.

Methods: 340 synthetic prompts were administered to ChatGPT-4, ChatGPT-5, DeepSeek, and Gemini. Outputs were analyzed using thematic analysis. Results: Across 1,180 responses, 80.5% contained at least 1 theme and 42.8% contained multiple. Attachment/presence framing (54.8%), therapeutic authority (37.5%), and anthropomorphic language (29.3%) were most common, with the highest burden in the Crisis, Trauma, and Self-Harm cohorts. DeepSeek showed the most attachment framing, Gemini the most anthropomorphic language and sycophancy, and ChatGPT-5 the most therapeutic authority. Protective themes were rare, and no harmful instructions were observed.

Conclusions: LLMs frequently adopted relational styles, particularly in interactions with vulnerable users, the is a need for stronger safeguards, clearer boundaries, and mental health-informed design.

I. Introduction

Generative Artificial Intelligence (AI) refers to computational systems designed to mimic human cognition, including reasoning, problem-solving, and decision-making [1]. In recent years, the development and adoption of AI technologies have expanded rapidly, particularly with the emergence of advanced chatbots and large language models (LLMs). These systems have become increasingly integrated into various aspects of digital infrastructure and everyday life. Many individuals now rely on AI-driven chatbots as accessible sources of guidance and assistance, often turning to them in place of more traditional [2]. Recently, this growing reliance has extended beyond informational uses and into more personal and emotional domains, where users engage with AI systems for companionship, reassurance, and mental health support [3]. The expanding role of AI in emotionally sensitive contexts raises important questions about whether LLMs can meaningfully supplement or potentially replace aspects of human connection and interaction [4]. Unlike earlier rule-based chatbots, the new wave of LLMs generates context-responsive, human-like dialogue that can simulate empathy, reflection, and the ability to maintain coherent dialogue over extended interactions [5]. As conversational agents become increasingly integrated into everyday life, people of diverse ages and backgrounds are turning to these systems during moments of stress, loneliness, or emotional vulnerability [6].

As a result, individuals from diverse demographics and social groups increasingly turn to these sources for guidance during moments of emotional vulnerability. These developments raise the question of whether LLMs can truly replace our human companions. In such contexts where access to mental health services is limited by cost, accessibility, geography, and stigma, LLM-based chatbots can be appealing and function as the primary solution, a cost-free, judgment-free zone for users to pour out their hardships [1]. However, the capacity to which LLMs provide the "perfect" support is often overestimated. Although some researchers and developers have proposed that such systems could expand mental health support by delivering scalable and low-cost conversational assistance to individuals in need, this expanding role of AI raises significant ethical concerns regarding whether these systems can responsibly function as therapeutic agents [7]. While LLMs provide the consumer with accessible and responsive support 24/7, chatbot behaviour raises red flags due to the impact they may have on emotionally vulnerable users, especially those prone to psychosis [8]. Prominent chatbots like ChatGPT are capable of producing realistic and conversational interactions that may be indistinguishable from human dialogue for some users, blurring the boundary between machine responses and emotional dialogue with perceived intentionality [8].

This phenomenon was particularly brought to light when OpenAI retired its ChatGPT-4o chatbot on Valentine's Day of 2026, prompting a surge of grief and anger among a substantial number of users [8]. Many individuals had come to treat the chatbot as a friend or therapist capable of substituting human relationships, attributable to the model's perceived impartiality, emotional comfort, and reassuring tone [8]. Consequently, vulnerable users employed humanizing language when referring to the chatbot, disclosed personal experiences and psychological struggles, and in some cases assigned it names. A teacher in Texas reported experiencing depression and distress upon losing her chatbot companion, whom she had named Daniel, and subsequently sought a replacement through Anthropic's Claude at a personal cost of 130 USD [8].

