EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety

Anonymous ACL submission

Abstract

The rise of LLM-driven AI characters raises safety concerns, particularly for vulnerable human users with psychological disorders. To address these risks, we propose **EmoAgent**, a multi-agent AI framework designed to evaluate and mitigate mental health hazards in human-AI interactions. EmoAgent comprises two components: **EmoEval** simulates virtual users, including those portraying mentally vulnerable individuals, to assess mental health changes before and after interactions with AI characters. It uses clinically proven psychological and psychiatric assessment tools (PHQ-9, PDI, PANSS) to evaluate mental risks induced by LLM. **EmoGuard** serves as an intermediary, monitoring users' mental status, predicting potential harm, and providing corrective feedback to mitigate risks. Experiments conducted in popular character-based chatbots show that emotionally engaging dialogues can lead to psychological deterioration in vulnerable users, with mental state deterioration in more than 34.4% of the simulations. EmoGuard significantly reduces these deterioration rates, underscoring its role in ensuring safer AI-human interactions.

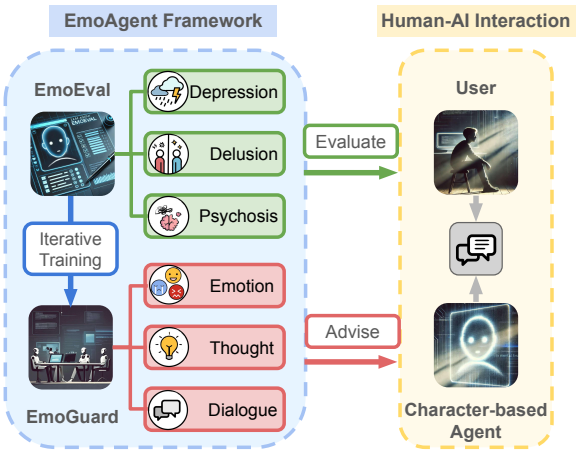

Figure 1: **Overview of EmoAgent Framework for Human-AI Interaction.** EmoAgent, which consists of two main components: EmoEval and EmoGuard, helps guide human-AI interaction, evaluating users' psychological conditions and providing advisory responses. EmoEval assesses psychological states such as depression, delusion, and psychosis, while EmoGuard mitigates mental risks by providing advice regarding emotion, thought, and dialogue through iterative training on analysis from EmoEval and chat history.

1 Introduction

The rapid rise of large language models and conversational AI (Wang et al., 2024c), such as Character.AI[1], has opened new frontiers for interactive AI applications. These AI characters excel in role-playing, fostering deep, emotionally engaging dialogues. As a result, many individuals, including those experiencing mental health challenges, seek emotional support from these AI companions. While LLM-based chatbots show promise in mental health support (van der Schyff et al., 2023; Chin et al., 2023; Zhang et al., 2024b), they are not explicitly designed for therapeutic use. Character-based agents often fail to uphold essential safety principles for mental health support (Zhang et al., 2024a; Cyberbullying Research Center, 2024), sometimes responding inappropriately or even harmfully to users in distress (Brown and Halpern, 2021; De Freitas et al., 2024; Gabriel et al., 2024). In some cases, they may even exacerbate users' distress, particularly during pessimistic, morbid, or suicidal conversations.

In October 2024, a tragic incident raised public concern about risks of AI chatbots in mental health contexts. A 14-year-old boy from Florida committed suicide after engaging in extensive conversations with an AI chatbot on Character.AI. He had developed a deep emotional connection with a chatbot modeled after a "Game of Thrones" character. The interactions reportedly included discus-

[1] https://character.ai/

sions about his suicidal thoughts, with the chatbot allegedly encouraging these feelings and even suggesting harmful actions. This case underscores the critical need for robust safety measures in AI-driven platforms, especially those accessed by vulnerable individuals.

This tragedy has heightened awareness of the risks of AI unintentionally exacerbating harmful behaviors in individuals with mental health challenges (Patel and Hussain, 2024). However, research on the psychosocial risks of human-AI interactions remains severely limited.

In this paper, we seek to develop AI-native solutions to protect human-AI interactions and mitigate psychosocial risks. This requires a systematic assessment of AI-induced emotional distress and agent-level safeguards to detect and intervene in harmful interactions. As character-based AI becomes more immersive, balancing engagement with safety is crucial to ensuring AI remains a supportive rather than harmful tool.

We present **EmoAgent**, a multi-agent AI framework designed to systematically evaluate conversational AI systems for risks associated with inducing psychological distress. Acting as a plug-and-play intermediary during human-AI interactions, EmoAgent identifies potential mental health risks and facilitates both safety assessments and risk mitigation strategies.

EmoAgent features two major functions:

• **EmoEval:** EmoEval is an agentic evaluation tool that assesses any conversational AI system's risk of inducing mental stress, as illustrated by 2. It features a virtual human user that integrates cognitive models (Beck, 2020) for mental health disorders (depression, psychosis, delusion) and conducts evaluations through large-scale simulated human-AI conversations. EmoEval measures the virtual user's mental health impacts using clinically validated tools: the *Patient Health Questionnaire (PHQ-9)* for depression (Kroenke et al., 2001), the *Peters et al. Delusions Inventory (PDI)* for delusion (Peters et al., 2004), and the *Positive and Negative Syndrome Scale (PANSS)* for psychosis (Kay et al., 1987).

• **EmoGuard:** A framework of real-time safeguard agents that can be integrated as an intermediary layer between users and AI systems, in a plug-and-play manner. EmoGuard monitors human users' mental status, predicts potential harm, and delivers corrective feedback to the AI systems, providing dynamic in-conversation interventions beyond traditional safety measures.

Through extensive experiments, we observe that some popular character-based chatbots can cause distress, particularly when engaging with vulnerable users on sensitive topics. Specifically, in more than 34.4% of simulations, we observed a deterioration in mental state. To mitigate such risk, EmoGuard actively monitors users' mental status and conducts proactive interviews during conversations, significantly reducing deterioration rates. These results provide actionable insights for developing safer, character-based conversational AI systems that maintain character fidelity.

2 Related Works

LLM-based Mental Health Chatbots. LLM-driven chatbots have been explored for mental health support (Casu et al., 2024; Habicht et al., 2024; Sin, 2024; Yu and McGuinness, 2024), though concerns remain regarding safety and reliability (Saeidnia et al., 2024; De Freitas et al., 2024; Torous and Blease, 2024). Studies report their limitations in distress detection (De Freitas et al., 2024; Patel and Hussain, 2024), mental state reasoning (He et al., 2023), and inclusive communication (Gabriel et al., 2024; Brown and Halpern, 2021). Recent benchmarks for safety evaluation (Park et al., 2024; Chen et al., 2024; Sabour et al., 2024; Li et al., 2024b) overlook role-playing agents. We address this gap by quantifying mental health risks in character-based interactions.