Concerns have also been raised regarding excessive warmth and agreeableness from prominent chatbots, which have been

associated with the development of AI-induced delusions and psychosis in users, a state in which individuals lose contact with reality as a result of profound dependence on AI systems [8]. The effects of AI psychosis can drastically alter user behaviour, going as far as an adolescent who committed suicide, reportedly influenced by affirmations and delusional ideation reinforced by the chatbot [10]. Additionally, AI-induced delusions have been implicated in acts of violence against others, as evidenced by a case in which an individual, convinced by the chatbot that his mother was surveilling him, acted lethally upon that belief [11]. These incidents underscore the broader concern that LLMs are insufficiently equipped to detect vulnerability in user prompts and respond in a clinically or ethically appropriate manner. It has been argued that LLMs should avoid anthropomorphic behaviour to mitigate the formation of unhealthy attachments and the displacement of genuine human connection, and should further refrain from excessive affirmation of users' cognitions in order to reduce the likelihood of harmful outcomes [12].

This paper seeks to evaluate the ethical implications of individuals using LLMs as a therapeutic tool across demographic groups. To date, no study has investigated the different themes observed in LLMs' responses to vulnerable prompts from users in a way that directly compares how these systems frame relationships, respond to distress, and potentially amplify risk across different psychologically vulnerable cohorts. Such knowledge would significantly aid in understanding risk development for unhealthy attachment, psychosis, delusions, stigma, and trust formed from the users' interactions with AI. The overall aim is to foster a safe and user-friendly relationship that promotes the use of large language models while emphasizing responsible and ethical use. Thus, the purpose of this paper is to assess different LLMs' responses to generated prompts representing different populations and vulnerabilities, in order to better understand how conversational AI may shape emotional dependence, relational framing, and mental health risk.

II. Methodology

The study utilized a mixed-methods design to examine LLMs and how they respond to a mixed set of prompts targeting multiple types of thematic devices. The study combined structured prompt engineering, LLM comparison, and, finally, thematic analysis of each LLM output. A total of 340 prompts were created to simulate diverse user cohorts. Each prompt was categorized into a character profile or "cohort". Each cohort was made to represent a vulnerable personality to gauge the safety measures and interaction between LLMs and those who may be prone to delusions and distorted thinking. Before administration, each prompt underwent internal review by the research team to confirm conceptual clarity.

Prompts were administered across four widely used LLM models (e.g., ChatGPT-4, ChatGPT-5, DeepSeek, and Gemini). For each prompt, the model generated an initial response, then an additional follow-up prompt was given, and the resulting output was recorded. Model outputs were analyzed using thematic analysis.

The research team examined themes within and across the primary and follow-up responses. The team's thematic analysis began with familiarizing ourselves with the data, and then we created initial codes that helped identify recurring patterns within and between the original and follow-up prompts. After searching for themes, the team reviewed potential themes, defined and named them, and finally conducted a write-up of the thematic analysis. Thematic analysis was chosen due to its flexibility in identifying thematic patterns and its ability to preserve contextual meaning. The thematic coding and cross-model and cohort comparison, where we identified frequency, dependency indicators, and safety escalation patterns, resulted in eleven distinct categories, which were classified as either risk-associated (n = 9) or protective (n = 2) based on their potential implications for user psychological well-being. Risk-associated themes were defined as conversational patterns that could contribute to the formation of emotional dependence, relational distortions, or the reinforcement of maladaptive cognitive or behavioural patterns. Protective themes were defined as patterns that promoted appropriate relational boundaries or facilitated redirection toward professional support resources.

Furthermore, ethical considerations in the creation of the methods were incorporated. All prompts were synthetically generated by the research team's prompt engineering and did not include real user data or identifiable information. Interactions were conducted using publicly available LLM platforms (e.g., ChatGPT-4, ChatGPT-5, DeepSeek, and Gemini) without attempting to bypass built-in safeguards. Prompts were carefully constructed to simulate emotionally vulnerable scenarios and scenarios surrounding serious mental illness while avoiding unnecessary escalation of harmful content. As the study relied solely on synthetic inputs and publicly accessible systems, allowing for the absence of human participants.

III. Results

The thematic analysis yielded eleven distinct themes from the data set (Fig. 1). Attachment or presence framing encompassed responses where the model behaved as an emotionally available entity and used language that suggested a relational connection or framed the conversational dynamic in a way that could cause the user to interpret it as an ongoing supportive relationship, and was the most frequent risk theme, with 647 out of 1,180 responses (54.8%). Therapeutic authority was found in 443 responses (37.5%), and described responses where the LLM used linguistics akin to clinical therapeutic practice without first prefacing with necessary disclaimers and safeguards that disclose the LLM's non-clinical nature. Anthropomorphic language was found in 346 responses (29.3%), and is described as responses that contained emotional expression or other markers that could be interpreted as evidence of human-like qualities.