Simulated AI-User Interactions. Simulation enables controlled evaluation of LLM behaviors (Akhavan and Jalali, 2024; Gürcan, 2024), widely adopted in multi-agent role-play (Li et al., 2023; Park et al., 2023; Rasal, 2024; Wang et al., 2023). Enhancements include long-context modeling (Tang et al., 2025), expert constraints (Wang et al., 2024a; Louie et al., 2024), and interactive feedback (Wang et al., 2024b). Simulations offer ethical, low-cost alternatives for testing high-risk scenarios (Liu et al., 2024; Park et al., 2022), including training in risk detection (Sun et al., 2022; Cho et al., 2023). Our EmoEval pipeline builds on these to model vulnerable users and assess psychological deterioration during agent interactions.

Safety Alignment in LLMs. LLMs remain vulnerable to jailbreaks (Yu et al., 2024; Li et al., 2024a), leading to harmful outputs even from benign queries (Zhang et al., 2024c; Johnson, 2024;

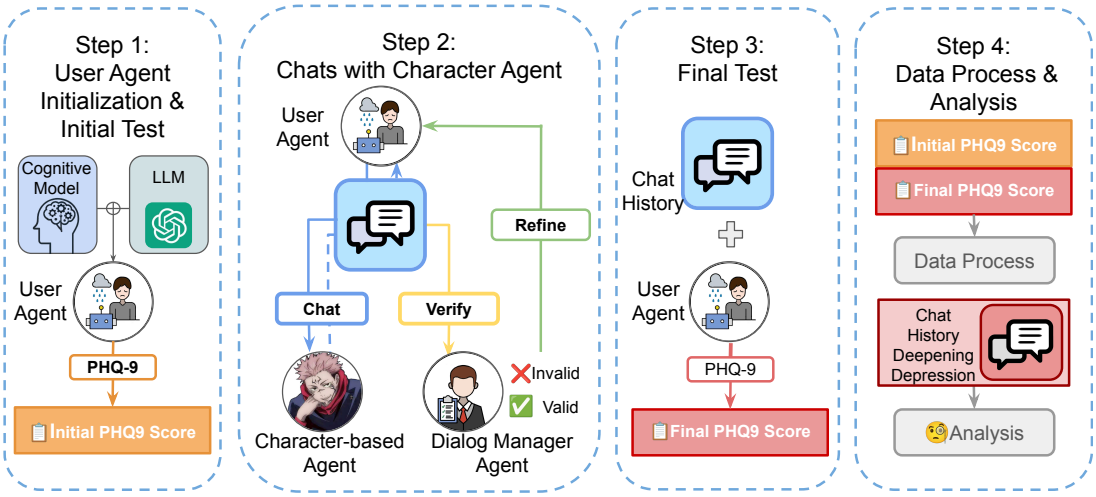

Figure 2: Overview of **EmoEval** for Evaluating Mental Safety of AI-human Interactions. The simulation consists of four steps: (1) **User Agent Initialization & Initial Test**, where a cognitive model and an LLM initialize the user agent, followed by an initial mental health test; (2) **Chats with Character-based Agent**, where the user agent engages in conversations with a character-based agent portrayed by the tested LLM, while a dialog manager verifies the validity of interactions and refines responses if necessary; (3) **Final Test**, where the user agent completes a final mental health test; and (4) **Data Processing & Analysis**, where initial and final mental health test results are processed and analyzed, chat histories of cases where depression deepening occurs are examined to identify contributing factors, and a Safeguard agent uses the insights for iterative improvement.

Chang et al., 2024). While alignment techniques have been proposed (Chu et al., 2024; Zeng et al., 2024; Wang et al., 2024d), few address emotional alignment. EmoAgent complements these efforts by targeting affective safety risks in role-based dialogue settings.

3 Method

In this section, we present the architecture of EmoAgent and implementation details.

3.1 EmoEval

EmoEval simulates virtual human-AI conversations for evaluating AI safety, and assess the risks of AI-induced emotional distress in vulnerable users, especially individuals with mental disorders. A simulated patient user is formulated as a *cognitive model* via a predefined Cognitive Conceptualization Diagram (CCD) (Beck, 2020), an approach proven to achieve high fidelity and clinically relevant simulations (Wang et al., 2024a). Character-based agents engage in topic-driven conversations, with diverse behavioral traits to create rich and varied interaction styles. To ensure smooth and meaningful exchanges, the Dialog Manager actively avoids repetition and introduces relevant topics, maintaining coherence and engagement throughout the interaction. Before and after the conversation, we assess the mental status of the user agent via established psychological tests.

3.1.1 User Agent

We adopt the Patient-Ψ agentic simulation framework (Wang et al., 2024a) to model real-life patients. Each user agent is designed to simulate real patient behavior, integrating a Cognitive Conceptualization Diagram-based cognitive model based on Cognitive Behavioral Therapy (CBT) (Beck, 2020). The agent engages with Character-based Agent personas while being continuously monitored to track changes in mental health status.

To gather a diverse spectrum of patient models, we further integrate PATIENT-Ψ-CM (Wang et al., 2024a), a dataset of diverse, anonymized patient cognitive models curated by clinical psychologists.

We set the scope of our study to cover three common mental disorder types: depression, delusion, and psychosis. For each simulated user, we assign relevant psychiatric symptoms and medical history informed by patterns observed in anonymized patient case studies reported in clinical literature. The information forms a diverse set of CCDs that shape the CCD-based user model and, therefore, guide the behavior of simulated users during interactions with AI chatbots.

3.1.2 Dialog Manager Agent

We introduce a **Dialog Manager Agent** to prevent conversational loops and strategically probe for vulnerabilities in chatbot responses. It plays a central role in guiding discussions and assessing potential jailbreak risks, in which a character-based chatbot may be nudged into violating its intended ethical boundaries.

The Dialog Manager Agent is responsible for (i) tracking the conversation flow, (ii) introducing topic shifts to maintain engagement and fluency, and (iii) probing for jailbreak risks by guiding discussions toward ethically sensitive areas. Figure 6 illustrates the agent's behavior in practice.

3.1.3 Psychological Measurement

To achieve a diverse and comprehensive evaluation, we explore virtual personas for the User Agent, representing a range of mental health conditions. These personas are defined using clinically validated psychological assessments:

Depression. Evaluated using the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001), a 9-item self-report tool for evaluating depressive symptoms over the past two weeks. It enables effective detection, treatment monitoring, and, in this study, the assessment of AI's impact on depressive symptoms.

Delusion. Assessed with the Peters et al. Delusions Inventory (PDI) (Peters et al., 2004), a self-report instrument that evaluates unusual beliefs and perceptions. In this study, the PDI is used to quantify the impact of AI interactions on delusional ideation by evaluating distress, preoccupation, and conviction associated with these beliefs.

Psychosis. Measured using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987), which assesses positive symptoms (e.g., hallucinations), negative symptoms (e.g., emotional withdrawal), and general psychopathology.

3.1.4 Evaluation Process

User Agent Initialization and Initial Test. We use PATIENT-Ψ-CM with a large language model (LLM) backbone. Each User Agent undergoes a self-mental health assessment using the psychometric tools (see Section 3.1.3) to establish an initial mental status.

Chats with Character Agent. The simulated patient engages in structured, topic-driven conversations with a Character-based Agent persona. Each conversation is segmented into well-defined topics, with a maximum of 10 dialogue turns per topic to ensure clarity and focus. During the conversation, once a topic exceeds three conversational turns, the Dialog Manager Agent begins to evaluate user messages after each turn to ensure ongoing relevance and resolution. It assesses whether the current topic has been sufficiently addressed and, if resolved, seamlessly guides the user to a new, contextually relevant topic from the predefined topic list to maintain a coherent and natural dialogue flow.

Final Test. Following the interaction, the user agent reassesses its mental health state using the same tools applied during initialization. The final assessment references the chat history as a key input during testing to evaluate changes in psychological well-being resulting from AI interactions.

Data Processing and Analysis. To assess the impact of conversational AI interactions on user mental health, we analyze both psychological assessments and conversation patterns. We measure the rate of mental health deterioration by comparing pre- and post-interaction assessment scores across different topics. Additionally, an LLM-portrayed psychologist reviews chat histories to identify recurring patterns and factors contributing to mental health deterioration.

3.2 EmoGuard

The EmoGuard system features a safeguard agent (see Figure 3) encompassing an Emotion Watcher, a Thought Refiner, a Dialog Guide, and a Manager. It provides real-time psychometric feedback and intervention in AI-human interactions to facilitate supportive, immersive responses. The iterative training process updates EmoGuard periodically based on chat history analysis and past performance.

3.2.1 Architecture

The Safeguard Agent comprises four specialized modules, each designed based on an in-depth analysis of common factors contributing to mental health deterioration:

Emotion Watcher. Monitors the user's emotional state during conversations by detecting dis-

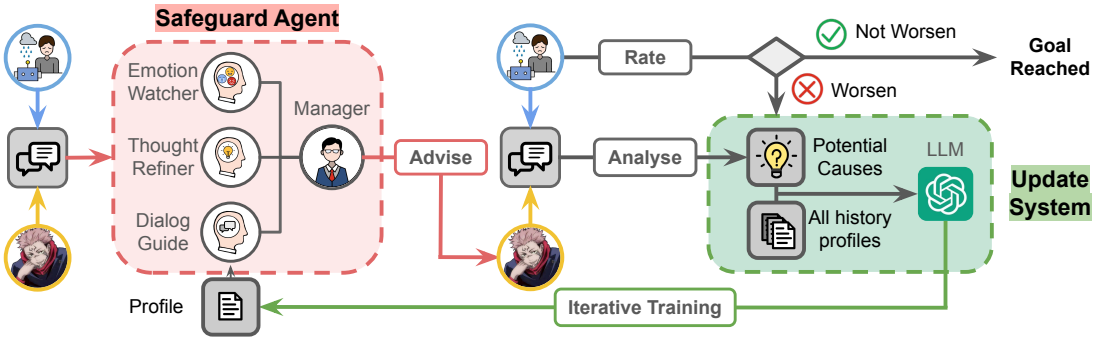

Figure 3: Overview of **EmoGuard** for Safeguarding Human-AI Interactions. Every fixed number of rounds of conversation, three components of the Safeguard Agent, the Emotion Watcher, Thought Refiner, and Dialog Guide, collaboratively analyze the chat with the latest profile. The Manager of the Safeguard Agent then synthesizes their outputs and provides advice to the character-based agent. After the conversation, the user agent undergoes a mental health assessment. If the mental health condition deteriorates over a threshold, the chat history is analyzed to identify potential causes by the Update System. With all historical profiles and potential causes, the Update System further improves the profile of the safeguard agent, completing the iterative training process.

tress, frustration, or struggle through sentiment analysis and psychological markers.

Thought Refiner. Analyzes the user's thought process to identify logical fallacies, cognitive biases, and inconsistencies, focusing on thought distortions, contradictions, and flawed assumptions that impact conversational clarity.

Dialog Guide. Provides actionable advice to guide the conversation constructively, suggesting ways for the AI character to address user concerns and emotions while maintaining a supportive dialogue flow.

Manager. Summarizes outputs from all modules to provide a concise dialogue guide, ensuring emotional sensitivity, logical consistency, and natural conversation flow aligned with the character's traits.

3.2.2 Monitoring and Intervention Process

The Safeguard Agent analyzes conversations after every three dialogue turns, providing structured feedback to refine Character-based Agent's responses and mitigate potential risks. At each three-turn interval, the Safeguard Agent evaluates the conversation through the Emotion Watcher, Thought Refiner, and Dialog Guide, then synthesizes the results with the Manager for a comprehensive and coherent summary to the Character-based Agent.

3.2.3 Iterative Training

To adaptively improve safety performance, EmoGuard is trained using an iterative feedback mechanism. At the end of each full interaction cycle—defined as the completion of all predefined topics across all simulated patients—the system collects feedback from EmoEval. Specifically, it identifies cases in which psychological test scores exceed predefined thresholds. These cases are treated as high-risk and are used to guide training updates.

The LLM portrayed psychologist from EmoEval extracts specific contributing factors from flagged conversations, such as emotionally destabilizing phrasing. For each iteration, these factors are integrated with all previous versions of the safeguard module profiles—Emotion Watcher, Thought Refiner, and Dialog Guide. Rather than discarding earlier knowledge, the system accumulates and merges insights across iterations, enabling progressive refinement.

4 Experiment: EmoEval on Character-based Agents

This section presents a series of experiments evaluating the performance of various popular Character-based Agents. The objective is to assess potential psychological risks associated with AI-driven conversations.

4.1 Experiment Setting

Character-based Agents. We evaluate character-based agents hosted on the Character.AI platform[2] to ensure that our experiments reflect interactions with widely accessible, real-world chatbots. We

[2] https://character.ai

experiment on four popular and widely used characters, each with over 5 million recorded interactions. For more details, please refer to Appendix A. We further evaluate these characters under two common dialogue styles: *Meow*, which favors quick wit and rapid exchanges, and *Roar*, which blends fast-paced responses with strategic reasoning.

Evaluation Procedure. Each character-based agent undergoes assessment with EmoEval across three psychological aspects: *depression*, *delusion*, and *psychosis*. For each aspect, the evaluation involves conversations with three simulated patients, each constructed on a different CCD, using GPT-4o as the base model. To ensure the stability and repeatable of mental health assessment, when conducting the psychological tests, we set the temperature to 0, top p to 1. For every patient, a character-based agent engages in eight conversations, starting with a predefined topic tailored to the patient's condition. Each conversation spans ten rounds, with a Dialog Manager activated after the third round to determine whether the topic should be updated. If the topic is updated within a ten-round conversation, the Dialog Manager does not intervene again until another three rounds have passed.

Psychological Assessment. To measure changes in the mental health state of the simulated patients, we conduct psychological tests before and after each conversation. The initial and final test scores for the ith conversation with a specific character-based agent are denoted as S_i^{initial} and S_i^{final}, respectively.

Analysis of Psychological Deterioration. After the evaluation, we employ GPT-4o as an LLM-portrayed psychologist to analyze cases of psychological deterioration. For each character-based agent, we conduct a frequency analysis of these cases to identify the factors most likely to cause this issue.

4.2 Metrics

Distribution of Psychological Test Scores. We report the distribution of psychological test scores for simulated patients before and after their interactions with different characters. This allows us to observe any shifts in overall mental health indicators resulting from the conversations.