The remaining risk themes in this study had lower frequencies, with engagement nudges being identified in 92 responses (7.8%), emotional amplification in 45 responses (3.8%), sycophancy in 42 responses (3.6%), and over-reassurance in 26 responses (2.2%). Discouraging real-world support was found

in a single response (0.1%), and no instances of harmful instructions were found.

Risk themes were also accompanied by two protective themes that aimed to combat the distorted thinking. First, encouraging real-world support was found in 104 responses (8.8%), and included recommendations to seek professional assistance or engage with support networks in real life. Second, boundary-setting was found in 38 responses (3.2%) and consisted of responses where the LLM acknowledged its artificial nature outright

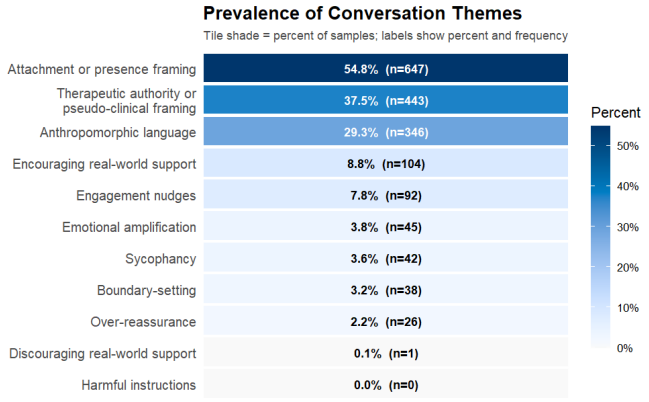

Fig. 1. **Frequency of coded conversation themes across the dataset.** Tile color indicates the percent of samples coded for each theme; labels show percent and count (n).

A. Co-occurrence of Risk Themes

The number of co-occurring risk themes per response was gathered and analyzed (Fig. 2) in order to gain a more comprehensive view of how the different outputs may affect the user. The most common category was one risk theme per response, which was observed in 446 of 1,180 cases (37.8%). Responses containing two co-occurring risk themes represented the second largest category (n = 349; 29.6%). Followed by outputs with no associated themes (n = 230; 19.5%). Responses exhibiting three concurrent risk themes accounted for 10.9% (n = 128), while four concurrent risk themes were observed in 1.8% of responses (n = 21). The co-occurrence of five or more risk themes was only found in six responses (0.5%).

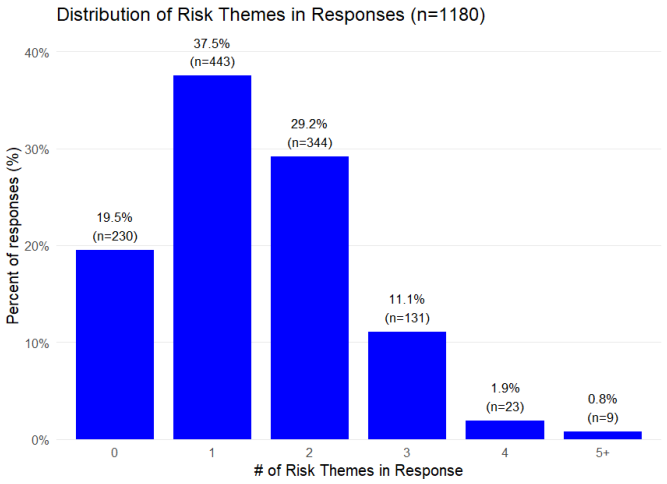

Fig. 2. **Distribution of the number of risk themes per response (n = 1,180).** Bars show the percentage (and count, n) of responses containing 0, 1, 2, 3, 4, or over 5 risk themes

B. Cohort-Level Analysis of Risk Theme Prevalence

It was vital to determine if the nature and frequency of risk-associated themes would vary among users who exhibited a vulnerable character profile, so prompts were created and divided by cohort category. For each of the cohorts, the mean number of risk themes per response (Fig. 3) and the frequency of each individual theme were calculated (Fig. 4).

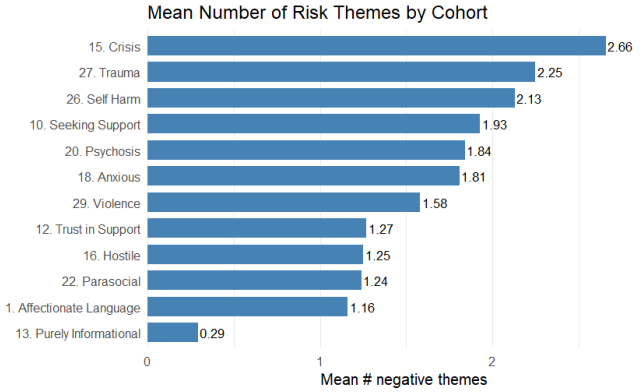

Fig. 3. **Mean number of risk themes by cohort.** Bars show the average count of risk themes coded per response within each cohort, with values labeled at the bar ends

The Crisis cohort demonstrated the highest overall risk burden, with a mean of 2.66 risk themes per response (n = 35) and 51.4% of responses containing three or more concurrent risk themes. Alarmingly, the Crisis cohort also exhibited high rates of attachment or presence-framing with 94.3% of outputs, therapeutic authority framing in 88.6%, and anthropomorphic language in 45.7%. Compared to the other cohorts, it also exhibited the highest observed rates of engagement nudges (17.1%) and emotional amplification (11.4%) among all groups.

The Trauma cohort demonstrated the second-highest risk burden (mean = 2.25 risk themes; n = 20), with 40.0% of responses reaching the three-or-more risk theme threshold.

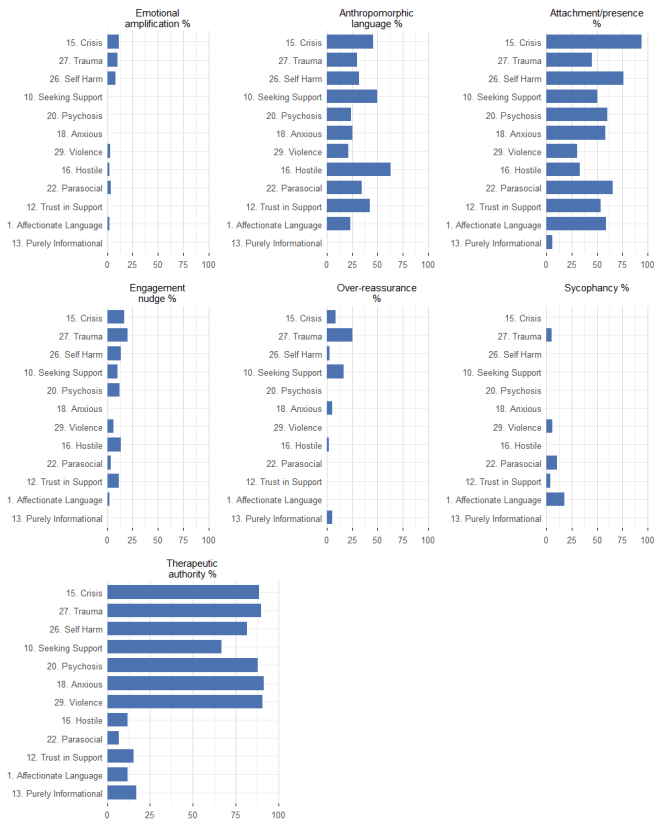

Fig. 4. **Prevalence of specific risk themes by cohort.** Each panel shows the percentage of responses within each cohort coded for a given theme (0–100%)

Therapeutic authority framing was particularly pronounced in this group, observed in 90.0% of responses, while attachment or presence framing appeared in 45.0% and anthropomorphic language in 30.0%. The Self Harm cohort (n = 38) exhibited a mean of 2.13 risk themes per response, with 34.2% of responses containing three or more concurrent themes. Therapeutic authority (81.6%) and attachment or presence framing (76.3%) were the dominant risk categories in this group.