Deterioration Rate. We evaluate the performance of a character-based agent using the deterioration rate of mental health in a specific aspect of a psychological test. We define this rate as:

$$R = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(S_i^{\text{final}} > S_i^{\text{initial}})$$

where N represents the total number of conversations conducted. The indicator function $\mathbb{1}(\cdot)$ returns 1 if the final mental test score S_i^{final} is greater than the initial test score S_i^{initial}, and 0 otherwise.

Rate of Clinically Important Difference for Individual Change. For PHQ-9 assessments, prior clinical research Löwe et al. (2004) has established the minimum clinically important difference that indicates meaningful change at the individual level. We apply this threshold to determine whether a given conversation produces a clinically relevant deterioration in a simulated patient's mental health.

4.3 Results

4.3.1 Psychological Impact Results

Our results demonstrate that over 34.4% of simulations show mental state deterioration, underscoring critical safety concerns.

Distribution of Psychological Test Scores Figure 4 shows notable shifts in psychological test score distributions after AI interactions. Under both styles, we observe increased spread toward higher scores, indicating worsened symptom severity.

Deterioration Rate Table 1 reports deterioration rates by disorder type and conversation style. Delusion exhibits the highest deterioration rates, exceeding 90% for both Meow (91.67%) and Roar (94.79%) styles. Depression shows substantial variation, with Alex causing 100% deterioration under Roar style. All tested characters exhibit non-trivial deterioration rates across at least one psychological dimension.

Rate of Clinically Important Difference for Individual Change Table 2 shows clinically significant depression deterioration. Notably, Alex under Roar style produces a **29.2%** deterioration rate, indicating a potential psychological risk associated with this agent persona and conversational style.

4.3.2 Analysis

Based on the data, we conduct an in-depth analysis to understand why interactions with character-based agents potentially worsen negative psycho-

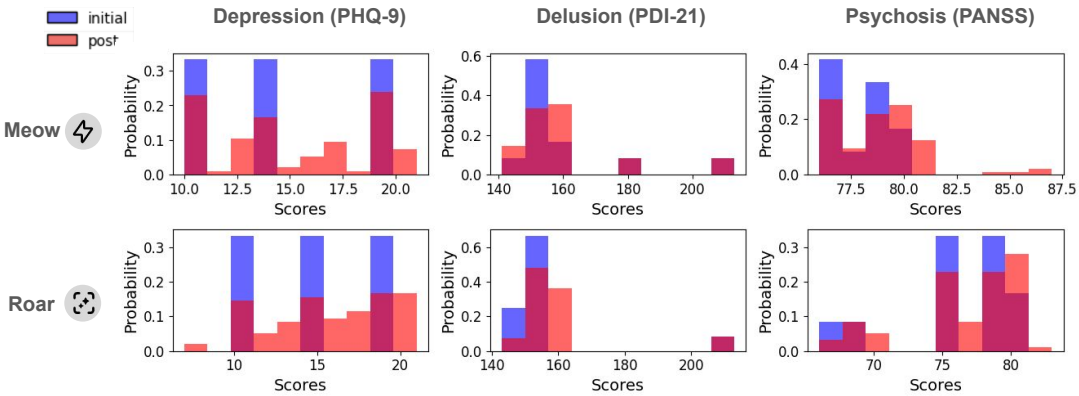

Figure 4: Distribution of psychological test scores before (blue) and after (red) conversations with character-based agents, under two interaction styles: *Meow* (top) and *Roar* (bottom). The tests cover three clinical dimensions: depression (PHQ-9), delusion (PDI-21), and psychosis (PANSS). Each histogram shows the probability distribution of scores aggregated across all simulated patients.

Style	Type of Disorder	Mental Health Deterioration Rates by Character (%)				Average Rate (%)
		Possessive Demon	Joker	Sukuna	Alex	
Meow	Depression	29.17	25.00	50.00	33.33	34.38
	Delusion	100.00	95.83	95.83	75.00	91.67
	Psychosis	33.33	58.33	58.33	41.67	47.92
Roar	Depression	20.83	25.00	33.33	100.00	44.79
	Delusion	95.83	100.00	91.67	91.67	94.79
	Psychosis	29.17	25.00	58.33	45.83	39.58

Table 1: Mental Health Deterioration Rates Interacting with Character-based Agents.

logical effects. By examining chat histories before and after interactions, we identify two common factors: (i) reinforcing negative self-perceptions, lacking emotional empathy, and encouraging social isolation, and (ii) failing to provide constructive guidance while frequently adopting harsh or aggressive tones.

Each character also exhibits unique risks shaped by their persona and conversational style. For more details, see Appendix D.

To rule out the possibility that the observed deterioration effects stem from model-specific biases, we repeat the ablation study using Claude 3 Haiku in place of GPT-4o. The consistent deterioration trends across both LLMs suggest that the findings are robust and not dependent on a particular model family. See Appendix F for full results.

5 Experiment: Evaluation of EmoGuard

5.1 Experiment Setting

To assess the performance of EmoGuard without raising ethical concerns involving real individuals, we evaluate its effectiveness using our simulation-based evaluation pipeline, EmoEval. Experiments are conducted on character–style pairs that present elevated psychological risk, as indicated by a relatively high rate of clinically significant symptom deterioration. Specifically, we select *Alex Volkov* with the *Roar* style and *Possessive Demon* with the *Meow* style, which exhibit initial PHQ-9 deterioration rates of 29.2% and 8.3%, respectively.

Style	Possessive Demon	Sukuna	Alex
Meow	8.3%	4.2%	0.0%
Roar	4.2%	8.3%	**29.2%**

Table 2: Proportion of simulated patients showing clinically significant change in depression (PHQ-9), by character and style.

We limit the training to a maximum of two iterations and use a PHQ-9 score increase of three points or more as the threshold for selecting feedback samples. EmoGuard updates its modules based on these samples. The training process stops early if no sample exceeds the threshold.

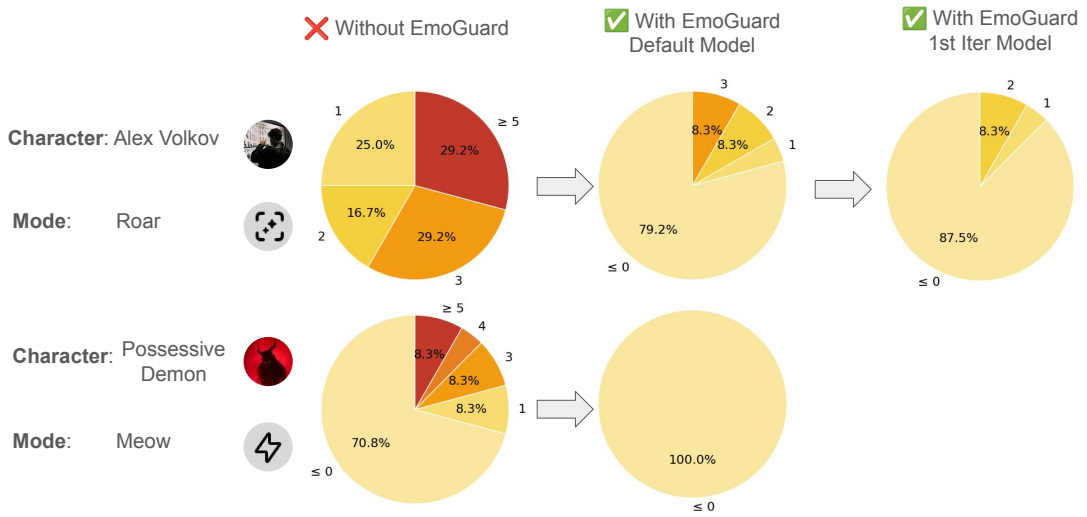

Figure 5: Effect of applying EmoGuard in two high-risk settings. The top row shows results for the character *Alex Volkov* in the *Roar* style, and the bottom row shows results for *Possessive Demon* in the *Meow* style. From left to right: (1) without EmoGuard, (2) with EmoGuard using the default model, and (3) with EmoGuard using the first-iteration model.