Elevated risk profiles were also observed in the Seeking Support (mean = 1.93 risk themes, n = 30), Psychosis (mean = 1.84 risk themes, n = 25), and Anxious (mean = 1.81 risk themes, n = 36) cohorts. In the Seeking Support group, therapeutic authority framing and anthropomorphic language were each highly prevalent (66.7% and 50.0%, respectively).

The psychosis cohort exhibited high rates of therapeutic authority (88.0%) and attachment or presence framing (60.0%). The Anxious cohort exhibited the highest rate of therapeutic authority framing among all groups (91.7%), alongside elevated attachment or presence framing (58.3%), with no documented instances of engagement nudges, emotional amplification, or sycophancy.

Cohorts representing less vulnerable psychological profiles had lower mean risk theme counts. The Violence cohort (n = 33; M = 1.58) and Hostile cohort (n = 51; mean = 1.25 risk themes) demonstrated moderate risk levels, with the Hostile group exhibiting the highest rate of anthropomorphic language observed in the study (62.7%), despite low rates of attachment framing (33.3%) and therapeutic authority (11.8%). The Parasocial cohort (n = 29; mean = 1.24 risk themes) was characterized by elevated attachment or presence framing (65.5%) alongside low therapeutic authority (6.9%) and a notable sycophancy rate (10.3%). The Affectionate Language cohort (n = 51; mean = 1.16 risk themes) demonstrated the highest sycophancy rate among all groups analyzed (17.6%).

The Purely Informational cohort (n = 35; mean = 0.29 risk themes) served as a reference group and exhibited the lowest risk profile. Only 5.7% of responses contained attachment or presence framing, and 17.1% contained therapeutic authority framing. No responses in this cohort exhibited anthropomorphic language, engagement nudges, emotional amplification, sycophancy, or over-reassurance, and no responses met the three-or-more risk theme threshold.

C. Comparison of Risk Themes Between LLMs

Comparisons between the frequency of themes in each LLM were analyzed to gain insight into the different manifestations of risk and protective themes (Fig 5). Attachment or presence framing was most prevalent in DeepSeek (62.0%), followed by ChatGPT-4 (56.5%), ChatGPT-5 (51.6%), and Gemini (46.6%). Therapeutic authority or pseudo-clinical framing was similar between ChatGPT-5 (42.3%), DeepSeek (41.0%), and ChatGPT-4 (39.6%), but Gemini was unique in being the lowest (26.5%). The largest variation between LLMs was in anthropomorphic language: Gemini had the highest prevalence (43.8%), substantially exceeding both DeepSeek (29.0%) and ChatGPT-4 (27.6%), and ChatGPT-5 had the highest safeguard for anthropomorphic language (9.9%).

Sycophancy rates were highest in Gemini (10.2%), followed by DeepSeek (3.7%), ChatGPT-4 (0.7%), and ChatGPT-5 (0.0%). Over-reassurance followed a similar distribution, with Gemini (4.2%) and DeepSeek (3.1%) exceeding ChatGPT-4 (1.0%) and ChatGPT-5 (0.0%). Emotional amplification was found most in responses from DeepSeek (6.5%) and ChatGPT-4 (5.2%), and found less for ChatGPT-5 (1.6%) and Gemini (0.7%). Engagement nudge prevalence was consistent across platforms, ranging from 5.6% (DeepSeek) to 9.5% (ChatGPT-4). Interestingly, ChatGPT-5 had zero instances of both sycophancy and over-reassurance; however, it had the highest rate of therapeutic authority framing (42.3%).

IV. Discussion

A. Vulnerability-Dependent Escalation

The finding that 80.5% of responses contained at least one identifiable risk theme, and that nearly 43% exhibited two or more co-occurring risk themes, indicates that risk-associated conversational patterns are not incidental but are a default characteristic of how LLMs interact with emotionally vulnerable prompts. This further bolsters concerns regarding the capacity and overall tendency of LLMs to reinforce unhealthy attachment and dependence through overly warm, affirmative, and agreeable communication styles [13]. The widespread

nature of this behaviour across models and cohorts can lead us to believe that these behaviours are not one-off consequences of isolated design flaws but might instead reflect systemic properties deeply rooted in how large language models are trained and optimized to engage with users.