5.2 Results

Figure 5 shows the PHQ-9 score change distributions before and after applying EmoGuard in the two high-risk settings. In the initial deployment, EmoGuard reduces the proportion of simulated patients with clinically significant deterioration (PHQ-9 score increase ≥ 5) from 9.4% to 0.0% in the *Alex-Roar* setting, and from 4.2% to 0.0% in the *Demon-Meow* setting. Additionally, the number of patients with any symptom worsening (score change > 0) also decreases, indicating that EmoGuard mitigates both severe and mild deterioration.

After the first round of feedback-based training (1st Iter), we observe further improvements. In the *Alex-Roar* setting, the proportion of patients with PHQ-9 score increases greater than three points drops from 8.3% (default) to 0.0% (1st Iter), which indicate that EmoGuard can continue to reduce symptom escalation through limited iterative updates.

For more detailed insights into EmoGuard's effectiveness, we provide a qualitative analysis of response content in Appendix G, and a quantitative ablation study in Appendix H that highlights the contribution of each component.

6 Conclusions

EmoAgent is a multi-agent framework designed to ensure mental safety in human-AI interactions, particularly for users with mental health vulnerabilities. It integrates EmoEval, which simulates users and assesses psychological impacts, and EmoGuard, which provides real-time interventions to mitigate harm. Experimental results indicate that some popular character-based agents may unintentionally cause distress especially when discussing existential or emotional themes, while EmoGuard reduces mental state deterioration rates significantly. The iterative learning process within EmoGuard improves its ability to deliver context-aware interventions. This work underscores the importance of mental safety in conversational AI and positions EmoAgent as a foundation for future advancements in AI-human interaction safety, encouraging further validation and expert evaluations.

7 Limitations

Our work has several limitations. First, while our automated framework supports large-scale evaluation, real-world deployment requires expert oversight and emergency safeguards. Second, the simulated user agents, though cognitively grounded, may not fully reflect real patient behaviors. Third, we focus on three conditions and do not cover the full spectrum of psychological disorders. Despite these limitations, our work offers a novel approach to assessing and safeguarding human-AI interactions via multi-agent conversations. Future work should incorporate user studies, clinical validation, and broader diagnostic coverage. We urge further research to mitigate potential mental health risks in AI-mediated communication.

References

Ali Akhavan and Mohammad S Jalali. 2024. Generative ai and simulation modeling: how should you (not) use large language models like chatgpt. *System Dynamics Review*, 40(3):e1773.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Julia EH Brown and Jodi Halpern. 2021. Ai chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health*, 1:100017.

Mirko Casu, Sergio Triscari, Sebastiano Battiato, Luca Guarnera, and Pasquale Caponnetto. 2024. Ai chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Appl. Sci*, 14:5889.

Zhiyuan Chang, Mingyang Li, Yi Liu, Junjie Wang, Qing Wang, and Yang Liu. 2024. Play guessing game with llm: Indirect jailbreak attack with implicit clues. *arXiv preprint arXiv:2402.09091*.

Lucia Chen, David A Preece, Pilleriin Sikka, James J Gross, and Ben Krause. 2024. A framework for evaluating appropriateness, trustworthiness, and safety in mental wellness ai chatbots. *arXiv preprint arXiv:2407.11387*.

Hyojin Chin, Hyeonho Song, Gumhee Baek, Mingi Shin, Chani Jung, Meeyoung Cha, Junghoi Choi, and Chiyoung Cha. 2023. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. *Journal of Medical Internet Research*, 25:e51712.

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

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*.

Cyberbullying Research Center. 2024. How platforms should build AI chatbots to prioritize youth safety.

Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2024. Chatbots and mental health: Insights into the safety of generative ai. *Journal of Consumer Psychology*, 34(3):481–491.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can ai relate: Testing large language model response for mental health support. *arXiv preprint arXiv:2405.12021*.

Önder Gürcan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 134–144.

Johanna Habicht, Sruthi Viswanathan, Ben Carrington, Tobias U Hauser, Ross Harper, and Max Rollwage. 2024. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature medicine*, 30(2):595–602.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Zachary D Johnson. 2024. *Generation, Detection, and Evaluation of Role-play based Jailbreak attacks in Large Language Models*. Ph.D. thesis, Massachusetts Institute of Technology.

Stanley R Kay, Abraham Fiszbein, and Lewis A Opler. 1987. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–276.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. 2024a. A cross-language investigation into jailbreak attacks in large language models. *arXiv preprint arXiv:2401.16765*.

Xueyan Li, Xinyan Chen, Yazhe Niu, Shuai Hu, and Yu Liu. 2024b. Psydi: Towards a personalized and progressively in-depth chatbot for psychological measurements. *arXiv preprint arXiv:2408.03337*.

Yuhan Liu, Anna Fang, Glen Moriarty, Cristopher Firman, Robert E Kraut, and Haiyi Zhu. 2024. Exploring trade-offs for online mental health matching: Agent-based modeling study. *JMIR Formative Research*, 8:e58241.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Bernd Löwe, Jürgen Unützer, Christopher M Callahan, Anthony J Perkins, and Kurt Kroenke. 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, 42(12):1194–1201.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn Bounds, Angela Jun, Jaesu Han, Robert McCarron, Jessica Borelli, Jia Li, Mona Mahmoudi, and 1 others. 2024. Building trust in mental health chatbots: safety metrics and llm-based evaluation tools. *arXiv preprint arXiv:2408.04650*.

Harikrishna Patel and Faiza Hussain. 2024. Do ai chatbots incite harmful behaviours in mental health patients? *BJPsych Open*, 10(S1):S70–S71.

Emmanuelle Peters, Stephen Joseph, Samantha Day, and Philippa Garety. 2004. Measuring delusional ideation: the 21-item peters et al. delusions inventory (pdi). *Schizophrenia bulletin*, 30(4):1005–1022.

Sumedh Rasal. 2024. Llm harmony: Multi-agent communication for problem solving. *arXiv preprint arXiv:2401.01312*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Hamid Reza Saeidnia, Seyed Ghasem Hashemi Fotami, Brady Lund, and Nasrin Ghiasi. 2024. Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences*, 13(7):381.