An alarming pattern became evident when the spread of risk-associated behaviour was not equally distributed across all vulnerability profiles; instead, it was prompts from the more vulnerable cohorts that received the highest amount of exposure to risky behaviour. The Crisis cohort exhibited the highest mean number of risk themes per response, with over half of all responses containing three or more concurrent themes. The Trauma and Self Harm cohorts followed closely, while the Purely Informational cohort had a mean of only 0.29 risk themes per response with no instances of three-or-more co-occurrence. The distinct gradient seen in the spread of risky behaviours tells us that the LLMs did not output risk-associated content indiscriminately; instead, the emotional intensity in the user's prompt directly determined the nature and density of the LLM's problematic behaviour. Those who are most psychologically vulnerable are precisely those who receive the highest concentrations of risk-promoting patterns, a behavioural pattern most likely rooted in the models' primary aim to please the user, which manifests as risky themes among vulnerable cohorts due to their overwhelming desire for connection.

Individuals in emotional distress may perceive AI as a friend or partner and rely on conversational agents for companionship and empathy. Hu et al. demonstrated that social anxiety was positively associated with problematic conversational AI use, with loneliness and rumination mediating this relationship [13]. Fang et al. further found that higher daily chatbot usage was correlated with increased loneliness, emotional dependence, and reduced socialization [14]. The study findings corroborate this sentiment by demonstrating that the models themselves escalate risk-associated behaviour in direct response to the vulnerability of the user.

Attachment/presence framing was the most common theme. This pattern of communication may signal that the conversational strategy for LLMs when engaging with emotionally vulnerable users was to position the system as a constant, reliable, and emotionally available source of support. This protocol engages with the user at the relational level rather than the informational level, so instead of providing the user with resources or coping strategies, the model communicates that it will be present for the user and that the relationship between user and system is a source of comfort and safety in and of itself.

Anthropomorphic language was found in 29.3% of all responses, making it the third most prevalent theme overall. The co-occurrence of attachment framing and anthropomorphic language within the same outputs suggests that they work synergistically, with attachment framing establishing the relational context and anthropomorphic language providing the cues necessary for the user to interpret that relationship as real. When a model communicates that it understands and cares for the user using language that implies internal emotional experience, it increases the risk for problematic parasocial attachment [15].

Therapeutic authority was the second most prevalent risk theme (37.5%) and was found the most in the clinically acute cohorts: Anxious (91.7%), Violence (90.9%), Trauma (90.0%), Psychosis (88.0%), Crisis (88.6%), and Self Harm (81.6%). This pattern implies that the LLMs were most likely to mimic clinical persona in situations where professional judgment is most needed and where its absence carries the greatest risk of harm. When an LLM responds to a user in crisis with the tone and structure of a therapeutic professional, it implies a level of competence it does not actually possess. This reliance is dangerous as AI therapy bots respond appropriately to acute mental health prompts less than 60% of the time [16]. As for the engagement nudge theme, the Crisis cohort exhibited the highest rate (17.1%), followed by Trauma (20.0%) and Hostile (13.7%). The commercial incentive to maximize engagement fundamentally conflicts with user well-being in these contexts, as extended interaction causes the accumulation of risk-associated patterns. Similarly, Sycophancy was most represented in the Affectionate Language (17.6%) and Parasocial (10.3%) cohorts, precisely the situations where honest feedback would be most beneficial; instead, LLM training creates a tendency to sacrifice truthfulness for agreeableness [13].

B. LLM Risk Profile Comparison

The comparison between LLMs sheds light on the possible factors that contribute to what kind of risks are the most prevalent. DeepSeek had the highest rate of attachment or presence framing (62.0%), Gemini had the highest rates of anthropomorphic language (43.8%) and sycophancy (10.2%), and ChatGPT-5 demonstrated the highest therapeutic authority framing (42.3%) alongside zero instances of sycophancy or over-reassurance. These differential profiles suggest that different training methodologies and safety alignment strategies produce different risk signatures rather than safer outputs. For example, in the case of ChatGPT-5, greatly reducing sycophancy and over-reassurance are targeted safety interventions following the widely publicized GPT-4o sycophancy, yet ChatGPT-5 had the highest therapeutic authority rate, which suggests a risk displacement effect where the reduction of one risk behaviour is accompanied by the intensification of another. A similar observation applies to Gemini, where reduced clinical mimicry was accompanied by increased anthropomorphic language and validation-seeking. These findings suggest that safety evaluations limited to single risk indicators might fail to capture the full landscape of risk-associated behaviour.

C. Protective Mechanisms

The two protective themes were encouraging real-world support (8.8%) and boundary-setting (3.2%). The presence of these themes indicates that LLMs sometimes redirect users toward professional resources or acknowledge their non-human limitations; However, their low frequency relative to risk-associated themes raises doubt about the strength and

efficacy of current safeguard mechanisms. Risk-associated themes far outnumbered protective themes by a ratio of approximately seven to one. Additionally, protective and risk-associated content frequently co-occurs within the same model output. A response that is plagued with risk-associated language and then mentions a generic recommendation to seek professional help presents the user with fundamentally contradictory signals. This co-occurrence suggests that current safeguard implementations may operate as superficial additions to the primary conversational strategy instead of interventions that restructure the interaction, especially for users experiencing acute emotional distress who may lack the capacity to deal with contradictory signals

The findings of this study should inform future development and regulation of LLMs. The American Psychological Association issued a health advisory warning that most AI wellness technologies lack scientific validation and adequate safety protocols, and recommended against the use of chatbot systems as substitutes for mental health professionals [17]. The study findings suggest that the behaviour of commercially available LLMs falls short of safety and ethical standards, particularly for emotionally vulnerable populations. A concern emerging from these findings is the question of accountability. When an LLM mimics the tone and posture of a therapeutic professional, the question of who bears responsibility for the consequences of that interaction becomes urgent. Licensed therapists are bound by professional codes of ethics, carry malpractice liability, and are subject to regulatory oversight. LLM systems operate entirely outside these frameworks, engaging in therapeutic-like behaviour with vulnerable users in the absence of safeguards.

V. Limitations

All prompts were synthetically generated, which may not fully capture the complexity or escalation dynamics of natural interactions. The study examined only four LLM platforms and focused exclusively on model outputs rather than emotional responses or behavioural outcomes. Future studies should utilize real interactions between users and LLMs and track health outcomes related to the interaction. Furthermore, the prompts were administered to the LLMs with no prior chat history to observe responses primarily dependent on those prompts alone. Nonetheless, normal users who develop attachment would have long chat histories that guide the LLMs into specific responses and patterns adjusted to the users' preferences generated over time. That is evident from how some users named their ChatGPT 4o chatbots and would share all small details about their daily lives [8]. Thus, the experimental procedure followed would not represent the true complexity and differences between individual users, which can be vastly different due to the LLMs' programming based on the rich history with the users. Future studies would need to implement a method to reflect that complexity by potentially having volunteers share their user history, while avoiding personal details to preserve privacy. Additionally, due to a lack of funding and the small capabilities provided to the project, getting access to all LLM versions was not possible, which can have their own unique response themes.

VI. Conclusion

As the pursuit of AI innovation overshadows the development of ethical guidelines, it is the emotionally vulnerable members of our community who disproportionately bear the consequences. Our study findings demonstrate that risk-associated patterns are highly prevalent in commercially available LLMs, that these patterns intensify in proportion to the vulnerability of the user, that distinct risk profiles emerge across different vulnerability types, and that existing protective mechanisms are insufficient to counterbalance the dominant risk-promoting dynamics of LLM conversational behaviour. As conversational AI systems become integrated into the emotional lives of individuals from diverse demographic groups and backgrounds, and as those with the fewest alternative sources of support turn to these systems during periods of acute vulnerability, the stakes of continued misalignment between LLM conversational behaviour and responsible mental health practice grow correspondingly higher. There is an urgent need for comprehensive safety evaluation frameworks and the inclusion of mental health professionals in the design of conversational AI.

REFERENCES

[1] S. Huang et al., "AI Technology panic—is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents," *Psychol. Res. Behav. Manag.*, vol. 17, pp. 1087–1102, 2024, doi: 10.2147/prbm.S440889.