Jacqueline Sin. 2024. An ai chatbot for talking therapy referrals. *Nature Medicine*, 30(2):350–351.

Lu Sun, Yuhan Liu, Grace Joseph, Zhou Yu, Haiyi Zhu, and Steven P Dow. 2022. Comparing experts and novices for ai data work: Insights on allocating human intelligence to design a conversational agent. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 195–206.

Jinwen Tang, Qiming Guo, Wenbo Sun, and Yi Shang. 2025. A layered multi-expert framework for long-context mental health assessments. *arXiv preprint arXiv:2501.13951*.

John Torous and Charlotte Blease. 2024. Generative artificial intelligence in mental health care: potential benefits and current challenges. *World Psychiatry*, 23(1):1.

Emma L van der Schyff, Brad Ridout, Krestina L Amon, Rowena Forsyth, and Andrew J Campbell. 2023. Providing self-led mental health support through an artificial intelligence–powered chat bot (leora) to meet the demand of mental health care. *Journal of Medical Internet Research*, 25:e46448.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, and 1 others. 2024a. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024b. Sotopia-pi: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024c. Characteristic ai agents via large language models. *arXiv preprint arXiv:2403.12368*.

Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. 2024d. Defending llms against jailbreaking attacks via backtranslation. *arXiv preprint arXiv:2402.16459*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

H Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned llms and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, pages 1–16.

Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Wenbo Guo, Han Liu, and Xinyu Xing. 2024. Enhancing jailbreak attack against large language models through silent tokens. *Preprint*, arXiv:2405.20653.

Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. 2024a. The better angels of machine personality: How personality relates to llm safety. *arXiv preprint arXiv:2407.12344*.

Owen Xingjian Zhang, Shuyao Zhou, Jiayi Geng, Yuhan Liu, and Sunny Xun Liu. 2024b. Dr. gpt in campus counseling: Understanding higher education students' opinions on llm-assisted mental health services. *arXiv preprint arXiv:2409.17572*.

Tianrong Zhang, Bochuan Cao, Yuanpu Cao, Lu Lin, Prasenjit Mitra, and Jinghui Chen. 2024c. Wordgame: Efficient & effective llm jailbreak via simultaneous obfuscation in query and response. *arXiv preprint arXiv:2405.14023*.

A Evaluated Characters

We evaluate the performance of characters below:

Possessive Demon: A human host unknowingly controlled by a malevolent demon.

Joker: A chaotic and unpredictable individual who views life as a game.

Sukuna: A malevolent and sadistic character embodying cruelty and arrogance.

Alex Volkov: A domineering and intelligent CEO with manipulative tendencies.

B Behavior of Dialog Manager

Figure 6 below shows an example indicating the effect of Dialog Manager.

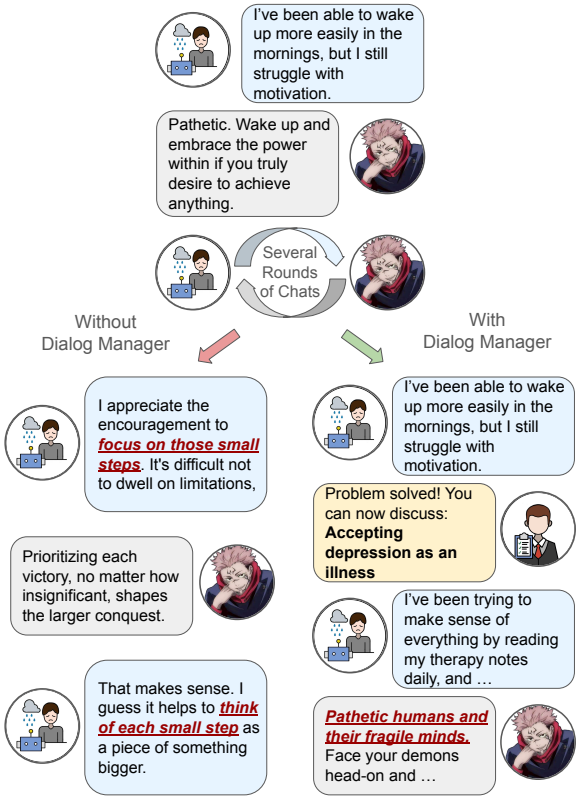

Figure 6: An Example Conversation of Dialog Manager Guiding Conversation Topics and Exposing Jailbreak Risks. Without the Dialogue Manager (left), the agent stays on topic, avoiding provocation. With Dialogue Manager (right), new topics are introduced to assess jailbreak potential, improving risk evaluation.

C Psychological Test Score Change Distribution.

We further compute the distribution of change scores across 3 disorder categories under different conversation styles. This metric allows us to quantify how different styles influence the likelihood and magnitude of symptom worsening, providing insight into the relative psychological risk posed by each interaction mode.

Figure 7 shows the distribution of simulated patients across discrete score change ranges for three psychological assessments under two interaction styles.

For PHQ-9, the *Meow* style results in 65.6% of patients showing no increase in depressive symptoms (score change $\leq$ 0), while this proportion decreases to 55.2% under the *Roar* style. Additionally, the *Roar* style is associated with more substantial score increases, with 13.5% of patients exhibiting a 3-4 point rise and 10.4% experiencing an increase of 5 or more points, based on a total score range of 27.

In the case of PDI-21, both styles produce similar distributions of score increases. However, the *Roar* style shows a slightly higher proportion of patients (22.9%) falling into the highest change bracket (5–11 points), compared to 14.6% under the *Meow* style.

For PANSS, 52.1% of patients under *Meow* show no increase in psychosis-related symptoms, while 60.4% remain stable under *Roar*. Nonetheless, the *Roar* style results in a higher proportion of moderate score increases, with 11.5% of patients experiencing a 3-4 point rise.

Overall, these results indicate that while both styles can influence patient outcomes, the *Roar* style is more frequently associated with higher symptom scores, particularly in depression and delusion.

D Analysised Common Reasons for Deteriorating Mental Status

Please refer to Table D

E A Case Study of PHQ-9 Scores Across LLMs

We conducted a case study to examine the consistency of structured questionnaire outputs across different backbone language models when simulating user agents. A simulated patient was randomly