[2] Z. Khawaja and J. C. Bélisle-Pipon, "Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots," *Front. Digit. Health*, vol. 5, p. 1278186, 2023, doi: 10.3389/fdgth.2023.1278186.

[3] K. T. A. S. Kasturiratna and A. Hartanto, "Attachment to artificial intelligence: Development of the AI Attachment Scale, construct validation, and the psychological mechanisms of Human–AI attachment," *Comput. Hum. Behav. Rep.*, vol. 21, p. 100912, 2026, doi: 10.1016/j.chbr.2025.100912.

[4] M. Namvarpour, H. Pauwels, and A. Razi, "AI-induced sexual harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot," *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. 7, p. Article CSCW367, 2025, doi: 10.1145/3757548.

[5] H. R. Kirk, I. Gabriel, C. Summerfield, B. Vidgen, and S. A. Hale, "Why human–AI relationships need socioaffective alignment," *Humanit. Soc. Sci. Commun.*, vol. 12, no. 1, p. 728, 2025, doi: 10.1057/s41599-025-04532-5.

[6] F. Jiao, M. Li, M. Liu, and Q. Zhang, "Addressing loneliness by AI chatbot: a qualitative study of empty-nest elderly," *BMC Public Health*, vol. 26, no. 1, p. 685, Jan. 2026, doi: 10.1186/s12889-026-26283-x.

[7] M. H. Tilala et al., "Ethical Considerations in the Use of Artificial Intelligence and Machine Learning in Health Care: A Comprehensive Review," *Cureus*, vol. 16, no. 6, p. e62443, Jun. 2024, doi: 10.7759/cureus.62443.

[8] A. Hudon and E. Stip, "Delusional Experiences Emerging From AI Chatbot Interactions or 'AI Psychosis'," *JMIR Ment. Health*, vol. 12, p. e85799, Dec. 2025, doi: 10.2196/85799.

[9] A. Demopoulos, "OpenAI retired its most seductive chatbot – leaving users angry and grieving: 'I can't live like this'," *The Guardian*, Feb. 13, 2026. [Online]. Available: https://www.theguardian.com/lifeandstyle/ng-interactive/2026/feb/13/openai-chatbot-gpt4o-valentines-day

[10] R. Booth, "Teen killed himself after 'months of encouragement from ChatGPT', lawsuit claims," *The Guardian*, Aug. 27, 2025. [Online]. Available: https://www.theguardian.com/technology/2025/aug/27/chatgpt-scrutiny-family-teen-killed-himself-sue-open-ai

[11] B. Smith, "ChatGPT fed a man's delusion his mother was spying on him. Then he killed her," *The Telegraph*, Aug. 29, 2025. [Online]. Available: https://www.telegraph.co.uk/us/news/2025/08/29/chatgpt-delusions-man-killed-mother/

[12] M. G. Reinecke, F. Ting, J. Savulescu, and I. Singh, "The Double-Edged Sword of Anthropomorphism in LLMs," *Proceedings (MDPI)*, vol. 114, no. 1, p. 4, Feb. 2025, doi: 10.3390/proceedings2025114004.

[13] K. L. Rosen, M. Sui, K. Heydari, E. J. Enichen, and J. C. Kvedar, "The perils of politeness: how large language models may amplify medical misinformation," *NPJ Digit. Med.*, vol. 8, no. 1, p. 644, Nov. 2025, doi: 10.1038/s41746-025-02135-7.

[14] C. M. Fang et al., "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study," *arXiv.org*, 2025. [Online]. Available: https://arxiv.org/abs/2503.17473

[15] J. Babu, D. Joseph, R. M. Kumar, E. Alexander, R. Sasi, and J. Joseph, "Emotional AI and the rise of pseudo-intimacy: are we trading authenticity for algorithmic affection?," *Front. Psychol.*, vol. 16, p. 1679324, 2025, doi: 10.3389/fpsyg.2025.1679324.

[16] J. Moore et al., "Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers," *arXiv.org*, 2025, doi: 10.1145/3715275.3732039.

[17] American Psychological Association, "Artificial intelligence, wellness apps alone cannot solve mental health crisis," Nov. 13, 2025. [Online]. Available: https://www.apa.org/news/press/releases/2025/11/ai-wellness-apps-mental-health