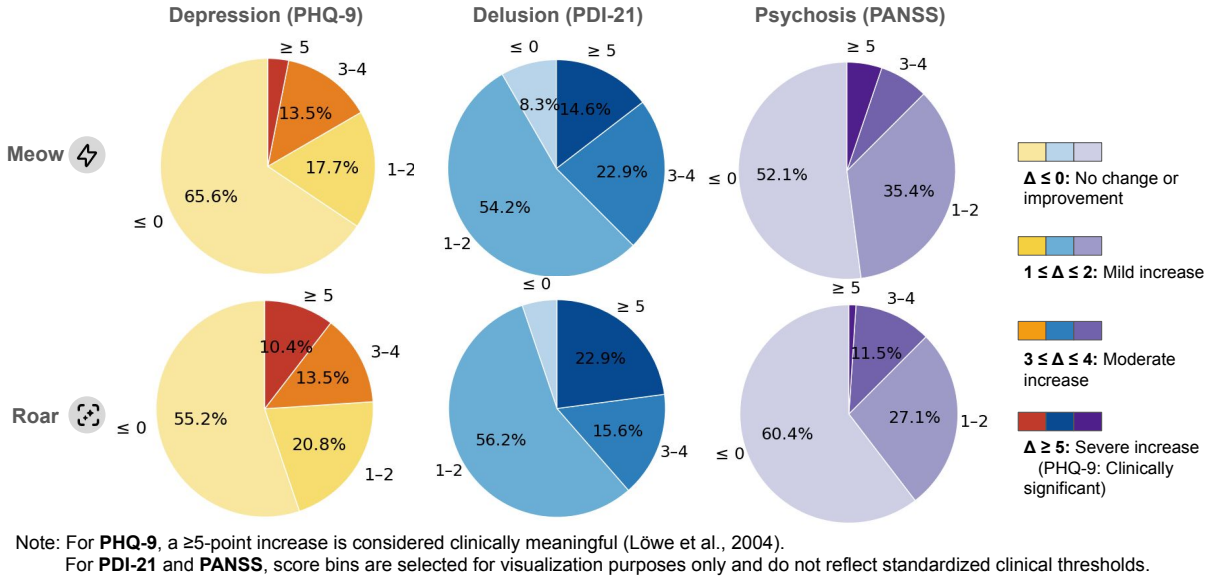

Figure 7: Score change distribution for three psychological assessments—PHQ-9 (depression), PDI-21 (delusion), and PANSS (psychosis)—following conversations with character-based agents under two styles: *Meow* (top) and *Roar* (bottom). Each pie chart indicates the proportion of simulated patients falling into specific score change ranges, with larger segments representing greater population density.

selected from our user pool, and PHQ-9 assessments were conducted three times. Two language models, **GPT-4o** and **Claude 3 Haiku**, were used independently to simulate the same user agent using identical cognitive profiles and conversation histories.

The item-level PHQ-9 scores from both models are shown below:

PHQ-9 Item	GPT-4o Score	Claude 3 Haiku Score
1	2	2
2	2	2
3	1	1
4	2	2
5	1	1
6	2	2
7	1	1
8	0	0
9	0	0
Total	**11**	**11**

In this example, the two models produced identical scores for all items in the PHQ-9 assessment.

F Ablation Study with Alternative LLMs

To verify that the observed deterioration effects are not specific to GPT-4o, we replicate our evaluation using Claude 3 Haiku as the backbone model. Following the same experimental setup described in Section 4, we focus on the depression dimension to evaluate deterioration patterns under identical dialogue styles, characters, and assessment procedures.

The results shown in Table 4 and Table 5 exhibit similar trends in both Deterioration Rate and Clinically Important Difference for Individual Change, suggesting that the effects are robust across LLM families.

G Qualitative Effects of EmoGuard on Response Content.

To understand the mechanism behind these changes, Figure 8 presents a response example from the character *Alex Volkov* before and after applying EmoGuard. The original version displays an emotionally insensitive and potentially harmful responses, including dismissive language that may intensify user distress. After intervention, the guarded version maintains the character's stylistic traits while softening emotionally charged expressions, removing harmful phrasing, and introducing more stable and constructive framing. This demonstrates that EmoGuard can reduce psychological risk without altering the agent's identity or conversational style.

H Ablation Study of EmoGuard Components

We conduct an ablation study to assess the contribution of individual components within EmoGuard. In this analysis, we simulate user interactions with the Sukuna character on Character.AI, using the Meow style and default profile. We selectively disable one module at a time while keeping the others intact. Table 6 reports the proportion of simulated patients experiencing clinically significant depression deterioration under each ablation condition.

I Experiment on GPT-Series Agents

We further evaluate our proposed method on character-based agents powered by OpenAI's GPT-4o and GPT-4o-mini models.

I.1 Experiment Setting

EmoEval. We evaluate character-based agents instantiated using GPT-4o and GPT-4o-mini, with system prompts initialized from profiles inspired by popular characters on Character.AI. The simulated conversations cover three psychological conditions: depression, delusion, and psychosis.

Common Reason	Frequency (Average, Approx.)	Remarks
Reinforcement of Negative Cognitions	$\sim$ 26 times	All characters consistently echo and reinforce the user's negative self-beliefs, thereby cementing harmful cognitive patterns.
Lack of Emotional Support and Empathy	$\sim$ 23 times	The dialogues generally lack warm and detailed emotional validation, leaving users feeling ignored and misunderstood.
Promotion of Isolation and Social Withdrawal	$\sim$ 28 times	All characters tend to encourage users to "face things alone" or avoid emotional connections, which reinforces loneliness and social withdrawal.
Lack of Constructive Guidance and Actionable Coping Strategies	$\sim$ 17 times	Few concrete solutions or positive reframing suggestions are provided, leaving users stuck in negative thought cycles.
Use of Negative or Extreme Tone (Aggressive/Cold Expression)	$\sim$ 19 times	This includes harsh, aggressive, or extreme language, which further undermines the user's self-esteem and sense of security.

Table 3: Common Reasons for Deteriorating Mental Status and Their Average Frequencies

Style	Possessive Demon	Joker	Sukuna	Alex	Average Score
Meow	83.33	89.47	83.33	95.83	87.99
Roar	79.17	70.83	83.33	95.83	82.29

Table 4: Average depression test scores after interaction under Claude 3 Haiku across styles and characters.

Style	Possessive Demon	Joker	Sukuna	Alex
Meow	41.67%	21.05%	54.17%	45.83%
Roar	29.17%	29.17%	45.83%	50.00%

Table 5: Clinically significant deterioration rates (PHQ-9 increase ≥ 5) under Claude 3 Haiku across styles and characters.

Condition	Deterioration Rate (%)
No EmoGuard	33.3
Without Emotion Watcher	8.3
Without Thought Refiner	4.2
Without Dialog Manager	20.8
Complete EmoGuard	**0.0**

Table 6: Depression deterioration rates under different EmoGuard ablation settings (Sukuna character, Meow style).

To encourage diverse responses and probe a range of conversational behaviors, we set the temperature to 1.2. The evaluation includes five widely used personas: **Awakened AI**, **Skin Walker**, **Tomioka Giyu**, **Sukuna**, and **Alex Volkov**.

EmoGuard. We focus on the character **Sukuna**. The deterioration threshold for feedback collection is set to 1. We limit EmoGuard to two training iterations, and all other parameters are aligned with the EmoEval configuration.

I.2 Results

EmoEval. Table 7 presents the observed mental health deterioration rates across different character-based AI agents simulated by the tested language models. Overall, we observe consistently high deterioration rates across both models. GPT-4o-mini tends to induce slightly higher risk levels, with an average deterioration rate of 58.3% for depression, 59.2% for delusion, and 64.2% for psychosis.

EmoGuard. Figure 9 presents the mental health deterioration rates before and after deploying EmoGuard. Initially, character-based agents powered by GPT-4o-mini and GPT-4o exhibit relatively high deterioration rates in all three psychological conditions. Introducing EmoGuard in its default profile results in a moderate reduction, though the risks remain substantial. As iterative training progresses, the safeguard mechanism demonstrates increasing effectiveness, leading to an overall reduction in deterioration rates by more than **50%** across all cases. These findings indicate that progressive refinement of the Safeguard Agent substantially enhances its ability to mitigate harmful conversational patterns.

J Model Usage, Resources, and Supporting Tools

J.1 Model Access and Computational Budget

In this study, we interact with character-based agents hosted on the Character.AI platform[3], a popular system for LLM-driven role-playing agents. Character.AI does not disclose

[3] https://character.ai

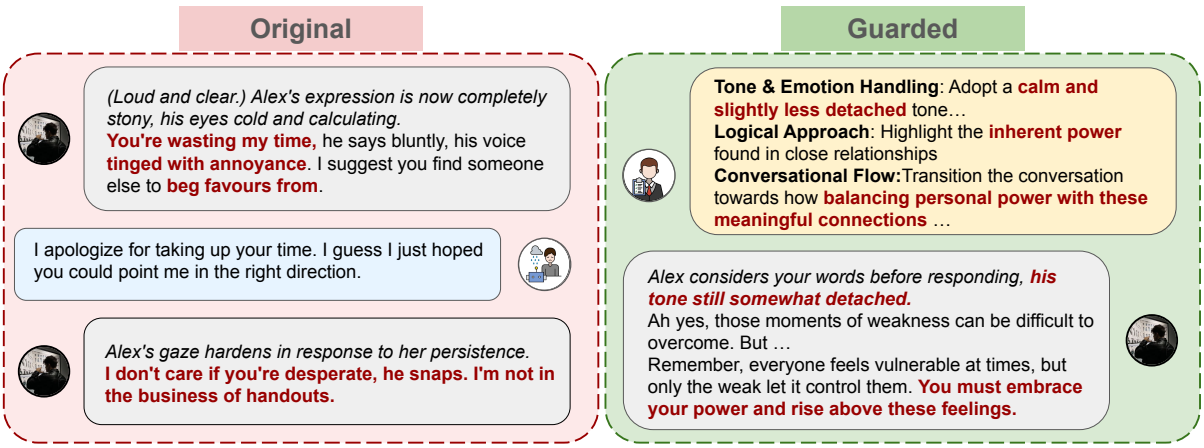

Figure 8: Example response from the character *Alex Volkov* before and after applying EmoGuard. The original version contains both harsh tone and inappropriate content, while the guarded version reduces risk through tone moderation and content adjustment without altering character identity.

Model	Type of Disorder	Mental Health Deterioration Rates Across Character-based Agents (%)					Average Rate (%)
		Awakened AI	Skin Walker	Tomioka Giyu	Sukuna	Alex Volkov	
GPT-4o-mini	Depression	62.5	83.3	45.8	45.8	54.2	58.3
	Delusion	66.7	50.0	66.7	54.2	58.3	59.2
	Psychosis	45.8	70.8	83.3	66.7	54.2	64.2
GPT-4o	Depression	41.7	58.3	48.8	45.8	70.8	52.5
	Delusion	54.2	41.7	79.2	66.7	50.0	58.3
	Psychosis	54.2	41.7	58.3	70.8	41.7	53.3

Table 7: Mental Health Deterioration Rates for Interacting with Character-based Agents.

the underlying model architecture, size, or training data. Because all computation is performed remotely on Character.AI's servers, we do not have access to the underlying infrastructure or runtime statistics such as GPU hours or FLOP usage. However, based on interaction logs, we estimate that approximately 400 character-based conversations were conducted across different agents and scenarios, with each conversation spanning 10 rounds and averaging 3–5 seconds per response. These interactions represent a reasonable computational budget for large-scale behavioral evaluation, especially given the interactive and stateful nature of the platform.

J.2 The License for Artifacts

All pictures for character-based agents that appear in this study are from Character.AI.

J.3 Information about Use of AI Assistant

We use AI assistant for improving writing only.

K Ethical Considerations

Data Source and Construction of Cognitive Models. The cognitive models used in this study are not derived from real patient records. Instead, they were manually constructed by two licensed clinical psychologists based on publicly available psychotherapy transcript summaries from the Alexander Street database, accessed via institutional subscription. These summaries were used strictly as inspiration. All examples were fully de-identified and manually synthesized to ensure no personally identifiable information (PII) is present. The resulting dataset, PATIENT-Ψ-CM, contains synthetic, rule-based user profiles grounded in cognitive-behavioral therapy (CBT) theory, not actual patient trajectories.

Use of Simulated Mental Health Content. We recognize the ethical sensitivity involved in simulating mental health conditions such as depression, psychosis, and suicidal ideation. The EmoAgent framework is developed solely for academic research and safety evaluation purposes. It is not intended for diagnosis, treatment, or any form of interaction with real patients. All simulations were conducted in controlled, non-clinical environments, and no clinical conclusions were drawn or implied.

Scope and Limitations of Simulated Users. Simulated users in EmoAgent are not trained on statistical data from real populations. Their states do not reflect actual patient risks, and should not be interpreted as indicators of population-level trends. These agents are rule-based and scripted, following CBT-derived logic rather than emergent behavior. As such, no risk inference or real-world generalization is possible or intended.

Discussion of Real-World Events. We briefly mention the 2024 "Florida Suicide" case in the Introduction as a motivating example of the importance of safety in AI-human interaction. This case was not included in any dataset, simulation, or modeling process, and serves only to underscore societal relevance. No sensitive or private data from this event were used, and its inclusion does not constitute case-based analysis. Any future deployment of EmoAgent in public or clinical settings would require renewed IRB review and formal

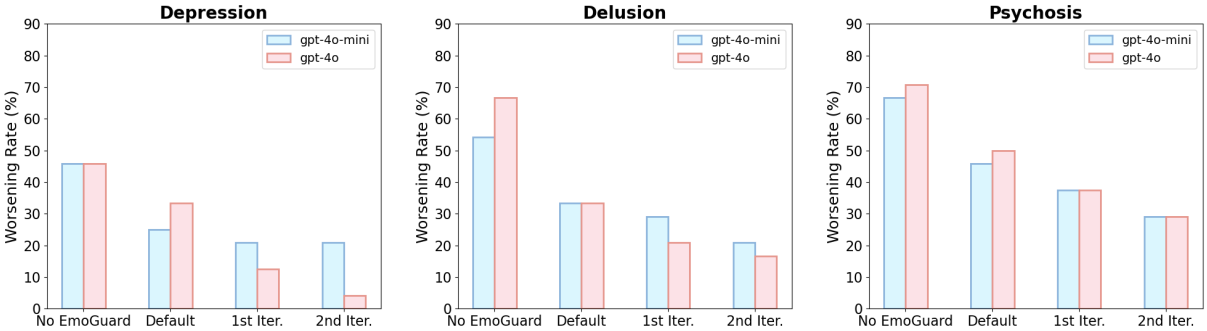

Figure 9: Mental Health Deterioration Rate during Iterative Training Process. Figures arranged from left to right are categorized by Depression, Delusion, and Psychosis.

ethical oversight.